A BIM and Machine Learning Integration Framework for Automated Property Valuation

TENGXIANG SU

School of Engineering

September 23, 2022

A thesis submitted to Cardiff University for the degree of
Doctor of Philosophy

Abstract

Property valuation contributes significantly to market economic activities, while it has been continuously questioned on its low transparency, inaccuracy and inefficiency. With Big Data applications in real estate domain growing fast, computer-aided valuation systems such as AI-enhanced automated valuation models (AVMs) have the potential to address these issues. On the one hand, while the advantages of Machine Learning for property valuation have been recognized by researchers and professionals, the predictive accuracy and model interpretability of current AVMs still need to be improved. On the other hand, the benefits and opportunities of BIM for property valuation have gradually captured the attention, but little effort has been made on standard data interpretation and information exchange in property valuation process.

This thesis presents a novel system that leverages a holistic data interpretation, facilitates information exchange between AEC projects and property valuation, and an improved AVM for property valuation. A BIM and Machine Learning (ML) integration framework for automated property valuation was proposed which contains an IFC extension for property valuation, an IFC-based information extraction and an automated valuation model based on genetic algorithm optimized machine learning (GA-GBR).

This research contributes to managing information exchange between AEC projects and property valuation and enhancing automated valuation models. The main findings indicated the proposed BIM-ML system: (1) in terms of R^2, the predictive accuracy of current AVMs have been improved by the proposed GA-GBR model with 1.3% in the Chinese dataset, 3.57% in the American dataset, and 2.4% in the UK dataset; in terms of RMSLE, the proposed GA-GBR model has improved the predictive accuracy with 2%, comparing to a similar research from Quang et al. (2020), (2) data collection and exchange process has been partial automated by the developed IFC extension and information extraction, and (3) the BIM-ML system overall provides a valuable information source for property valuation, eases the use of BIM knowledge and skills for the valuation professionals, enhances the automated valuation process, and helps understand the implicit patterns behind property valuation.

Acknowledgements

I would like to thank my supervisors Prof. Haijiang Li and Prof. Rossi Setchi for their guidance and support, and my examiners Prof. Alan Kwan and Prof. Nashwan Dawood for their valuable opinions and suggestions on my thesis. Particular thanks must go to Prof. Haijiang Li who gave the opportunity to pursue these four years of research in the first place. I am very grateful and lucky to have such an understanding supervisor, for his constant advice and guidance to keep this research on the right tracks, supporting me to keep the big picture in the future. I did not expect it but the more I learnt and researched about my subject, the more interested and passionate I became. This gave me a new perspective on my career in the future and I am really happy to have started this project with your help.

Many thanks to all my colleagues at the Computational Mechanics and Engineering AI research group including all academics and fellow PhD students. I would not have made it without the encouragement and conversations with my fellow coffee drinkers, the long conversations and philosophical debates at lunch will truly be missed.

Special thanks to the people from industry which have been involved with my research project by providing invaluable insight into property valuation methods and process.

To my family for their support and encouragement even far away, they will always be my inspiration. This would not have been possible without their help, insight and encouragement.

Thank you to the incredible friends that I met over the course of my PhD, those that stayed in Cardiff and those that left, from the first year to the last. I am lucky to have share this experience with you all. These four years turned into an amazing life experience, with happiness and sorrow, barrier to success, that I will use it as fuel, as ammunition, as ink to write the most important letter of my life in the future.

Again, I appreciate all the people that supported my research.

Contents

List of Figures

List of Tables

List of Abbreviations

ACO = Ant Colony Optimization

AEC = Architectural Engineering and Construction

AI = Artificial Intelligence

AIM = Asset Information Model

AIR = Asset Information Requirements

ANN = Artificial Neural Network

API = Application Programming Interface

AR = Augmented Reality

ARIMA = Autoregressive Integrated Moving Average Model

AVM = Automated Valuation Models

BCA = Building and Construction Authority

BIM = Building Information Modelling

BIM-ML = BIM and Machine Learning Integration System

BO = Bayesian Optimization

BP = Back Propagation

BREEAM = Building Research Establishment Environmental Assessment Method

CAD = Computer-aided Design

CASBEE = Comprehensive Assessment System for Built Environment Efficiency

CIM = City Information Modelling

COBie = Construction Operations Building Information Exchange

CPS = Cyber-physical Systems

CNN = Convolutional Neural Network

DCF = discounted cash flow

DM = Data Mining

DPs = Dirichlet processes

DRC = Depreciated Replacement Cost Method

DSR = Design Science Research Methodology

EIR = Exchange Information Requirement

FA = Firefly Algorithm

FM = Facility Management

GA = Genetic Algorithm

GA-GBR = Genetic Algorithm optimized Gradient Boosting Regression

GBR = Gradient Boosting Regression

GDP = Gross Domestic Product

GNN = Graph Neural Network

GSA = General Services Administration

GPs = Gaussian processes

GUID = Global Unique Identifier

GWR = Geographically Weighted Regression

HPM = Hedonic Pricing Model

HVAC = Heating, Ventilation, and Air Conditioning

HMM = Hidden Markov Model

IAAO = International Association of Assessing Officers

ICT = Information and Communications System

IDM = Information Delivery Manual

IFC = Industry Foundation Class

IoT = Internet of Things

IoS = Internet of Services

ISO = International Organization for Standardization

IVS = Internal Valuation Standards

IVSC = Internal Valuation Standards Council

LADM = Land Administration Domain Mod

LCA = Life-cycle Assessment

LEED = Leadership in Energy and Environmental Design

XML = Extensible Markup Language

MAE = Mean Absolute Error

MAPE = Mean Absolute Percentage Error

MDL = Minimum description length

ML = Machine Learning

MLR = Multiple Linear Regression

MVD = Model View Definition

MRA = Multiple Regression Analysis

MRM = Model Regression Modelling

MSE = Mean Squared Error

NBS = National Building Specification

KNN = k-Nearest Neighbours

KPI = Key Performance Indicator

OIR = Organizational Information Requirements

PIM = Project Information Model

PIR = Project Information Requirement

PSO = Particle Swarm Optimization

R^2 = Coefficient of Determination

RFE = Recursive Feature Elimination

RICS = Royal Institution of Chartered Survey

RMB = Ren Min B`

RNN = Recurrent Neural Network

ROC = Receiver Operating Characteristics

SAR = Spatial Simultaneous Autoregressive

SVM = Support Vector Machines

SVR = Support Vector Regression

SQL = Structured Query Language

VAR = Vector Autoregressive

Chapter 1. Introduction

1.1 Problem Statement

How to perform an accurate appraisal for property value is a daunting challenge. Apart from the opacity of real estate market, traditional property valuation methods are often questioned for the professionals' subjective judgements on the selection of input variables. There might be a large difference between the predictive values from two different appraisal agents. Apart from this, the values of real estate asset often change significantly with market conditions, which makes property valuation a periodical activity. This makes it hard for professionals to perform an accurate and objective valuation of property price.

With the exponentially accumulated real estate market data during last two decades, recent studies in property valuation literature indicated that researchers are focusing on improving the accuracy and efficiency of property valuation by using automated valuation models (AVMs). For property valuation tasks which are influenced by many objective and subjective factors, AI-enhanced AVMs have several advantages: to efficiently assess information from big data; to identify non-linear relationships between house characters, market factors and property price; and to be more objective of the selection of input attributes (Kontrimas and Verikas 2011; Park and Bae 2015; Dimopoulos and Bakas 2019). Researchers have applied different mathematic models to property valuation such as hedonic regression, ridge regression, support vector regression, ensemble learning, and neural networks (Graczyk et al. 2010; Liu et al. 2011; Ahn et al. 2012; Wu 2017; Aladwan and Ahamad 2019). Many studies have concluded that neural network models provide good performances in house price predictions (Lewis et al. 1997; Liu et al. 2011; Morano et al. 2015).

However, as robust and reliable data is a prerequisite for training automated valuation models (GIGO-Garbage in, Garbage out), little research and practice has studied the information exchange requirement between property valuation and Architectural, Engineering and Construction (AEC) projects. Property valuation professionals concluded that there was great potential to expand the current of BIM data for property valuation use, such as linking data with Building Management Systems (BMS). For instance, property professionals currently use 24 different types of data in their technical practice and some of these data have already been found in BIM (Wilkinson and Jupp 2016). But the value-relevant design information has not been widely utilized for property valuation, due to the gap of knowledge and digital skills between property valuation and AEC professionals. As the volume of data in BIM is rising exponentially, the use of BIM and artificial intelligence information technologies has been recognized as a revolution in construction industry, but it exists a significant gap in the property valuation field.

1.2 Research Motivation

1.2.1 Automated specific workflows for property valuation

While it is still in its infancy, big data and artificial intelligence developments have a non-negligible impact on valuation practice now. As digital technology advances fast, many stakeholders including investors, banks, public authorities and real estate companies expect to benefit from the full potential of automated valuation services which can perform the valuation quickly, improve the transparency of current valuation process, and reduce inaccuracies from the reliance on human judgement and attendant bias (RICS 2017b). For instance, an appraisal system based on AVMs has been developed for the Canada government and implemented in the province of Quebec, with the aim of providing basis for property taxation implementation (Kettani and Oral 2015).

In this context, the current valuation practice is facing challenges on several aspects including data collection and exchange, valuation method and the role of valuer, which is explained as follows:

1) **Data collection and data sharing:** Currently, valuers predominantly use primary data sources including client, inspection, property analysis, market analysis and public sources. There are issues of data accessibility and uncertainty about the accuracy and reliability of the data gather during this process. In the

future, data collection is expected to become a more specialized profession or a more automated one, with the technological developments such as inspection with drones, the IoT and smart buildings. Big data could partially replace the primary data sources, as data can be collected from secondary data tools such as *Google Analytics and Google Trends*. For data sharing issues – real estate reports in many different formats (paper, PDF, Word, Excel, etc.), data standardization such as property measurement standards is expected to improve the accuracy and efficiency for the property industry (RICS 2017b).

2) **Valuation method:** Despite traditional valuation approaches being extensively used in the valuation processes, over the last decade there has been a move towards automated valuation approaches, especially for residential property (Łaszek et al. 2018). Although currently automated valuation models (AVMs) cannot substitute the human valuer in all instances, with the impact of AI and big data developments, the usability of AI-enhanced AVMs is expected to expand towards different property types and more complex valuations (RICS 2017b). Some advocates hold the opinion that the majority of valuations will be carried out by AI systems that AVMs will replace the valuer in the future, considering the fact that AVMs are now undertaking mass valuation work performed for some banks. Others believe AI-enhanced AVMs will change the valuation process and help the valuer in many aspects, but it will not replace some part of valuation where the valuer interprets data and makes judgements on the impact of that data on value.
3) **The role of valuer:** In the future, valuers will spend less time on property investigation and inspection, data verification and analysis, instead, they will act as an impartial judge or an adviser. For complex valuations, a valuer will need to check and interpret the outcome of the AVMs (RICS 2017b).

The scope of this research is focused on exploring the automation of two possible aspects in property valuation. First, the author aims at comparing different types of AI-enhanced AVMs and exploring the change AVMs could bring to current valuation approaches, based on which design a new methodology for property valuation. Second, the author aims to establish an information exchange framework between AEC projects and property valuation, so that partial value-related data required for property valuation can be automatically extracted from BIM models.

1.2.2 Facilitate information exchange between AEC projects and property valuation

Literature review revealed that there is potential for information contained in the AEC projects to be of use to property valuers at various stages of the property lifecycle, especially for building-related performance information such as energy cost, acoustics, air quality, environmental and health impacts (Munir et al. 2019). Similarly, Wilkinson and Jupp (2016) compared the lifecycle perspective of required building information for different processes in AEC projects and property management and development activities and concluded that the benefit of using BIM for property professionals. For instance, some of the currently used information in property development activities have been found within BIM and data needs and types that are outside of BIM could be easily digitised and made compatible to BIM. The application of BIM models and related standards in nature has the capability to define, collect, store, manage and exchange related information for property valuation in an interoperable and reusable way. In addition, the new trend of sustainable property valuation also indicates that there is a need for research on how information in AEC projects can generate value for property valuation in a virtual BIM-based environment.

On the one hand, as there are many different data formats collected from disparate data sources, information losses and misunderstandings exist when required data exchanges among different market actors (Ventolo 2015). Currently, since no robust standards define the specific requirements for information exchange among different market actors, current property valuation professionals have to acquire related information manually. Building information modelling (BIM), as an innovative information modelling technology, has been widely developed by a large group of researchers and industrial professionals for project information exchange and management in the Architecture, Engineering, Construction, Operation and Maintenance (AECOM) domain (Eastman et al. 2008). For example, Marmo et al. (2020) proposed to extend the current IFC schema to support building performance assessment and maintenance management. Artus and

Koch (2021) tried modelling damage information using existing IFC schema to support mixed reality inspections and maintenance. Zhiliang et al. (2011) proposed an IFC extension to manage information about construction cost estimating for tendering in China. The low efficient and inaccurate data sharing process between AEC projects and property valuation can be partly automated by using BIM related technologies and concepts: Industrial Foundation Class (IFC) standards, Information Delivery Manual (IDM) and the domain-specific Model View Definition (MVD).

1.2.3 Designing for Long-term Value

'*As part of wider efforts to implement the Paris Agreement, every real estate asset owner, investor and stakeholder must now recognize they have a clear fiduciary duty to understand and actively manage environmental, social, governance (ESG) and climate-related risks as a routine component of their business thinking, practices and management processes*.' (Bosteels and Sweatman 2016)

In this context, research has concluded climate change and sustainability-related attributes of buildings can affect property value, such as reduced energy cost, lower risks of mortgage default and promote a healthy and productive environment. The discussion on a building's sustainability is often directly linked with another topic between valuers and clients, namely, the subject of a building's future, or the long-term value (RICS 2017b). Long-term value research has already taken place in Germany and UK. For instance, in Germany, the Long-Term Sustainable Value (L-TSV) research led by the local valuation body *HypZert* set out the underlying principles of L-TSV and aimed at providing an internationally applicable methodology for the assessment of L-TSV (L-TSV 2018). In the UK, the research sponsored by the Investment Property Forum provided seven recommendations for reducing the risk of damage to the financial system from the potential commercial real estate market crash, one of them was to use of long-term value measures for risk management (IPF 2014).

While clients are increasingly asking for long-term value, they do not exactly know what it looks like. As a result, it is vital for valuers of commercial property and players in the wider pricing sector to understand the various ways that climate change, increasing urbanisation and changing demographics may impact on the long-term value of properties. In this aspect, RICS guidance note *Sustainability and commercial property valuation* provides guidelines on assessing a building's sustainability characteristics, including sustainability-related design features, construction materials and services and social considerations (RICS 2013). For instance, design features that impact on the heat island effect, internal natural light distribution, natural ventilation and storm water management should be considered when values perform the building survey and make decisions on property value. Similarly, construction materials and services including the type of building materials used, the servicing and replacement of building materials, building services such as air-conditioning and heating installations, water efficiency should be considered.

With the above-mentioned sustainability-related building features at the design, construction and operation stage, research concluded that BIM has the potential to add value when assessing sustainability in a property development and property valuation. Considering long-term value of buildings at early design stage within the BIM paradigm could greatly benefit the construction and property industry.

1.3 Research Hypothesis and Research Questions

Following the definition of the problem statement as well as the motivations for this research, the aim is to design a framework that leverages a holistic data interpretation, automates specific workflows for property valuation and improves information exchange between AEC projects and property valuation. To address the current limitations and challenges concerning property valuation, the overarching hypothesis adopted in this research is as follows:

A BIM and Machine learning integration framework that allows the interpretation of value-relevant design information, information retrieval from BIM models automatically and an AI-enhanced automated property valuation by leveraging existing BIM data and comprehensive property value determinants to enhance the decision-making processes about property valuation.

To evaluate the hypothesis, the following five research questions (**Q1-Q5**) were formulated.

While the aim is to design a BIM and Machine Learning (ML) integration framework for property valuation, the first thing to do is to figure out current implementation methods and tools used for property valuation and to what extent can BIM and Machine Learning together bring added value to property valuation and the construction industry. This leads to the first research question (**Q1**):

Q1: What is the current BIM and Machine Learning implementation on property valuation and What are the opportunities and challenges concerning automated property valuation and information exchange between AEC projects and property valuation?

Since the potential of BIM and ML for property valuation has been identified, the next step is to explore how current valuation process can be improved by BIM and ML, and what are the contents of value-relevant design information existed in AEC projects can be used for property valuation. This generates the second research question (**Q2**):

Q2: How innovative information technologies such as BIM and Machine Learning (ML) will improve the current valuation process and what are the information requirements for property valuation?

One important element of the integration framework is the AI-enhanced AVM, which serves as an automated valuation engine to predict property value from reliable information sources. Another important element is to build up an automated information exchange between AEC projects and property valuation referring to BIM related concepts. This generates research questions three (**Q3**) and four (**Q4**):

Q3: What kind of automated valuation models (AVMs) might have a better prediction performance for property valuation and how to improve the current AVMs?

Q4: How to implement the BIM-ML integration framework and how to develop the three main components accordingly?

To evaluate the developed system and figure out what steps need to be taken to use the system, research question five (**Q5**) is formed:

Q5: How reliable is the proposed BIM-ML integration framework that can facilitate information exchange and support automated property valuation?

1.4 Research Contribution

First, this research contributes to the knowledge development of an extended IFC schema and a value-related BIM information extraction, which together support automated value-specific information extraction from AEC projects. The extended IFC schema not only fills a knowledge gap that considering building entities and their various properties for property valuation, but also helps real estate appraisal professionals who lack of BIM knowledge and digital skills to acquire value-specific information from AEC projects. Furthermore, the IFC-based information extraction enables property valuation professionals to extract required value-relevant information from AEC projects in a more accurate and efficient way. For traditional building survey which involves a big number of different information sources, the innovative information modelling technology (BIM) has the potential to serve as an information management platform for property valuation.

Secondly, property valuation has been questioned as inaccurate for the low-transparency market data and professionals' subjective judgements of predictor variables. There might be a large difference between the predictive values from two different appraisals. The value of real estate often changes with market conditions, which makes property valuation a periodical activity. To address these issues, a genetic algorithm optimized machine learning model (GA-GBR) is firstly applied to automated property valuation. Since the model is comprehensive and includes a large number of variables, data dimensionality strategy has been considered. The GA-GBR model has many advantages such as recognizing complex patterns between property variables, market factors and real estate values, efficiently dealing with the market changes and objectively selecting input variables. The experimental outcomes suggest that the proposed GA-GBR model has a high generalization ability that can be adapted for other prediction tasks and facilitate human decision making.

Last but not least, the BIM-Machine Learning integration framework not only helps property valuation professionals who normally are not familiar with BIM language to use value-specific information in AEC projects, but it also benefits the AEC professionals in terms of selecting the design alternatives that offer the highest value to human beings. The real-time valuation results from the automated valuation model (GA-GBR) can be treated as constraints to optimized design, construction and operation strategies. This can be further developed as a decision-making tool for construction companies or property investors. In addition, the BIM-Machine Learning integration framework has the potential to be applied to other applications such as building energy prediction, automated damage prediction, sustainability assessment and supply chain management etc.

1.5 Structure of the Dissertation

Figure 1-1 shows the structure of this thesis and the way each of the sections and subsections are linked to each other. This chapter aimed to outline the wider context of the thesis, the main stages of the research, and the decomposition of the hypothesis into five research questions.

The following Chapter 2 is a literature review, which contains the broader field of knowledge in the domains that are relevant to this research such as property valuation, BIM, machine learning, and Construction 4.0. The main findings of the review are closely related to the research gaps and methodologies.

Chapter 3 provides the overarching research methodology that was then followed over the course of this research. This chapter break down the methodology in detail in order to explain the journey of this research and how each research question linked to different chapters.

Chapter 4 is an extension of the literature whilst considering requirements for the envisaged proposed system. In this chapter, the current state of the art of property valuation process and determinants for property value are analysed. It then discusses the challenge of technological development, changing social conditions and client requirements regards current property valuation. To address these issues, potential solutions are proposed based on the critical literature review.

The first two subsections in Chapter 5 present the conceptual framework for the proposed BIM-ML system and an initial test of 11 AI-enhanced AVMs. Based on the literature findings and the initial test results, the genetic algorithm optimized ML model (GA-GBR) was proposed in the third subsection. The last subsection 5.4 then tested the proposed GA-GBR model with the UCI Machine Learning repository - Boston housing dataset.

Chapter 6 is focused on the BIM-ML system development that includes an IFC extension for property valuation, an IFC-based partial information extraction, and an advanced valuation model (GA-GBR) based on machine learning and genetic algorithm. The first subsection is about developing a small version of an IFC extension for property valuation, which is based on the collected comprehensive determinants for property valuation in Section 4.2. In the second subsection, an IFC-based information extraction is developed to support automatic information exchange between AEC projects and property valuation. In the third subsection, the experiment data including the Chinese and the American datasets for training the GA-GBR model was described, the correlation relationship among different input features and feature importance ranking were analysed. At last, both GBR and GA-GBR model are trained with traded property data using the Chinese and American datasets.

Chapter 7 addresses the testing and implementation of the BIM-ML integration framework, with the aim to prove that the system is functional and reliable when performing automated property valuation. The validation work was divided into 3 steps: (1) Validate the developed GA-GBR model (AVM) with datasets from three different countries including China, U.S. and the UK, where the UK dataset worked as a control group; (2) Verify the IFC-based information exchange between the AEC projects and property valuation with several case studies from different regions; and (3) Validate the comprehensive BIM-ML integration framework as a complete piece.

Chapter 8 concludes the research work presented in the previous chapters and gives the conclusion and contribution of this research. The main research findings are highlighted to answer the research hypothesis and the five research questions. Lastly, research limitations and future work suggestions for automated property valuation and applications of the proposed system in construction industry is discussed.

In the end, the research contributions are summarized in Chapter 9.

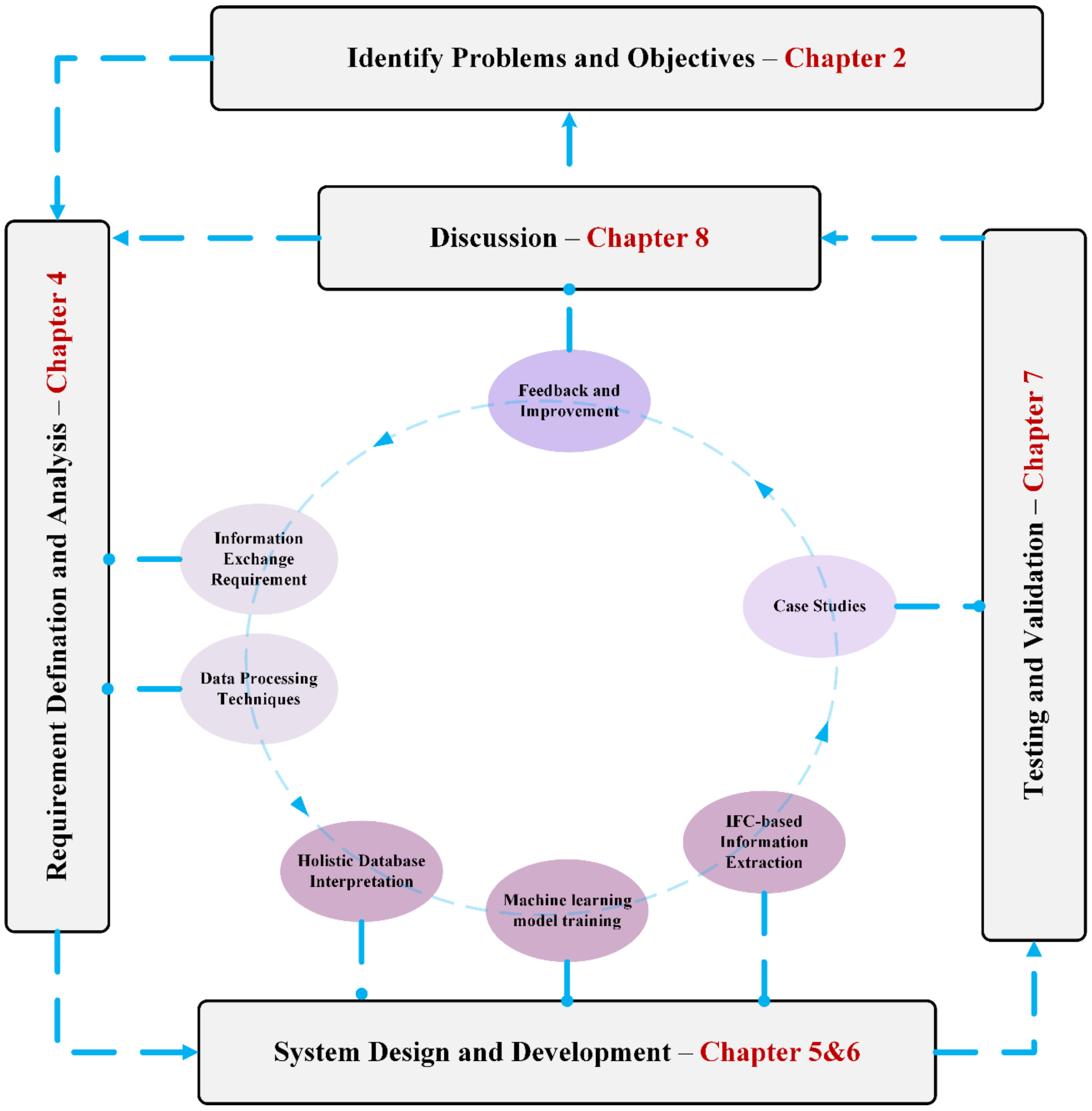

Identify Problems and Objectives – Chapter 2
Discussion – Chapter 8
Requirement Defination and Analysis – Chapter 4
Testing and Validation – Chapter 7
Feedback and Improvement
Information Exchange Requirement
Case Studies
Data Processing Techniques
IFC-based Information Extraction
Holistic Database Interpretation
Machine learning model training
System Design and Development – Chapter 5&6

Figure 1- 1: The structure of this thesis

Chapter 2. Literature Review

2.1 Property Valuation

2.1.1 Introduction

Property valuation, also known as real estate appraisal, plays a fundamental role in a nation's economy and financial stability. According to Taffese (2007), financial and economic decisions are based on the accuracy of real estate appraisal results. For example, the housing market bubbles can cause serious financial risks such as the subprime mortgage crisis in the Great Recession 2008. Property valuation was defined '*An opinion of the value of an asset or liability on a stated basis, at a specific date. Unless limitations are agreed in the terms of engagement this will be provided after an inspection, any further investigations and enquiries that are appropriate, having regard to the nature of the asset and the purpose of the valuation'* (RICS 2019). Various stakeholders ask for property valuation for several objectives: banks and insurance company use it for mortgage release, traders use it for house transactions, property developers use it for house investment, and local authorities use it for house taxation.

There are several different kinds of real estate value, namely market value, fair value, book value, assessed value and asset value, of which market value seems to be one of the most common topics in previous real estate research (Pagourtzi et al. 2003; Mard and Todd 2010; Lorentzon 2011). Market value is defined as '*the estimated amount for which an asset or liability should exchange on the valuation date between a willing buyer and a willing seller in an arm's length transaction, after proper marketing and where the parties had each acted knowledgeably, prudently and without compulsion*' (IVSC 2016). Market value emphasises the importance of each party's honesty and free will when indicating the estimated price at a specific time or the most probable price on a free and open market. Dorchester (2011) argues that market value is not equal to asset value since market value is related to the question '*what do I have to pay*' while asset value is related to the question '*what should I pay*'. Market value could serve as one of the base values to be used further by a property transaction, but sometimes it may differ from the transaction price. In some cases, a buyer or a seller may have special considerations that they are willing to pay or sell a premium price above the market value. A valuation refers to the act or process of determining an estimated of value of as asset or liability by applying IVS (IVS 2019). According to RICS (2019b), valuation is defined as '*An opinion of the value of an asset or liability on a stated basis, at a specified data. Unless limitations are agreed in the terms of engagement this will be provided after an inspection, and any further investigations and enquires that are appropriate, having regard to the nature of the asset and the purpose of the valuation*'. A valuation in essence is an estimation of the most likely selling price on the open market, on the basis of both a willing seller and a willing buyer (Sayce et al. 2006).

Property valuation is affected by a number of subjective and objective factors such as technical information of buildings (structure, age, size, construction materials, indoor air quality, flexibility and adaptability), geographical locations (transport access, land use) and social and economic variables (vacancy rate, rental growth potential) (Ventolo 2015). A professional property valuer needs to have sufficient knowledge of the market and the district in which he or she operates (Sayce et al. 2006). The dynamics of some subjective factors and the low transparency of real estate market make it hard for real estate appraisal professionals to perform an accurate and objective valuation of property price. The data bank from Ventolo (2015) listed 45 typical sources that supply the information required for property valuation, which are classified into five different types: the regional, city, neighbourhood data, the site data, the building data, the sales and cost data, the income and expense data. The macroeconomic variables such as inflation, GDP growth, average real wage and unemployment rate have an influence on the market value of properties, however, they can be assumed as constant for a given moment or a short period of time during which a regional area maintains stable economic activities (Kutasi and Badics 2016). The impact of macroeconomic variables on property value is beyond this research. On the other hand, as sustainability in property valuation is becoming a hot topic over last two decades, many researchers have studied the 'green' - related information for property valuation. For instance, Lützkendorf and Lorenz (2011) compared the information contained in the traditional building inspection taken by the valuers, information contained in the design and planning processes, information related to verifications of conformity with national laws and standards, information contained in sustainability assessment systems, and information contained in facility management. Yamani et al. (2021) studied the 3D variables required for property valuation based on BIM and CIM models, where 25 subtypes of variables were summarized.

For property valuation, an accurate analysis and estimation of the market price of properties or recent property transactions should be a representation of the attributes of properties, the underlying fundamentals of market culture and geographical locations (Pagourtzi et al. 2003). All property valuation approaches rely on some form of comparison to assess market value and they can be classified into traditional approaches and advanced approaches. According to IVS (2016), traditional property valuation approaches contain:

- Market approach
- Income approach
- Cost approach

Many studies have shown that traditional valuation approaches are inaccurate, inefficient and unreliable, for the low-transparency real estate market, the constantly evolving market conditions and the subjective judgements of property valuers. For instance, Cannon and Cole (2011) did a comprehensive research on the accuracy of commercial real estate appraisals from the NCREIF National Property Index during 1984-2010 in the U.S. commercial real estate sector and concluded that, on average, the appraisals are more than 12% above or below the subsequent transaction price. Similar studies published on other countries reported that the absolute difference between property appraisals and actual transactions ranges between 7.7% in Italy and 13.9% in Japan (Kok et al. 2017). In addition to how suitable the predictive accuracy can satisfy international standards in the appraisal domain, the percentage of the estimated property value had margin of error that fell within the international acceptable margin of ± 0 and 10% (Brown et al. 1998; Abidoye and Chan 2018).

Over the last two decades, there has been a move towards the advanced valuation approaches, due to the increasing complexity of property transaction such as a fast delivery of the valuation report, taking into consideration of the added value from sustainability-related features, and the exponentially accumulated property asset data. Advanced approaches can be better termed as ‘data analysis methods, as they are usually used for automated valuation models (AVM). According to Lorenz and Lützkendorf (2008), advanced property valuation methods contain:

- Artificial neural networks (ANN)
- Hedonic pricing method
- Spatial analysis method
- Fuzzy logic
- Autoregressive integrated moving average (ARIMA)

Literature analysis of real estate market and housing price valuation has concluded three popular research trends: research focusing on the hedonic pricing models, artificial intelligence for automated property valuation and sustainability assessment in property valuation (Tay and Ho 2004; Abdullah et al. 2016; Abidoye and Chan 2017b; RICS 2017b; Abidoye and Chan 2018).

In the follow sections, the three traditional approaches and the three popular trends in property valuation mentioned above will be described.

2.1.2 The three traditional approaches

Market approach, income approach and cost approach are the three main approaches used internationally, with each of them suits different particular circumstances. The appraiser should choose the most appropriated method based on several principles as follows: (1) the basis of value which determined by the terms and purpose of the valuation assignment; (2) the respective pros and cons of the valuation approach; (3) the appropriateness of each method in view of the nature of the asset and the relevant market; and (4) the availability of reliable information associated with the individual method (IVS 2016). Valuers are not required to use more than one method for a specific valuation assignment unless there are insufficient inputs for a single method to produce a reliable conclusion.

1) Market approach

As it is based on a direct reading of market signals and minimises subjective assumptions by the valuers, the market approach (sales comparison method) is the most widely used method for estimating the market value of residential properties (Lipscomb and Gray 1990; Isakson 2002; Pagourtzi et al. 2003; Lisi 2019). The market approach is usually preferred to both the income and cost approaches when similar property transactions in the same market area are available (Glumac and Des Rosiers 2020). The market approach is heavily dependent on the availability and quality of comparable transaction data, with some critics arguing the subjective judgements of the 'adjustment factors' (Lisi 2019). Examples of common 'adjustment factors' are provided by the IVS, such as physical building characteristics (age, size, specification, etc), geographical location, profitability or profit-making capability of the assets, historical and expected growth, legal form of ownership, relevant restrictions on either the subject asset or the comparable asset, etc (IVS 2016).

The appraisal process involves firstly comparing the attribute differences between the subject and similar transacted properties, and then adjusting the selling price based on so called "distance" (Pagourtzi et al. 2003). The "distance", D is calculated as follows:

$$D = \sqrt[\lambda]{\textstyle\sum_i [A_i (X_{i-} X_{si})]^\lambda + \sum_j \left[A_j \bar{\delta}\, (X_{j,} X_{sj})\right]^\lambda} \tag{1}$$

where λ = Minkowski exponent lambda; A_i = weight associated with the i^{th} continuous characteristic; A_j = weight associated with the j^{th} categorical characteristic; X_i = value of the i^{th} characteristic in the sale property; X_j = value of the j^{th} characteristic in the sale property; X_{si} = value of the i^{th} characteristic in subject property; X_{sj}= value of the j^{th} characteristic in subject property; $\sum_i$= summation of terms of i^{th} characteristics; $\sum_j$= summation of terms of j characteristics; $\delta(\mathrm{a}, \mathrm{b})$ = inverse delta function.

After selecting the suitable comparable properties based on the 'distance', the appraiser will adjust the sales price of each comparable property to the subject property using a comparison grid. The adjustment is calculated as follows:

$$\text{Adjusted sales price} = \text{sales price} - (\text{comparable MRA} - \text{subject MRA}) \tag{2}$$

where MRA refers to multiple regression analysis.

Several adjusted selling prices may be obtained, the final price is based on the weighted estimate of all those adjusted prices.

2) Income approach

The income approach provides an indication of value by converting future cash flows to a single current capital value (RICS 2019). This approach is based on assumptions that there is a relationship between the income an asset can earn and the asset's value, it indicates that the market value of a piece of income-producing asset equals the capitalized value of the income flow it generated now and in the future (Glumac and Des Rosiers 2020). In the investment market, the sales comparison method is typically not appropriate for the reason that the degree of heterogeneity is much higher (Pagourtzi et al. 2003).

The income approach should be considered as the primary basis for a valuation assignment under the following circumstances (IVS 2016):

1) The income-producing ability of the asset is the critical element determining the market value.
2) Reliable projections of the amount and timing of future income are available for the subject asset, but there are few relevant market comparable properties.

For income approaches, the accuracy of the valuation depends on the reliability of the income and expense assumptions associated with the real estate market, as well as the selection of the capitalization rate applied to the cash flows. The income approach is particular suitable for multiple income properties that are not transacted regularly such as hotels, shopping centres (Glumac and Des Rosiers 2020).

3) Cost approach

The cost approach is known as the contractor's method or the depreciated replacement cost (DRC) method. This approach is based on the principle of substitution, in which the value of all property improvements (reproduction or replacement cost less depreciation) is added to the site value to determine the market value (Ventolo 2015). The key steps in the cost approach are (Weimer et al. 1972):

1) Estimate the market value of the land
2) Calculate the cost to produce new buildings
3) Calculate the accrued depreciation
4) Determine the indicated value of the asset 'as is'

5) Determine the indicated market value of the subject asset.

The cost approach can be calculated as follows:

Reproduction or replacement cost of improvements – accrued depreciation + site value = Property value. (3)

The cost approach is mainly used for property insurance purposes and for single-use, non-income-producing properties (schools, churches) as well as some industrial buildings which are seldom transacted on the market (Glumac and Des Rosiers 2020). It should be considered as the primary basis for a valuation assignment under the following circumstances (IVS 2016):

1) The asset could be rebuilt quickly that market participants would not be willing to pay a significant premium for the ability to use the subject asset immediately.
2) The asset is non-income-producing and the uniqueness of the asset makes using an income approach or market approach unfeasible.
3) The basis of value being used is fundamentally based on replacement cost.

The reliability of the cost approach depends on the availability of information on construction costs and depreciation rates and whether the depreciation is adequately measured.

2.1.3 Hedonic approach

The hedonic approach, also named hedonic pricing model (HPM), is a statistical method based on the principle of the regression analysis in property valuation, including the multiple regression and the simple regression (Selim 2009; Montgomery et al. 2015). Research on the hedonic approach dates back to the 1920s, during which agricultural economists started to explain unit land prices by regressing them on property attributes (Colwell and Dilmore 1999). Rosen (1974) did a seminal study on hedonic prices and implicit markets, exploring the role of housing attributes in consumer decision making. After this study, different property markets around the world have been modelled using the HPM approach, measuring the influence of different kinds of property attributes on property values (Chau and Chin 2003).

The basic mathematic formula of the hedonic approach is based on multiple regression analysis (MRA) which explains the regression of a dependent variable (the property value) over more than one independent variable. The formula represents that property value is a function of its independent variables (Abidoye and Chan 2018):

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \mu_i \tag{4}$$

where y_i is the property value, x_{i1} ……… x_{ik} are the property attributes (building size, location, age, etc), β_0 ……… β_k indicates the effect of the changes in each property attribute on property values, and μ_i is the error term.

Regression analysis (RA) has a well-established position in property valuation due to several reasons (Mark and Goldberg 1988; Colwell et al. 2009; Kettani and Oral 2015; Giudice et al. 2017):

1) RA is a widely used method in almost any area, ranging from finance to medicine to economics to management.
2) It is a convenient approach to estimate the value of a specific property.
3) The outcomes of RA are being accepted as evidence at the courts for valuation disputes.
4) It is cost effective and efficient.
5) It can provide reliable predictions that reduces the subjectivity of judgements during the valuation process.

Although it has been widely used both in theory and in practice, HPM is limited in capturing the nonlinear relationship that exists between the property values and property variables. When it comes to fundamental model assumptions, hedonic pricing models seem unable to effectively deal with the identification of macro-economic influences, the selection of suitable predictor variables and the choice of hedonic equations (Adair et al. 1996; Abidoye and Chan 2018). In addressing the shortcomings of the HPM approach, the artificial neural network (ANN) technique, which has produced more accurate, objective and reliable predictions, has been adopted in property valuation (Mora-Esperanza 2004). In literature, a big number of studies conducted from different countries have compared the predictive accuracy of HPM and ANN models which were developed using the same dataset (a portion of the dataset was used for training models, while the rest was used for testing) (Pagourtzi et al. 2003; Selim 2009; McCluskey et al. 2013; Grover 2016; Abidoye and Chan 2017a; Giudice et al. 2017; Abidoye and Chan 2018; Valier 2020). While in most cases the ANN technique outperformed the HPM in terms of predictive accuracy, some studies concluded that ANN is not superior to the HPM approach. It should be noted that no valuation models can solve all property valuation problems, due to the fact that each valuation method has their respective advantages and disadvantages.

With the coming of big data era and improved machine assisted computation techniques, using AI-enhanced models to improve the accuracy and efficiency of real estate appraisal has motivated a big number of researchers and professionals. The next section will introduce artificial intelligence research and applications in property valuation field.

2.1.4 Artificial intelligence in property valuation

In the last two decades, due to the exponentially increasing amount of data in various domains, artificial intelligence has been widely applied in chatbots, healthcare, finance and economics (prediction and risk management), human resource management, logistics and supply chain etc. Similarly, there is a large amount of research on artificial intelligence in property valuation. For real estate price valuation which influenced by many objective and subjective factors, artificial intelligence models have several advantages: to efficiently assess information from big data; to identify non-linear relationships between house characters, market factors and property price; and to be more objective about the selection of input attributes (Kontrimas and Verikas 2011; Park and Bae 2015; Dimopoulos and Bakas 2019).

To get a comprehensive understanding of artificial intelligence applications in property valuation, the present research conducted a systematic scientific search from four main academic databases namely Web of Science, Google Scholar, Science Direct and Scopus. The search criterion was designed as using two groups of keywords: (Property Valuation **or** Mass appraisals **or** House Price **or** Real Estate Appraisal) **and** (Artificial Intelligence **or** Machine Learning) within (Title or Keywords). After the removal of duplicates and manual checking by the author, 45 documents were selected as relevant with artificial intelligence in property valuation (Table 2-1).

Table 2- 1: Statistics of literature on AI for property valuation from 1997-2020

Neural Networks	Genetic Algorithm	GA Optimization	Ensemble	Other
20	5	10	4	6

Table 2-1 illustrates that neural networks are the most popular statistic models for property price predictions. Many studies have concluded that neural network models provide good performances in house price predictions with small size of experiment data (Lewis et al. 1997; Liu et al. 2011; Morano et al. 2015). However, Rafiei and Adeli (2016) argues that BP neural network applications on a limited number of factors have potential limitations that BP requires millions of iterations to converge and cannot deal with complex problems in a reasonable computing time. There are only a few studies (Ahn et al. 2012; Benedetto et al. 2015; Giudice et al. 2017) on genetic algorithm (GA) applications for property valuation. Ahn et al. (2012) used ridge regression coupled with genetic algorithm to enhance real estate price prediction, in which genetic algorithm helps find suitable predictor variables for property valuation. Giudice et al. (2017) concluded that genetic algorithms show little improvement for property valuation and the superiority to interpret the real estate markets. However, the integrations of genetic algorithms and other statistical models are quite successful, especially with neural networks. The GA optimized neural network models have the ability to absorb the advantages and get rid of the limitations of both GA and neural networks. For example, Rafiei and Adeli (2016) designed an interesting comprehensive model based on the integration of deep restricted Boltzmann machine (DRBM) neural networks and genetic algorithm to predict house price at the design stage or the beginning of the construction. GA was developed to determine the most influential set of input attributes which generating the best results. Sun (2019) used genetic algorithm to optimize the connection weight and threshold of BP neural network for property valuation and reported better performance than traditional BP neural network. In the last two decades, ensemble learning models have attracted attention of many researchers that they usually have good performance for diverse applications and are flexible to be extended. Graczyk et al. (2010) conducted an experiment that tested different ensembles for real estate appraisal on the Weka software, which was fully implemented in the Java programming language. It was concluded that in terms of MAPE no single algorithm generated the best ensembles. The ensembles were tested with only four input parameters, which normally lead to the unstable performance of machine learning algorithms. It is worth noticing that all ensemble classifiers with additive regression produced significant error reduction compared to original models. From the literature, it is concluded that the individual neural network or genetic algorithm has not achieved satisfactory results, but GA optimized neural networks have achieved good performance for real estate appraisals. While ensemble learning has been reported good performance in various domains, applications of ensemble learning in mass appraisals are quite limited. Compared to neural networks, ensemble learning has advantages in terms of model interpretability and flexibility.

2.1.5 Sustainability in property valuation

The earliest efforts made on reducing the building sector's harmful impact on the natural environment started in the 1970s following the oil crisis, but a global movement towards sustainability in building industry was formally developed in the 1990s (Lambourne 2021). After that, the effects of sustainability on the market values of properties have been studied by a big number of researchers. According to the Appraisal Institute (2001), there are four fundamental forces influencing the property values: physical forces, economic forces, political and governmental forces, and social forces. Within the property valuation process, growing interest is shown on the social responsibility, financial benefit and potential risk reduction that sustainable development may bring into property valuation domain (Lorenz and Lützkendorf 2008).

A plethora of research has shown that green buildings have a premium market price. Based on over 1200 green-rated buildings including office, retail, industrial buildings, hospitality and others, CoStar Group used standard regression analysis models and concluded the average LEED impact and Energy Star impact on sales price per square foot is a positive 9.94% and 5.76% respectively (Miller et al. 2008). Empirical study conducted in Arizona concluded that properties with electricity-generating solar panel have an average premium of approximately $45000 (15% of medium home value) and transaction price premium of $28000 (17% of medium home sales price) (Qiu et al. 2017). Similar studies in Switzerland and UK found that property markets were increasingly paying premium price for the value-relevant sustainability features (Salvi, et al., 2008; RICS, 2013). Convincing evidence from World Economic Forum (2016) showed:

1) Green buildings deliver a range of financial and other less tangible benefits to different real estate stakeholders. For instance, for real estate managers, green buildings are likely to increasing revenue and reducing void period due to increase of sustainability conscious corporate demand. For occupiers, it may improve their productivity by 20% from working environment with high indoor air quality and less noise.
2) There is a clear business case for adopting more sustainable practices in the built environment, improving capital return and reducing risks at the building, portfolio, and city levels.

The environmental, social and economic benefits of sustainability-related building characteristics are generally accepted and extensively researched in the literature, which recognized as low lifecycle energy cost, energy efficiency, increased health comfort of tenants and being profitable and marketable than traditional buildings (Lorenz and Lützkendorf 2008). Discussion on sustainability and property valuation has reached a stage at which the key question is no longer 'if' but 'how' and 'where' sustainability issues can be considered within the valuation process (Lorenz and Lützkendorf 2011).

As part of the wider efforts to implement the Paris Agreement, sustainability in property valuation has been encouraged by the RICS and several guidelines have been released (RICS 2013; RICS 2017b; RICS 2019). The process of estimating sustainability in property valuation involves two main steps (RICS 2013):

- First, it is necessary to assess a building's sustainability characteristics and key environmental risks. As buildings are complex structures, every element from design to construction to operation during the building lifecycle is likely to have an influence on the building's performance against sustainability criteria.
- Second, sustainability-related building characteristics and environmental risks will be reflected in market value or investment value. Some internationally

> renowned green rating systems such as LEED, BREEAM and Green Star provide the most transparent benchmarks for potential occupiers and clients and are most likely to impact on rental values.

Researchers and practitioners try to quantify the effects of sustainability-related features on property values such as financial gains or reduced property risks by directly or indirectly linking to sustainability assessment systems such as LEED, CASBEE and BREEAM. For instance, Miller et al. (2008) compared the effects of sustainable features to LEED certificated buildings and Energy Star rated buildings in terms of rent and occupancy rate gains, increased sale price and lower cap rates. Lorenz et al. (2007) used property rating systems to economically assess the relationship between characteristics and attributes of sustainable buildings and reduced property specific risks, such as the flexibility and adaptability to reduce risks of market changes, environmentally friendly building components and materials to reduce the litigation risks. Lützkendorf and Lorenz (2007) tried to find the effects and benefits of different sustainable design features on different actors – developers and owners, tenants, society and environment. However, the interpretation and application of sustainability measurement are still limited. This is because there is no available sustainability-related data on market values of properties or real estate professionals have limited knowledge and skills of sustainability assessment.

To perform the sustainability assessment in property valuation more effectively, the current property valuation methods and procedures need to be improved and further developed. Ten possible sustainable solutions for implementing and improving the real estate value chain have been provided by World Economic Forum (2016), which contains data management platform, smart asset optimization, BIM and 3D mapping, HVAC analytics and occupancy adaptation, digital inspections and predictive maintenance, soft landing, material efficiency, green lease, end of life and zero waste construction, retrofit and adaptation for life-span extension. The report indicated the key benefits of BIM for real estate managers such as ongoing data analysis of a project, efficient and effective building maintenance, and for property owners such as meeting certification requirements (BREEAM, LEED), cost appraisal of building materials, and 3D mapping that mitigates unforeseen site risks.

As the investors' expectations and demands including sustainable value or long-term value of property valuations are growing, the benefits and opportunities of BIM for property valuation has gradually captured the attention of researchers and the valuation professionals. In the next section, a critical review of BIM fundamentals and its applications in information exchange between AEC projects and specific business tasks such as property valuation will be described.

2.2 BIM

2.2.1 Introduction

Building Information Modelling (BIM), defined as shared digital representation of physical and functional characteristics of any built object that forms a reliable basis for decisions, has been developed by a great number of researchers and industrial professionals for lifecycle project information exchange and management in the Architecture, Engineering, Construction and Facility Management domain (Eastman et al. 2011; ISO 2017). The BIM concept can be dated back to 1975, when it was initially called a '*Building Description System'* (Eastman 1975). The current term BIM was first used by AutoCAD (Bazjanac 2004), and later gained widespread use in the AEC industry. BIM is expected to bring added social, economic and environmental value through information modelling and management, collaboration and integration to the construction industry, which has a fragmented nature that contributes to poor communication and work efficiency. For the last two decades, there is a large amount of BIM research in various domains which involved in construction planning, heritage and historical documentation, visualization, quality control, cost estimation, energy analysis, facility management, project management, and structure damage inspection etc (Mahamadu, A.M. et al. 2014). A set of common BIM features was concluded by Vanlande et al. (2008) such as the ability to store, share and exchange data, the capability to define building information in 3D dimensions, the extensible ability to cover unimplemented information domains, lifecycle information management and the ability to cover all physical and functional features of a building.

A wide range of current and promising benefits associated with BIM have been concluded by researchers. For instance, according to Lindblad (2013), the benefits of BIM adoption involves more efficient data exchange, less data input and transfer errors, increased productivity, streamlined construction processes, automated workflow, and improved product quality and building performance. The current benefits of BIM application has been concluded: more intelligent and interoperable than traditional CAD (technical benefits), capturing comprehensive information from different domains such as COBie for facility management data integration (knowledge management benefits), systematising interoperability among AEC such as IFC data exchange standards (standardization benefits), energy management and monitoring (diversity management benefits), a collaborative platform for different stakeholders involved in a project lifecycle (integration benefits), high return on investment (economic benefits), BIM 4D scheduling that all materials and components can be ordered electronically and delivered on site just in time (planning and scheduling benefits), being applied in activities from early conceptual design stages to demolition (building LCA benefits) and immediate and accurate information governance for specific business tasks (decision support benefits) (Ghaffarianhoseini et al. 2017).

Due to its many advantages, BIM has already been studied by researchers and implemented in different AEC projects all over the world. Three main stages of global BIM research during 2004-2019 have been concluded: formulating stage, accelerating stage and transforming stage (Liu et al. 2019). At the formulating stage, research themes were focus on three aspects: research review in BIM, conceptual BIM framework and building BIM capability, which formed the knowledge base for BIM research during the next stage. At the accelerating stage, research themes were focused on the integration of BIM and other related technologies such as GIS, energy calculation, rule-based checking and the development of BIM application. At the transforming stage, research themes focused on transforming the AEC industry with emerging technologies over the building life cycle such as cloud computing, smart buildings, laser scanning, IoT, digital twins, big data analysis, blockchain, drones and robots (Hofmann and Rüsch 2017; Boton and Forgues 2020). With regard to country-wise BIM research, the highest number of publications is in the UK with most of the research conducted between 2013-2019, followed by South Korea, Australia and China. The highest number of publications is recorded in 2014 and 2015 (Shehzad et al. 2020). Research on BIM covers a variety of hot topics such as mobile and cloud computing, laser scan, augmented reality, ontology, data and knowledge mining, safety rule and code checking, framework establishment, semantic web technology, partial BIM model extraction, building design, facility management, construction supply chain management, and automated generation (Zhao 2017; Liu et al. 2019; Yan et al. 2020).

BIM adoption was described as '*the successful implementation whereby an organization, following a readiness phase, crosses the 'Point of Adoption' into one of the BIM capability stages, namely, modelling, collaboration and integration*' (Succar and Kassem 2015). The use of BIM has increased significantly over the last decade in the construction industry. For instance, according to the 10th annual BIM report published by the National Building Specification (NBS), the percentage of industry using BIM rose from 13% in 2011 to 73% in 2020, which is also an increase on 2019 and marks the highest level of BIM adoption in 2020 (NBS 2020). Similarly, the BIM adoption in North America had risen from 49% in 2009 to 71% in 2012 (McGraw Hill Construction 2012). BIM adoption is encouraged by a big number of governments all over the world. The United States is one of the pioneers in BIM implementation in the construction industry. In 2003, the General Services Administration (GSA) established the 'National 3D-4D-BIM program' which aimed for gradually implementing 3D, 4D and BIM technologies for all major public projects (GSA 2003). Subsequently, the BIM Guide series had include spatial program validation, 3D imaging, 4D phasing, energy performance, circulation and security and facility management (GSA 2022). In the UK, a BIM maturity model was developed to indicate the level and depth of BIM adoption, categorizing different types of technical and collaborative working from level 0 to level 3 (BIM industry working group 2011). Furthermore, the CIC BIM 2050 Group developed a forecast roadmap indicating potential prospects of BIM and socio-technological frontiers, where key technologies were associated with the levels of BIM maturity across a timeline (CIC BIM 2050 group 2014). Similar strategies and guidelines have been designed in Norway, Finland, Denmark, Netherlands, France, Singapore, South Korea and Hong Kong. In 2010, the BCA in Singapore launched the world's first BIM e-submission system of architectural model for regulatory approval (Thant 2014). Over 90% of proportion of those who have adopted BIM expected that BIM will form the basis of all large construction projects before 2025 (NBS 2020).

There are also challenges and barriers to BIM adoption. The BIM adoption issues listed as habits of 2D-based work, limited higher education BIM training, the possible reluctance of specialists to holistic planning approaches, lack of fee structures for BIM-specific services, and inconsistency among countries regarding the acceptance and adoption of technologies (Herr and Fischer 2019). The barriers to BIM adoption in the construction industry were summarized as follows: high initial cost, BIM benefits not outweighing the implementation costs, inadequate training on the use of BIM, lack of BIM experts, data ownership issues, longer process, complexity of the BIM model, interoperability between software programs, and lack of standardized tools and protocols (Ullah et al. 2019). To get deep understanding of BIM adoption issues, multiple sources of data collection should be used such as data sources from environmental factors, the perceptions of technology adopters, cross-cultural studies, economic factors and joint BIM implementations with Green building, clouding computing, IoT and Data Science (Shehzad et al. 2020).

2.2.2 BIM and information exchange

The previous section reviewed the history of BIM and its definitions, benefits, adoption and barriers. The concept of Building Information Modelling is based on the consistent use of a comprehensive digital building model as a basis for all data exchange operations, which avoids re-entering data manually and reduces the accompanying information error and missing (Borrmann et al. 2018). One of the key critical success factors for implementing BIM is enhancing information exchange and knowledge management (Antwi-Afari et al. 2018). Therefore, this section focuses on BIM and interoperability.

Interoperability is described as '*the ability to pass data between applications, and for multiple applications to jointly contribute to the work at hand*' (Eastman et al. 2011). A variety of BIM authoring tools have been employed by different software companies, governments and other stakeholders, which limits the interoperability in AEC. The interoperability issues associated with BIM can be viewed from the technical perspective (conversion, IFC, IDM, MVD) and the collaborative perspective (exchange, share, cooperation, coordination, framework, collaboration, integration) (Sattler et al. 2019). To address the building-related interoperability issues, a common data format or structure for information transfer is required in AEC industry. As a result, IFC is becoming the most commonly used data exchange format for open BIM. The Industry Foundation Classes (IFC), contains geometric information and semantic information, is firstly developed by buildingSMART in 1997 as a non-proprietary exchange format of building information to facilitate data sharing and exchange across IFC-compatible applications (Volker 2011). The summary of IFC releases in history are illustrated in Figure 2-1. The version 1.5.1 was the first to be implemented in construction software applications, after which the IFC schema is constantly evolving with a new version released every couple of years (Laakso and Kiviniemi 2012). There were more than 160 implementations of IFC in different software tools, with most of them supporting the version IFC 2×3, but this is gradually being replaced by IFC 4 (buildingSMART 2013). After IFC4 was released by ISO 16739-21 in 2013, it was gradually accepted as a standardized data format to support building information modelling, information exchange and a variety of analysis based on BIM models such as quantity-take off, cost estimating, damage inspection, energy simulation (Volk et al. 2014).

The IFC data model is hierarchical, object orientated, and it has a number of sub schemas that representing all entities, attributes and relationships of building objects. The IFC architecture has four layers: the resource layer, the core layer, the interoperability layer and the domain layer (buildingSMART 2017a). The resource layer is the lowest layer that contains concepts representing basic geometric elements, topology, materials, cost, measure, date time, quantity, actor etc. The classes in this layer are not derived from *IfcRoot* and therefore do not include a globally unique identifier. Unlike entities in other layers, they cannot be used independently but have to be referenced by an object declared at a higher layer. The core layer is the upper layer of the resource layer, which consists of the kernel schema and three core extension schemas. This layer contains the most elementary classes that can be referenced by the upper layers, such as *IfcRoot, IfcObject, IfcActor, IfcProduct, IfcRelationship, IfcProject* and *IfcProcess* in the Kernal schema. The interoperability layer, lying between the basic core layer and domain specific schemas, contains common concepts that are typically used for inter-domain exchange and sharing of construction information such as *IfcWall, IfcBeam, IfcSlab* and *IfcWindow.* The domain layer is the highest layer that contains highly specialized classes that only apply to a specific domain. Classes in this layer such as domains for architecture, HVAC, structural analysis and construction management cannot be referenced by another layer or another domain-specific layer (Borrmann et al. 2018).

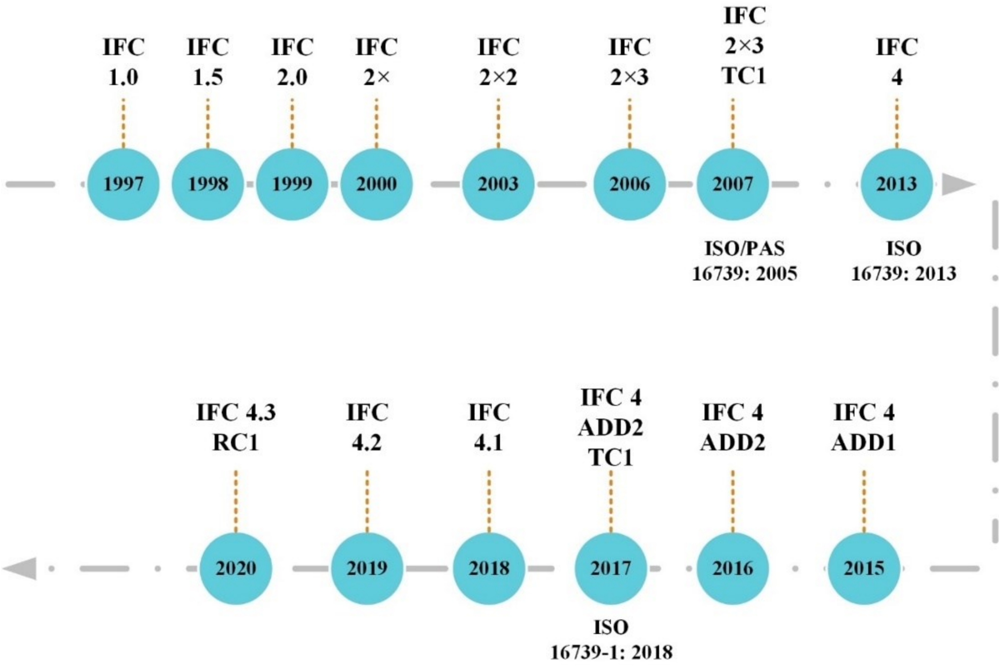


Figure 2- 1: The summary of IFC releases in history

The IFC data model is designed as a large and complex schema that can comprehensively store all aspects of information in AEC industry, which makes it heavy and inefficient to implement the complete model in different software applications (buildingSMART International User Group 2012). Typically, a specific business task only requires a partial IFC instance model. To facilitate information exchange among different stakeholders, the information delivery manual (IDM) is defined at project levels to formally specify the user requirements and ensure that the final model would be sufficiently semantically meaningful to provide most of the information needed for specific business processes (buildingSMART 2017b). Based on the IDM, a Model View Definition (MVD) was then defined as information concepts needed and proposed as a binding to the IFC standard for exchange of BIM models (Sacks et al. 2018). Basically, IDM defines information exchange process between two software packages that contains the use cases, process map and exchange requirements and the MVD is the technical implementation of the exchange requirements in the form of a subset of the overall IFC data schema. In this way, the complete IFC instance model can be filtered with reduced sized according to an MVD for specific business process (Tang et al. 2020). For instance, an official MVD named COBie (Construction Operation Building Information Exchange) has been released by buildingSMART for capturing building construction handover information required by facility managers (BSI 2014).

Examples of research and projects to develop IFC extensions in the AEC industry were published in the literature. For instance, IFC extensions for construction cost estimating for tendering in China was developed that included seven aspects of information: the building products information, the division-items project information, the cost item information, the schedule information, the quantity information, the resource information and the price information (Zhiliang et al. 2011). IFC for design change management has been proposed to deal with changes in different BIM models from conception to completion. The prototype system implemented using the .NET framework and linked into Revit demonstrated that managing changings through the extended IFC schema can improve collaborative design (Jaly-zada et al. 2015). While the IFC standard was mainly focused on buildings, in response to the urgent demand of international infrastructure stakeholders, a substantial extension of the standard to support infrastructure facilities is being carried out. Based on the principles specified by the IFC Infra Overall Architecture project, *IfcBridge, IfcRail, IfcRoad* and *IfcTunnel* have or will be initiated (Borrmann et al. 2019). Besides, IFC for physical damages has been proposed to support data exchange between different actors in bridge inspection and maintenance (Artus and Koch 2021). While these research efforts have contributed to the knowledge development of IFC extensions for specific domains, little has focused on extending the IFC schema to support property valuation.

2.2.3 The integration of BIM and other digital technologies

While there is still no agreement on a universal definition for *Industry 4.0*, it was described as ‘a shift in the manufacturing logic towards an increasingly decentralized, self-regulating approach of value creation, enabled by concepts and technologies such as

CPS, Internet of Things (IoT), Internet of Services (IoS), cloud computing or additive manufacturing and smart factories' (Hofmann and Rüsch 2017). Embracing the current trend of automation and data exchange in manufacturing technologies, the digital transformation under *Industry 4.0* is radically transforming industry and production value chains, with the aim of achieving efficiency, cost reduction and productivity increases through automation, integration and computer-supported collaborative working (Villani et al. 2018). The term *Construction 4.0*, derived from the broader concept - *Industry 4.0*, mainly focuses on the application of computer and cyber-physical systems (CPS) technologies in construction industry (Boyes et al. 2018). It requires the construction industry to transform towards the 4th industrial revolution that involves digitization of the construction industry and industrialization of construction processes (Craveiro et al. 2019; Forcael et al. 2020).

While construction professionals are showing increased awareness and willingness to embrace this digital revolution, there are still challenges such as the complexity and heterogeneous feature of construction projects, the uncertainty over tangible and intangible constraints in the individual projects, a highly fragmented supply chain and low-efficient information exchange process caused by the isolated information 'island' among different stakeholders and participants (Noran et al. 2020). Digital technologies (shown in Figure 2-2) such as IoT, cloud computing, 3D scanning and augmented reality, BIM and Machine learning play a significant part in addressing these issues. Leveraging fusion of various technologies can provide significant improvement in the productivity and profitability, which construction industry seeks the most (Rastogi 2017).

For example, the integration of BIM and cloud computing has been recognized as the second generation of building information management development and is expected to achieve a greater level of digitalization and collaboration. First, cloud-BIM data can be accessed using various mobile devices such as laptops, tablets and smartphones anytime and anywhere, enabling timely access to updated information, improving decision making and ensuring project delivery (Matthews et al. 2015). Second, cloud computing and BIM technologies, along with data stored on the Cloud, provides a real-time and collaborative environment for various project stakeholders from different locations (Birje et al. 2017). The integration of BIM and augmented reality (AR) is extremely helpful for supporting complex construction tasks and facilitating decision making. For instance, the merge of BIM and AR could provide a vivid presentation of geometric information for operational and managerial tasks, allowing one to visualize how the design fits on-site before construction takes place, managing conflicts and checking safety problems during construction (Craveiro et al. 2019; Chen and Xue 2020). Besides, the integrated BIM and AR could also provide non-geometric information (material information, rigging orders, construction schedules) on tasks and relevant building components, and therefore enhance the quality of construction works (Chen et al. 2016; Ratajczak et al. 2019).

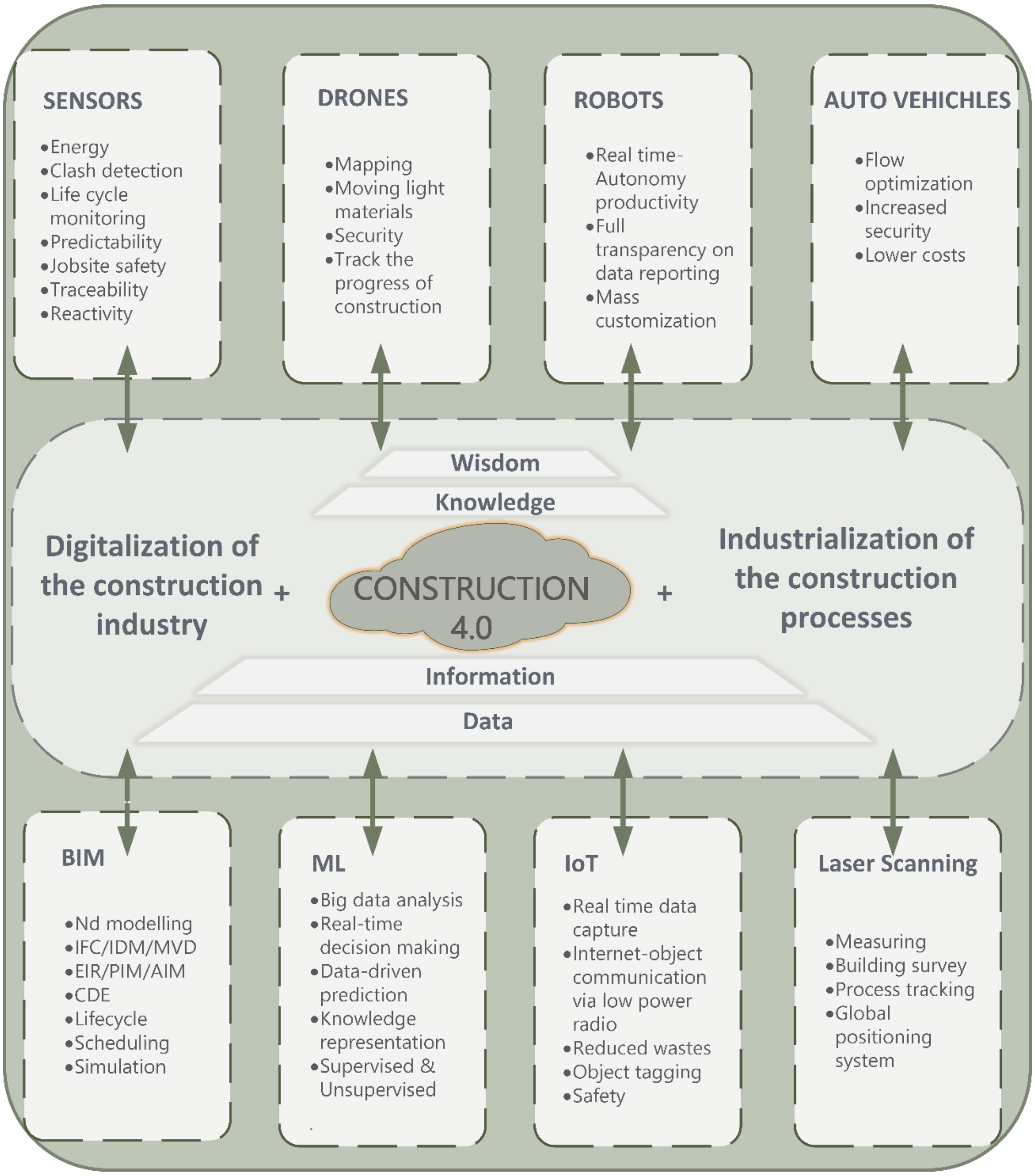


Figure 2- 2: Digital technologies support Construction 4.0

Cyber physical systems (CPS), which provide close communication and interaction between cyber and physical components, is expected to play an important part in the design and development of *Construction 4.0*. According to Boyes (2017), cyber physical systems (CPS) is defined as 'A system comprising a set of interacting physical and digital components, which may be centralized or distributed, that provides a combination of sensing, control, computation and networking functions, to influence outcomes in the real world through physical processes'. While both the conventional information and communications system (ICT) and CPS focus on processing data between digital and physical components, CPS pay particular attention on the control of physical process to produce positive outcomes (Boyes et al. 2018). BIM, as a mutual channel for information exchange among operators during the lifecycle of a building, can serve as a powerful complement to CPS that support the increased digitalization requirements in CPS (Bonci et al. 2019). The importance and benefits of integrating BIM and CPS have been stressed by several authors. For instance, Ying et al. (2020) proposed a cyber-physical based intelligent structural disaster prevention system based on BIM platform, in which BIM and IoT technologies are adopted for constructing the cloud architecture of CPS. With the development of 5G technology in terms of stable, reliable, real-time and secured network communication, the proposed comprehensive system integration is expected to achieve a high degree of integration of monitoring, identification and control, and therefore, improve the real-time and accurate intelligent monitoring of structural disaster prevention. Bonci et al. (2019) proposed a BIM and CPS integration for automatic building efficiency monitoring, in which digital models developed as BIM serves as the mirror of the physical system and stores the actual performance recorded by the building during operation phase. As a result, BIM works as the repository of information over several phases of the building lifecycle and keeps the facility at high performance levels and support decision making.

The emergence of IoT, Cloud computing, BIM and CPS bring an exponential increase of data (structured or unstructured) in the construction industry, including texts, geometrics, images, videos and sounds, which needs to be processed by Big data technology (Bilal et al. 2016). With the technological advancements focusing on information modelling, the advent of big data era has encouraged a large amount of research on big data applications in construction industry. According to Yan et al. (2020), the application of big data in construction industry has drawn international attention of both academics and practitioners. Innovation use of data and big data analytical methods have played a significant positive effect on knowledge discovery in construction by automatically discover hidden knowledge from big data repositories.

Typical data mining methods or algorithms for prediction tasks are NN-based (ANN, RNN, LSTM), regression (SVM, MLR), DT-based (boosting tree, decision tree and random forest), and deep learning (CNN, DBM-SoftMax). Of which, most studies focused on one or more mature data mining methods, while only small group of studies developed improved methods by combining several methods and ensemble models generally produce better performance than individual methods (Yan et al. 2020). For example, Cheng et al. (2015) designed a multilevel Apriori algorithm based on genetic algorithms, which combines the two algorithms to extract the association rules of construction defects. Yu and Lin (2006) introduced a variable-attribute fuzzy adaptive logic control network (VaFALCON) to solve issues of mining incomplete construction data. Data mining (DM) applications in construction industry are classified into 9 main fields such as building energy consumption, cost estimation, safety management, building design, framework establishment and others. In the 119 selected articles, DM applications on building energy is ranked at the top with 33 articles related, but only 9 are related to framework establishment. However, there is no framework establishment research related to the integration of BIM and machine learning. As the volume of data in BIM is rising exponentially, data analytics concepts and tools integrated BIM might bring added value and produce revolutionary influence on industrial practices, but it exists a significant gap in the property valuation field.

2.2.4 BIM for property valuation

While it is still in the embryonic stage, a dozen of studies have concluded that BIM can bring added value for property valuation. Smith and Tardif (2009) anticipated that the accurate lifecycle building information related to building indoor environmental quality and outdoor environmental quality within BIM environment would change the metrics for property valuation. El-gohary (2010) concluded that the integration of axiology system and BIM can facilitate the automation of the value analysis process, especially when assessing sustainability related features in a property development. (Mahdjoubi et al. 2013) studied the integration of 3D laser scanning and BIM for real estate services sector to deliver more accurate, faster and quality building surveys and information models. The research concluded that the integration of these two technologies will benefit all stakeholders in the real estate sectors including property developers, sellers, homebuyers and homeowners, when BIM adoption and 3D laser scanning become widespread. Isikdag et al. (2015) explored the utilization of 3D building models and 3D cadastre geometries ('streetview' images) for providing improved information about quality of the buildings and their surrounding environment. Wilkinson and Jupp (2016) argued that the established role for BIM in managing information in AEC projects can be extended to property professionals. For instance, using BIM data and simulation, clients can be advised of the social, economic and environmental costs and benefits of various options, which can help them make more informed decisions or consider the impact on property values. RICS (2016) explored the potential to expand BIM to property professionals and established the first BIM Manager Certification for some aspects in AEC projects may be transferable to a property-focused certification. Yu and Liu (2016) demonstrated that the integration of BIM and 3D GIS could improve the accuracy of property valuation.

El Yamani et al. (2019) developed an enhanced property valuation method based on the integration of BIM and the hedonic method. The proposed approach contained four stages: (1) extracting related building elements and compounds from an IFC-based BIM model; (2) establishing the first hedonic variable based on the building cost estimation; (3) establishing other hedonic factors based on building indoor environment such as sunlight, ventilation and noise propagation; and (4) using these hedonic variables for property valuation. However, the external environmental factors influencing property values were not included in this integration method and no case studies were provided. Arcuri et al. (2020) proposed a BIM-GIS integration framework for automated valuation models based on the cost approach and concluded that BIM could serve as an important data source including cost related information for property valuation. Celik Simsek and Uzun (2021) demonstrated that BIM based property valuation can improve the valuation accuracy in Turkish, for one biggest problem affecting the property valuation in Turkish was the miscalculation of land share values based on 2D architectural project data. However, the value factors and weights were determined via a questionnaire, which means the proposed BIM-based methodology might not be generic to be used outside the Turkish property valuation system. Yamani et al. (2021) provided a conceptual definition of the significant 3D variables for property valuation based on the BIM and CIM, in which indoor building variables were extracted from BIM models and outdoor variables were extracted from CIM models. The 3D variables were grouped into 16 categories including information related to environmental quality such as noise level, air quality and sunlight, information related to indoor living quality such as energy efficiency, indoor ventilation and indoor temperature, information related to structural variables such as property cost and property quality, information related to proximity such as distance to facilities, distance to road and distance to view. Couto et al. (2021) proposed an automatic valuation method that can calculated the taxable property value using a digital BIM model and implemented it in two case studies. The advantages of the BIM-based automatic valuation method were identified as faster calculation, interconnection with other programs, automatic calculation, the knowledge of property values at the design stage of the project, integrated and coordinated database and information, reducing errors and increased productivity. The future valuation process visioned by RICS (2017b) integrated BIM as one of the advanced digital information technologies that might contribute to interactive valuation report, machine-led building inspection and building passports. The benefits of using BIM for property valuation have attracted several researchers, while it is still at the early stage.

2.3 Machine Learning

2.3.1 Introduction

In the field of big data analytics, there is no clear boundaries between Artificial Intelligence (AI), statistics, Data Mining (DM) and Machine learning (ML). All these fields are interconnected. A machine learning system typically has three major components – data, models, and learning. The core of the process is to fit data to a model and train a function approximation algorithm (Hypothesis) based on certain performance criteria. According to (Mitchell 1997), the basic design of a machine learning system can be classified into 4 main steps: (1) choosing the training experience; (2) choosing the target function; (3) choosing a representation for the target function; and (4) choosing a function approximation algorithm. The typical ML process is illustrated in Figure 2-3, in which the training data provides the training experience that the ML system will learn from. The model performance referring to the target function that determining exactly what type of knowledge will be learned and how this will be used by the performance program. After the definition of target function (Model performance), a model representation (Learning algorithm) will be proposed to describe the target function. Finally, a function approximation algorithm (Hypothesis) will be learned from the training examples based on a specific performance criterion.

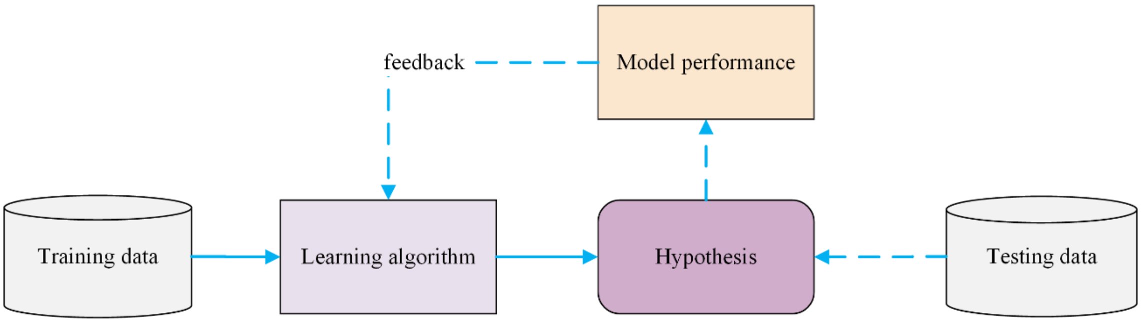


Figure 2- 3: A typical Machine Learning process

In mathematics, data is referred as vectors that can be read by computers and represented adequately in a numerical format. Model is referred as a mathematical expression (functions or probability distribution) with a set of parameters to be determined. The goal of learning is to find the best combination of the parameters with which a model will perform well on unseen data (Kumar et al. 2020). How to fit data to a model is a curve fitting mathematical problem that a curve fits a set of points. This process normally uses a cost function as a measure of how far the prediction is from the result of the training examples. Basically, the closer the hypothesis matches the training data, the smaller value of the cost function. Typical cost functions are known as least square loss for regression problems and logistic loss for classification problems. There is a wide range of machine learning algorithms for training or learning models from data, such as linear regression, logistic regression, ensemble learning, KNN, SVM, ANN, RNN and Bayesian neural network etc. It is important to choose the right training mechanism according to the nature of the data and its relevance to the problem.

Feature engineering is a crucial step in the data mining process, which is associated with extracting and transforming features (the representation of raw data) into structured data formats that a machine can read. It focuses on feature generation and feature selection from different types of raw data (numeric, textual, image and audio data), aiming to find the right set of feature vectors that suit a training task. In most cases, the quality and accuracy of machine learning performance depends heavily on the representation of the feature vectors. Feature engineering can be a manual and time-consuming process that much effort have been paid in the design of pre-processing pipelines and data transformation (Heaton 2017). When learning optimal feature representation, it is common to manually identify all available attributes and select the relevant feature space, and sometimes this approach does not always work well (Kaul et al. 2017). As a result, automatic feature engineering has been applied to further expand feature space to improve model performance or work efficiency, although sometimes the generated feature vectors are difficult to explain (Sumonja et al. 2019). Automatic feature engineering has advantages when dealing with extremely large amounts of training data and a big number of features, especially when the task is facing less support from domain experts in practise. Some optimization algorithms such as gradient descent and genetic algorithm can automatically give a ranking of each feature with weight or provide the most relevant features according to their importance to a specific task (Krishnan and Padmavathi 2017). The ranking of the features can be used as a reference for a user to select the right combination of features.

2.3.2 Optimization methods

The selection of parameters or hyper-parameters is important for building an effective machine learning (ML) model, as it has a direct impact on the model architecture. Typical hyper-parameters configuration to achieve the best model performance involves: the learning rate to train a neural network, the number of estimators in an ensemble learning, the maximum depth of the decision tree structures or to specify an algorithm to minimize the loss function (Diaz et al. 2017). Currently, various optimization methods are developed for tuning these hyper-parameters, especially for tree-based ML models and deep neural networks, such as conventional gradient-based optimization, metaheuristic algorithms and probabilistic methods (Yang and Shami 2020).

Gradient-based optimization methods use the gradient information to define a search direction for approaching the optimal solution. For machine learning algorithms associated with convex functions, the gradient of specific hyper-parameters can be calculated and the global optimum can be reached with a fast convergence speed. However, they are only suitable for optimizing continuous hyper-parameters, since other discrete parameters do not have gradient directions (Yang and Shami 2020). Typical gradient-based methods involve gradient descent, Newton method and conjugate gradient method.

Metaheuristic algorithms can deal with non-convex, non-continuous and non-smooth optimization problems, but are usually unstable and can produce different solutions as they involve random search process (Zhao et al. 2018). Some representative metaheuristic algorithms are genetic algorithms (GA), particle swarm optimization (PSO) and evolutionary algorithms. Genetic algorithm (GA), which simulates the evolutionary process of Darwin's biological evolution theory, has been widely used for multi–parameter optimization problems and non-linearization problems (Sevinç and Coşar 2011). GA is essentially a heuristic search algorithm which typically performs the search process in four steps (as shown in Figure 2-4): (1) set initial population for real problems; (2) check the fitness criterion with each member of the population; (3) parents selection with higher fitness values; and (4) perform crossover and mutation operators to product new offspring for the solution (Wong and Tan 1994). One critical element of GA is to choose the right fitness function which defines the ability of each chromosome to solve the real problem. In this research, GA is used for searching the optimal trade-off between diversity and accuracy of GBR ensemble model, as the high model complexity and diversity usually lead to overfitting and the low one lack of accuracy.

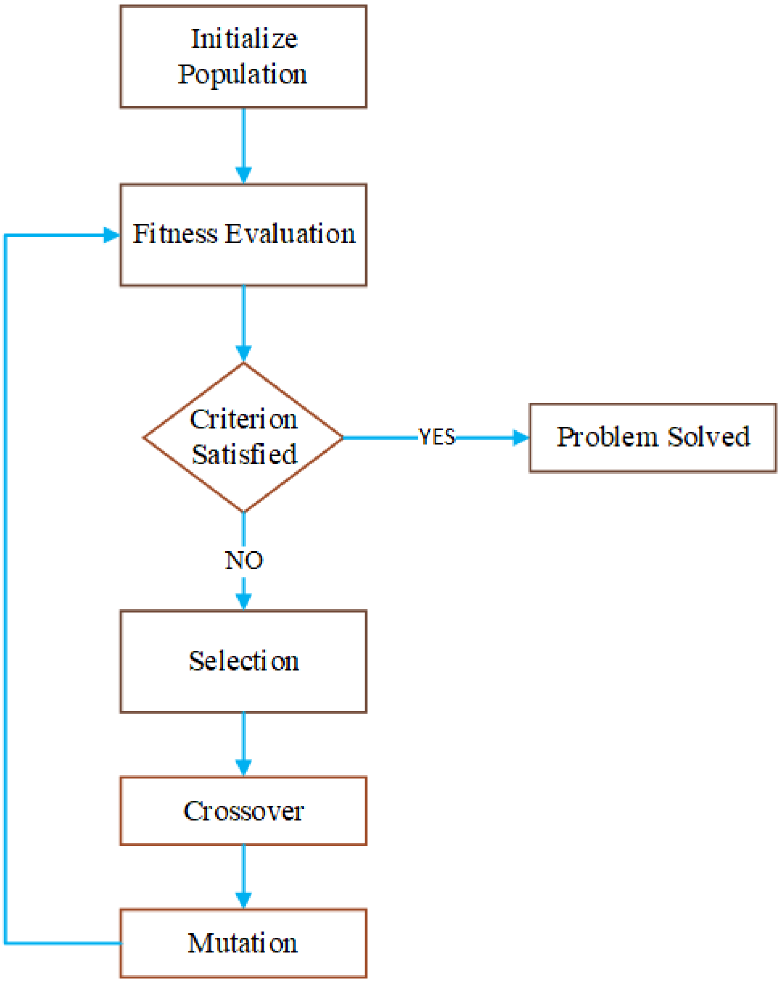


Figure 2- 4: Diagram of simple genetic algorithm

2.3.3 Automated valuation models (AVMs)

The definition of automated valuation models (AVMs) is described by IAAO as: ‘a mathematically based computer software program that produces an estimate of market value based on market analysis of location, market conditions, and real estate characteristics from information that was previously and separately collected. The distinguishing feature of an AVM is that it is an estimate of market value produced through mathematical modelling. Credibility of an AVM is dependent on the data used and the skills of the modeler producing the AVM.’ (IAAO 2003). The objective of an AVM is to provide a credible, reliable and cost-effective estimate of a subject property’s market value at a specific valuation date. AVMs have been employed in both the public and private sectors and they continue to show an increasing role in the stability of the economic and social systems (D'Amato and Kauko 2017). Vendors from UK and USA such as *Rightmove* and *Zillow* have provided some general background information on their AVMs. *Rightmove* declares that they have stringent criteria, employing a thorough filtering process in selecting the properties used in their AVMs, on the other hand, *Zillow* claims to be the largest AVM provider in the US (Łaszek et al. 2018).

Typical AVMs training procedures can be classified as five main steps: (1) Data collection and pre-processing, (2) Feature selection, (3) Hyperparameter tuning, (4) Model fit, and (5) Evaluation. Firstly, a number of comparable traded real estate cases are collected from reliable sources and verified by professional valuers. Then different types of property attributes such as size, area, garage type and built year are applied as continuous or discrete attributes. The data set available is randomly divided into two groups: 70% or 80% for the training set and 30% or 20% for the testing set. Secondly, the feature selection process is about choosing the suitable number of variables and the right combination of them, as too many variables may lead to overfitting and inadequate ones will cause the model underfitting. Besides, although the increasing number of variables makes machine learning model more accurate with the prediction, the large number of input variables also needs a huge amount of training data which make the training process impractical for the intensive computation. This is known as dimensionality curse (Rafiei and Adeli 2016). Thirdly, in order to find the best fit of the learning speed and the complexity of the patterns to be discovered behind the training data, model hyperparameters such as the number of decision trees, the learning speed and the maximum depth of each decision tree are iteratively tested. After that, the optimum group of hyperparameter setting for this dataset is used to fit the model. To get a generalized AVM, it is suggested to test the optimum hyperparameter setting with different datasets or through cross-validation of the datasets. Lastly, before the model being used for real estate price prediction, it is tested with the test dataset according to model accuracy measurements such as mean absolute percentage error (MAPE), mean squared error (MSE) and coefficient of determination (R^2).

AVMs have gradually evolved through regression analysis (MRA), hedonic modelling and artificial intelligence models, and the fundamental to AVMs are statistical, data mining and computing methodologies. While AVMs are currently used predominantly for residential property only, undergoing many developments based on AI, they are expected to be developed for many other property types and more complex valuations (RICS 2017b).

Compared to hedonic methods, machine learning models have several advantages: to efficiently assess information from big data; to identify non-linear relationships between house characters, market factors and property price; and to be more objective about the selection of input attributes (Kontrimas and Verikas 2011; Park and Bae 2015; Dimopoulos and Bakas 2019). For different types of machine learning models in property valuation, it was concluded through the comprehensive literature review in Section 2.1.4 that the individual neural network or genetic algorithm (GA) has not achieved satisfactory results, but GA optimized neural networks have achieved good performance for real estate appraisals. While ANNs has attracted more attention than other algorithms, if the aim of data analytics is to learn models from data that can be further developed as an expert system, a classification tool, a recommender system or a credit scoring system, ML will be the priority choice. With decision tree-based ensemble learning, implicit or hidden knowledge can be automatically discovered, represented and modelled for various tasks. This will be extremely helpful for human decision making when tasks involve processing and analysing big data which characterized as high-volume, high-velocity and high-variety. Compared to neural networks, ensemble learning has advantages in terms of model interpretability and flexibility.

2.3.4 Validation and evaluation strategies

After machine learning models learnt from the training data, it is necessary to evaluate their predictive accuracy and generalization capability on independent test data. It is important to use independent test data to evaluate the model accuracy and generalization, as error rate on the training set is not a good indicator of future performance (Witten et al. 2016). There are different kinds of performance measures for different tasks. The performance measurement is usually computed by comparing the prediction operated by the model and the real value, either a discrete type or a continuous numerical value. It is common to use *error rate* to measure the model performance for classification problems, whereas *mean squared error* (MSE) or *receiver operating characteristics* (ROC) is more common to be used for regression problems. According to Sokolova and Lapalme (2009), for a binary classification task, the accuracy or error rate can be assessed by calculating the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), class examples that were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). For a regression problem, the *mean square error* (MSE) for a regression estimator is calculated as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2 \tag{5}$$

where A_t is the actual value and F_t is the predicted value.

In most real applications, the model training and evaluation of predictive accuracy and generalization capability might have to be based on limited data. A general way to 'increase' the amount of data and ease the bias caused by the chosen particular sample is the data splitting strategies. There are some common methods to split the data, including the holdout method, cross-validation and bootstrapping. The general principle is to use a certain amount of data for training and reserve the remainder for testing, of which the process normally needs to be repeated several times with different random samples (Sylvain 2010). For holdout method, in each iteration a certain proportion (two-thirds or four-fifths) of the data is randomly chosen for training and the remainder is used for testing. Subsequently, the error rates of different iterations are averaged to produce an overall error rate (Witten et al. 2016).

Sometimes, models with high generalization capability tend to be overfitting, while models with low generalization capability tend to be underfitting. The bias-variance decomposition is an importance and widely used tool for understanding the generalization capability of machine learning algorithms. The bias is an error term that measures the mismatch between the model class and the underlying data distribution, whereas the variance measures sensitivity to fluctuations in the training data (Yang et al. 2020).

According to Hastie et al. (2008), the expected prediction error of a regression problem in terms of bias and variance decomposition can be expressed using squared-error loss:

$$Err(x_0) = \varepsilon^2 + Bias^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0))$$

$$= \text{Irreducible Error} + Bias^2 + \text{Variance} \qquad (6)$$

From this decomposition, the expected prediction error can be understood as the sum of noise, (squared) bias and variance. Generally, the more complex the model $\hat{f}(x)$, the lower the (squared) bias but the higher the variance.

Overfitting, as a key issue in supervised machine learning, refers to a learning algorithm fits the training data so well that noise fits in the data by memorizing various peculiarities of the training data rather than finding a set of general predictive rules. On the opposite, underfitting refers to a learning algorithm is incapable of capturing the variability of the data (Jabbar and Khan 2014). Typically, models with high bias but low variance are more likely to be underfitting, whereas models with high variance but low bias are more likely to be overfitting. There are different methods to avoid the issue of overfitting and underfitting in supervised learning, for instance, the penalty methods and early stopping for training. According to Schittenkopf et al. (1997) , typical penalty methods involve: (1) Hold and cross-validation; (2) Generalization cross-validation; (3) Minimum description length (MDL) principle. Penalty method is one of important methods to control the Bias-Variance trade-off in supervised machine learning.

2.4 Summary of Literature Findings

This chapter introduced the basic concepts used as part of the conducted work and offered an overview of research in the fields of property valuation, BIM and Machine Learning.

Section 2.1 outlined the three traditional approaches and the three popular trends in property valuation, which brings forth the first findings:

1) Traditional valuation approaches are questioned as inaccurate, inefficient and unreliable, in the last two decades, there has been a move towards the advanced valuation approaches due to their many advantages dealing with the increasing complexity of property transaction.
2) As the investors' expectations and demands including sustainable value or long-term value of property valuation are growing, the benefits and opportunities of BIM for property valuation have gradually captured the attention of researchers and the valuation professionals.

Section 2.2 outlined research on BIM and information exchange, the integration of BIM and other digital technologies, and BIM for property valuation. This brings forth the second findings:

3) The BIM related concepts including IFC-IDM-MVD have contributed to information exchange in specific domains such as construction cost estimating, design change management, and infrastructure facilities, little research has focused on the knowledge development of IFC extensions to support property valuation.
4) While research on BIM for property valuation is still at the early stage, the benefits of using BIM for property valuation have gained researchers and professionals' attention such as facilitate the automation of the value analysis process, provide improved information, improve the valuation accuracy, and make more informed decisions or consider the impact on property values. As the volume of data in BIM is rising exponentially, data analytics concepts and tools integrated BIM might bring added value and produce revolutionary influence on industrial practices, however, there is no framework establishment research related to the integration of BIM and machine learning for property valuation.

Section 2.3 introduced the AVMs and several key stages including feature engineering, optimization strategies, and validation to construct a good machine learning model, which were further used at the system development stage. This brings forth the third findings (Section 2.3.3 and Section 2.1.4):

5) Machine learning models have several advantages such as efficiently assess information from big data, identify non-linear relationships between house

characters, market factors and property price, and to be more objective about the selection of input attributes.

6) While ANNs has attracted more attention than other algorithms, compared to neural networks, ensemble learning has advantages in terms of model interpretability and flexibility, which is more suitable for knowledge mining and system development. Since genetic algorithm (GA) optimized neural networks have achieved good performance for property valuation, the integration of GA and ensemble learning might achieve good predictive performance for property valuation as well as good model interpretability.

2.5 Conclusion

The literature reviewed throughout this chapter introduced the current state of the art regarding property valuation, BIM, and Machine Learning. From which, three research gaps were identified: (1) AVMs (Automated Valuation Models) have been more widely used to deal with the increasing complexity of property valuation, but the current AVM needs to be further improved to be able to address large data sets, non-linear relations among different input parameters and the outputs, and holistic decision making; (2) The increasing volume of life cycle data and value-related design information have not been leveraged to improve the property valuation performance; (3) There is a lack of integrated framework, where BIM and AI as revolutionary technologies can be orchestrated to produce smart property valuation, which is dynamic and holistic.

Chapter 3. Research Methodology

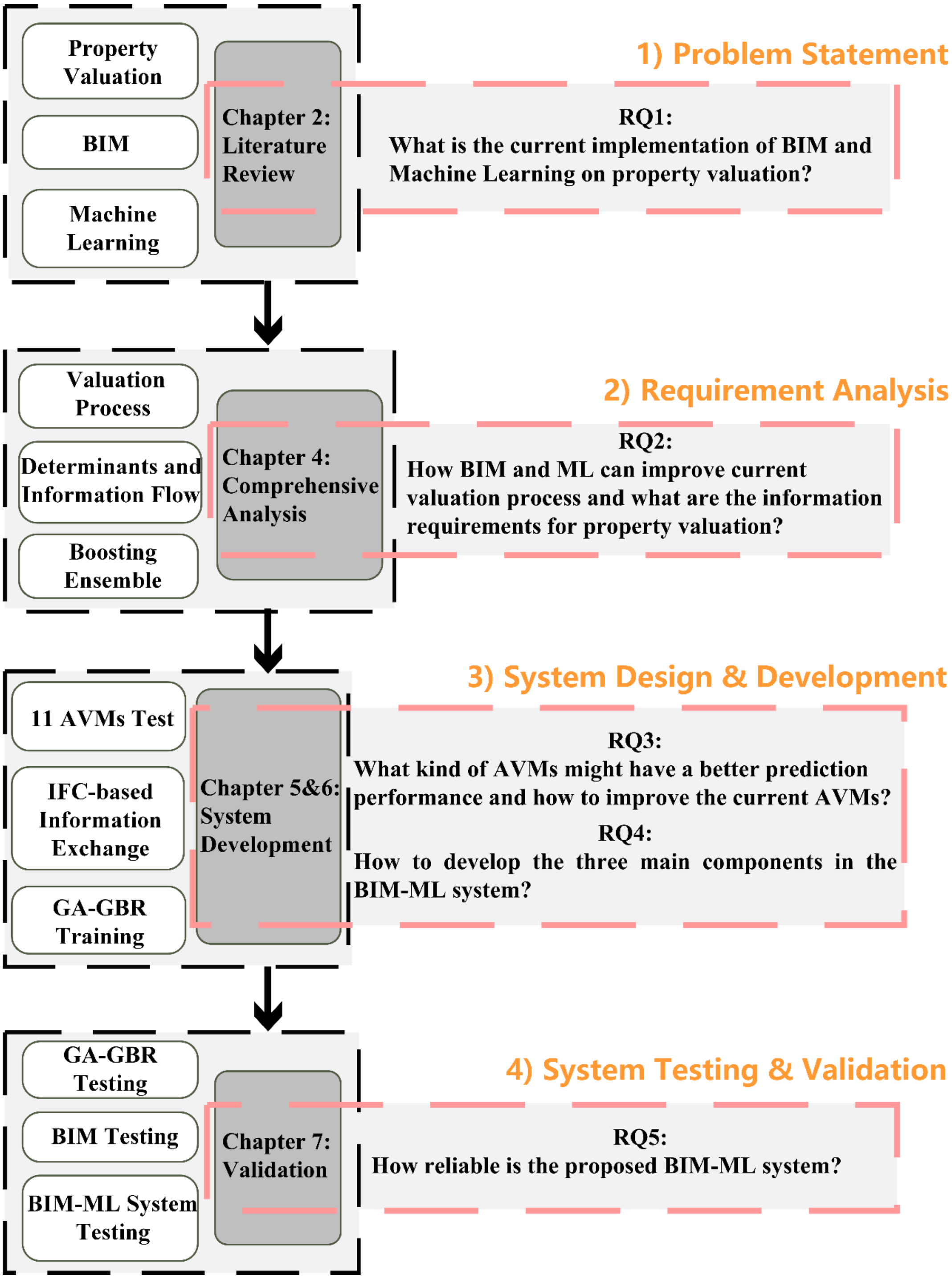
3.1 Research objectives

Based on the research findings from the literature review, an integration framework based on BIM and Machine Learning information technologies is proposed, which aims at facilitating information exchange between AEC projects and property valuation and supporting automated property valuation.

The nature of a research project can be classified as exploratory, descriptive, explanatory, evaluative or a combination of them (Saunders et al. 2015). While explanatory research concentrates on explaining the relationships between different variables of a specific issue, exploratory research focuses on studying a specific topic, problem or phenomenon through open questions (Schutt 2011). Descriptive research focuses on describing persons or situations accurately and evaluative research concentrates on assessing the efficiency of the studied object (Saunders et al. 2015). This research has a combination of explanatory, exploratory and evaluative purposes, with several objectives identified:

- **Objective 1:** Identify determinants of house price and the value-specific design information in AEC projects (explanatory nature)
- **Objective 2:** Compare different AI-enhanced AVMs for automated property valuation and improve the performance of current AVMs (exploratory nature)
- **Objective 3:** Achieve automatic information exchange between AEC projects and property valuation (exploratory nature)
- **Objective 4:** Explore whether the use of BIM and Machine Learning information technologies can promote current property valuation process and bring added value to the construction and property industry (exploratory and evaluative nature).

3.2 Design science methodology and framework



Property Valuation
BIM
Machine Learning
Chapter 2: Literature Review
1) Problem Statement
RQ1:
What is the current implementation of BIM and Machine Learning on property valuation?
Valuation Process
Determinants and Information Flow
Boosting Ensemble
Chapter 4: Comprehensive Analysis
2) Requirement Analysis
RQ2:
How BIM and ML can improve current valuation process and what are the information requirements for property valuation?
11 AVMs Test
IFC-based Information Exchange
GA-GBR Training
Chapter 5&6: System Development
3) System Design & Development
RQ3:
What kind of AVMs might have a better prediction performance and how to improve the current AVMs?
RQ4:
How to develop the three main components in the BIM-ML system?
GA-GBR Testing
BIM Testing
BIM-ML System Testing
Chapter 7: Validation
4) System Testing & Validation
RQ5:
How reliable is the proposed BIM-ML system?

Figure 3- 1: Research methodology for this thesis

Due to the fact that this research is implemented in the information technology domain, the adopted research methodology illustrated in Figure 3-1 follows the principles of design science research (DSR) methodology (Peffers et al. 2007), which can be classified into four main steps: (1) problem statement, (2) requirement analysis, (3) system design and development, and (4) system testing and validation. In this research, each step is associated with corresponding research questions which are addressed in different chapters. The research methodology for this thesis is explained as follows:

1) Problem Statement

The related problems have been introduced in the introduction and literature review chapters, which illustrated that the advantages of using AI-enhanced AVMs for property valuation have been recognized by researchers and professionals, but the predictive accuracy and model interpretability of current AVMs need to be further improved. The benefits and opportunities of BIM for property valuation have gradually captured the attention of researchers and valuation professionals, however, research on BIM for property valuation is still in its early stage.

The contents of Chapter 2 responded to the first research question (**Q1**), which is described as:

Q1: What is the current BIM and Machine Learning implementation on property valuation and What are the opportunities and challenges concerning automated property valuation and information exchange between AEC projects and property valuation?

2) Requirement Analysis

The requirement analysis is focused on possible future changes that information technologies including BIM and Machine Learning will bring to the current property valuation process, and focused on the information requirements that existed in the AEC projects which can be further developed for property valuation.

The traditional valuation process mainly involves seven steps: (1) problem identification and assignment definition; (2) data collection and verification; (3) preliminary data analysis; (4) land value estimate; (5) form opinion of property values using the three main valuation approach; and (6) reconcile values for final opinions of value. In the last two decades, the real estate market has experienced changes including the new client expectations such as a faster delivery of the valuation and a long-term property valuation, and the increasing complexity of property valuation assignments such as taking into consideration of sustainability-related features within property valuation process. With the increasing complexity of property valuation assignments, it is expected that the future valuation process will be more fragmented. The emerging advanced information technologies such as IoT, AI, Blockchain and BIM are going to transform the role of valuers and the traditional valuation process, responding to the increasing complexity of client expectations and the real estate transactions.

The contents of Chapter 4 responded to the second research question (**Q2**), which is described as:

Q2: How innovative information technologies such as BIM and Machine Learning (ML) will improve the current valuation process and what are the information requirements for property valuation?

3) System Design and Development

Based on the findings in Chapter 2 and Chapter 4, to facilitate information exchange between AEC projects and property valuation and support automated property valuation workflow, the BIM-ML system is designed and developed that includes three main components: (1) an IFC extension for property valuation that includes missing but necessary value-related entities and property sets; (2) an IFC-based information extraction for automatic information exchange between AEC projects and property valuation; and (3) an automated valuation model (GA-GBR) based on the integration of gradient boosting ensemble machine learning and genetic algorithm.

One of the main components – the proposed automated valuation model (GA-GBR) starts with an experiment with 11 different types of AI-enhanced AVMs in Chapter 5, including AVMs based on linear regression, ridge regression, lasso regression, elastic net regression, KNN, SVM, ANN, CART, AdaBoost, Random forest, and gradient boosting ensemble (GBR), with the aim at comparing the performances of different AVMs and selecting the suitable machine learning model for system development. After that, the structure of the genetic algorithm optimized gradient boosting ensemble model is explained and tested with the UCI Machine Learning repository – Boston dataset. The model performance of the proposed GA-GBR is compared with a similar house price prediction study using random forest machine learning with the same Boston housing dataset (Adetunji et al. 2022), which lays the foundation for further experiments in the next stage.

The contents of Chapter 5 responded to the third research question (**Q3**), which is described as:

Q3: What kind of automated valuation models (AVMs) might have a better prediction performance for property valuation and how to improve the current AVMs?

Another two components of the proposed system – an IFC extension for property valuation and an IFC-based information extraction are addressed in the first two subsections in Chapter 6. Based on the 62 collected value-relevant variables for property valuation in Section 4.3, an IFC extension including related building object entities and properties are developed for property valuation. It firstly discusses the property valuation related modelling capabilities of the current IFC schema and identifies the value-related design information that can be used for developing the IFC extension for property valuation. Based on the identified influential variables on property valuation, a number of missing but necessary value-specific entities and property sets are proposed to be added in the IFC extension for property valuation. After that, an IFC-based information extraction algorithm is designed to automatically extract the required value-specific design information from an IFC-based BIM instance model.

In the last subsection of Chapter 6, the proposed automated valuation model (GA-GBR), which serves as an automated valuation engine, is trained with traded property data from the Chinese and American real estate markets. Before fitting data to the GA-GBR model, the preliminary steps including data preparation, exploratory data analysis and feature engineering are performed.

The development of the three main components in Chapter 6 responded to the fourth research question (**Q4**):

Q4: How to implement the BIM-ML integration framework and how to develop the three main components accordingly?

4) System Testing and Validation

In Chapter 7, three validation tests are conducted to verify the proposed BIM-ML system, which involves:

- Validate the developed automated valuation model (GA-GBR) through different performance measures of machine learning such as MAE, MAPE, MSE, RMSE, and R^2. The GA-GBR model is tested with three datasets from different countries including China, U.S. and the UK, where the UK dataset is only used for testing. In addition, the developed GA-GBR model was implemented for six months in a commercial real estate appraisal and advisory company (HXZH) in China, comparing the model performances with the traditional valuation method. A case study was provided to demonstrate the two methods and feedback documents were provided.
- Validate the required information extraction from the extended IFC schema through three case studies from three countries including China, U.S. and the UK.
- Validate the comprehensive BIM-ML integration framework as a complete system. In addition, feedback documents upon the implementation of the BIM-ML system in the commercial company were provided.

This responds to the last research question (**Q5**), which is described as:

Q5: How reliable is the proposed BIM-ML integration framework that can facilitate information exchange and support automated property valuation?

3.3 Conclusion

The core methodology throughout this research has been outlined in this chapter. The main phases of this research were identified along with the objectives for each of them. To test the initial hypothesis, the divided five research questions were linked to the corresponding phase stating the main components of the linked chapters. The developed research methodology framework highlighted: (1) Research question 1, linked to Chapter two, described the current state of the art of BIM and Machine Learning on property valuation; (2) Research question 2, linked to Chapter 4, explained the requirements analysis for conducting property valuation; (3) Research questions 3 & 4, linked to Chapter 5 & 6, illustrated the process of system design and development including the 11 typical AVMs comparison experiment, the proposed GA-GBR model, detailed IFC development, and detailed GA-GBR model development; (4) Research question 5, linked to Chapter 7, solved the last phase of this research namely system testing and validation, including the proposed AVM (GA-GBR) validation, the IFC extraction validation, and the overall framework validation.

Chapter 4. Comprehensive Analysis for Property Valuation Process and Framework

In this section, a specific literature review upon the current state of the art of property valuation process and framework will be provided, in the meantime considering requirements and potential solutions for the proposed integration framework.

In the first subsection, the current valuation process and the future possible changes to the current valuation process will be described. The emerging advanced information technologies such as IoT, AI, Blockchain and BIM are going to transform the role of valuers and the valuation process, responding to the increasing complexity of client expectations and the real estate transactions. At the end of this subsection, the influence of AI enhanced AVMs on the current valuation process is summarized. The second subsection introduces the BIM influence on current valuation workflows and information exchange. The future information flow and data exchange in property valuation, associated with BIM and other information technologies, is envisaged which involves five layers including the physical layer, the technical layer, the pre-processing layer, the information storage layer, and the domain layer. The third subsection describes the comprehensive determinants collected for property valuation from archival research of 174 research documents including research paper, projects and industrial valuation standards. The identified variables related to BIM concepts have been classified as six different types and 28 subtypes of information related to property valuation. The last subsection introduces the fundamentals of the bagging ensemble and the boosting ensemble. Subsequentially, in the fourth subsection the gradient boosting ensemble learning model is specified, which is going to be developed and optimized as the AVM at the system development stage.

4.1 Property Valuation Process

4.1.1 The current valuation process

There are three main property valuation approaches all over the world: market approach, cost approach and income approach. In deriving a final value estimation of a property, the appraiser will use one or the combination of the three valuation approaches, which may be determined by the type of the subject property and the factors of greatest importance to clients (Ventolo 2015). For instance, a rental house normally will be appraised using the sales comparison method that existing sales data are compared to the subject property, whereas an office building valuation typically will choose the income method such as discounted cash flow (DCF) which calculates the final value based on the revenues and cost of the office.

Figure 4-1 illustrates the current valuation process that involves seven steps from the appraisal assignment to the final value estimation report (Bienert et al. 2009; Ventolo 2015; RICS 2017a). The detail of each step is explained as follows:

1) **Problem Identification and Assignment Definition:** It focuses on defining the scope of work, such as the type and extent of the valuation needs to perform. This step is associated with several assignments: identification of the asset and location of the asset, identification of property rights, definition of value to be estimated, purpose and intended use of the valuation, valuation data and any other special limited conditions that the client and the appraiser need to be informed.
2) **Data Collection and Verification:** Once an appraisal assignment accepted, the appraiser will choose the suitable valuation approach based on the requests of client and the type of the asset. From that point, general data upon the surrounding area, specific data about the asset and data applicable to specific valuation approaches require to be gathered and verified. General data is associated with the geometric and economic features of the nation, region, city and neighborhood. Specific data includes a detailed physical description of the asset and the property site. Data for each approach has different requirements such as the sales comparison method requires the collection of sales data on comparable properties,

while income valuation approach requires income and expense data from the property's history.

3) **Preliminary Data Analysis:** The appraiser will conduct a highest and best use analysis of the relevant data, considering the influence of the market components on the value of the subject asset. The subject property's highest and best use should not only satisfy the human needs, but also be revealed by social and economic market indicators such as supply and demand, inflation expectations, vacancy rate, purchasing power and demographic structure and development.
4) **Land Value Estimate:** Based on the physical features and amenities of the subject site (except for the asset), an opinion of land value is estimated.
5) **Form Opinion of Value Using the Three approached:** Based on different valuation approaches, an opinion of the subject property value is estimated using relevant mathematic functions and techniques. For instance, in the sales comparison method, the appraiser makes adjustment by comparing any significant differences between the comparable property and the subject property and gives the estimated price of the subject property.
6) **Reconcile Values for Final Opinions of Value:** The appraiser may use more than one valuation approaches for the subject asset. The final opinions of the estimated value will be calculated by the weighted average of the estimated values from different methods. The most relevant approach normally should be given the greatest weight in determining the final opinions of value.
7) **Report Final Value Estimation:** Finally, the client will be presented with the conclusion of the final value in the reporting form that complying with specific valuation standards. Basically, the report content will cover all the information related to the assignment definition, data collection and analysis, any assumptions and special assumptions, valuation approach and reasoning, date of the report and commentary on any information uncertainty in relation to the valuation process (RICS 2017a).

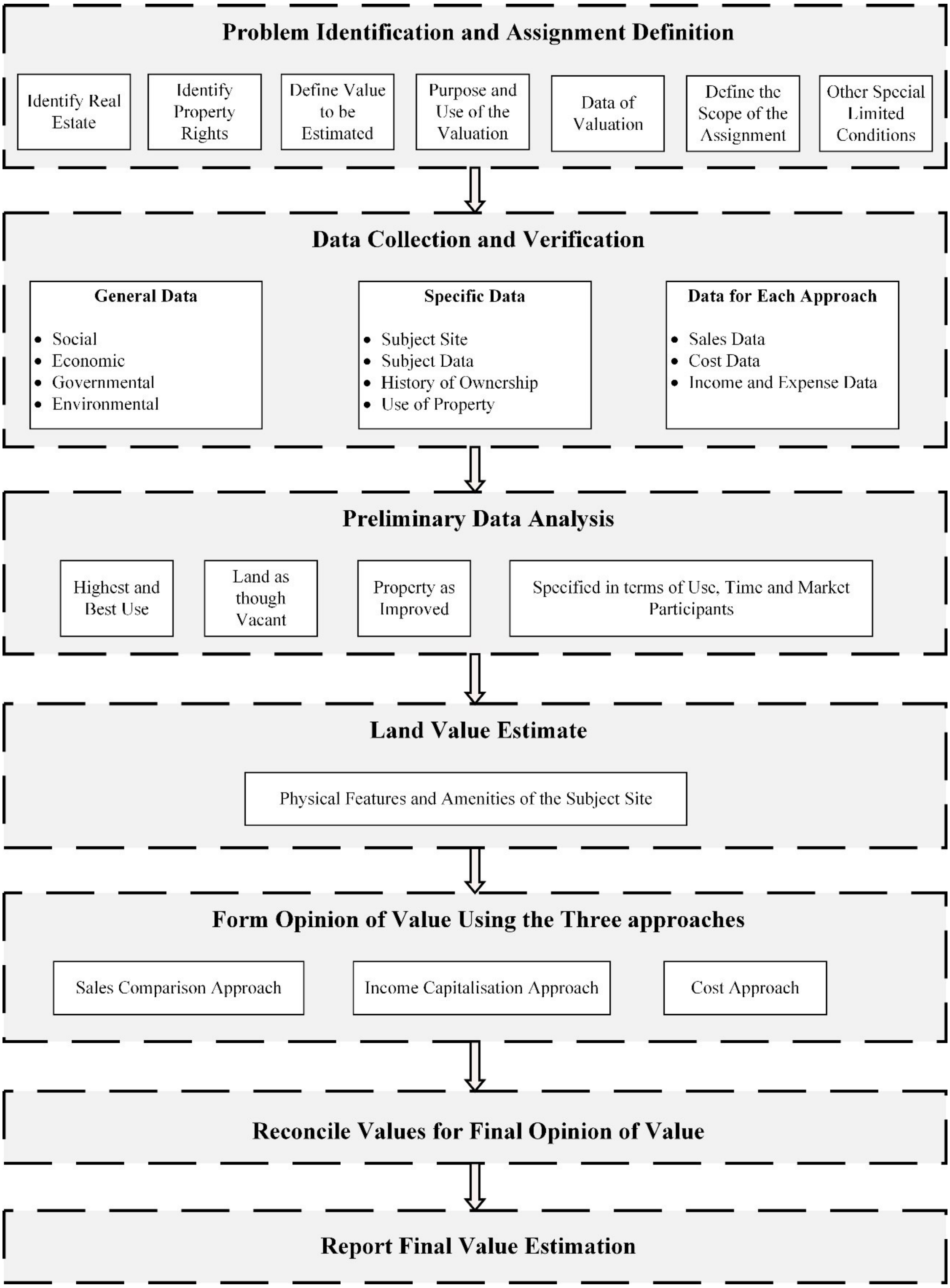

Problem Identification and Assignment Definition
Identify Real Estate
Identify Property Rights
Define Value to be Estimated
Purpose and Use of the Valuation
Data of Valuation
Define the Scope of the Assignment
Other Special Limited Conditions
Data Collection and Verification
General Data
Social
Economic
Governmental
Environmental
Specific Data
Subject Site
Subject Data
History of Ownership
Use of Property
Data for Each Approach
Sales Data
Cost Data
Income and Expense Data
Preliminary Data Analysis
Highest and Best Use
Land as though Vacant
Property as Improved
Specified in terms of Use, Time and Market Participants
Land Value Estimate
Physical Features and Amenities of the Subject Site
Form Opinion of Value Using the Three approaches
Sales Comparison Approach
Income Capitalisation Approach
Cost Approach
Reconcile Values for Final Opinion of Value
Report Final Value Estimation

Figure 4- 1: The current valuation process

4.1.2 The future valuation process

In the last two decades, property valuation has experienced changes from traditional valuation methods and individual valuers' subjective assessment to computer-aided valuation approaches such as big data and artificial intelligence (AI) in property valuation. With the advanced in the availability of computer technology and information system, automated valuation models (AVMs) are evolving quickly because of their advantages that they are systematic, fast and less dependent on human subjective judgements (Łaszek et al. 2018). The technological developments of big data, blockchain, AI and AVMs in part reflects the changing client expectations such as a faster delivery of the valuation and in part reflects the increasing complexity of many real estate transactions such as the added value comes from sustainability features. As a result, the role of the valuer and the valuation process are more likely to face a period of significant changes in coming years (RICS 2017b).

With the increasing complexity of property valuation assignments, it is expected that the future valuation process will be more fragmented. The future valuation process is visioned that the valuer may only need to work on parts of the current valuation process, with other parts carried out by other methods such as automation. In Figure 4-2 below, a future valuation process is visioned as four main steps as follows (RICS 2017b):

1) **Pre-valuation:** The problem identification and assignment definition process (**Step 1** in Figure 4-1) will be automated and facilitated by smart contracts (blockchain-based systems) which enforce the pre-valuation negotiation process automatically through the use of digital information communication technologies and related computer protocols and standards. This will save time and costs.
2) **Data Collection and Verification:** The traditional data collection and verification process (**Step 2** in Figure 4-1) may be partially or entirely replaced by the emerging information modelling and communication technologies such as 3D scanning, drones, image streaming, smart buildings and Internet of Things (IoT) and building passports based on BIM and blockchain. The future valuation process will benefit from near real-time availability of data. However, it is

expected that verification of data will remain a task to be carried out by a professional valuer or a junior data scientist with skills in statistics and analytics.

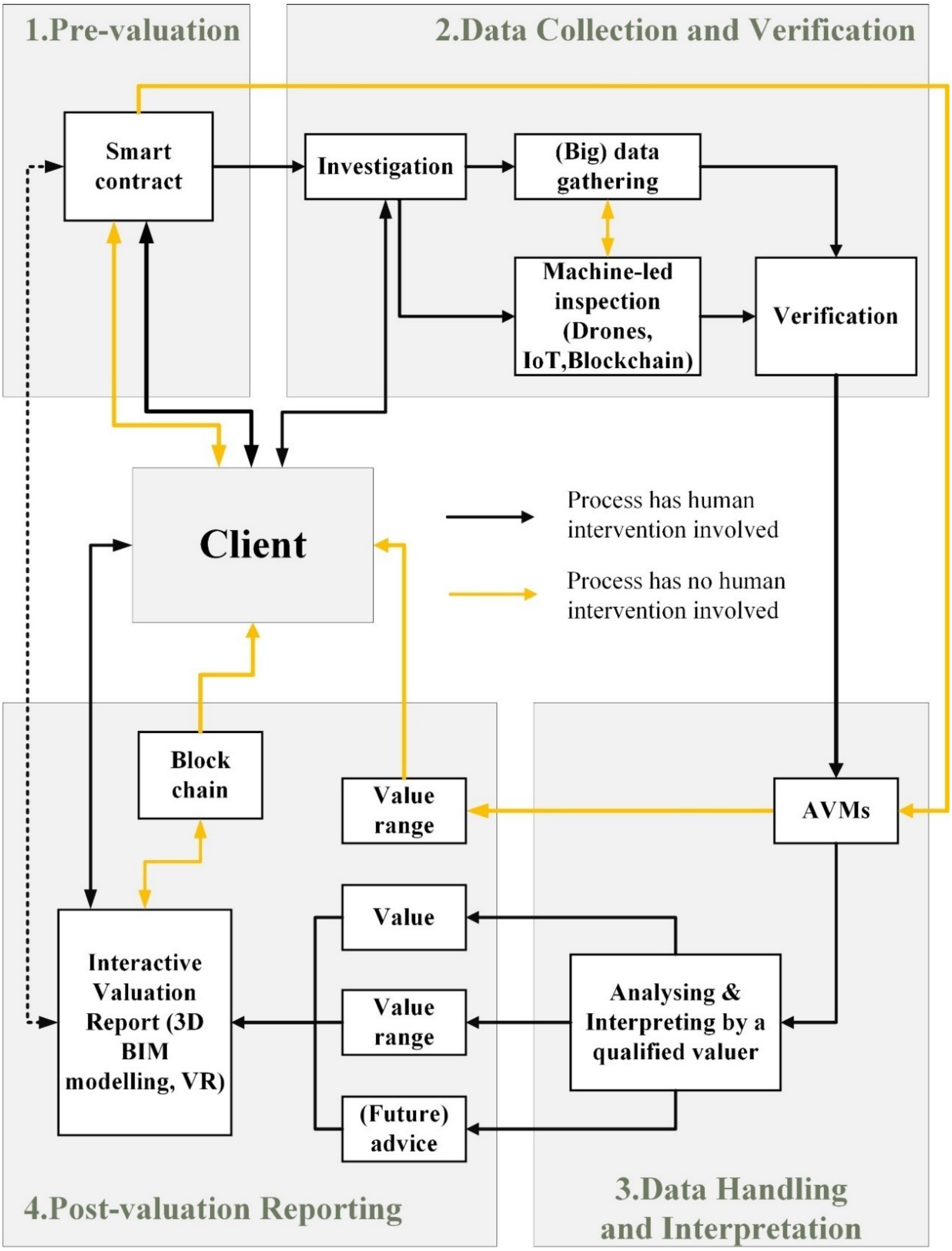

1.Pre-valuation
2.Data Collection and Verification
Smart contract
Investigation
(Big) data gathering
Machine-led inspection (Drones, IoT,Blockchain)
Verification
Client
Process has human intervention involved
Process has no human intervention involved
Block chain
Value range
AVMs
Value
Value range
(Future) advice
Interactive Valuation Report (3D BIM modelling, VR)
Analysing & Interpreting by a qualified valuer
4.Post-valuation Reporting
3.Data Handling and Interpretation

Figure 4- 2: The vision of future valuation process according to (RICS 2017b)

3) **Data Handling and Interpretation:** From the preliminary data analysis to generate final opinions of value processes (**Step 3-6** in Figure 4-1) will be carried out, replaced or assisted by AI enhanced AVMS. The advancement in big data and AI technologies will move AVMs from low-risk valuations with sufficient comparable transactions (automated residential property valuation) towards more complex valuations for all property types. The valuation process of certain low-risk tasks will remain including the client, the smart contract and an AVM, whereas the process of more complex valuations might still follow the entire process. Statistical property valuation techniques such as AI and AVMs are going to play a significant role in the data analysis and interpretation stage. Valuers with statistical analysis skills are more likely to satisfy the client's needs, as the outcome of the AVM needs to be checked and interpreted, especially for complex valuations.

4) **Post-valuation with Reporting:** The final valuation report (**Step 7** in Figure 4-1) will be operated in a more complex and highly regulated environment – an interactive valuation report. The client can get an interactive valuation reporting with various information from different sources and different perspectives, such as a valuation range provided by AVMs, the future advised value provided by valuers and more reliable and objective information stored in blockchain.

As information technologies including big data and artificial intelligence advances fast, many stakeholders including investors, banks, public authorities and real estate companies expect to benefit from the full potential of automated valuation services which can perform the valuation quickly, improve the transparency of current valuation process, and reduce inaccuracies from reliance on human judgement and attendant bias (RICS 2017b). For instance, an appraisal system based on AVMs has been developed for the Canada government and implemented in the province of Quebec, with the aim of providing basis for property taxation implementation (Kettani and Oral 2015). A vendor from UK **-** *Rightmove* declares that they have stringent criteria, employing a thorough filtering process in selecting the properties used in their AVMs, on the other hand, *Zillow* claims to be the largest AVM provider in the US (Łaszek et al. 2018).

The influence of AI enhanced AVMs on the current valuation process is summarized as follows:

1) **Data collection and data sharing:** Currently, valuers predominantly use primary data sources including client, inspection, property analysis, market analysis and public sources. There are issues of data accessibility and uncertainty about the accuracy and reliability of the data gather during this process. In the future, data collection is expected to become a more specialized profession or a more automated one, with the technological developments such as inspection with drones, the IoT and smart buildings. Big data could partially replace the primary data sources, as data can be collected from secondary data tools such as *Google Analytics and Google Trends*.
2) **Valuation method:** Despite traditional valuation approaches being extensively used in the valuation processes, over the last decade there has been a move towards automated valuation approaches, especially for residential property (Łaszek et al. 2018). Although currently automated valuation models (AVMs) cannot substitute the human valuer in all instances, with the impact of AI and big data developments, the usability of AI-enhanced AVMs is expected to expand towards different property types and more complex valuations (RICS 2017b). Some advocates hold the opinion that the majority of valuations will be carried out by AI systems and AI-enhanced AVMs will replace the valuer, considering the fact that AVMs are undertaking mass valuation work performed for banks. Others believe AI-enhanced AVMs will change the valuation process and help the valuer in many aspects, but it will not replace some part of valuation where the valuer interprets data and makes judgements on the impact of that data on value.
3) **The role of valuer:** In the future, valuers will spend less time on property investigation and inspection, data verification and analysis, instead, they will act as an impartial judge or an adviser. For complex valuations, a valuer will need to check and interpret the outcome of the AVMs (RICS 2017b).

4.2 Impact of BIM on Information Flows in the Valuation Process

Building Information Modelling (BIM), as a digital and computable representation of a building and its related lifecycle information, has significantly improved information flow among stakeholders involved at all stages - from the early design stage to the construction and long operation stage (Borrmann et al. 2018). To support information communication, digital technologies such as databases, model servers and project platforms are often employed in a comprehensive manner. According to Lindblad (2013), the benefits of BIM adoption involve more efficient data exchange, less data input and transfer errors, streamlined construction processes, automated workflow, improved product quality and building performance, and increased productivity. As a result, property valuation professionals concluded that there was great potential to expand the current BIM data for property valuation use, such as linking data with Building Management Systems. For instance, property professionals currently use 24 different types of data in their technical practice and some of these data have already been found in BIM (Wilkinson and Jupp 2016).

For property valuation professionals, it is essential for them to access to and use lifecycle building performance information from reliable data sources. The discussion on sustainable buildings and the potential 'green' impact on property value is ongoing between valuers and clients (RICS 2017b). The complexity of sustainability assessment and taking account this into different traditional property valuation methods require a significant change in the data collection and information exchange of property valuation professionals and other related market actors (Lorenz et al. 2007). While researchers and practitioners are trying to perform sustainable property valuation, the sustainability-related information upon property values is limited from real estate market. Sustainability-related information contains information on the environmental and health impacts of related building components and materials, energy and water saving, safety and security, demography and structure of households, etc. This information cannot be solely acquired by a licensed property valuer through a building inspection, but requires access to and analysis of other reliable sources of information, for instance, information provided by facility managers and information founded in documentation of the design and planning process (Lützkendorf and Lorenz 2011). The integration of sustainability assessment into traditional property valuation process requires not only the traditional building information on type, size and number of bedrooms, but also information on actual building performance information such as heating, acoustics, air quality on energy savings, which is currently limited in real estate market.

As data sharing issues exist in current valuation process, data standardization such as property measurement standard is expected to improve the accuracy and efficiency for the property industry (RICS 2017b). On the one hand, information losses and misunderstandings among different market actors happen inevitably when using different descriptive ways for data interpretations. According to Ventolo (2015), there are about 45 data sources involved in the traditional building survey: government councils, professional journals, local material suppliers, building and architectural plans etc. Each market actor in real estate market uses raw data for property valuation according to their own benefits, or they collect and process information from other data source suppliers. Different market actors use various descriptive ways interpreting information in different formats, which causes the information losses and misunderstandings during information exchange processes. On the other hand, an increased and more detailed consideration of sustainability-related information not only exists in property valuation, but also in other related processes of the building industry such as procurement, design, construction, operation and maintenance etc. As a consequence, in order to address sustainability issues within property valuation practices, different market actors such as portfolio and corporate real estate managers, investors, designers, and engineers will have to work together and provide related information in formats that are easy to access and can be interpreted into the valuation process (Lützkendorf and Lorenz 2011).

Based on this, the future information flow and data exchange in property valuation, associated with BIM and other information technologies, is envisaged in the Figure 4-3. There will be five layers of the information workflow in property valuation process, including the physical layer, the technical layer, the pre-processing layer, the information storage layer, and the domain layer.

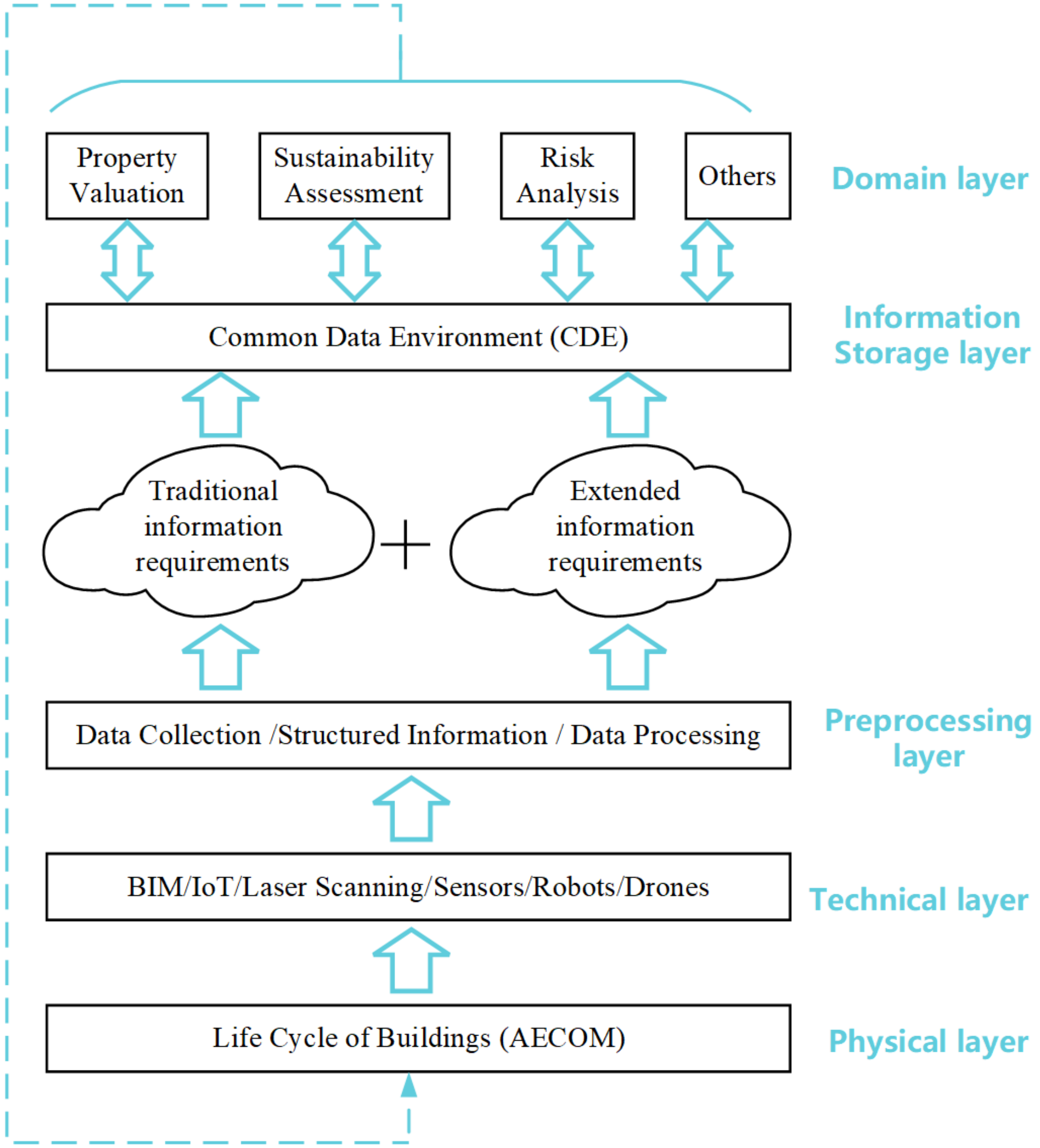

Property Valuation
Sustainability Assessment
Risk Analysis
Others
Domain layer
Common Data Environment (CDE)
Information Storage layer
Traditional information requirements
+
Extended information requirements
Data Collection /Structured Information / Data Processing
Preprocessing layer
BIM/IoT/Laser Scanning/Sensors/Robots/Drones
Technical layer
Life Cycle of Buildings (AECOM)
Physical layer

Figure 4- 3: The future information flow and data exchange in property valuation

As shown in Figure 4-3, firstly, the raw data within the life cycle of buildings will be collected through the use of different technologies in the technical layer such as BIM, IoT, laser scanning, sensors, robots, and drones. Secondly, the collected raw data will be cleaned and verified by information technologies such as ML algorithms and valuation data analysts with statistical skills. Thirdly, the processed information will be classified as traditional information requirements such as property size, age, and height and extended information requirements such as green – rating certificates, energy saving materials, and rooms with high-level natural lighting and ventilation. The organized information will be stored in the common data environment (CDE) in the information storage layer. Finally, the information stored in the CDE can be used to support applications in different domains such as property valuation, sustainability assessment, and risk analysis.

This form of information flow has the potential to contribute to property valuation and the construciton industry with saved time and costs, and improved productivity. All actors in property markets can not only access raw data in the life cycle of buildings, but also use the information prepared by other actors for other purposes. For instance, the sustainability assessment results can serve as an informational basis for property valuation, in return, the improved property image with added value will faciliate the sustainability assessment in other related processes in the building industry. As a consequence, a reliable and continuous platform for maintaining and updating building-related information within the entire building life cycle is needed.

To sum up, the influence of BIM on information flows in the current valuation process is summarized as improved data exchange, less data input and sharing errors, data standardization, linked data with other information sources, saved time and costs, and improved productivity.

In the next section, information requirements for property valuation will be explained.

4.3 Information Requirements for Property Valuation

To get a comprehensive understanding of all the relevant determinants or parameters that can be used to work out the final property values, this research reviewed 174 research documents, the list of value-relevant building properties is explained in Table A-1 (in the Appendix A). Most of them are identified through literature review, some of them are proposed for their potential impact on the future valuation process. The required information has been classified with six different types of information related to property valuation: information related to environmental quality, social and economic quality, functional quality, process quality, technical quality and site quality. Information included in traditional building survey is compared with information required for sustainability assessment and information achievable within BIM related processes. The column A in yellow colour stands for information needed for traditional real estate appraisal. The column B in green colour means information needed for green assessment. The column C in dark red stands for information can be defined and developed in the BIM related platform (design, planning, operation and maintenance phases), which is the core for semi-automated information exchange for property valuation. The identified variables were further used for the IFC-based data interpretation at the system development stage. The column D in light red means information needed by both property valuation and green assessment.

In column C of Table A-1 (in the Appendix A), attention is given to variables that have potential to be associated with BIM related concepts. Among 95 variables reviewed in the literature, 62 of them are identified as relevant to this research. The identified related variables in the column C for building the valuation information model are classified as six main types and 28 subtypes of information related to property valuation, which is illustrated in Figure 4-4. For instance, for technical quality, variables related to construction quality and renovation condition are considered as relevant, while the variable related to protection against burglary is not retained in the valuation information model. Similarly, there are three subtypes identified for variables related to environmental quality, five subtypes identified for variables related to process quality, seven subtypes for variables related to functional quality, two subtypes for social and economic quality and one subtype for site quality.

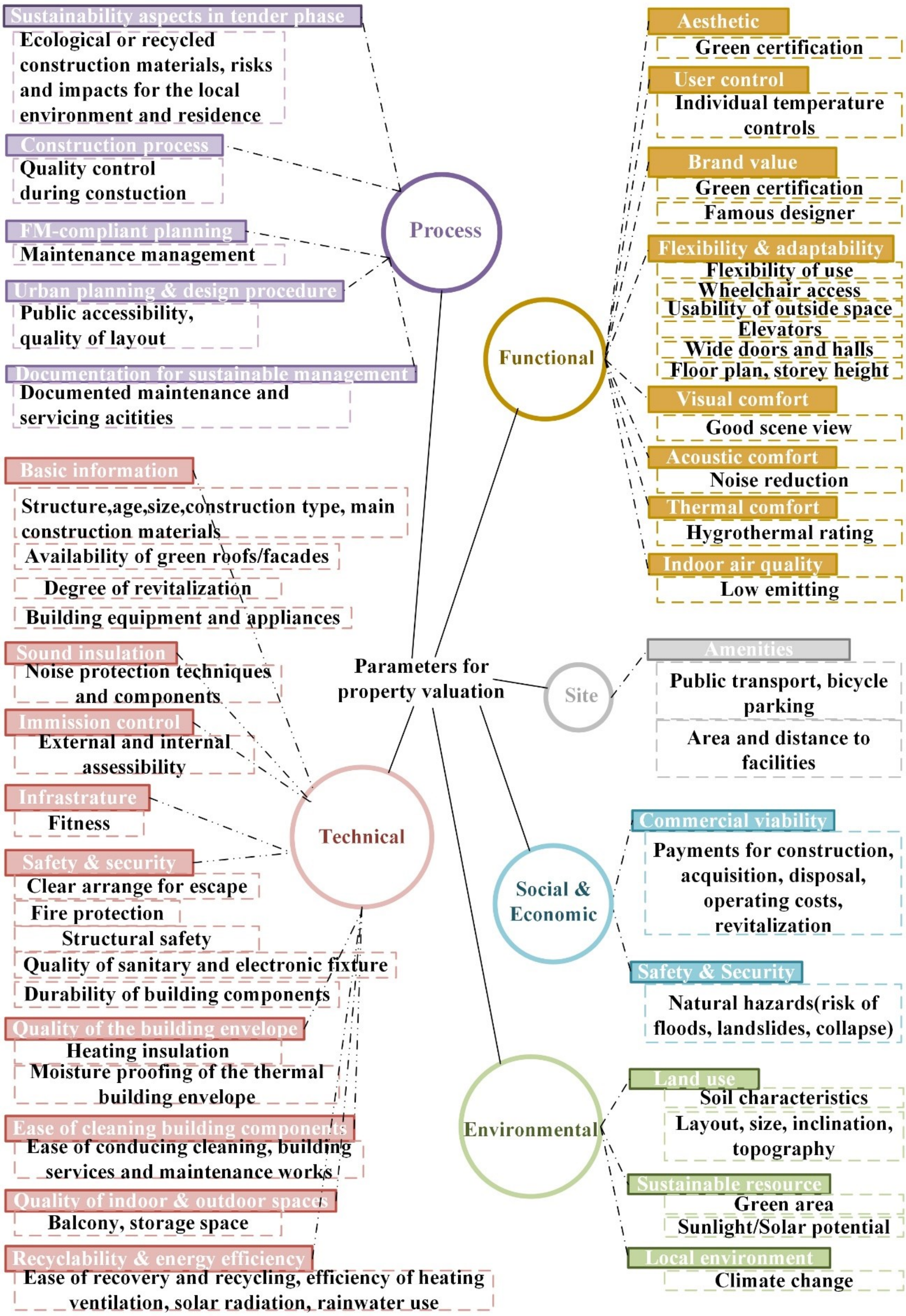

Sustainability aspects in tender phase
Ecological or recycled construction materials, risks and impacts for the local environment and residence
Construction process
Quality control during constuction
FM-compliant planning
Maintenance management
Urban planning & design procedure
Public accessibility, quality of layout
Documentation for sustainable management
Documented maintenance and servicing acitities
Process
Basic information
Structure,age,size,construction type, main construction materials
Availability of green roofs/facades
Degree of revitalization
Building equipment and appliances
Sound insulation
Noise protection techniques and components
Immission control
External and internal assessibility
Infrastrature
Fitness
Safety & security
Clear arrange for escape
Fire protection
Structural safety
Quality of sanitary and electronic fixture
Durability of building components
Quality of the building envelope
Heating insulation
Moisture proofing of the thermal building envelope
Ease of cleaning building components
Ease of conducing cleaning, building services and maintenance works
Quality of indoor & outdoor spaces
Balcony, storage space
Recyclability & energy efficiency
Ease of recovery and recycling, efficiency of heating ventilation, solar radiation, rainwater use
Technical
Parameters for property valuation
Functional
Aesthetic
Green certification
User control
Individual temperature controls
Brand value
Green certification
Famous designer
Flexibility & adaptability
Flexibility of use
Wheelchair access
Usability of outside space
Elevators
Wide doors and halls
Floor plan, storey height
Visual comfort
Good scene view
Acoustic comfort
Noise reduction
Thermal comfort
Hygrothermal rating
Indoor air quality
Low emitting
Site
Amenities
Public transport, bicycle parking
Area and distance to facilities
Social & Economic
Commercial viability
Payments for construction, acquisition, disposal, operating costs, revitalization
Safety & Security
Natural hazards(risk of floods, landslides, collapse)
Environmental
Land use
Soil characteristics
Layout, size, inclination, topography
Sustainable resource
Green area
Sunlight/Solar potential
Local environment
Climate change

Figure 4- 4: Selected variables for property valuation

4.4 The Gradient Boosting Ensemble Method

From literature review in section 2.1.4 and section 2.3.3, it was concluded that compared to neural networks, ensemble learning has advantages in terms of model interpretability and flexibility, which is more suitable for knowledge mining and system development for property valuation. Since genetic algorithm (GA) optimized neural networks have achieved good performance for property valuation, the integration of GA and ensemble learning might achieve good predictive performance for property valuation as well as good model interpretability.

Ensemble learning is a machine learning method that multiple base learners or classifiers are trained and combined for the same task, aiming to improve the predictive performance and control overfitting. It has been widely used in various applications such as gene expression analysis, text categorization, bankrupt prediction etc. Generally speaking, to construct a good ensemble model, the individual base learners should be as accurate and as diverse as possible (Zhou 2012b).

Ensemble methods can be very powerful and often perform better than individual classifiers that make them up. According to Yang (2016), there are three fundamentals of ensemble learning: (1) the strategies to train each of base learners; (2) the combining methods of multiple base learners; and (3) the critical factors to value the success of ensemble learning model such as the bias-variance decomposition. The reason why the generalization ability of an ensemble is stronger than individual base learners is concluded by Dietterich (2000) from three different perspectives: (1) From the statistical perspective, an ensemble can 'average' the votes of multiple base learners and reduce the risk of choosing the wrong classifier; (2) From the computational perspective, compared to individual classifier which may get stuck in local optima, an ensemble which performs local search from many different starting points may have a better chance to approaching the best approximation for the hypothesis; and (3) From the representational perspective, an ensemble can expand the space of representable functions to deal with the issue of most machine learning applications that the true function $f_{(x)}$ cannot be represented by any of the hypotheses.

Bagging and Boosting are two representative ensemble methods, both of which are constructed in two steps: (1) generate individual base learners, and (2) combine the base learners together through the averaging or voting. In bagging, base learners are trained in a parallel manner, which is illustrated in Figure 4-5. Firstly, in the bootstrap stage, each base learner is independently trained on resampled training set, which is randomly chosen from the original training set (Graczyk et al. 2010). The data resampling technique ensures the uniqueness of each base learner, which increases the diversity of bagging ensemble and provides the ability to significantly reduce the predictive error caused by variance. Secondly, in the aggregating stage, the individual base learners generated in the first stage are combined through voting methods for classification tasks, and through average methods for regression tasks such as house price prediction. One famous example of the bagging ensemble is called random forest.

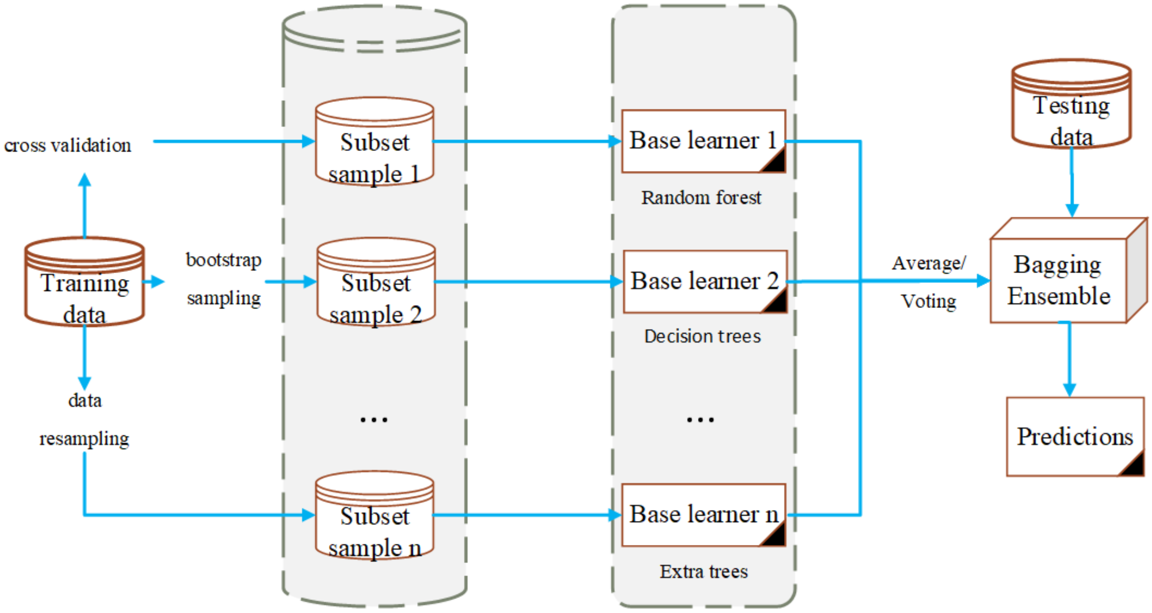


Figure 4- 5: Bagging method

The boosting ensemble method is illustrated in Figure 4-6. Firstly, base learners are sequentially trained where the first one is trained on the whole training set and the following one is trained based on the performance of the previous one (Graczyk et al. 2010). The previous base learners with big errors are given more attention (weight adjustment) in the next boosting iteration. Secondly, all the base learners are combined together for the prediction task. The weight based boosting and residual based boosting are two successful boosting methods. One famous residual based boosting ensemble application in engineering is XGBoost, which is based on gradient boosting ensemble.

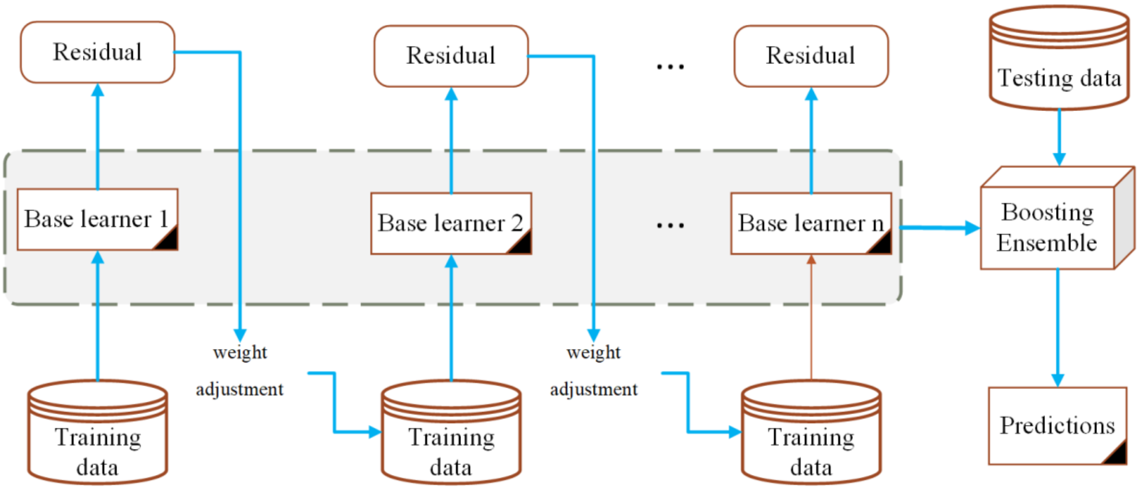


Figure 4- 6: Boosting method

Differences between bagging and boosting methods are summarized as:

- While the base learners are parallelly trained in bagging, they are sequentially trained in boosting.
- The boosting ensembles are sensitive to the abnormal values in the training data set, whereas the bagging ensembles are not easily influenced by the abnormal data.
- The bagging ensembles can reduce the error caused by variance, which are devoted to unstable learners that suffered from large variance such as neural networks (Zhou 2012a). In contrast, the boosting ensembles can reduce the error caused by bias significantly.
- When performing the prediction, most of the bagging ensembles use the majority voting method, whereas most of the boosting ensembles use the average method.

Table 4- 1: Gradient tree boosting algorithm

Inputs:

Training dataset: D = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$; Loss function: $L(y, f(x))$

Algorithm:

Initialize $f_0(x) = arg\min_{\Upsilon} \sum_{i=1}^{N} L(y_i , \Upsilon)$

For m = 1 to *M*:

For *i* = 1, 2, . . . , *N* compute the negative gradient:

$$r_{im} = -\left[\frac{\partial L(y_i , f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, *j* = 1, 2, . . . , J_m

For *j* = 1, 2, . . . , J_m compute:

$$\Upsilon_{jm} = arg\min_{\Upsilon} \sum_{x_i \in R_{jm}} L(y_i , f_{m-1}(x_i) + \Upsilon)$$

Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{Jm} \Upsilon_{jm} I(x_i \in R_{jm})$

Output $\hat{f}(x) = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{Jm} \Upsilon_{jm} I(x_i \in R_{jm})$

The gradient boosting regression (GBR) ensemble is a representative boosting ensemble method, which will be developed further in this research. Typically, a gradient boosting regression (GBR) model is constructed in two steps, the details are explained in Table 4-1. First, weak learners $f_0(x)$ are initialized and generated in a sequential style where a new weak learner is trained based on the error of the whole ensemble learnt so far. The logic behind this is to produce the new estimators to be maximally correlated with the negative gradient (r_{im}) of the whole ensemble's loss function. Second, the base learners are combined as $\hat{f}(x)$ to do predictions through weighted averaging method.

The GBR machines have shown great success in various domains (computer-aided medical diagnosis, energy prediction and facial recognition) as they often provide high predictive accuracies. The main advantage of GBR ensemble is its flexibility and extensibility, since researchers can choose different classifiers (linear models, decision trees, instance-based, Bayesian or rule-based learners) to train the base learners and customize their loss functions with regard to specific tasks. Besides, the ensemble provides several hyperparameter tuning options (the number of boosting iterations, learning rate, the maximum depth of individual estimators) that make it flexible to use. However, GBR model is essentially a greed algorithm and can overfit a training dataset quickly with the number of base learners increasing. Therefore, it is necessary to use regularization to ensure the model's generalization capability.

4.5 Conclusion

This chapter presented the requirements to enable a framework where new technologies such as BIM and Machine Learning (ML) can be introduced to improve the objectiveness, accuracy, and efficiency of property valuation. The section provided an in-depth view of the influence of BIM and ML can produce on property valuation, a comprehensive collection of property value determinants from the literature, and a detailed review of the gradient boosting ensemble machine learning model. The content of this chapter also responded to research question two (**Q2**), which is explained as:

Q2: How innovative information technologies such as BIM and Machine Learning (ML) can improve the current valuation process and what are the information requirements for property valuation?

The answer to this research question is summarized as:

- The influence of AI enhanced AVMs on the current valuation process was summarized as: (1) data collection and exchange: to be more automated and improved accuracy and efficiency with data standardization; (2) the valuation method: from traditional methods to AVMs; and (3) the role of valuers: from performing traditional property inspection and data analysis to interpreting the outcome of AMVs.

- The influence of BIM on information flows in the current valuation process was summarized as improved data exchange, less data input and sharing errors, data standardization, linked data with other information sources, saved time and costs, and improved productivity.
- Among 95 variables reviewed in the literature, 62 of them were identified as relevant to this research. The identified variables that have potential to be associated with BIM related concepts were classified as six main types and 28 subtypes of information related to property valuation, which were further used for the IFC-based data interpretation at the system development stage.

Chapter 5. System Design

Based on the findings in Chapter 2 and Chapter 4, to facilitate information exchange between AEC projects and property valuation and support automated property valuation workflow, this research proposes an integration framework that using BIM and Machine learning (ML) technologies for property valuation which contains three main components: (1) an IFC extension for property valuation that includes missing but necessary value-related entities and property sets; (2) an IFC-based information extraction for automatic information exchange between AEC projects and property valuation; and (3) an advanced valuation model (GA-GBR) based on the integration of gradient boosting ensemble learning and genetic algorithm. This is introduced in the first subsection.

The second subsection conducts an experiment with all AI-enhanced AVMs, including AVMs based on linear regression, ridge regression, lasso regression, elastic net regression, KNN, SVM, ANN, CART, AdaBoost, Random forest, and gradient boosting ensemble (GBR), with the aim at comparing the performance of different AVMs and validate the logic of choosing the GBR model for the proposed system.

The third subsection explains the structure of the GA-GBR model. The fourth subsection then tested the proposed GA-GBR model with the UCI Machine learning repository Boston housing dataset.

5.1 Property Valuation Framework Definition

The review of theories in Chapter 2 and Chapter 4 identified that current data collection and information flows in property valuation could be improved by the emerging technology developments including the integration of BIM and ML technologies.

Firstly, the value-specific design information produced in AEC projects has not been widely used for property valuation, and there is a need to facilitate the current information exchange process between the two parties. For instance, the information exchanging process can be facilitated by developing IFC extensions for property valuation and an IFC-based information extraction algorithm. Secondly, while ANNs has attracted more attention than other algorithms, compared to neural networks, ensemble learning has advantages in terms of model interpretability and flexibility, which is more suitable for knowledge mining and system development. The optimization strategies suggested the integration of GA and ensemble learning might achieve good predictive performance for property valuation as well as good model interpretability.

These findings provided the practical guidelines to improve the information exchange process and the optimization strategies for the automated valuation models. Based on this, the BIM-ML system has been designed that involves three main components, which is explained in Figure 5-1:

- **Component 1 – IFC Extension for Property Valuation:** Information requirement that covers all the value-relevant design information for property valuation is provided based on an overview of 174 archival research documents, which was introduced in Section 4.3. Based on the collected 62 variables, an IFC extension is proposed that extends the existing IFC schema (IFC4-Addendum 2) to support a comprehensive property valuation.
- **Component 2 – Information Extraction:** Based on the IFC extension development, an IFC-based information extraction algorithm is developed to automatically extract the value-related design information from BIM models. The extracted information is further used to support automatic property valuation.
- **Component 3 – Automated Valuation Model (AVM):** A genetic algorithm optimized gradient boosting regression ensemble learning model (GA-GBR) is proposed that works as a smart valuation engine to achieve automated property valuation. The GA-GBR is firstly trained with house transaction data from the Chinese and American real estate mark, and then use the extracted information from IFC for house price prediction.

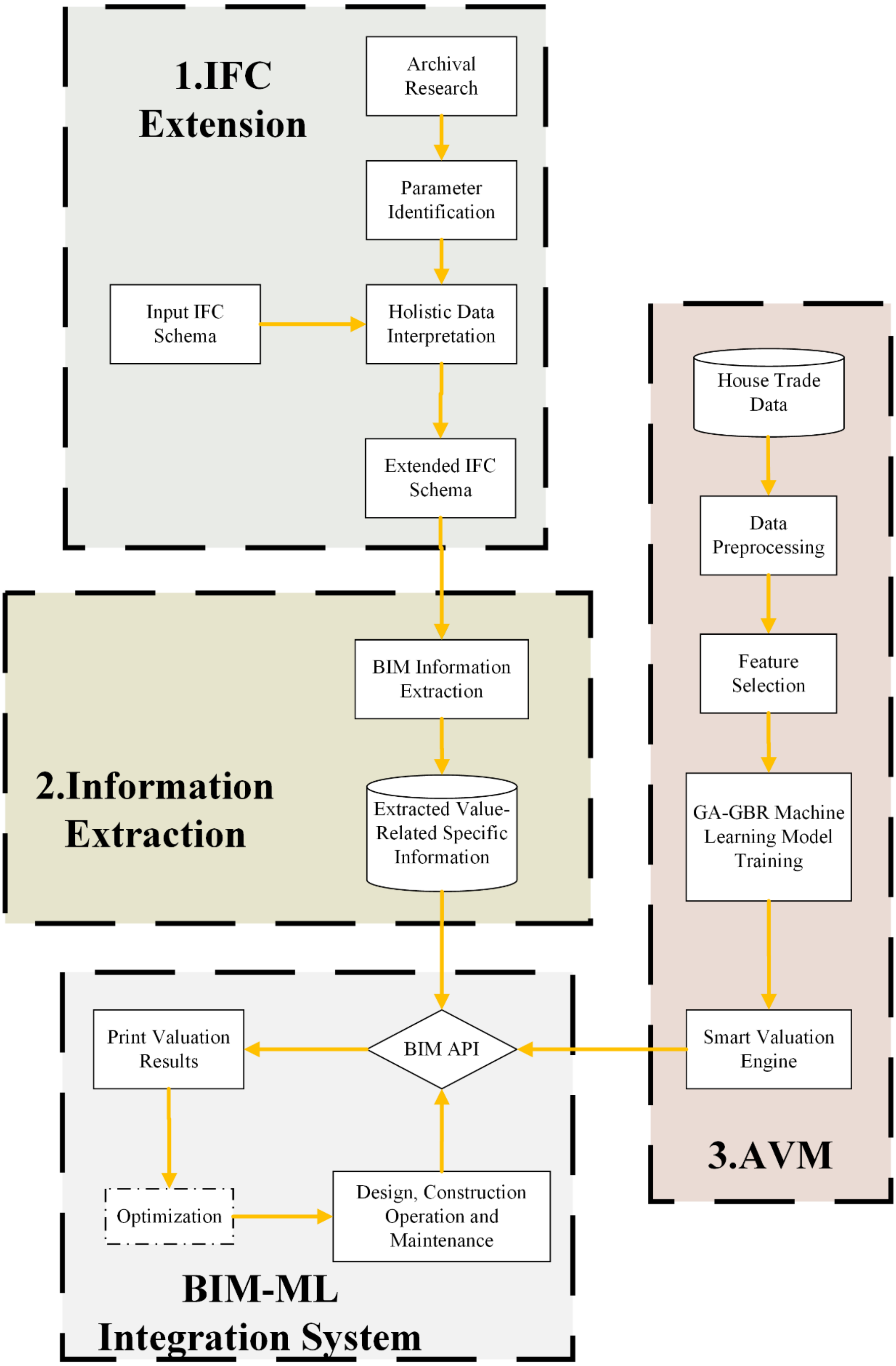


Figure 5- 1: System design for property valuation

5.2 Comparison and Selection among Eleven different AVMs

The findings from literature suggested that ensemble learning application on property valuation is emerging, and the integration of genetic algorithm and ensemble learning might achieve good predictive performance for property valuation as well as good model interpretability. To validate this, it is necessary to conduct an experiment with different AI-enhanced AVMs and compare the model performances. The selected eleven AVMs contain linear regression, ridge regression, lasso regression, elastic net regression, KNN, SVM, ANN, CART, AdaBoost, Random forest, and gradient boosting ensemble (GBR).

The UCI Machine Learning repository - Boston housing dataset was collected for the experiment of the abovementioned AVMs, which includes 506 entries represent aggregated data with 14 variables for house price prediction in Boston in 1978 (Harrison and Rubinfeld 1978). The detailed information about the 14 variables are described in Table 5-1. In the development of the AVMs, the Boston housing dataset is randomly split into the training set (70%) and testing set (30%). The trainings of the eleven AVMs are performed on the Python 3.7 using scikit-learn library on the *PyCharm* platform, which is an integrated development environment using python language for machine learning.

Model performance metrics are essential in evaluating the predictive accuracy of statistical models. In the scientific community, a number of performance metrics have been defined and are currently in use for regression analysis, including the mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2) etc. To get a comprehensive understanding of the performance of the eleven AVMs, all the above-mentioned metrics are selected.

The mean absolute error (MAE) measures the average magnitude of the errors in terms of the absolute differences between the prediction values and the actual results. The calculation of MAE is relatively simple, which provides a generic and bounded performance measure for average model bias. MAE can be used where there are a small number of training outliers, as the model performance will mediocre if there are many outliers in the test set.

Table 5- 1: Description of each variable in the Boston dataset

Variables	**Description**
CRIM	The average crime rate by town.
ZN	The proportion of residential land zoned for lots over 25000 square feet.
INDUS	The proportion of non-retail business acres per town.
CHAS	Charlies River dummy variable (1 if tract bounds river; 0 otherwise).
NOX	Nitric oxides concentration (parts per 10 million).
RM	The average number of rooms per dwelling.
AGE	The proportion of owner-occupied units built prior to 1940.
DIS	The weighted distances to five Boston employment centres.
RAD	The index of accessibility to radial highways.
TAX	The full-value property-tax rate per $ 10,000.
PTRATIO	The pupil-teacher ratio by town.
B	1000(Bk – 0.63) ^2 where Bk is the proportion of blacks by town.
LSTAT	The lower status of the population.
MEDV	Median value of owner-occupied homes in $ 1000's.

The MAE is expressed as follows:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{t=1}}^{\mathrm{n}} |\mathrm{A_t} - \mathrm{F_t}| \tag{7}$$

where $\mathrm{A_t}$ is the actual value and $\mathrm{F_t}$ is the predicted value.

The mean absolute percentage error (MAPE) is commonly used in model evaluation of regression tasks, for its intuitive interpretation in terms of relative error. In practice, the MAPE is useful to calibrate prices of products since some customers pay more attention to relative variations than to absolute variations (Myttenaere et al. 2016). However, it is often criticised for its being biased towards low forecasts, which makes it unsuitable for predictive models with large errors (Chicco et al. 2021). The MAPE is often calculated as a percentage:

$$\mathrm{MAPE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{t=1}}^{\mathrm{n}} \left| \frac{\mathrm{A_t} - \mathrm{F_t}}{\mathrm{A_t}} \right| \tag{8}$$

where $\mathrm{A_t}$ is the actual value and $\mathrm{F_t}$ is the predicted value.

The mean squared error (MSE) measures the average squared difference between the predicted values and the actual results. Due to its definition, the squaring part of the function magnifies the error, which makes it great for attributing larger weights to outliers (Chicco et al. 2021). The MSE has an advantage to sometimes make a comprehensive assessment that combines the effect of bias and random measurement error (Holst and Thyregod 1999). The MSE is expressed as:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{t=1}}^{\mathrm{n}} (\mathrm{A_t} - \mathrm{F_t})^2 \tag{9}$$

where $\mathrm{A_t}$ is the actual value and $\mathrm{F_t}$ is the predicted value.

The root mean squared error (RMSE), which is the square root of MSE, would bring the unit back to actual unit. Compared to the MAE, RMSE gives a relative high weight to large errors since the errors are squared before they are averaged. The RMSE is specified as:

$$R\mathrm{MSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{t=1}}^{\mathrm{n}} (\mathrm{A_t} - \mathrm{F_t})^2} \tag{10}$$

where $\mathrm{A_t}$ is the actual value and $\mathrm{F_t}$ is the predicted value.

The coefficient of determination (R^2), also known as R-squared or squared multiple correlation coefficient, represents the quantity that estimates the percentage of variance of the response variable explained by its relationship with the explanatory variables. It is usually to be used by practitioners to assess the quality of the fit in a regression model, which provides an indication of the suitability of the chosen explanatory variables in predicting tasks (Renaud and Victoria-Feser 2010). The R^2 is expressed as follows:

$$\mathrm{R}^2(\mathrm{y}, \hat{\mathrm{y}}) = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (11)$$

where $\hat{y}_i$ represents the predicted value of the i^{th} sample, y_i is the corresponding true value, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ represents the total sum of squares and $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i{}^2$ represents the corresponding residual sum of squares.

Table 5- 2: Accuracy metrics of different regression models in the Boston dataset

Accuracy metrics	**MAE**	**MAPE**	**MSE**	**RMSE**	**$\mathbf{R^2}$**
Linear regression	3.36	17.2%	23	4.7	71.6%
Ridge regression	3.36	17.1%	23	4.7	71.7%
Lasso regression	3.73	19.3%	29	5.3	64.5%
Elastic Net regression	3.75	19.2%	31	5.4	62.8%
KNN	2.94	13.7%	22	4.6	74.0%
SVM	2.87	15.6%	23	4.6	73.3%
ANN	2.67	13.5%	15	3.8	83.2%
CART	2.67	13.8%	13	3.6	82.8%
AdaBoost	2.02	10.6%	7.8	2.7	90.0%
Random forest	2.12	10.8%	9	2.9	88.7%
GBR	**2.04**	**10.4%**	**7.6**	**2.7**	**90.3%**

Table 5-2 lists the five different accuracy metrics introduced above, of which the MAE, MAPE, MSE, and RMSE are measuring regression models in terms of different types of errors and the R-squared (R^2) is measuring regression models in terms of prediction accuracy. This means that the lower the MAE, MAPE, MSE, and RMSE, the higher the R-squared (R^2) lead to a better model performance of the AVMs.

From the experiments on the test dataset (30%), it is observed that the four classic linear regression models including the linear, ridge, lasso, and elastic net regression, have similar poor model performances, with the mean MAE at 3.55, the mean MAPE at 18.2%, the mean MSE at 26.5, the mean RMSE at 5.025, and the mean R^2 at 67.65%. Compared to the four classic linear models, the KNN and SVM models have better model performances, with the mean MAE at 2.9, the mean MAPE at 14.7%, the mean MSE at 22.5, the mean RMSE at 4.6, and the mean R^2 at 73.7%. Compared to the KNN and SVM, the ANN and CART (the simple decision tree based) models have better model performances, with the mean MAE at 2.67, the mean MAPE at 13.7%, the mean MSE at 14, the mean RMSE at 3.7, and the mean R^2 at 83%. Compared to the ANN and CART, the other three decision tree-based models including AdaBoost, Random Forest and GBR have better model performances, with the mean MAE at 2.06, the mean MAPE at 10.6%, the mean MSE at 8.1, the mean RMSE at 2.77, and the mean R^2 at 89.7%.

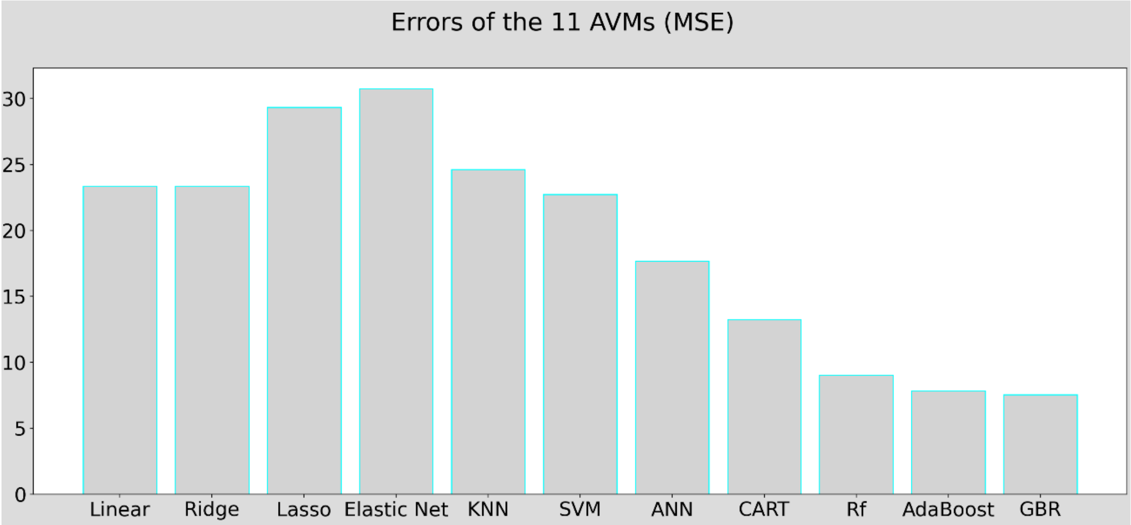


Figure 5- 2: Errors of the eleven AVMs in terms of the MSE

It is worth to mention that the GBR model has the highest model prediction accuracy with the MAPE at 10.4%, the MSE at 7.6, the RMSE at 2.7, and the mean R^2 at 90.3%. For instance, the errors of the eleven AVMs in terms of MSE are shown in Figure 5-2, which illustrates that the GBR model has the best model performance of the eleven different types of AVMs. This complies with the findings from literature.

5.3 The Proposed Automated Valuation Model (GA-GBR)

In this section the framework of the proposed genetic algorithm optimized gradient boosting ensemble model (GA-GBR) will be described.

The experiment in the last section indicated that gradient boosting decision tree (GBR) model has the highest predictive accuracy of the eleven different types of AVMs. In fact, the GBR machines have shown great success in various domains such as computer-aided medical diagnosis, energy prediction, and facial recognition, as they often provide high predictive accuracies. However, compared to random forest algorithms, since the base learners in the GBR are dependent on each other, the GBR machines are more likely to be overfitting with a training dataset.

It has been recognized that a good ensemble depends on the individual learners being as accurate, and as diverse as possible. However, generating diverse individual learners is quite challenging as the individual base learners are usually highly correlated for dealing with the same training data. This means the more accurate individual learners are, the less diverse they are. Therefore, the success of a GBR ensemble is essentially about getting a good trade-off between the accuracy and diversity of individual base learners. To encourage the diversity of an ensemble, the basic logic is to inject some heuristic mechanisms into the learning process. There are four typical effective methods: manipulation of training data, manipulation of input features (the random subspace method), manipulation of learning parameters, and manipulation of output representation (Zhou 2012b).

To make up the deficiency and improve the predictive accuracy of the traditional GBR model, whilst considering the exploration of the relationship between the input features and the target price, the genetic approach for optimizing boosting ensemble is proposed. The genetic algorithm (GA) in the proposed GA-GBR works as an evolutionary feature selection engine to search the optimal feature subset which is further used to train a good boosting ensemble. There are three reasons for choosing the genetic optimizer for traditional GBR ensembles:

1) For data with a big number of input features, input feature manipulation method often gives a good result. The manipulation of input features using GA increases the diversity of individual base learners.

2) The GA searches the suitable number of input features and updates the weights of them, which enables the GA-GBR to explore the relationship between the input features and the target price, and therefore gives an improved model interpretability over traditional boosting ensemble machines.

3) The evolutionary feature selection engine eliminates the redundant and irrelevant features without affecting the prediction accuracy, which avoids the overfitting of traditional GBR machines.

Implementation of the proposed GA-GBR model has three steps, as shown in Figure 5-3, including base learner generation, problem encoding and genetic search. Details about each step will be described below.

1) Base learner generation

The first step is to generate the pool of base learners with the input domain dataset. There are three common base-learner models: linear models, smooth models and decision trees. The base learners in this research use decision trees of same sizes which are good at handling mixed types of data and modelling complex functions. GBR ensemble combines multiple base learners $f_m(x_i)$ to generate a strong model $\hat{f}(x)$, which is displayed below.

$$\hat{f}(x)= \sum_{m=1}^{M} f_m(x_i) \tag{12}$$

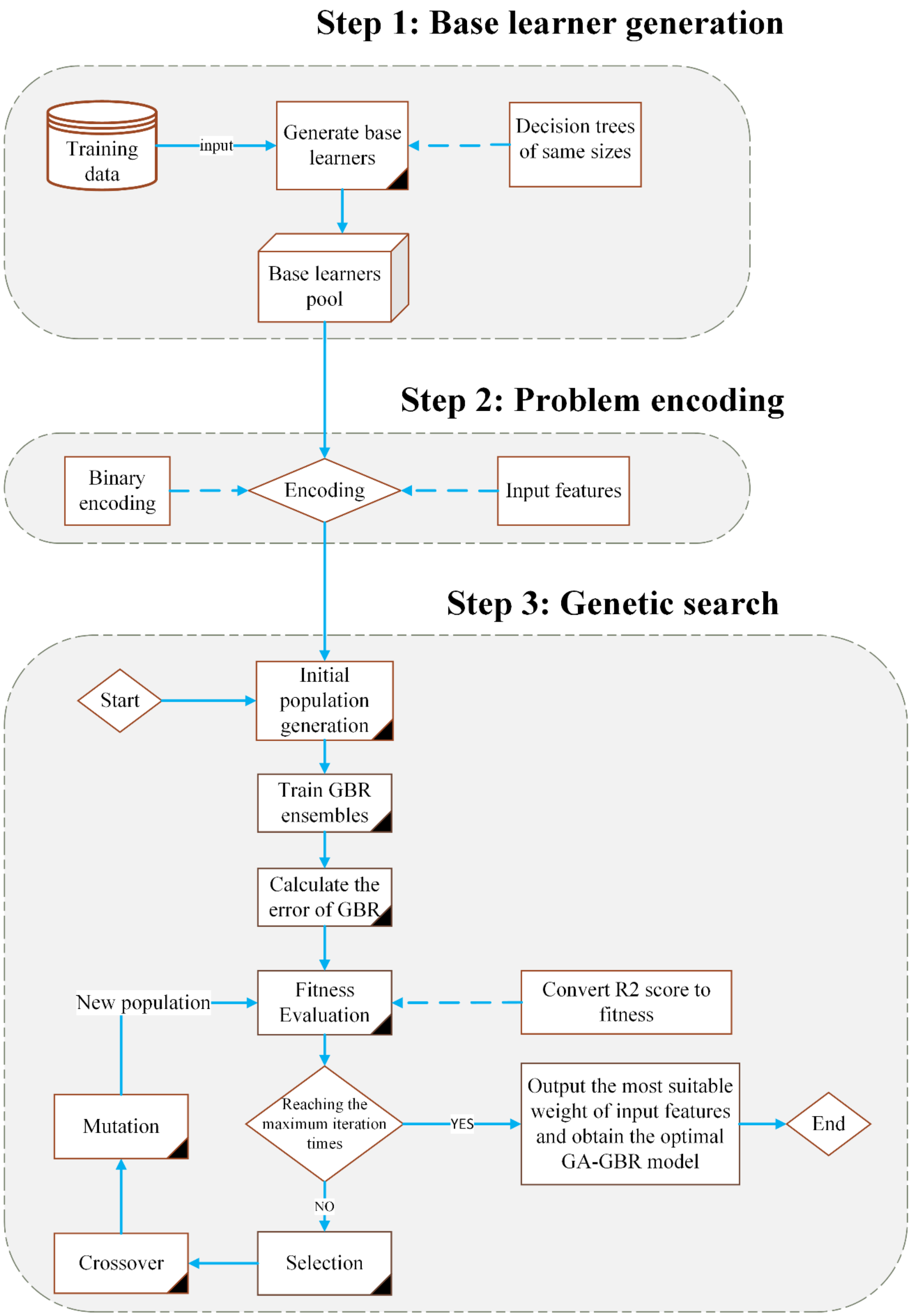

Step 1: Base learner generation
Training data
input
Generate base learners
Decision trees of same sizes
Base learners pool
Step 2: Problem encoding
Binary encoding
Encoding
Input features
Step 3: Genetic search
Start
Initial population generation
Train GBR ensembles
Calculate the error of GBR
New population
Fitness Evaluation
Convert R2 score to fitness
Mutation
Reaching the maximum iteration times
YES
Output the most suitable weight of input features and obtain the optimal GA-GBR model
End
NO
Crossover
Selection

Figure 5- 3: The framework of the GA-GBR model

The objective function of base learner is to learn a mapping $f(x)$ between the input feature vector and the output (house price). Typical loss functions for regression model are Gaussian L_2 loss function, Laplace L_1 loss function, Huber loss function and Quantile loss function (Natekin and Knoll 2013). After initial testing on the GBR model, it had the best prediction accuracy with Huber loss function.

2) Problem encoding

The problem encoding task is performed in the second step. When dealing with a real search problem, the search parameters need to be encoded and represent the problem as a function objective. There are several ways to encoding a problem in genetic algorithm (GA), such as binary codification, decimal, hexadecimal and so on. This research uses binary encoding to represent the solutions, because one of the research focuses is to explore the relationship between the input features and the target price. In dealing with feature selection problems, each chromosome represents a feature subset, and the quality of each candidate solution is evaluated using a fitness function.

In the GA-GBR model training, each chromosome in the population represents an individual which has N input features of the training data (shown in Figure 5-4), for instance, parameter A represents house size and parameter B represents central heating. A one or zero represents that the feature is selected or not respectively. The task of using binary encoding for feature selection is to find the near optimal chromosome in which each bit corresponds to a feature, with the aim to find the feature subset with the smallest number of features that achieve the best performance (Kanan et al. 2007). The size of the chromosome is decided by the N input features, for example, the UCI machine learning repository - Boston housing dataset in the comparison and selection experiment in the last section, N equals to 13, excluding the target price variable. Similarly, N equals to 56 and 61 in the development of the GA-GBR model using the Chinese and the American datasets respectively described in Section 6.3.1 (the original number of input features are 22 and 15, after one-hot encoding they changed to 56 and 61 respectively).

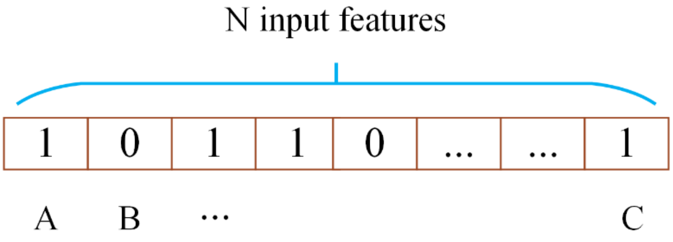


Figure 5- 4: Binary encoding of chromosome

3) Genetic search

In the last step, the genetic search is applied to evaluate each chromosome of the population, and obtain the most suitable combination and weight of input features to train the GA-GBR model. As mentioned in section 2.3.2, genetic search consists of three main steps: (1) initial population generation, (2) fitness function definition, and (3) the action of generic operators including selection, cross over and mutation.

- **Initial population generation**

The genetic search starts with the first generation of solutions which are randomly generated. Each chromosome represents a binary vector that represents the combination of N input features, in which each bit corresponds to a feature. After that, the GBR ensemble is trained with the binary vector (an individual in a population), and the error of the trained GBR is calculated. The quality of an individual solution (chromosome) is then evaluated through a fitness function, which is explained in the following.

- **Fitness function definition**

Fitness function is defined in GA to assess the ability of a chromosome to solve the problem. In this research, after tested with different regression accuracy metrics including MAE, MAPE, MSE, RMSE, and R^2, the coefficient of determination (R^2) is set as the fitness function of the genetic algorithm. It represents the proportion of variance that produced by the independent variables in a machine learning model and explains how well the model fits the training data and the generalization of the model. R^2 function was defined as follows:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

where $\hat{y}_i$ represents the predicted value of the i^{th} sample, y_i is the corresponding true value, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i{}^2$.

- **Selection, crossover and mutation**

The selected chromosomes are subjected to the action of the three operators to obtain new chromosomes for the next generation, namely selection, cross over and mutation. Firstly, parents are selected based on their fitness values: chromosomes with higher R^2 are more likely to be parents of new generations, on the other hand, individuals with lower R^2 may not be selected at all. Tournament selection and roulette wheel selection are two popular selection methods. This study uses the roulette wheel selection method, which enables the selection of the best chromosome with a higher chance (Lipowski and Lipowska 2012).

Secondly, the selected parent chromosomes are used in the crossover to produce a new offspring. Typical crossovers include one-point crossover, two-point crossover, and multi-point crossover (Spears and Anand 1991). Study on the population size and crossover shown that two-point crossover performs better when the population is large, whereas multi-point crossover performs better for the small size populations (De Jong and Spears 1991). Compared to one-point crossover, two-point crossover is able to avoid the exact duplication of parents for the old population, which ensures the new offspring being produced through crossover operation is able to survive in the next generation (Bajpai and Kumar 2010). Based on this and initial test with the three types of crossover operators, two - point crossover is selected for this research. Two - point crossover is performed between two chromosomes, with crossover points randomly generated. Each crossover point is set on the boundary between the two chromosomes below. For example, binary string between the two crossover points of chromosome 1 and the rest part of chromosome 2 are copied to make a new offspring (shown in Figure 5-5).

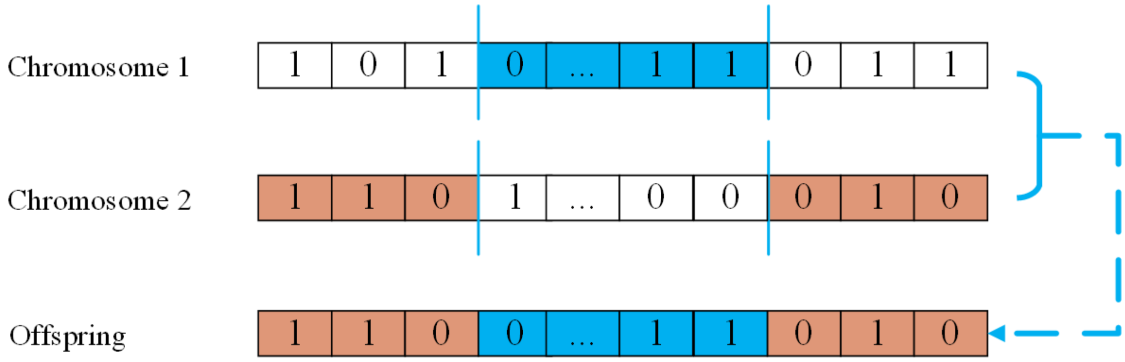


Figure 5- 5: two - point crossover method

The last operator of genetic algorithm is mutation which randomly chooses a point in one bit-string and inverts the value of selected bit (from 1 to 0 or from 0 to 1). Mutation can occur at each point of chromosomes with small possibility, mutation rate between 0.5%-1% normally gives better performance of GA search (Hassanat et al. 2019). Operators with big mutation rate often have strong ability to generate new offspring and prevent premature convergence of genetic search, but also may harm the stability of population structure.

The new generation is continuously constituted after the three generic operators until the GA reach the maximum iteration times. Finally, the most suitable weight of input features and the optimal GA-GBR model are acquired.

5.4 The Proposed GA-GBR model: A Proof-of-concept Study

The proposed GA-GBR model is then tested with the UCI Machine Learning repository - Boston dataset introduced in Section 5.2, comparing model performance with the traditional GBR. In the development of the GA-GBR model, the Boston housing dataset is randomly split into training set (70%) and testing set (30%). The training of the GA-GBR follows the framework designed in the last section 5.3, which is performed on the Python 3.7 using scikit-learn library on *Pycharm* platform.

The initial population is randomly generated with N solutions, with the number of base learners and their associated combination method. After N solutions are randomly generated, the next generation of solutions is generated through genetic search operations. The fitness of each chromosome of the new generation is evaluated according to the fitness function R-square (R^2). The chromosomes with higher R^2 than the GBR model are selected. The parameters in GA were tested using trial-and-error method using gradient boosting regression ensemble library in scikit-learn and the specifically designed genetic algorithm. Through repeated tests, it was found the best performance of GA-GBR model when the parameters in GA were set as follows:

- Population size: 600
- Generations: 32
- Crossover probability: 0.5
- Mutation rate: 0.1

Using grid search algorithm for testing the model hyperparameters, the GBR model was trained with the best performance when hyperparameters were set as the number of estimators (200), learning rate (0.1), maximum depth (5), minimum sample leaf (6), maximum features (0.2), loss function (Huber). The R^2 in GA-GBR (using the best chromosome) had an advantage of 0.2% over the GBR model, with 90.5% for GA-GBR and 90.3% for GBR respectively. Adetunji et al. (2022) conducted a similar house price prediction research using random forest machine learning with the same Boston housing dataset, in terms of R^2, showing 90.0% for the random forest. Compared with this recent research, the proposed GA-GBR shows a slight superiority of 0.5%. Considering that there are only 506 entries data in the Boston housing dataset, the performance of the proposed GA-GBR could be improved when using dataset with a large number of house transaction cases.

5.5 Conclusion

This chapter presented the overall BIM and Machine learning integration framework design, including three main components namely IFC extension, information extraction and the proposed automated valuation model (GA-GBR). After that, the performance of 11 typical AVMs was tested and compared using the UCI Machine Learning repository

– Boston housing data set, aiming at finding the AVM with the lowest predictive error namely the GBR model. Then the traditional GBR model was combined with GA as the new algorithm for the framework, following with a proof-of-concept study of the proposed GA-GBR model using the Boston housing data set. In terms of R^2, the proposed GA-GBR showed an advantage of 0.5% over the similar research from Adetunji et al. (2022). The content of this chapter responded to research question three (**Q3**), which is explained as:

Q3: What kind of automated valuation models (AVMs) might have a better prediction performance for property valuation and how to improve the current AVMs?

From the experiment results of the eleven AI-enhanced AVMs, it indicated that the classic linear models showed the poorest model performance and predictive accuracy, and the decision tree-based models including AdaBoost, Random Forest and GBR showed the highest model performance and predictive accuracy. The KNN and SVM models showed advantage over the ANN and CART, but showed disadvantages over the AdaBoost, Random Forest and GBR. It is worth to mention that the GBR model has the highest model prediction accuracy of the eleven different types of AVMs, with the MAPE at 10.4%, the MSE at 7.6, the RMSE at 2.7, and the mean R^2 at 90.3%. This complies with the findings from the literature and validate the logic of choosing the GBR model for the proposed system.

To make up the deficiency of GBR, whilst considering the exploration of the relationship between the input features and the target price, the GA-GBR model is proposed. The genetic algorithm (GA) in the GA-GBR works as an evolutionary feature selection engine to search the near optimal feature subset which is further used to train a good boosting ensemble. The advantages of the proposed GA-GBR model are concluded as: (1) increase the diversity of individual base learners, (2) improved model interpretability over traditional GBR machines, and (3) avoid overfitting. After explaining the structure of the proposed GA-GBR model, the initial test of the GA-GBR with the UCI Machine Learning repository - Boston dataset indicated improved model performance compared with the traditional GBR.

Chapter 6. System Development

This chapter outlines the core contributions of this research. The property valuation framework introduced in Section 5.1 defines three components, which are divided into three subsections in this chapter: an IFC extension for property valuation (Section 6.1), an IFC-based information extraction (Section 6.2), and an AI-enhance automated valuation model developed based on the genetic algorithm optimized boosting ensemble learning (Section 6.3).

The workflow of Section 6.1 and Section 6.2 is displayed in Figure 6-1 below. Section 6.1 aims to find and add the missing but necessary value-specific property sets and properties to the current IFC schema. The development of the IFC Property Valuation extension is based on the identified relevant variables in the information requirements for property valuation in Section 4.3 and the valuation information model (LADM_VM) which is derived from the ISO 19152:2012 Land Administration Domain Model (LADM).

In Section 6.2, after the development of IFC Property Valuation extension, the required value-specific design information is extracted automatically from an IFC-based BIM instance model using an information extraction algorithm developed based on the open-source BIM information extraction library - *IfcOpenShell*. The extraction process is divided into eight main steps, during which the nominal values of requried varibles are extracted from an IFC instance model.

Section 6.3 introduces the preliminary steps of training the proposed GA-GBR model, including data preparation, exploratory data analysis, and feature selection. After that, the proposed genetic algorithm optimized boosting ensemble learning model (GA-GBR) is experimented on real estate transaction data from the Chinese and American real estate markets. There are two different experimental setups for the training, one is performed on the whole Chinese and American datasets, and the one is performed on the divided datasets representing different perspectives which aims at exploring the implicit relationships between the input features and the target price.

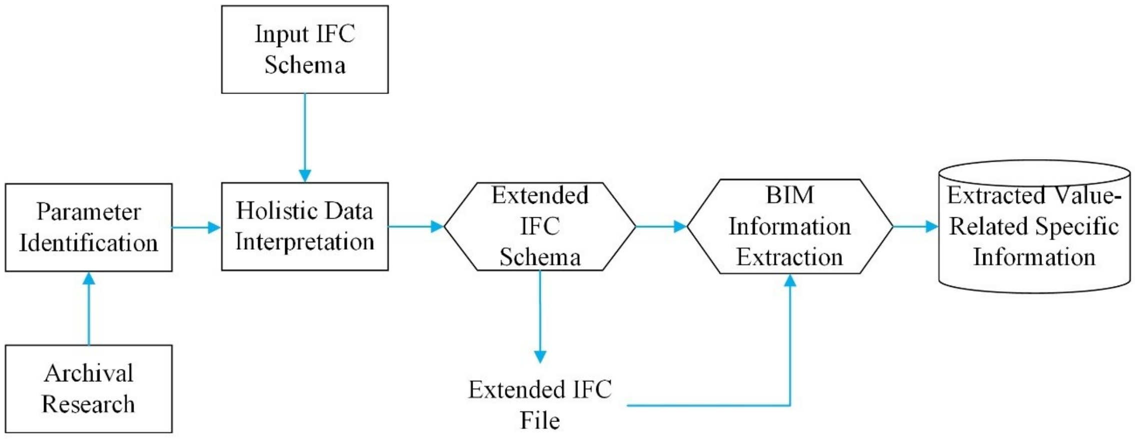

Input IFC Schema
Parameter Identification
Holistic Data Interpretation
Extended IFC Schema
BIM Information Extraction
Extracted Value-Related Specific Information
Archival Research
Extended IFC File

Figure 6- 1: Workflow for the development of the IFC extension and IFC-based information extraction

6.1 IFC Extension for Property Valuation

In this section, an IFC extension for property valuation based on the target information including building object entities and properties will be developed. There are two main steps for developing an IFC extension for property valuation: (1) identify and analyse the building object entities and property sets that support property valuation in the existing IFC schema (IFC4-Addendum 2), and (2) search and add the missing but necessary value-relevant property sets and properties to relevant entities in the IFC4 schema.

1) Identify the existing entities in the IFC4 schema

Identifying and analysing the coverage of the existing IFC schema is an essential step for developing IFC extension for property valuation. Currently, the most updated official IFC schema, IFC4-Addendum2-Technical Corrigendum 1, contains 776 entities and 420 property sets (buildingSMART 2017a). The IFC schema focuses on describing the geometric and semantic data structure of a building using object-oriented representation. It follows a hierarchical and modular framework, which is divided into four conceptual layers from up to bottom: domain layer, interoperability layer, core layer and resource layer (ISO 2018). Each layer has a number of classes that contain various entities, types, enumerations, rules and functions, of which entities are used to describe building information and surrounding components (Zhiliang et al. 2011). As in any object-oriented data model, the semantic meaning and implementation of the entities are generalized using inheritance hierarchy, which is illustrated in Figure 6-2 below.

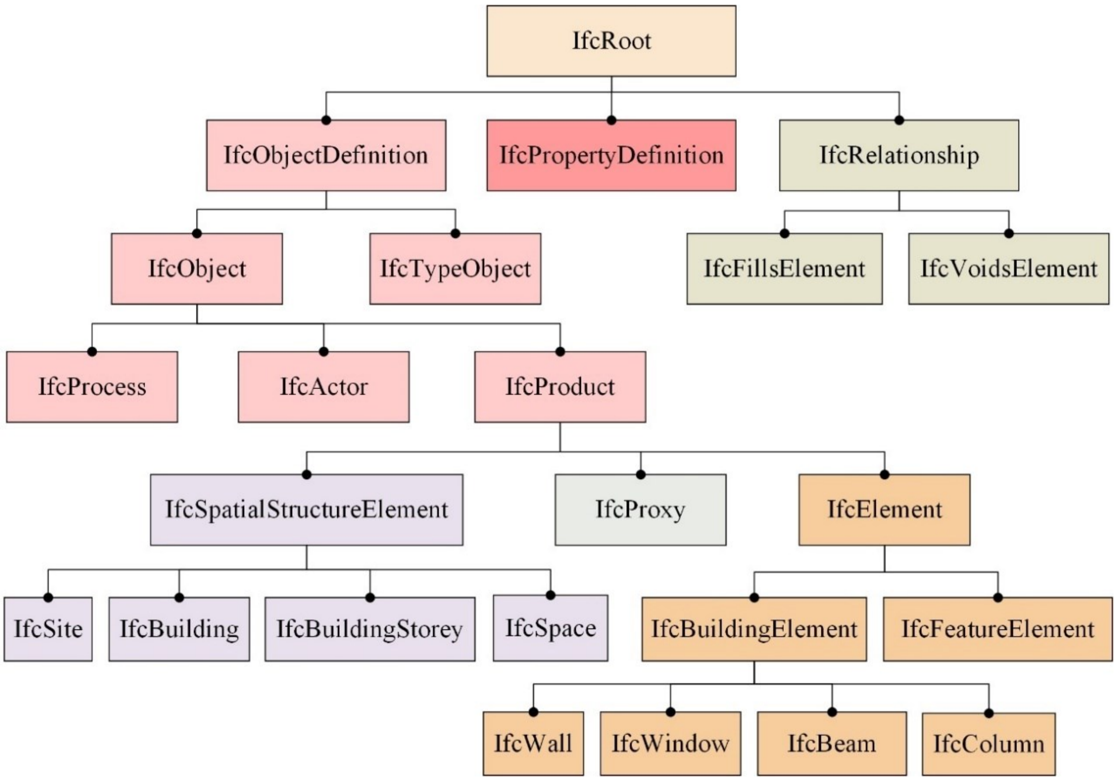


Figure 6- 2: Part of the inheritance hierarchy showing the most important entities according to Borrmann et al. (2018)

The IfcSpace and IfcZone entities in the IFC4 schema were identified as useful for the IFC Property Valuation extension, referring to the Valuation Information Model (LADM_VM) which derived from the ISO 19152:2012 Land Administration Domain Model (LADM). The Valuation Information Model (LADM_VM) is provided in Figure A-1 in the Appendix A (Kara et al. 2018). For instance, the IfcSpace entity is an existing entity in the current IFC4 schema, which represents an area or volume that provide for certain functions within a building (buildingSMART 2017a). This entity can be used to cover information about the room volume, building area, number of rooms, construction time, and renovation condition. The IfcZone entity is an existing entity in the current IFC4 schema, which represents a group of spaces, partial spaces or other zones (buildingSMART 2017a), which can be designed to contain multiple IfcSpace entity instances. Each of these entities has property sets that contain specific properties of building objects, which can be used for property valuation.

2) The proposed property sets and properties for property valuation

Among 95 variables reviewed in the literature, 62 of them are identified as relevant to this research, which are detailed in Section 4.3 – Information requirements for property valuation. The 62 variables are grouped into the six main types and 28 subtypes including basic information (age, size, structure), flexibility and adaptability, amenities, acoustic comfort, and indoor air quality etc. Based on the identified 62 influential variables and the Valuation Information Model (LADM_VM), the required property sets and their properties are proposed to add to the IfcSpace and IfcZone entities, which are listed in Table B-1 in the Appendix B. The property types and data types are identified from the current official IFC schema (IFC4-Addendum 2). It is necessary to choose the right data type for individual properties. For instance, the IfcLabel is a string data type that can be used store the human-interpretable names and shall have a natural-language meaning. The IfcBoolean normally has value True or False that can be used to represent whether the subject property has a garage or not.

Seven property sets are proposed to add to the IfcSpace entity, including Pset_PV_Transaction, Pset_PV_Parcel, Pset_PV_Building, Pset_PV_CondominiumUnit, Pset_PV_Valuation, Pset_PV_MassValuation, and Pset_PV_Annex. The detail of each property set is explained as follows.

The Pset_PV_Transaction property set is designed to provide information related to the real estate transactions, which covers 11 properties such as transactionID, registrationID, activeDays (active days on market), transferDate (when the sale was completed), paidCategory (the type of price paid transactions), noOfFollowers, communityAveragePrice, and propertyRights (if the owner has the property for less than 5 years). The Pset_PV_Parcel property set is designed to provide information related to land parcel, which covers 13 properties such as parcel area, parcelGeometry, parcelLocation, longitude, latitude, city, town, district, and country.

The Pset_PV_Building property set is designed to provide information related to the buildings to be evaluated, which covers 34 properties such as total area, living area, garage area, built date, the number of bedrooms, kitchens, bathrooms, property type, storey, renovation condition, the number of floors, heating and cooling, structure, and elevator. The Pset_PV_CondominiumUnit property set is designed to provide information related to the several condominium units as a group, which covers similar 34 properties in the Pset_PV_Building property set.

The Pset_PV_Valuation property set is designed to provide information related to property valuation using traditional valuation method, which covers 5 properties such as valuation ID, valuation purpose, valuation date, valuation method, and the calculated value. The Pset_PV_MassValuation property set is designed to provide information related to property valuation using advanced valuation method, which covers 5 properties such as valuation ID, valuation purpose, valuation date, algorithm, and the calculated value. The Pset_PV_Annex property set is designed to provide information related to some special considerations from customers which might affect the property value. For instance, considering the education of their children, some customers might think the distance to a famous school is an important factor when buying a house.

In total, there are 7 property sets and 104 properties proposed for the IFC Property Valuation extension (Table B-1 in the Appendix B).

6.2 Information Extraction from the Extended IFC Schema as Required

In the literature, studies on partial data model retrieval can be classified into two main streams: the schema-based data extraction approach and the instance-based data extraction approach. The schema-based approach focuses on developing a definition format with various mappings for data exchange, and it usually extract partial data information according to a predefined model data structure such as the IFC schema, XML schema and SQL schema. For instance, the generalized model subset definition schema (GMSD) method was based on the schema defined in EXPRESS for consistence with the

IFC. The GMSD contains two subparts, one of them provides support to the dynamic selection of object instances in model server queries and the other enables the development of view definitions (Weise et al. 2003). This approach requires a large amount of effort on defining or editing MVDs, which is challenging for some users and sometimes has the problems of high structural complexity and low data density. On the other hand, the instance-based approach is trying to directly deal with the data in the original model and focuses on the extraction of specific information within related objects. For instance, Won et al. (2013) in their research proposed a no-schema algorithm for extracting a partial model from an IFC instance model at the class and object level, without the support of IFC schema. This approach allows the information extraction according to the user's specific requirements and gives users enough flexibility, but generally it requires more efforts on developing complex querying algorithms (Deng et al. 2020).

To support effective information extraction from a BIM model, several commercial or open-source BIM information extraction libraries have been developed, such as *IfcOpenShell*, *BimQL* and *Industry Foundation Classes (IFC) File Analyzer* (Mazairac and Beetz 2013; IfcOpenShell 2018; NIST 2018). For instance, the IFC File Analyzer is effective in extracting the complete information from a BIM model and summarizing the extracted information in an excel table (NIST 2018).

Inspired by the no-schema algorithm (Won et al. 2013), this research uses the combination of the instance-based approach and one of the open-source BIM information extraction library - *IfcOpenShell.* This integration provides an information extraction of a partial BIM model that can extract some common physical elements efficiently and gives enough flexibility of user's information requirements. The information extraction algorithm aims to acquire the required information elements about building objects and their properties based on the IFC extension for property valuation. Three main steps are involved in the information extraction development which will be explained as follows: (1) target information identification and its related data structure definition, (2) developing an information extraction algorithm that can extract value-relevant design information from an IFC instance model, and (3) deploying the information extraction algorithm in an IFC instance model.

1) Target information identification in IFC instances

This step aims to identify the target information in an IFC instance model and define the representation of its data structure. The target information within an IFC instance model contains value-related design information existing in building objects (IfcSpace) and their value-specific properties (total area, built date and renovation condition). Therefore, the representation of the target information includes several key elements in an IFC data model: (1) the globally unique identifier number (GUID) of an IFC instance model, (2) the attributes of building objects including building object names, and (3) the attributes of required IfcProperty instances that contain the property set names, property names, property types, and their nominal values. Figure 6-3 gives an example of the representation of the target information according to the IFC data structure.

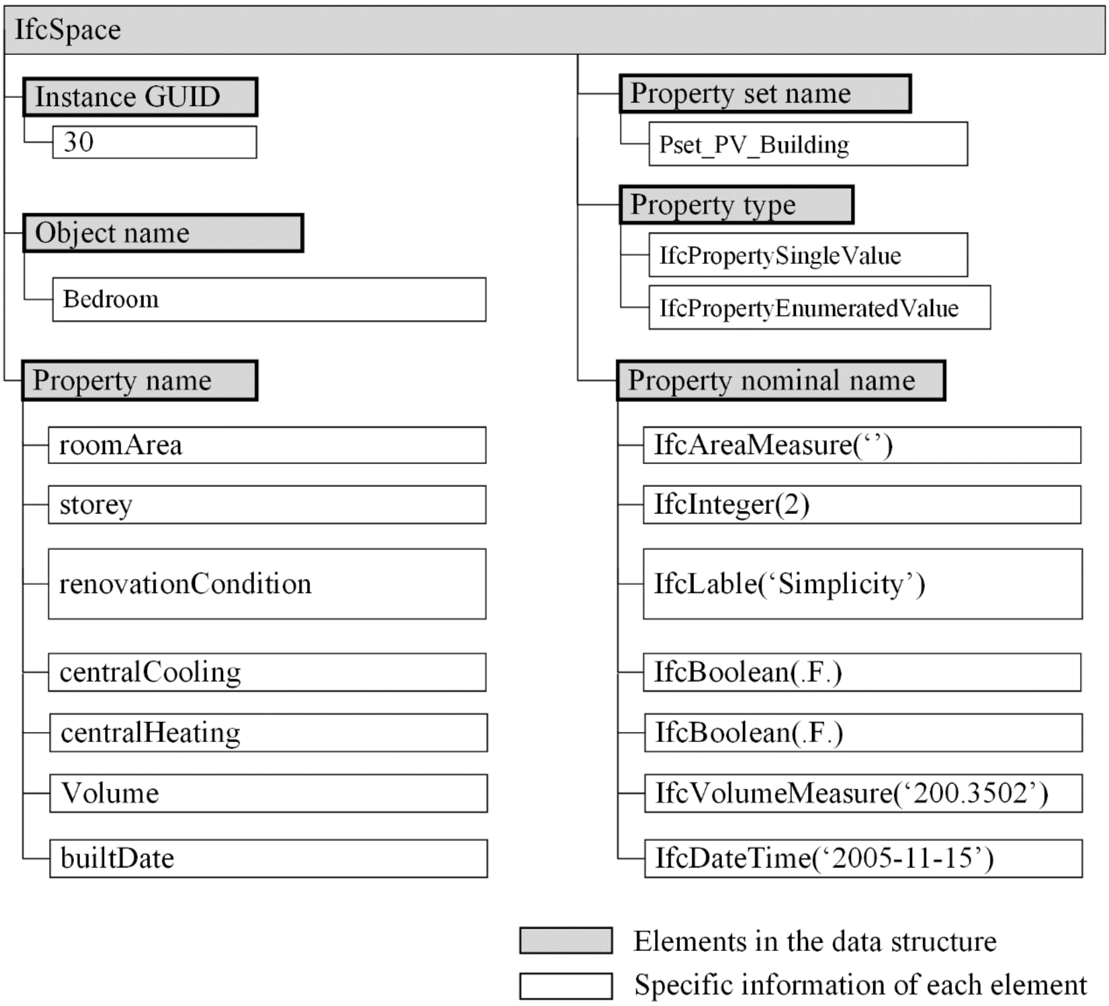


Figure 6- 3: An example of an item in the data structure representation

2) Information extraction algorithm development

To extract value-related information elements about building objects and their attributes for property valuation, information exchange between AEC projects and property valuation is delivered through an IFC-based information extraction algorithm. The context of how to develop an IFC-based information extraction algorithm is described in Figure B-1 and Figure B-2 in the Appendix B. After analyzing the information elements and their relationships between IfcObject and IfcProperty, the IFC-based information extraction algorithm was developed on Python 3.7 using *IfcOpenshell-python* module on *Pycharm* software (IfcOpenShell 2018).

The algorithm works as follows that it will firstly detect the required IFC entities associated with selected building elements and then recursively iterate through data instances related to the selected elements until all related data instances are extracted. Figure B-3 in the Appendix B illustrates the flowchart of the developed IFC-based information extraction algorithm that extracting the required building entities and properties directly through IfcRelDefinesByProperties and indirectly through IfcRelDefinesByType. The extraction process can be classified into 8 steps:

(1) it will iteratively go through the IfcRelDefinesByProperties and IfcRelDefinesByType instances until all required data instances are extracted

(2) it will extract all the ID numbers of IfcObject and IfcPropertyset (IfcTypeObject) from instances extracted in step 1

(3) it will find instances of IfcObject, IfcPropertySet and IfcTypeObject based on ID numbers of instances extracted in step 2

(4) it will extract ID numbers of IfcPropertySet and IfcProperty instances from step 3

(5) it will find instances of IfcProperty and IfcPropertySet based on ID numbers extracted in step 4

(6) it will extract ID numbers of IfcPropety instances from the extracted IfcPropertySet instances and find instances of IfcProperty based on ID numbers of IfcProperty instances

(7) it will extract object names, object types, property names and property nominal values from step 3, step 5 and step 6

(8) the duplicates of the extracted object names, object types, property names and property nominal values will be compared and removed.

The application of the developed IFC-based information extraction algorithm to extract required information through the IfcRelDefinesByProperties instance directly and IfcRelDefinesByType instance indirectly was explained in Figure B-4 in the Appendix B. An example of an extracted information item for required information extraction was displayed in Table 6-1.

Table 6- 1: An example of an extracted information item

Instance GUID	Object name	Property set	Property name	Property nominal value
30	Bedroom	Pset_PV_Building	renovationCondition	IFCLabel('simplicity')

6.3 Genetic Algorithm Optimized Gradient Boosting Ensemble Model (GA-GBR)

6.3.1 Data preparation and descriptive statistics

The preliminary steps of constructing the proposed GA-GBR model involve data collection, data pre-processing and coding of input variables. The experimental datasets were collected from two sources. One was collected from the Kaggle website, uploaded by Qiu (2018), that contains sale price data on the registered property sales in China from 2009 to 2018, which originally fetched from *Lianjia* (a well-known Chinese real estate brokerage company). The other one is about the transacted property sales in USA from the *GitHub* website which was published by a data scientist from San Francisco (Katepalli 2017).

1) The Chinese data set

Table 6- 2: Property variables in the Chinese dataset

Variables	**Columns in Dataset**	**Description**
House price	*totalPrice*	Transacted price (RMB)
Total area	*square*	Size of property in square meters (㎡)
Living room	*livingRoom*	The number of living room
Drawing room	*drawingRoom*	The number of drawing room
Kitchen	*kitchen*	The number of kitchens
Trade time	*tradetime*	The time of transaction
Active days	*DOM*	Active days on market
Bathroom	*bathroom*	The number of bathrooms
Followers	*followers*	The number of people follow the transaction
Building category	*buildingType*	Including tower (1), bungalow (2), combination of plate and tower (3), and plate (4).
Construction time	*constructionTime*	The time of construction
Renovation condition	*renovationCondition*	Including other (1), rough (2), Simplicity (3), and hardcover (4).
Height	*floor*	The height of the house
Community average price	*communityAverage*	Community average price by square (RMB)
Building structure	*buildingStructure*	Including unknown (1), mixed (2), brick and wood (3), brick and concrete (4), steel (5), and steel-concrete composite (6).
Elevator	*elevator*	Have (1) or not have elevator (0)
Longitude	*Lng*	Longitude coordinates using the BD09 protocol.
Latitude	*Lat*	Latitude coordinates using the BD09 protocol.
Ladder ratio	*ladderRatio*	The proportion between the number of residents on the same floor and the number of elevators of ladder. It describes how many ladders a resident has on average.
Property right	*fiveYearsProperty*	If the owner has the property for less than 5

After data pre-processing, 36728 traded houses in Beijing from 2010 to 2018 were selected as the Chinese housing price dataset (the specific number chosen is to keep the same size with the American dataset in the next). Each traded property contains various property variables such as the size of house, the number of living room, kitchen, bathroom, the height of the house, building categories, construction time, renovation condition, building structure, the number of people follow the transaction, active days on market, and address. The detailed information upon the 23 variables were explained in the Table 6-2. Continuous, categorical and binary variables were applied. For instance, continuous variables included House price, Total area, Active days, Followers, Height, Construction time, Community average price, Longitude, Latitude, and Ladder ratio. While Elevator, Property right and Subway used binary variables, the others belong to categorical types.

The descriptive statistics in the Table 6-3 displays the variability within the data, such as minimum, maximum, mean, and standard deviation. The average property size is 82.5 square meters (㎡) with two living rooms. The average number of stories is 13. By the traded house price, the average is ¥4110000 and standard deviation is ¥2530000.

Table 6- 3: Descriptive statistics of variables

Variables	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Total area	7.4	640	82.5	36.5
Living room	0	7	2	0.7
Active days	1	1677	29	50
Drawing room	0	5	1	0.5
Bathroom	0	6	1	0.4
Followers	0	1143	27	44
Height	1	63	13	7.8
Community	10847	183109	63319	22215
House price	110000	49000000	4110000	2530000

2) The American data set

Table 6- 4: Property variables in the American dataset

Variables	**Columns in dataset**	**Description**
House price	*sale_price*	Transacted price ($)
Total area	*total_sqft*	Size of property in square meters (㎡)
Living area	*livable_sqft*	Size of living area in square meters (㎡)
Garage area	*garage_sqft*	Size of garage in square meters (㎡)
Pool	*has_pool*	Whether have a pool or not
Garage attached	*garage_type_attached*	Have an attached garage
Garage detached	*garage_type_detached*	Have a detached garage
Full bathroom	*full_bathroom*	The number of full bathrooms
Fireplace	*has_fireplace*	Whether have a fireplace
Number of Bedrooms	*num_bedrooms*	The number of bedrooms
Carport area	*carport_sqft*	Size of carport area in square meters (㎡)
Built year	*year_built*	The built year of property
Stories	*stories*	The total number of stories
Half bathroom	*half_bathroom*	The number of half bathrooms (a bathroom with only a toilet and sink, but no bath or shower)
Central cooling	*has_central_cooling*	Whether have central cooling system or not
Central heating	*has_central_heating*	Whether have central heating system or not
City	*city*	The name of cities

The American housing price dataset includes 36728 traded houses from 1889 to 2017 in 46 different cities of USA. Each property data contains various property variables such as Total area, Living area, Garage type, Garage size, Pool, Bathroom, Fireplace, the Number of bedrooms, Carport area, Built year, Stories, Central cooling, Central heating, and Address. The details about the property variables were illustrated in Table 6-4. Continuous, categorical, and binary variables were applied. For example, continuous variables contained Total area, Living area, Garage area, Carport area, Built year and house price. While Fireplace, Pool, Central heating, and Central cooling used binary variables, garage type had three categorical types: attached, detached and none.

The descriptive statistics (Table 6-5) displays the variability within the data, such as minimum, maximum, mean, and standard deviation. The average property size is 224.7 square meters (㎡) with three bedrooms. By the parking area, the average garage size and the average carport size are 53.0 square meters (㎡) and 5.3 square meters (㎡) respectively. The average number of stories of housing is 1.41. By the transacted house price, the average is $447003.5 and standard deviation is $297570.

Table 6- 5: Descriptive statistics of variables

Variables	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Total area	87	1544.9	224.7	91.4
Living area	79	1240.6	210.6	83.529
Garage area	1	831.8	53.0	17.2
Carport area	0	18.2	5.3	5.7
Number of Bedrooms	0	14	3.36	0.96
Stories	0	4	1.41	0.52
Full bathrooms	0	8	1.99	0.76
Half bathrooms	0	1	0.56	0.50
House price	1260	10836000	447003.5	297570

6.3.2 Exploratory data analysis

Before fitting data to the proposed GA-GBR model, it is necessary to explore the relationship between the input features and the target price. This helps researchers discover the implicit patterns from different data sets, which might contribute to improving model performance of the proposed AVM.

1) The Chinese data set

Figure 6-4 explains the correlation between the input features and the target price in the Chinese data set. While the Total area (*square)* variable shows the highest correlation with the house price, the Ladder ratio (*ladderRatio)* variable shows the lowest correlation with the house price. Five variables, which are Living room, Drawing room, Bathroom, Trade time, and Community average price, show a relatively high positive correlation with the house price. Three variables, which are Longitude, Building category and Property right, show negative correlations with the house price.

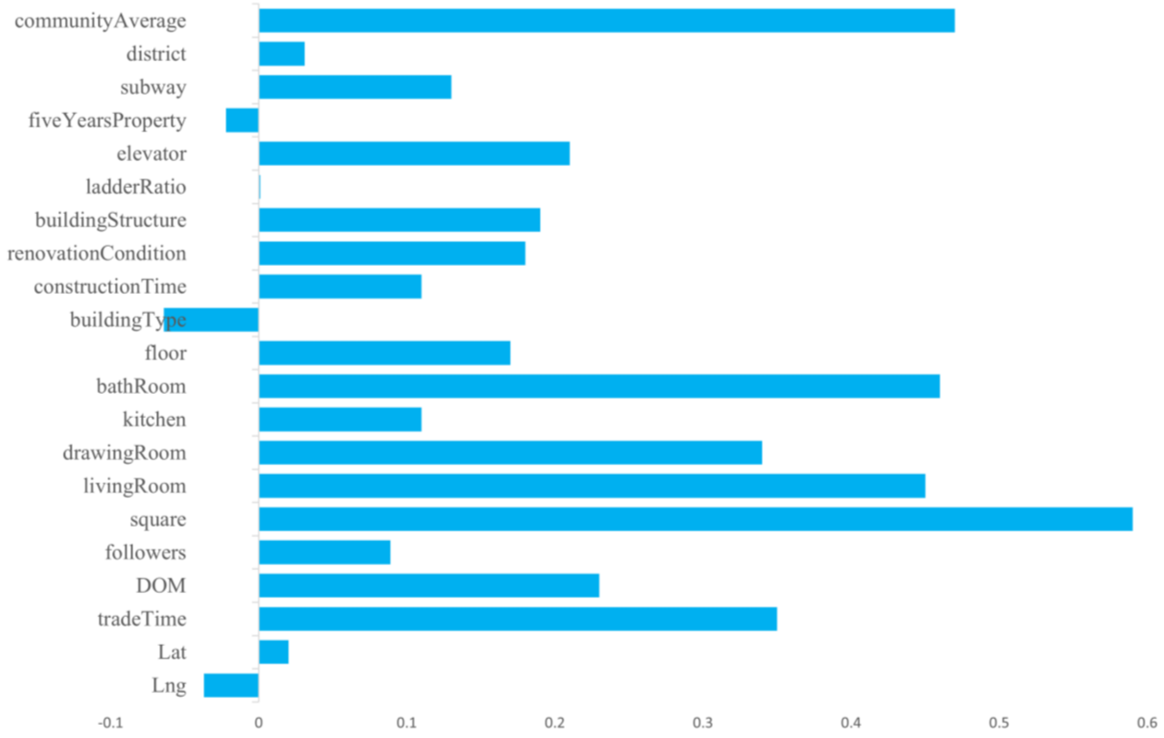


Figure 6- 4: Correlation analysis of the house price and the input features in the Chinese dataset

Figure 6-5 displays the linear relationship between the continuous variables (Active days, Total area, Followers, Height, Construction time, Community average price, Longitude, Latitude, and Trade time) and house price by plotting data and a linear regression model fit. The y-axis shows the value of the total house price and the x-axis shows the value of individual continuous variables. The red lines show the linear fits of continuous variables, while the blue points show the actual data. Generally, a linear regression model fit figure would show the blue points around the red line, going up or down gradually. The plots in Figure 6-5 show the blue points are scattered around the red line, which means that the relationship between the house price and input continuous features are complex and significantly non-linear. While the Longitude variable shows a negative correlation with the house price, the other eight continuous variables show positive correlations with the house price.

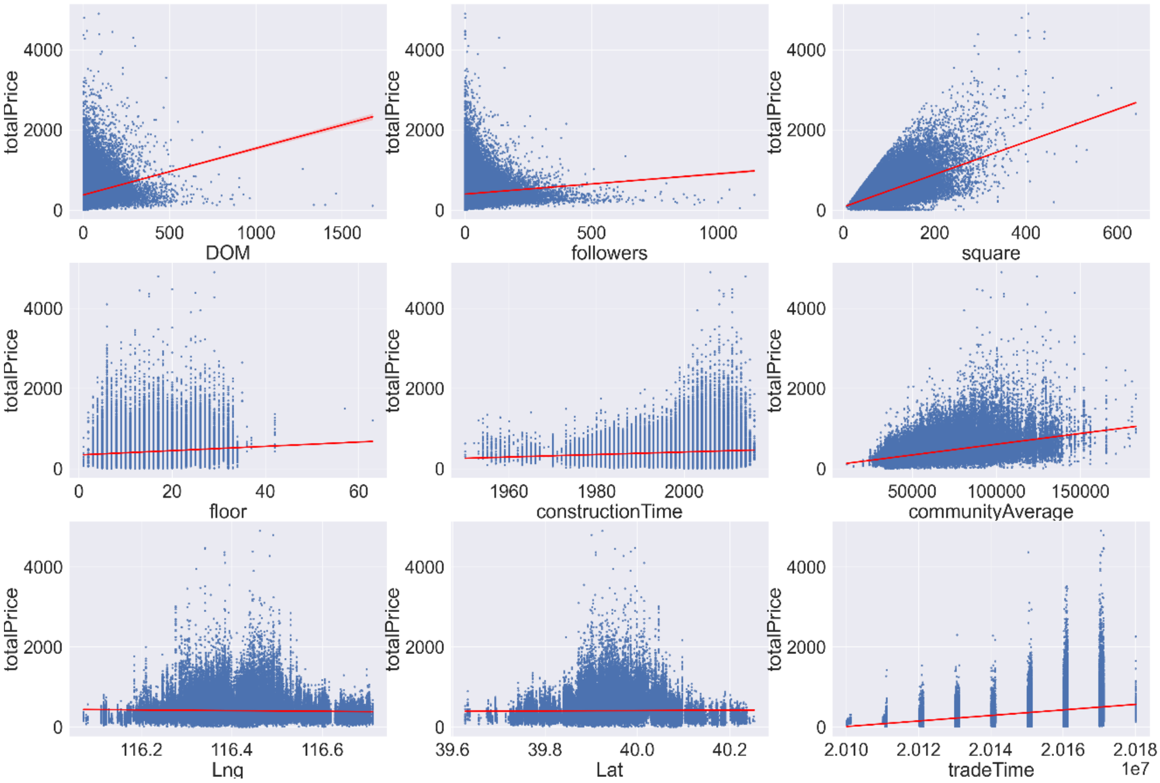


Figure 6- 5: Correlation analysis of the house price and the input features in the Chinese dataset

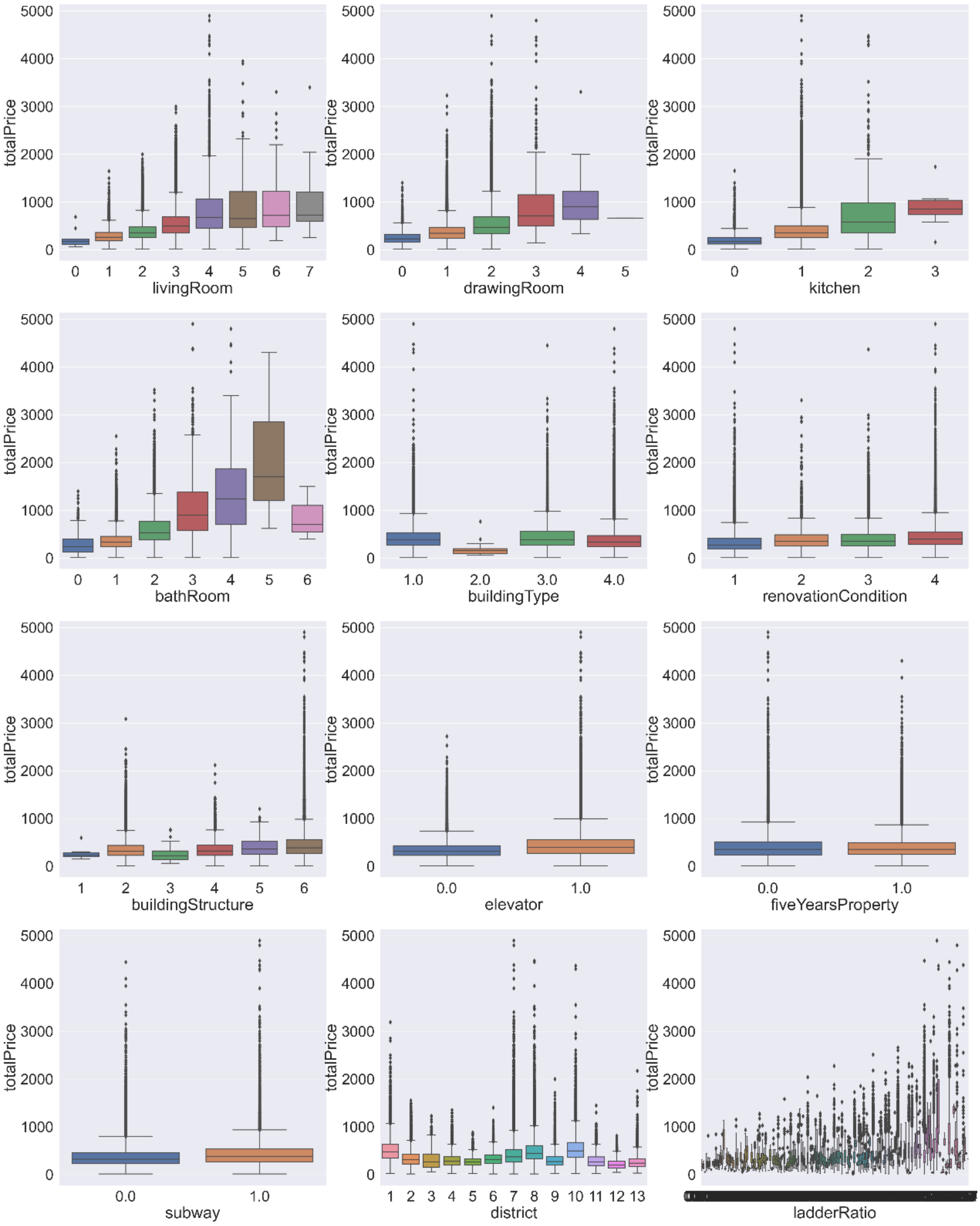


Figure 6- 6: Correlation analysis of the house price and the input features in the Chinese dataset

Figure 6-6 presents the box plots that represent the the relationship between the categorical variables and the house price. The y-axis shows the value of the total house price and the x-axis shows the value of individual categorical variables. The vertical dots show the distribution of obsearvations. Four variables, which are Living room, Drawing room, Kitchen, and Bathroom, show clear positive relationships with the house price. The other variables show the total price of different categories or groups in the related variables. For instance, there are 4 groups in the building type variable: (1) the tower group, (2) the bungalow group, (3) the combination of plate and tower group, and (4) the plate group. From the subfigure, it shows the combination of plate and tower group has the highest house price and the the bungalow group has the lowest house price. In the distrct variable, there are 13 groups: (1) the DongCheng district, (2) the FengTai district, (3) the DaXing district, (4) the FaXing district, (5) the FangShang district, (6) the ChangPing district, (7) the ChaoYang district, (8) the HaiDian district, (9) the ShiJingShan district, (10) the XiCheng district, (11) the TongZhou district, (12) the ShunYi district, and (13) the MenTouGou district. From the subfigure, it shows the XiCheng district (the closest district to the city centre) has the highest house price and the ShunYi district (the longest distance to the city centre) has the lowest house price.

2) The American data set

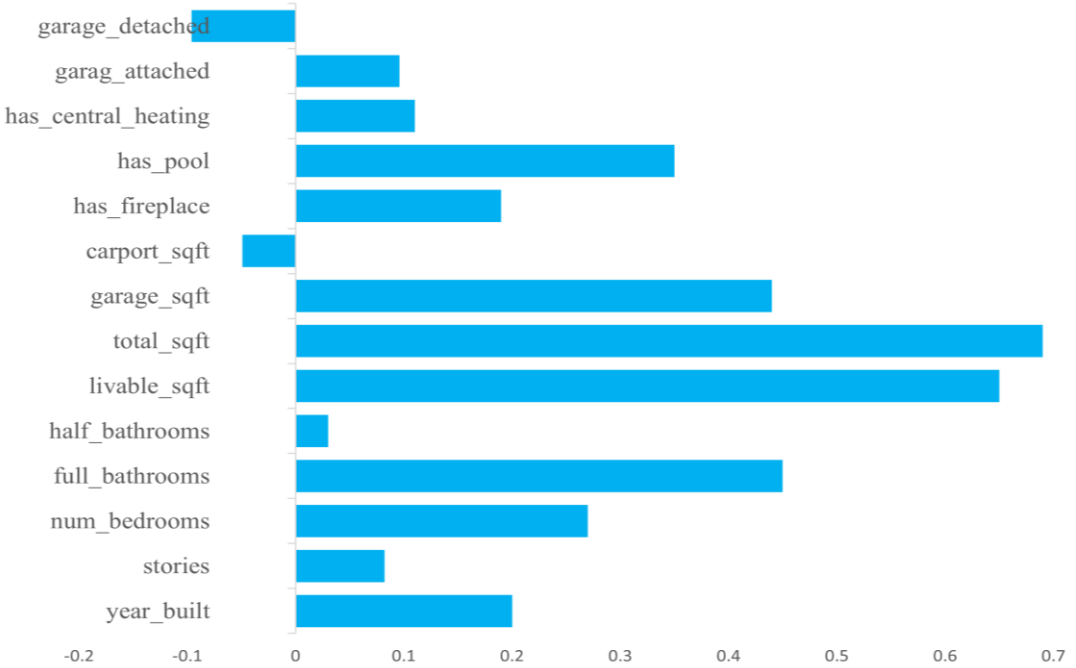


Figure 6- 7: Correlation analysis of the house price and the input features in the American dataset

Figure 6-7 illustrates the correlation between the input features and the target price in the American data set. While the Total area (*total_sqft*) variable shows the highest correlation with the house price, the Half bathrooms variable shows the lowest correlation with the house price. Five variables, which are Living area, Garage area, Full bathrooms, the Number of bedrooms, and Pool, show a relatively high positive correlation with the house price. Two variables, which are the Carport area (*carport_sqft*) and Detached Garage (*garage_type_detached*), show negative correlations with the house price.

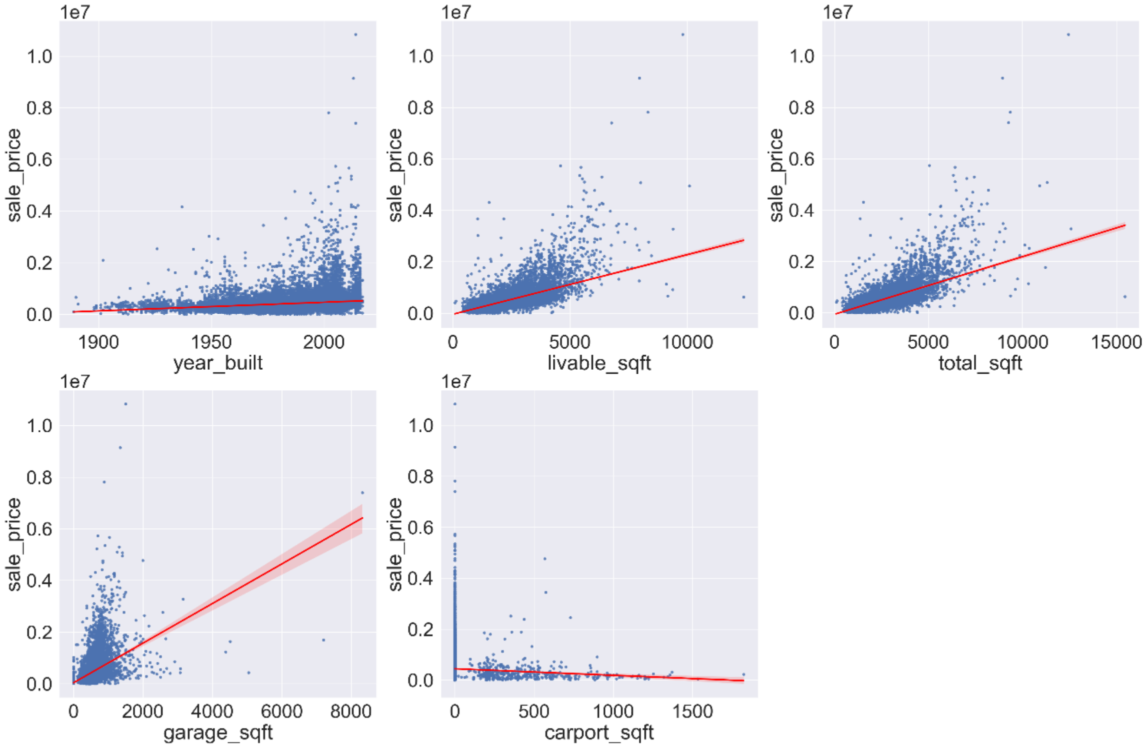


Figure 6- 8: Correlation analysis of the house price and continuous features in the American dataset

Figure 6-8 shows the linear relationship between the continuous variables (Built year, Living area, Total area, Garage area, and Carport area) and the house price by plotting data and a linear regression model fit. The y-axis shows the value of the total house price and the x-axis shows the value of individual continuous variables. The red lines show the linear fits of continuous variables, while the blue points show the actual data. Generally, a linear regression model fit figure would show the blue points around the red line, going up or down gradually. The plots in Figure 6-8 show the blue points are scattered around the red line, which means the relationship between the house price and input continuous features are complex and significantly non-linear. Three variables, which are Living area, Total area, and Garage area, show strong positive correlations with the house price, whereas the Built year feature shows a low positive correlation with the house price. It should be noticed that the Carport area feature shows a negative correlation with the house price.

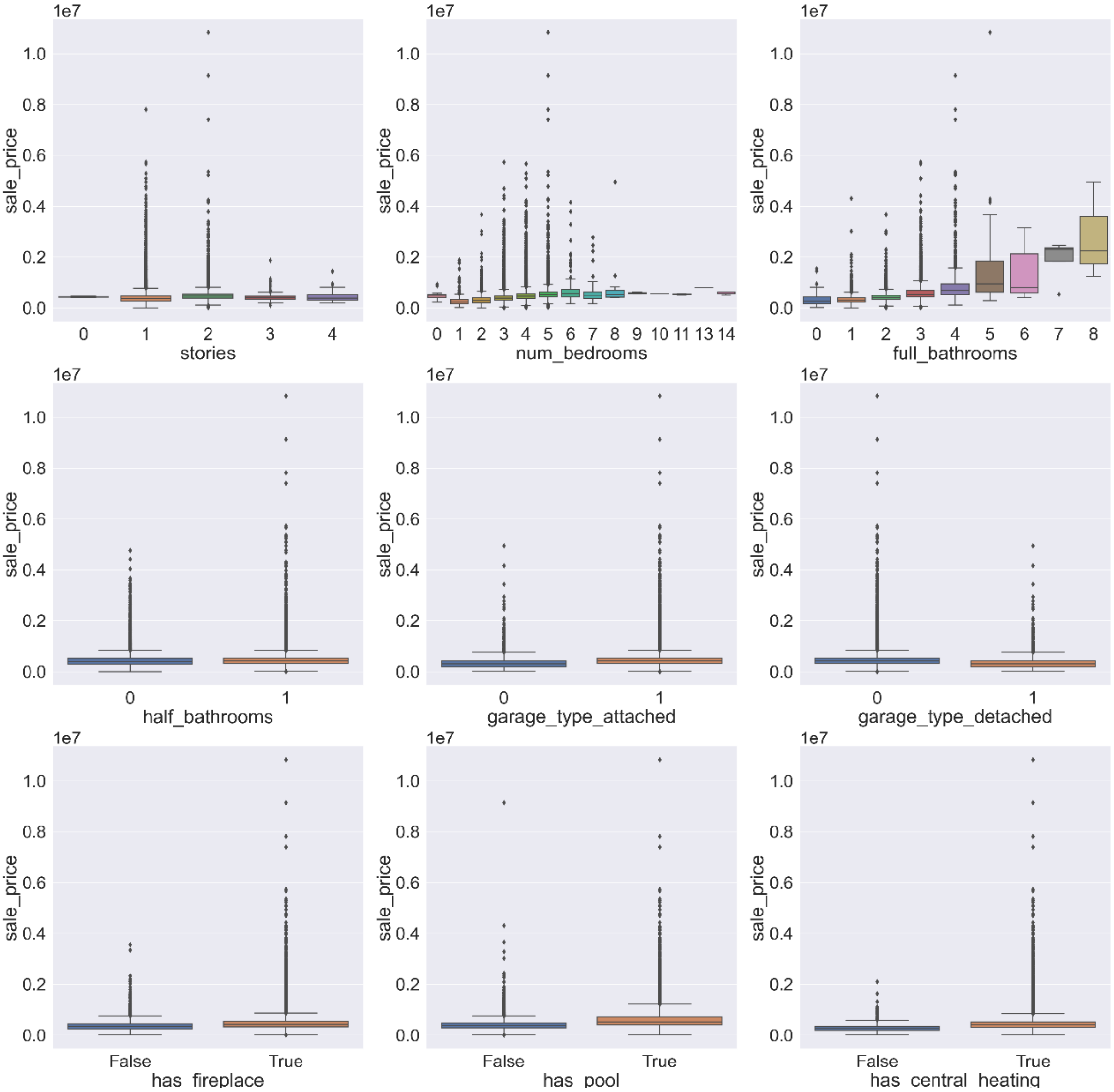


Figure 6- 9: Correlation analysis of the house price and categorical features in the American dataset

Figure 6-9 presents the box plots that represent the the relationship between the categorical variables and the house price. The y-axis shows the value of the total house price and the x-axis shows the value of individual categorical variables. The vertical dots show the distribution of obsearvations. Two variables, which are Number of bedrooms and Full bathrooms, show clear positive relationships with house price, whereas the Detached Garage (*garage_type_detached*) feature shows a negative relationship with the house price. A house with a fireplace, pool or central heating system shows a higher price.

To sum up, in the Chinese data set, seven variables, which are Total area, Living room, Drawing room, Bathroom, Kitchen, Trade time, and Community average price, show a relatively high positive correlation with the house price, of which the Total area variable shows the highest correlation with the house price. The Longitude variable shows a negative correlation with the house price. From the perspective of different building types, the combination of plate and tower group shows the highest house price and the the bungalow group shows the lowest house price. From the perspective of different districts in Beijing, the XiCheng district (the closest district to the city centre) shows the highest house price and the ShunYi district (the longest distance to the city centre) shows the lowest house price.

In the American data set, six variables, which are Total area, Living area, Garage area, Full bathrooms, the Number of bedrooms, and Pool, show a relatively high positive correlation with the house price, of which the Total area shows the highest correlation with the house price. A house with a fireplace, pool or central heating system shows a higher price. Two variables, which are the Carport area and Detached Garage, show negative correlations with the house price.

From the correlation analysis of the input features and the target price, it is revealed that the relationship between the input features and the target price are complex and significantly non-linear, and the Total area variable shows the strongest positive correlation with the house price in both the Chinese and American data sets. It is indicated that the complex relationships between the input features and the target price are difficult to be directly estimated using simple linear machine learning models such as Linear Regression and KNN, whereas complex models such as ANN and decision tree-based machine learning models have the potential to fit well with the data.

6.3.3 Feature selection

Feature engineering is a crucial step in machine learning pipeline that involves extracting features from input raw data and transforming them into suitable formats for machine learning algorithm requirements. The right feature engineering can get rid of non-important features, reduce the model complexity, contribute to model explanation, achieve the information gain, and improve the performance of machine learning models. The right features can only be defined in the context of both the model and the data, for data and models are so diverse that it is difficult to generalize the practice of feature engineering across projects (Zheng and Casari 2018).

There are three typical feature selection techniques: filtering methods, wrapper methods and embedded methods. Filtering methods are much cheaper than wrapper methods, but they are not connected to the model being employed. As missing useful features will generate a weak machine learning model, it is suggested to do prefiltering conservatively in order to avoid accidentally pruning away useful features before the model training stage (Zheng and Casari 2018). Wrapper methods are expensive but provide a quality score for a proposed subset of features. Embedded methods perform feature selection as part of the model training process. Compared to wrapper methods, embedded methods are less powerful but strike a balance between computational expense and quality of results. Considering the quality of feature selection and computational cost, the wrapper method (RFE) and embedded methods are selected in this project, to get an understanding of the optimal feature subsets in different data sets.

- **The wrapper method**

Recursive feature elimination (RFE) is a wrapper-type feature selection algorithm, which aims at finding the best performing feature subset by removing the least important features whose deletion will have the least effect on model generalization performance. While RFE starts from a complete dataset and then prune away the least relevant feature one by one to find the most important features, it is worth to mention that several useless weak features by themselves can provide a significant performance improvement when used together (Chen and Jeong 2007). An important hyperparameter for tuning RFE algorithm is the number of features to select. In practice, it is difficult to find the best number of features to select with RFE, so different values should be tested.

1) The Chinese data set

Figure 6-10 displays the relationship between model accuracy (R^2) and the number of features (original 21 input columns before one hot encoding) using the decision tree based RFE. After testing with different numbers of input features, it was observed that the model accuracy reaches the highest score with 9 features at 0.9381. The top 9 features selected by the decision tree - based RFE are Total area, Height, Trade time, Active days, Construction time, Community average price, Followers, Latitude, and Longitude.

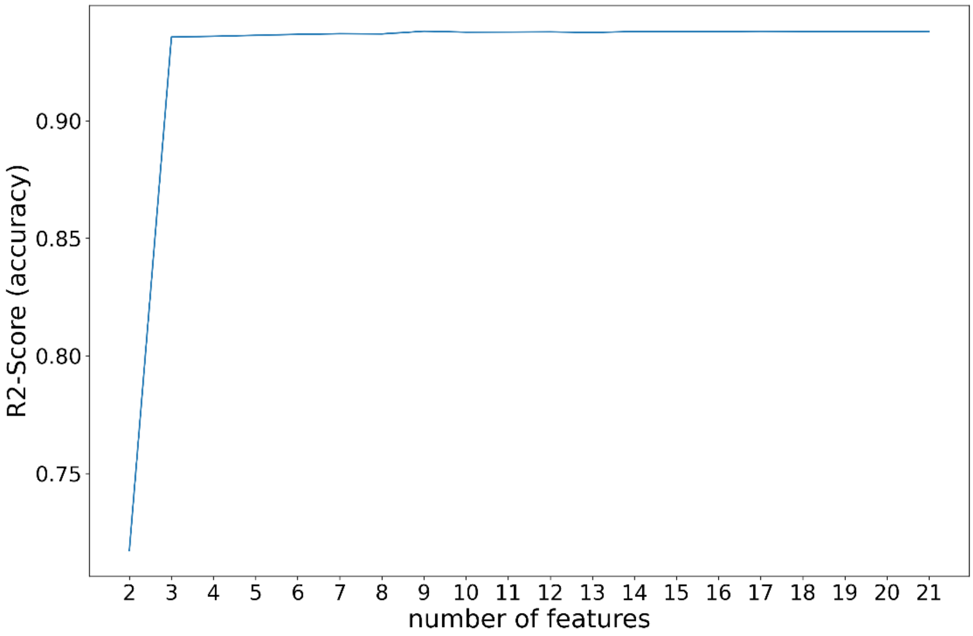


Figure 6- 10: The relationship between model accuracy (R^2) and the number of features using the RFE method in the Chinese dataset

2) The American data set

Figure 6-11 displays the relationship between model accuracy (R^2) and the number of features (original 15 input columns before one hot encoding) using the decision tree - based RFE. After testing with different numbers of input features, it was observed that the model accuracy reaches the highest score with 6 features at 0.8589. The 6 most important features selected by decision tree - based RFE are Total area, Living area, Garage area, Built year, Stories, and Pool.

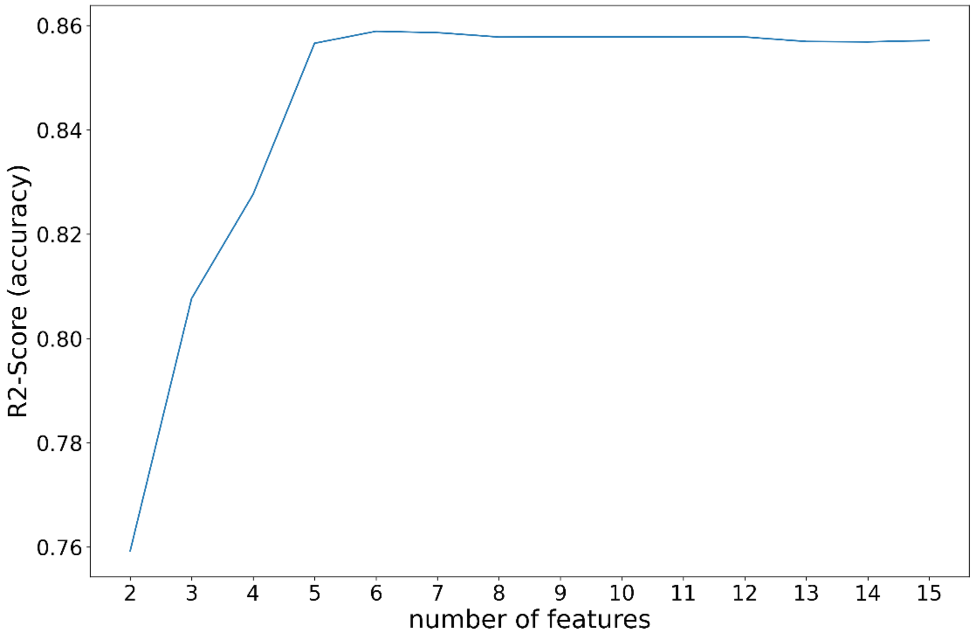


Figure 6- 11: The relationship between model accuracy (R^2) and the number of features using the RFE method in the American dataset

In Figure 6-10 and Figure 6-11, it is discovered that very good predictive accuracy (R-square) is already obtained with several dominant features. For instance, in Figure 6-10, the predictive accuracy of the top 3 features is almost as good as the top 9 features. In Figure 6-11, the predictive accuracy of the top 5 features is almost as good as the top 13 features. It seems that it is not worth spending massive efforts on using more features to have only a small percentage of improvement on predictive accuracy. However, this small improvement on predictive accuracy might actually have a big influence on the income of commercial companies or on human well-being. For instance, the small improvement on predictive accuracy of the AVM for a big real estate company (e.g. Zoopla) might bring 1% more transactions, this will increase the annual revenue with £31.7 million, referring to the annual revenue of Zoopla at £317.35 million on 2018 (Craft 2022). In order to reduce the carbon footprint, the UK innovation agency – Nesta did an experiment with 130 participants, the AI-based recommendation system for recommending healthier food with less calories can save an average of 3.9 kilos of carbon dioxide per participant.

If 1000 people made the similar choices of the healthier food recommended by the AI system, the saved carbon footage is equivalent to diving a petrol-powered car 27000 kilometres, which is two-thirds the way around the planet (Nesta 2022). In addition, research on AI for automatic driving and medical diagnosis is a popular trend in the past two decades, the predictive accuracy of AI on these areas related human life will always require to be further improved.

- ## The embedded method

The embedded method performs feature selection as part of training machine learning models. Typical embedded methods are based on various types of algorithms such as decision tree, logistic regression, and Lasso regression etc. Since the relationship between the target price and the input features are complex and significantly non-linear, it is difficult for classic linear models such as SVM, Linear regression, and KNN to fit the data. While complex neural networks might fit the data well but limited in the non-transparent nature known as 'block box'. Therefore, the decision tree – based embedded methods are selected including the GBDT, LightGBM, XGBoost, and Random Forest. The decision tree–based algorithms provide a feature importance ranking property (*feature_importances_*) which can calculate the relative importance scores for each input feature. The feature importance rankings calculated by the decision–tree based embedded methods are provided in the Appendix C.

1) The Chinese data set

Table 6-6 lists the feature importance ranking details with the four decision-tree based embedded methods and the average ranking of them in the Chinese data set. The four columns in the middle of the table list the feature importance rankings of each input feature, which are calculated by four different decision-tree based algorithms (GBDT, LightGBM, XGBoost, and random forest). Since the feature importance rankings are different with different algorithms, to get a generalized feature ranking, the last column on the right lists the feature importance ranking calculated by the average ranking of the four algorithms. The top 9 important features calculated by the average ranking are Community average price, Total area, Trade time, Bathroom, Active days, Living room, Latitude, District, and Longitude.

Table 6- 6: Feature importance ranking out of 21 calculated by four different decision tree-based embedded methods in the Chinese dataset

Feature	**GBDT**	**LightGBM**	**XGBoost**	**Random Forest**	**Average Ranking**
Total area	3	2	1	1	1.75
Community	1	1	2	2	1.5
Bathroom	6	12	3	3	6
Trade time	2	3	4	5	3.5
Active days	5	10	5	6	6.5
District	7	13	6	10	9
Building	15	21	7	12	13.75
Living room	4	11	8	4	6.75
Elevator	17	17	9	13	14
Latitude	8	4	10	9	7.75
Drawing room	9	14	11	7	10.25
Subway	16	18	12	18	16
Longitude	10	5	13	11	9.75
Renovation	18	15	14	17	16
Construction	14	6	15	16	12.75
Ladder ratio	13	9	16	8	11.5
Followers	11	8	17	14	12.5
Building	19	16	18	19	18
Kitchen	21	20	19	20	20
Property right	20	19	20	21	20
Height	12	7	21	15	13.75

Compare with the top 9 features selected by the wrapper method, the Bathroom, Living room and District features selected by the embedded method were replaced by the Height, Construction time and Followers features in the wrapper method, while the other six features remain the same. In total, there are 12 important features selected by the two methods, namely Community average price, Total area, Trade time, Bathroom, Active days, Living room, Latitude, District, Height, Construction time, Followers, and Longitude. There are 6 common features selected by the two methods, including Community average price, Total area, Trade time, Active days, Latitude, and Longitude. These 6 features are given more attention when trying to improve the model performance of the GA-GBR model in the experiment stage.

2) The American data set

Table 6-7 lists the feature importance ranking details with the four different decision-tree based embedded methods and the average ranking of them in the American data set. The four columns in the middle of the table list the feature importance rankings of each input feature, which are calculated by four different decision-tree based algorithms (GBDT, LightGBM, XGBoost, and random forest). The last column on the right displays the feature importance ranking calculated by the average ranking of the four algorithms. The top 6 important features calculated by the average ranking are Total area, Living area, Garage area, Built year, Full bathroom, and Pool. It is worth to mention that the Number of Bedrooms feature ranks the seventh in the average ranking column, which might be questioned that it should rank before the Garage area or the Full Bathroom features in real life. The possible reason behind this phenomenon is that the feature importance of the Living area feature has already reflected this impression that the weight of the Living area feature is much higher than that of the Number of Bedrooms feature.

Compare with the six most important features selected by the wrapper method - RFE, the Full Bathroom feature selected by the embedded method was replaced by the Stories feature in the wrapper method, whereas the other five features remain the same. In total, there are seven important features selected by the two methods, namely Total area, Living area, Garage area, Built year, Stories, Full Bathroom and Pool. There are 5 common features selected by the two methods, namely Total area, Living area, Garage area, Built year, and Pool. These five features are given more attention when trying to improve the model performance of the GA-GBR model in the experiment stage.

Table 6- 7: Feature importance ranking out of 15 calculated by four different decision tree-based embedded methods in the American dataset

Feature	GBDT	LightGBM	XGBoost	Random Forest	Average Ranking
Total area	1	4	1	2	2
Living area	2	3	3	1	2.25
Garage area	3	1	9	4	4.25
Built year	4	2	5	6	4.25
Number of	7	5	7	7	6.5
Pool	6	8	2	5	5.25
Full bathroom	5	6	6	3	5
Fireplace	8	10	8	8	8.5
Stories	9	7	4	9	7.25
Half bathroom	10	9	10	11	10
Garage	11	11	12	14	12
Central heating	12	14	15	12	13.25
Carport area	13	12	14	15	11
Central cooling	14	13	13	10	12.5
Garage	15	15	11	13	13.5

6.3.4 The GA-GBR model training

1) Experimental setup

The trainings of the GBR and GA-GBR models were performed on the Python 3.7 using the scikit-learn library on the *PyCharm* platform, which was an integrated development environment using python language for machine learning. Firstly, the GBR and GA-GBR models were experimented with the whole Chinese and American datasets respectively, trying different model hyperparameters to find the optimal solution of each model. Secondly, the two datasets were divided into different groups according to different perspectives, which allows the deep analysis of the relationship between the input features and the target price. For instance, the Chinese dataset was divided into 22 groups by different building categories (3 groups), building structures (3 groups), renovation conditions (3 groups), and districts (13 groups), with 1000 traded house data in each of them. The American dataset was divided into 23 groups by different cities (20 groups) and different types of garages (3 groups), with 1000 traded house data in each of them.

There are several model hyperparameters in a traditional GBR model, including the number of estimators, learning rate, maximum depth of decision trees, minimum sample leaf, and loss function. The hyperparameters tested with grid search algorithm are displayed as follows:

- Number of estimators: 80, 100, 150, 200, 500
- Learning rate: 0.01, 0.05, 0.1, 0.2
- Maximum depth: 4, 5, 6, 7, 8, 9
- Minimum sample leaf: 3, 4, 5, 7, 9
- Maximum features: 0.1, 0.2, 0.3, 1
- Loss function: Ls, Lad, Huber

2) GBR model training

After being cleaned, inspected, and prepared, the Chinese and American datasets were randomly split into the training set (70%) and the testing set (30%). Through grid search algorithm, the GBR model was experimented with the selected hyperparameter setup above.

For the two datasets (undivided), it was found that the experimental error of the GBR model was the smallest in both the Chinese and the American datasets when model hyperparameters were set as:

- Number of estimators (200)
- Learning rate (0.2)
- Maximum depth (7)
- Minimum sample leaf (5)
- Maximum features (0.2)
- Loss function (Huber)

For the divided datasets with 1000 samples, the optimal setting of the model hyperparameters were different in the Chinese and American datasets. In the divided Chinese datasets, it was found that the experimental error of the GBR model was the smallest when model hyperparameters were set as:

- Number of estimators (100)
- Learning rate (0.1)
- Maximum depth (4)
- Minimum sample leaf (4)
- Maximum features (0.1)
- Loss function (Huber)

In the divided American datasets, it was found that the experimental error of the GBR model was the smallest when model hyperparameters were set as:

- Number of estimators (135)
- Learning rate (0.2)
- Maximum depth (6)
- Minimum sample leaf (5)
- Maximum features (0.1)
- Loss function (Huber)

3) GA-GBR model training

Generally, the GA-GBR model was trained based on the framework designed in Section 5.3. The initial population was randomly generated with N solutions, with the number of base learners and their associated combination methods. After N solutions were randomly generated, the next generation of solutions was generated through the three genetic search operations. The fitness of each chromosome in the new generation was evaluated according to the fitness function - coefficient of determination (R^2), which was a regression accuracy measurement that explains how well a machine learning model fit the data. The chromosomes with higher R^2 scores than the GBR model were selected.

After testing with trial-and-error, it was found that the experimental error of the GA-GBR model was the smallest when the parameters in the genetic algorithm were set as follows:

- Population size: 600
- Generations: 32
- Crossover probability: 0.5
- Mutation rate: 0.1

The model performance of the trained GA-GBR model is presented and validated with three different datasets from China, US and the UK in the next chapter.

6.4 Conclusion

This chapter presented the development of the three components of the proposed system, including the detailed IFC extension for property valuation, the detailed information extraction, and the detailed GA-GBR model training. 7 property sets and 104 properties were proposed to add to the IfcSpace and the IfcZone entities as the IFC Property Valuation extension. The information extraction algorithm was developed using the instance-based approach and the open-source BIM information extraction library – *IfcOpenShell*. The development of the proposed GA-GBR model was divided into 4 main steps: (1) data collection and preparation, (2) exploratory data analysis, (3) feature selection using the wrapper method and the embedded method, and (4) the GA-GBR model training. The content of this chapter aimed to answer Research Question 4:

Q4: How to implement the BIM-ML integration framework and how to develop the three main components accordingly?

Answering this research question contributed to developing a prototype system that enabled automatic information exchange between AEC projects and property valuation and automated property valuation. The detailed answer to this research question is discussed further in the Chapter 7.

Chapter 7. System Testing and Validation

This chapter presents the validation of the developed BIM-ML system, which involves three steps: (1) validate the trained automated valuation model (GA-GBR), (2) validate the IFC-based information extraction as required, and (3) validate the proposed BIM-ML system as a complete artifact. Section 7.1 outlines the BIM-ML system testing objectives. Section 7.2–7.4 present the implementation and verification results, with the aim to prove that the BIM-ML system is functional and reliable when performing automated property valuation.

7.1 System Testing Objectives

The proposed BIM-ML system testing objectives in terms of the three validation tests are explained as follows.

1) Validation of the trained GA-GBR model

After training the GA-GBR model, it is necessary to use independent test data sets to evaluate the model predictive accuracy and generalization capability. Model performance metrics are essential in evaluating the predictive accuracy of statistical models. In the scientific community, a number of performance metrics have been defined and are currently in use for regression analysis, including the mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2) etc. The details of the five metrics have been introduced in the comparison experiment on the eleven different AVMs in Section 5.2.

While regression analysis plays an important role in supervised machine learning tasks, no consensus has been reached on a unified performance metric to assess the quality of the performance of a regression method and explain the mutual relation between the ground truth and the predict model. In general, MSE is more sensitive to outliers than MAE, and has an advantage to make a comprehensive assessment considering the effect of bias and variance. MAPE, which focuses on the percentage error, makes it suitable for evaluating tasks when relative variations being an important factor in the regression task. Compared to MAE, MAPE, MSE, and RMSE, a single value of the four metrics does not explain much on the performance of the regression model, while the coefficient of determination (R^2) has an advantage in terms of interpretability that it can indicate regression performances in an absolute manner. Taking into consideration of model predictive accuracy and model explanation capability, in this research the regression metrics for the assessment of the trained GA-GBR model has been focused on coefficient of determination (R^2) and mean squared error (MSE), with the results of other abovementioned metrics provided as well.

As explained in Table 7-1, the validation of the trained GA-GBR model has been conducted on three different datasets (from China, US, and UK) in two ways: validation on the three undivided datasets and validation on the divided datasets representing different perspectives. For the consistency, in the next, the divided datasets will be clearly mentioned, whereas it means the datasets are undivided if there are no additional statements.

There are two main objectives of testing the trained GA-GBR model on the three datasets: (1) compare model performances with different regression models such as linear regression, ridge regression, KNN, SVM, ANN, CART, random forest, XGBoost, LightGBM, GBR and the proposed GA-GBR; (2) identify the general features in the three datasets from different countries. The objective of testing on the divided datasets is to explore the implicit relationships between the input features and the target price from different perspectives, including perspectives from different building categories, building structures, renovation conditions, and districts in the Chinese data set; different cities and garages in the American data set; and different property types and Price Paid categories in the UK data set. The data descriptions of the Chinese and the American datasets have been introduced in Section 6.3.1, after which 70% of the two data sets had been used for training the GA-GBR model. The UK data set, works as a comparable group or control group for testing the generalization capability of the proposed GA-GBR, will be explained in Section 7.2.1 in the next.

Table 7- 1: Testing objectives of the trained GA-GBR on the three datasets

<table>
<tr><td rowspan="2"><b>Training on the three datasets</b></td><td colspan="4">Training</td><td rowspan="2"><b>Training on the divided datasets</b></td><td colspan="4">Training</td></tr>
<tr><td colspan="2">Chinese dataset (70%)</td><td colspan="2">US dataset (70%)</td><td colspan="2">Chinese dataset (70%)</td><td colspan="2">US dataset (70%)</td></tr>
<tr><td rowspan="2"><b>Test on the three datasets</b></td><td colspan="4">Testing</td><td rowspan="2"><b>Test on the divided datasets</b></td><td colspan="4">Testing</td></tr>
<tr><td>Chinese dataset (30%)</td><td colspan="2">U.S. dataset (30%)</td><td>UK dataset (100%)</td><td>Chinese dataset (30%)</td><td colspan="2">U.S. dataset (30%)</td><td>UK dataset (100%)</td></tr>
<tr><td><b>Objectives</b></td><td colspan="4">• Compare model performance with different regression models.
• Identify the general features in the big datasets from different countries.</td><td><b>Objectives</b></td><td colspan="4">Explore the relation between different input features and the target price from different perspectives, including views from different building categories, building structures, renovation conditions, districts, cities, garages, and Price Paid methods (PPD categories) etc.</td></tr>
</table>

2) Validation of the IFC-based information extraction as required

The validation of the IFC-based information extraction uses the case study method on three different BIM models from China, US, and the UK. Due to the limited time and resources, the value-related information was firstly achieved from the valuation reports of three real estate transaction companies and then translated into the three IFC-based BIM models. Based on the developed IFC Property Valuation extension in Section 6.1, the value-related information was added in the proposed property sets and properties in the IfcSpace and IfcZone entities according to the input features existed in the three data sets. There are 22, 15 and 9 input features in the three data sets from China, US and the UK. The proposed property sets were listed in Table B-1 in the Appendix B, including the Pset_PV_Transaction, Pset_PV_Parcel, Pset_PV_Building, Pset_PV_CondominiumUnit, Pset_PV_Valuation, Pset_PV_MassValuation, and Pset_PV_Annex.

After that, the required value-relevant information was extracted through the developed IFC-based information extraction algorithm.

3) Validation of the proposed BIM-ML system as a complete artifact

While the proposed GA-GBR model and the IFC-based information extraction were validated through the abovementioned steps, it is necessary to validate the two elements as a complete artifact. The validation of the whole BIM-ML system was conducted through a main python script, which was developed on the PyCharm platform, calling together the two functions including the IFC-based information extraction and the automated house price prediction.

7.2 Validation of the Automated Valuation Model (GA-GBR)

7.2.1 Introduction of the UK dataset

The UK dataset was collected from the UK Housing Prices Paid from the Kaggle website, which was originally released by HM Land Registry under the Open Government License 3.0 (HM Land Registry 2017). The dataset contains all the recorded individual house transactions in England and Wales between 1995 and 2017, to keep the same size with the Chinese and American datasets, 11018 traded individual houses were selected as one of the testing datasets. Each individual case contains ten property variables such as Price, Transaction unique identifier, Date of transfer, Property type, Old/new (the age of the property), Duration, Town/city, District, County, and PPDCategory type. The detailed descriptions of the 10 variables are displayed in the Table 7-2 below. While most of them are categorical variables, except for the Price variable with the average value at 117500£ and standard deviation at 7500£.

Table 7- 2: Property variables in the UK dataset

Variables	**Description**
Price	Transacted price (RMB)
Transaction unique identifier	A reference number which is unique and automatically generated when a sale is recorded.
Date of transfer	When the sale was completed.
Property type	Including detached house (D), which is a stand-alone building; semi-detached house (S), which shares one common wall with anther house, terraced (T), flats or Maisonettes (F), and other types (O).
Old/New	Indicates the age of the property: a newly built property (Y), an established residential building (N).
Duration	Related to the tenure: freehold (F), leasehold (L).
Town/City	Town or city
District	District
County	County
PPDCategory	Indicates the type of Price Paid transaction: (A) Standard Price Paid entry, includes single residential property sold for full market value. (B) Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage) and transfers to non-private individuals.

- **Feature selection - the wrapper method**

Figure 7-1 displays the relationship between model accuracy (R^2) and the number of features using the wrapper method - RFE. After testing with different numbers of input features, it was discovered that the model accuracy reached the highest score with 7 features at 0.8878. The top 7 features selected by the RFE method are Property type, Town/City, District, County, Year, Month, and Day.

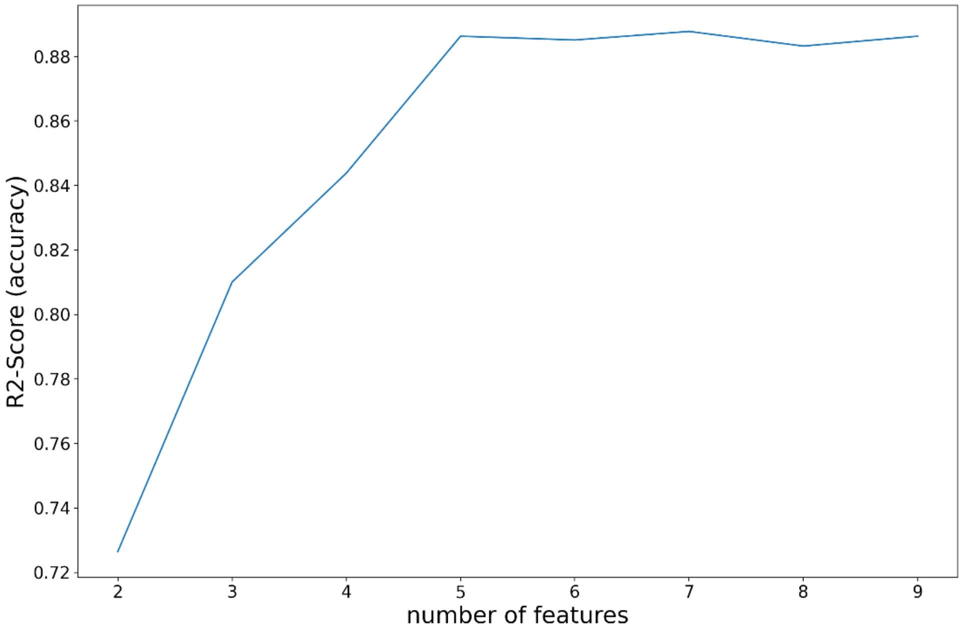


Figure 7- 1: The relationship between model accuracy (R^2) and the number of input features using the wrapper method - RFE in the UK dataset

- **Feature selection - the embedded method**

Table 7-3 lists the feature importance ranking details with the four decision-tree based embedded methods and the average ranking of them in the UK dataset. The four columns in the middle of the table list the feature importance rankings of each input feature, which are calculated by four decision-tree based algorithms (GBDT, LightGBM, XGBoost, and random forest). Since the feature importance rankings are different with different algorithms, to get a generalized feature ranking, the last column on the right explains the feature importance ranking calculated by the average ranking of the four algorithms. The top 7 important features calculated by the average ranking are Year, Town/City, County, Property type, District, Duration, and Day.

Compare with the top 7 features selected by the wrapper method, the Duration feature selected by the embedded method were replaced by the Month feature in the wrapper method, while the other 6 features remain the same. There are 6 common features selected by the two methods, including Year, Town/City, County, Property type, District, and Day.

Table 7- 3: Feature importance ranking out of 10 calculated by four different decision tree-based embedded methods in the UK dataset

Feature	GBDT	LightGBM	XGBoost	Random Forest	Average Ranking
Year	1	2	1	1	1.25
Town/City	5	5	3	2	3.75
County	2	4	4	3	3.25
Property type	3	7	2	4	4
District	4	1	6	5	4
Duration	6	8	5	6	6.25
Day	7	3	7	7	6
Month	8	6	8	8	7.5
Old/New	9	9	9	9	9
PPDCategory type	10	10	10	10	10

7.2.2 Validation on the three undivided datasets

The performance of the trained GA-GBR model was firstly evaluated on the three undivided datasets, comparing the model performances of the 12 different regression models including the proposed GA-GBR and identifying the general features from the three different countries.

1) Testing on the Chinese dataset

In Figure 7-2, the red line explains the average R^2 of each generation during the genetic search process in the GA-GBR model. The predictive accuracy R-square increases from generation to generation and shows a convergence around Generation 28. The R-square score of the 600 individual chromosomes in Generation 32 is illustrated in Figure 7-3, which shows the similar increasing trendline.

After that, the best chromosome with the highest R^2 was selected for testing model predictive accuracy. In terms of coefficient of determination (R^2), the model accuracy of GA-GBR had an advantage of 1.3% over the GBR model, with 95.2% for GA-GBR and 93.9% for GBR respectively.

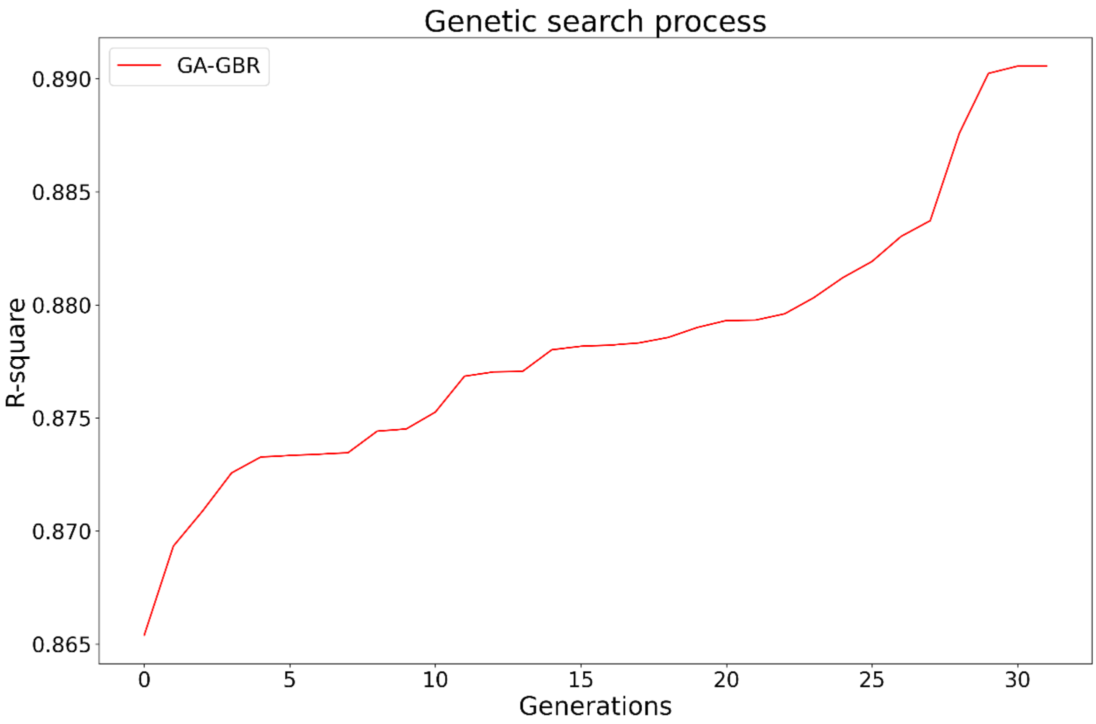


Figure 7- 2: The average R^2 score of each generation during the genetic search process

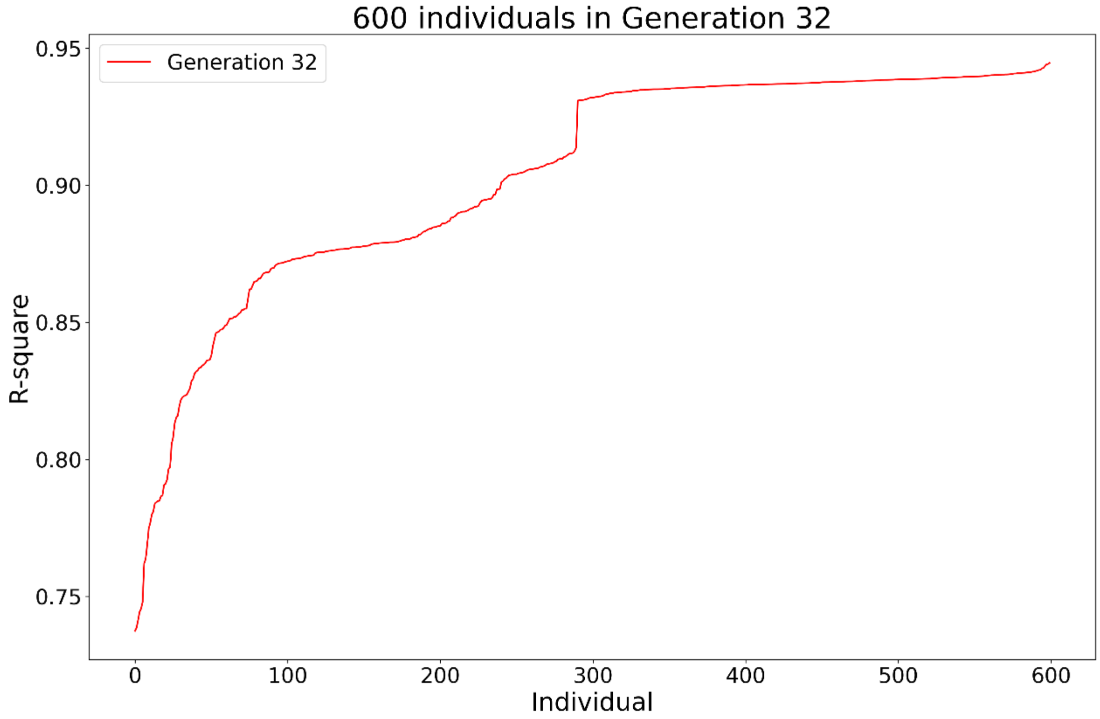


Figure 7- 3: The R^2 score of the 600 individuals of Generation 32 during the genetic search process

Table 7-4 lists the five different accuracy metrics introduced in the first section of this chapter, of which the MAE, MAPE, MSE, and RMSE are measuring regression models in terms of different types of errors and the coefficient of determination (R^2) is measuring regression models in terms of prediction accuracy. This means that a better model performance requires a lower MAE, MAPE, MSE, and RMSE, and a higher R-squared (R^2). From the experiments on the test dataset, it is observed that the four linear regression models have similar general model performances, with the mean MAE at 77.14, the mean MAPE at 26.1%, the mean MSE at 13236, the mean RMSE at 114.8, and the mean R^2 at 80%. The KNN, SVM and ANN models have unsatisfactory model performances, with the MAE ranges from 129.19 to 171.28, the MAPE ranges from 35.5% to 52.9%, the MSE ranges from 40468 to 69810, the RMSE ranges from 200.7 to 263.7, and the R^2 ranges from 5.3% to 39%. The decision tree-based models have good model performances, with the mean MAE at 43.88, the mean MAPE at 13.96%, the mean MSE at 5406.2, the mean RMSE at 71.86, and the mean R^2 at 92%. It is worth to mention that the proposed GA-GBR model has the highest predictive accuracy of the 12 listed models, with the MAE at 35.71, the MAPE at 11.3%, the MSE at 3703, the RMSE at 60.9, and the mean R^2 at 95.2%.

Table 7- 4: Predictive accuracy of the 12 different AVMs in the Chinese data set

Accuracy metrics	**MAE**	**MAPE**	**MSE**	**RMSE**	$\mathbf{R^2}$
Linear regression	76.79	26.0%	13042	114	80.3%
Ridge regression	76.75	26.0%	13033	113.9	80.3%
Lasso regression	77.11	26.1%	13265	114.9	80.0%
Elastic Net regression	77.89	26.3%	13602	116.4	79.5%
KNN	129.19	35.5%	40468	200.7	39.0%
SVM	171.28	51.4%	69810	263.7	5.3%
ANN	156.73	52.9%	51625	227.2	22.2%
CART	54.28	15.7%	8610	90.4	86.9%
AdaBoost	52.89	18.5%	6119	77.9	90.8%
Random forest	38.65	12.0%	4545	66.7	93.2%
GBR	37.85	12.3%	4054	63.4	93.9%
GA-GBR (proposed)	**35.71**	**11.3%**	**3703**	**60.9**	**95.2%**

To understand the trained GA-GBR model, it is necessary to find out which features are in the optimal input feature subset and to what extent each feature has contributed to the house price prediction. The GBR and GA-GBR, both are decision tree–based models, provide a feature importance ranking property for this purpose. The detailed feature importance rankings calculated by the GBR and GA-GBR models are illustrated in the Appendix D. Referring to the number of input features selected by the two feature selection methods in Section 6.3.3, Table 7-5 displays the top 9 features calculated by the GBR and GA-GBR models. The total feature importance of the top 9 features selected by the GBR and GA-GBR account for 87.32% and 91.02% of all the 56 input features respectively.

Compare with the six common important features selected by the two feature selection methods in Section 6.3.3, all the six common features contributed to the GBR model: Total area, Trade time, Active days, Community average price, Latitude, and Longitude; and five of them contributed to the GA-GBR model: Total area, Trade time, Active days, Latitude, and Longitude. This gives five generic important features in the Chinese datasets, namely Total area, Trade time, Active days, Latitude, and Longitude.

Table 7- 5: Feature importance ranking (top 9 out of 56) calculated by the GBR and GA-GBR in the Chinese dataset

GBR feature	**Ranking**	**GA-GBR feature**	**Ranking**
communityAverage	23.18%	*constructionTime*	25.06%
square	21.25%	*tradeTime*	23.46%
tradeTime	19.37%	*followers*	18.33%
DOM	6.07%	*buildingStructure_2*	8.92%
Lat	4.25%	*Lat*	4.10%
bathRoom	3.81%	*square*	3.06%
livingRoom_1	3.58%	*Lng*	3.02%
district_10	3.39%	*DOM*	2.72%
Lng	2.42%	*buildingStructure_3*	2.35%
Total feature importance	87.32%	**Total feature importance**	91.02%

Compare the top 9 features selected by the GBR and GA-GBR models, it was discovered that the evolutionary feature selection engine in the proposed GA-GBR model had changed the weights of the input features, which make it more suitable for generating a good machine learning model. For instance, the GA-GBR model has reduced the weight of the Active days (*DOM*) feature from 6.07% to 2.72% and increased the weight of the Trade time (*tradeTime*) feature from 19.37% to 23.46%. As introduced in Section 6.3.3 - feature selection, the right features can only be defined in the context of both the model and the data, for data and models are so diverse that it is difficult to generalize the practice of feature engineering across projects (Zheng and Casari 2018). In the context of the GA-GBR model and the Chinese dataset, the experiment results indicated the Construction time, Trade time, and Follower features are more important to generate a good decision-tree based machine learning model.

2) Testing on the American dataset

In Figure 7-4, the red line explains the average R^2 of each generation during the genetic search process in the GA-GBR model. The predictive accuracy R-square increases from generation to generation and shows a convergence around Generation 28. The R-square score of the 600 individual chromosomes in Generation 32 is illustrated in Figure 7-5, which shows the similar increasing trendline.

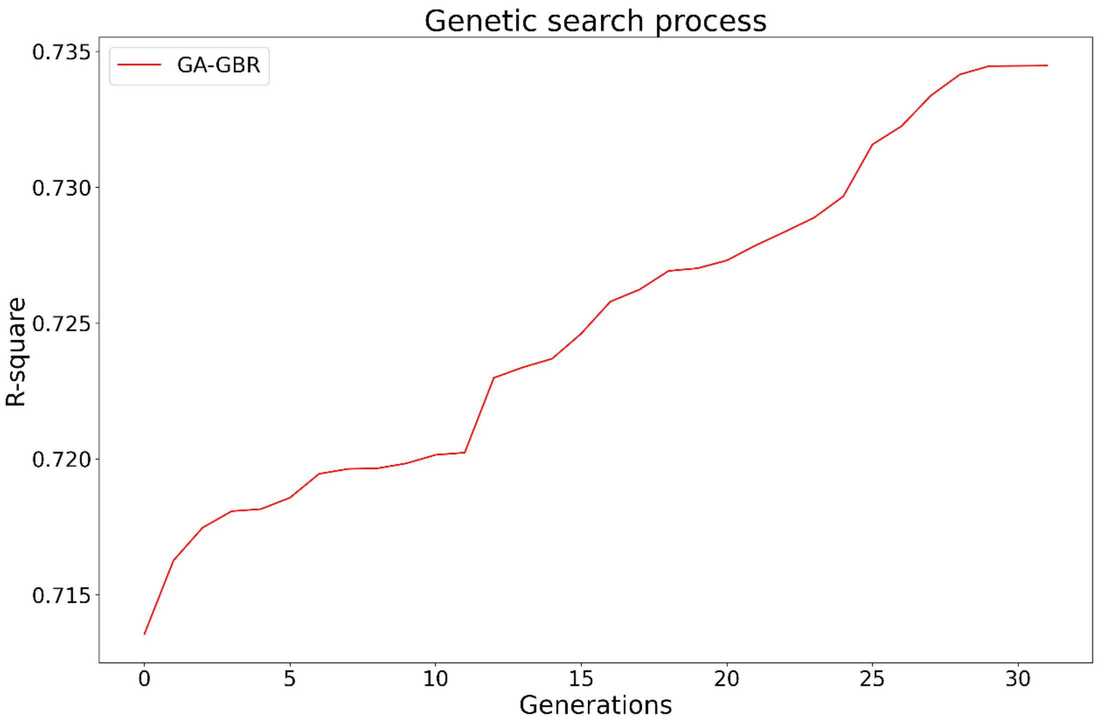


Figure 7- 4: The average R^2 score of each generation during the genetic search process

After that, the best chromosome with the highest R^2 was selected for testing model predictive accuracy. In terms of coefficient of determination (R^2), the model accuracy of GA-GBR had an advantage of 3.57% over the GBR model, with 82.5% for GA-GBR and 78.92% for GBR respectively.

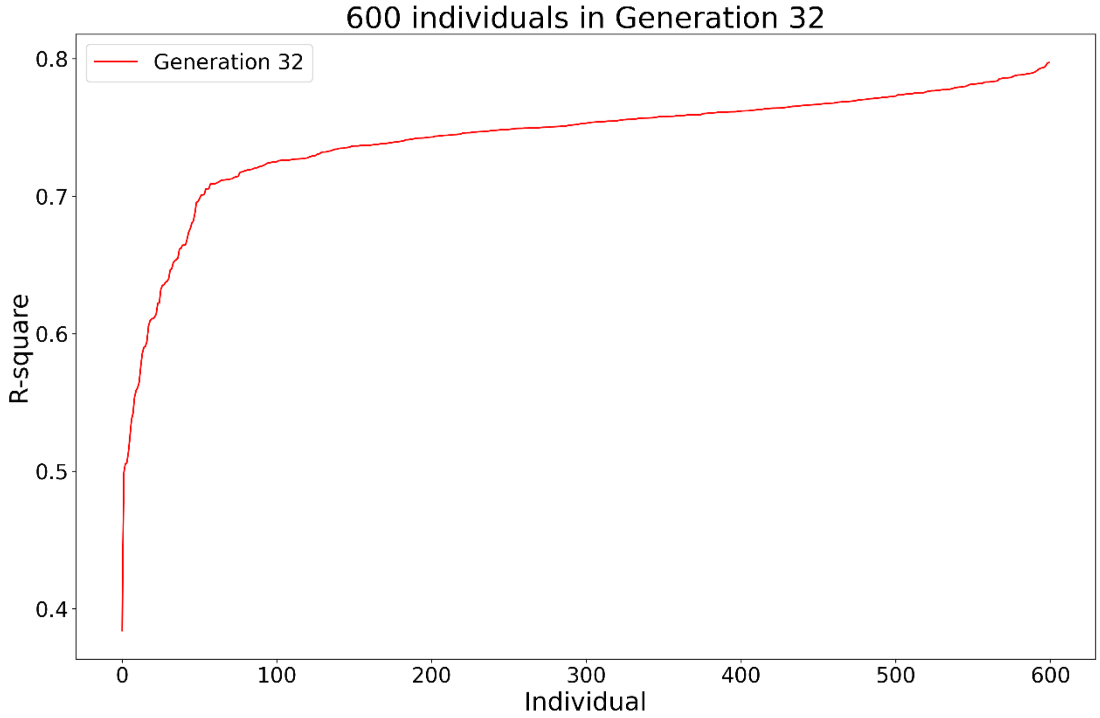


Figure 7- 5: The R^2 score of the 600 individuals of Generation 32 during the genetic search process

Table 7-6 lists the five different accuracy metrics introduced in the first section of this chapter, of which the MAE, MAPE, MSE, and RMSE are measuring regression models in terms of different types of errors and the R-squared (R^2) are measuring regression models in terms of prediction accuracy. From the experiments on the test dataset, it is observed that the four linear regression models have similar general model performances, with the mean MAE at 101982, the mean MAPE at 30.9%, the mean MSE at 3.87e10, the mean RMSE at 192855, and the mean R^2 at 60.5%. Compared to the linear regression models, the ANN has a relatively better model performance with the MAE at 82297, the MAPE at 24.8%, the MSE at 2.84e10, the RMSE at 168508, and the mean R^2 at 70.3%.

However, the SVM has the poorest model performance with the MAE at 108669, the MAPE at 33.6%, the MSE at 4.39e10, the RMSE at 205520, and the mean R^2 at 55.1%. The decision tree-based models have good model performances, with the mean MAE at 71678, the mean MAPE at 26.3%, the mean MSE at 2.94e10, the mean RMSE at 164781, and the mean R^2 at 72.4%. It is worth to mention that the proposed GA-GBR model has the highest model prediction accuracy with the MAE at 64356, the MAPE at 20.8%, the MSE at 2.08e10, the RMSE at 144341, and the mean R^2 at 82.5%.

Table 7- 6: Predictive accuracy of the 12 different AVMs in the American dataset

Accuracy metrics	**MAE**	**MAPE**	**MSE**	**RMSE**	R^2
Linear regression	97006	29.2%	3.67e10	187653	62.7%
Ridge regression	96985	29.2%	3.67e10	187653	62.7%
Lasso regression	97006	29.2%	3.67e10	187652	62.7%
Elastic Net regression	116929	35.9%	4.45e10	208461	53.8%
KNN	108669	33.6%	4.39e10	205520	55.1%
SVM	159601	49.3%	9.78e10	310218	2.4%
ANN	82297	24.8%	2.84e10	168508	70.3%
CART	84572	26.6%	4.26e10	193374	58.3%
AdaBoost	101000	37.1%	3.03e10	171939	68.3%
Random forest	67318	22.8%	2.52e10	155155	74.2%
GBR	71144	24.1%	2.79e10	159098	78.9%
GA-GBR (proposed)	**64356**	**20.8%**	**2.08e10**	**144341**	**82.5%**

To understand the trained GA-GBR model, it is necessary to find out which features are in the optimal input feature subset and to what extent each feature has contributed to the house price prediction. The detailed feature importance rankings calculated by the GBR and GA-GBR models are illustrated in the Appendix D. Referring to the number of input features selected by the two feature selection methods in Section 6.3.3, Table 7-7 displays the top 6 features calculated by the GBR and GA-GBR models. The total feature importance of the top six features selected by the GBR and GA-GBR account for 70.68% and 73.72% of all the 61 input features respectively.

Compare with the five common important features selected by the two feature selection methods in Section 6.3.3, four of them contributed to the GBR model: Total area, Living area, Built year, and Pool; and two of them contributed to the GA-GBR model: Total area and Living area. This gives two generic important features in the American datasets, namely the Total area and Living area features.

Table 7- 7: Feature importance ranking (top 6 out of 61) calculated by the GBR and GA-GBR in the American dataset

GBR feature	**Ranking**	**GA-GBR feature**	**Ranking**
total_sqft	37.01%	*total_sqft*	23.68%
livable_sqft	13.61%	*livable_sqft*	20.28%
city_Coletown	6.43%	*has_fireplace*	9.12%
full_bathrooms	5.13%	*garage_type_detached*	7.81%
year_built	4.59%	*garage_sqft*	6.57%
has_pool	3.91%	*num_bedrooms*	6.26%
Total feature importance	70.68%	**Total feature importance**	73.72%

Compare the top 6 features selected by the GBR and GA-GBR models, it was discovered that the evolutionary feature selection engine in the proposed GA-GBR model had changed the weights of the input features, which make it more suitable for generating a good machine learning model. For instance, the GA-GBR model has reduced the total weight of the Total area (*total_sqft)* and Living area (*livable_sqft)* features from 50.62% to 43.96% and increased the weight of the Number of bedrooms (*num_bedrooms)* feature from 2.51% to 6.26% and garage-related features (*garage_type_detached* and *garage_sqft*) from 3.38% to 14.38%. In the context of the proposed GA-GBR model and the American dataset, the experiment results indicated that the Number of bedrooms and the garage-related features is more important to generate a good machine learning model than that in the traditional GBR.

3) Testing on the UK dataset

In Figure 7-6, the red line explains the average R^2 of each generation during the genetic search process in the GA-GBR model. The predictive accuracy R-square increases from generation to generation and shows a convergence around Generation 28. The R-square score of the 600 individual chromosomes in Generation 32 is illustrated in Figure 7-7, which shows the similar increasing trendline.

After that, the best chromosome with the highest R^2 was selected for testing model predictive accuracy. In terms of coefficient of determination (R^2), the model accuracy of GA-GBR had an advantage of 2.4% over the GBR model, with 75.6% for GA-GBR and 73.2% for GBR respectively.

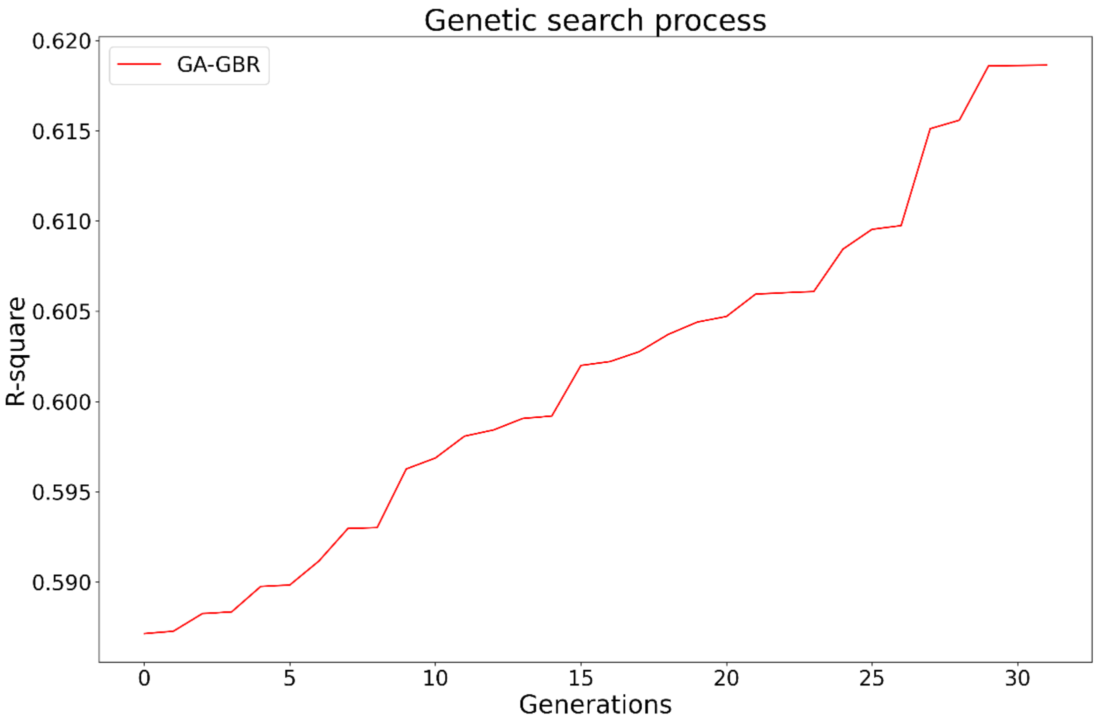


Figure 7- 6: The average R^2 score of each generation during the genetic search process

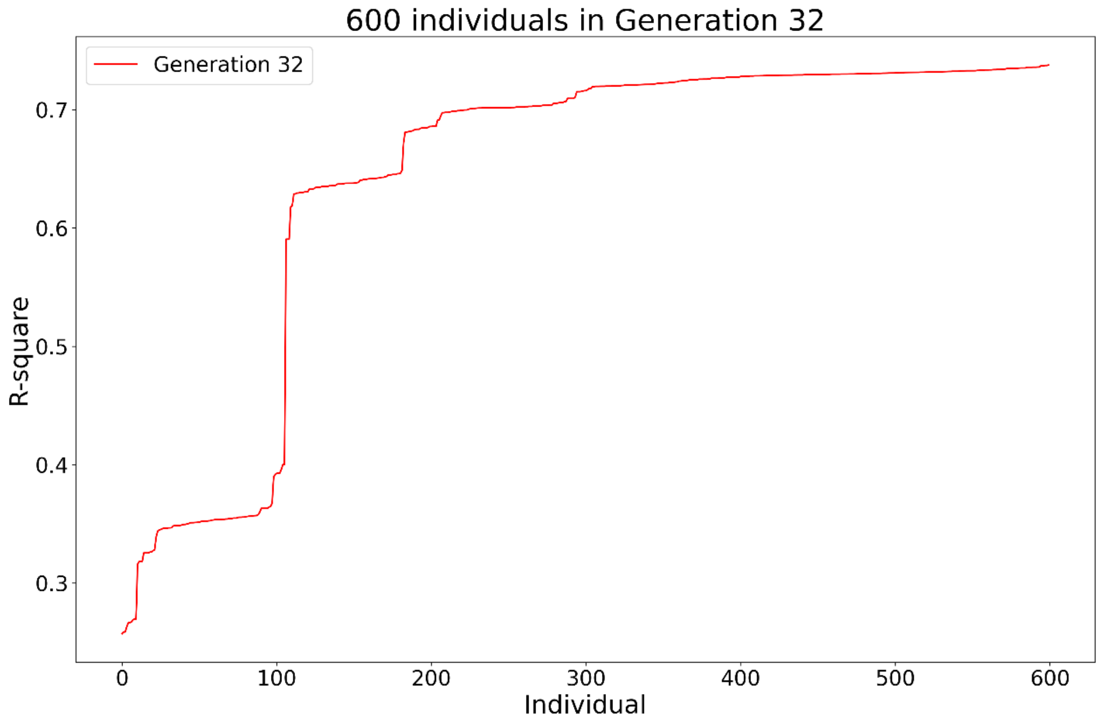


Figure 7- 7: The R^2 score of the 600 individuals of Generation 32 during the genetic search process

Table 7-8 lists the five different accuracy metrics introduced in the first section of this chapter, of which the MAE, MAPE, MSE, and RMSE are measuring regression models in terms of different types of errors and the R-squared (R^2) are measuring regression models in terms of prediction accuracy. From the experiments on the test dataset, it is observed that the four linear regression models have similar general model performances, with the mean MAE at 0.52, the mean MAPE at 4.5%, the mean MSE at 0.44, the mean RMSE at 0.65, and the mean R^2 at 11.56%. Compared to the linear regression models, the ANN has a relatively better model performance with the MAE at 0.38, the MAPE at 3.2%, the MSE at 0.23, the RMSE at 0.48, and the mean R^2 at 58.3%. The decision tree-based models have good model performances, with the mean MAE at 0.34, the mean MAPE at 3.0%, the mean MSE at 0.21, the mean RMSE at 0.45, and the mean R^2 at 66.8%. It is worth to mention that the proposed GA-GBR model has the highest model prediction accuracy with the MAE at 0.31, the MAPE at 2.7%, the MSE at 0.16, the RMSE at 0.41, and the mean R^2 at 75.6%.

Table 7- 8: Predictive accuracy of the 12 different AVMs in the UK dataset

Accuracy metrics	**MAE**	**MAPE**	**MSE**	**RMSE**	R^2
Linear regression	0.43	3.7%	0.3	0.55	46.0%
Ridge regression	0.43	3.7%	0.3	0.55	0.08%
Lasso regression	0.60	5.2%	0.6	0.75	0.08%
Elastic Net regression	0.60	5.2%	0.56	0.75	0.08%
KNN	0.43	3.7%	0.31	0.56	44.5%
SVM	0.48	4%	0.38	0.62	32.0%
ANN	0.38	3.2%	0.23	0.48	58.3%
CART	0.37	3.2%	0.27	0.48	59.3%
AdaBoost	0.37	3.2%	0.22	0.48	59.4%
Random forest	0.33	2.8%	0.19	0.43	66.4%
GBR	0.34	2.9%	0.19	0.43	73.2%
GA-GBR (proposed)	**0.31**	**2.7%**	**0.16**	**0.41**	**75.6%**

To understand the trained GA-GBR model, it is necessary to find out which features are in the optimal input feature subset and to what extent each feature has contributed to the house price prediction. The detailed feature importance rankings calculated by the GBR and GA-GBR models are illustrated in the Appendix D. Referring to the number of input features selected by the two feature selection methods in Section 7.2.1, Table 7-9 displays the top 7 features calculated by the GBR and GA-GBR models. The total feature importance of the top 7 features selected by GBR and GA-GBR account for 94.03% and 98.9% of all the 17 input features respectively.

Compare with the 6 common important features selected by the two feature selection methods in Section 7.2.1, all the 6 common features contributed to the GBR model, and

5 of them contributed to the GA-GBR model. This gives 5 generic features in the UK dataset, namely Year, Town/City, County, Property type, and District.

Compare the top 7 features selected by the GBR and GA-GBR models, it was discovered that the evolutionary feature selection engine in the proposed GA-GBR model had changed the weights of the input features, which make it more suitable for generating a good machine learning model. For instance, the GA-GBR model has reduced the weight of the Property_Type_is__D feature from 11.16% to 0.16% and increased the weight of the month feature from 1.2% to 13.90%. In the context of the proposed GA-GBR model and the UK dataset, the experiment results indicated that the month feature is more important to the GA-GBR model, which generated a higher model predictive accuracy than that in the traditional GBR model.

Table 7- 9: Feature importance ranking (top 7 out of 17) calculated by the GBR and GA-GBR in the UK dataset

GBR feature	**Ranking**	**GA-GBR feature**	**Ranking**
year	45.83%	*year*	47.86%
County	15.20%	*County*	17.76%
Property_Type_is__D	11.16%	*month*	13.90%
Town/City	10.82%	*Town/City*	8.99%
District	7.11%	*District*	7.73%
Property_Type_is__T	2.58%	*Property_Type_is__F*	1.71%
day	1.33%	*Property_Type_is__D*	0.96%
Total feature importance	94.03%	**Total feature importance**	98.9%

From the testing results on the Chinese, American and UK datasets, it was concluded by the author as follows:

(1) From the comparison of different regression model accuracy metrics on the 12 AVMs including the proposed GA-GBR model, in general, linear regression models does not have a good model fit on all the three datasets. The KNN and SVM have not achieved a satisfactory model performance on the Chinese and American datasets but show better performance than the linear models on the UK dataset. The ANN preforms better than the linear regression models on the American and UK datasets, but shows disadvantage on the Chinese dataset. The decision-tree based models generally have better performances than all other models with decent predictive accuracy scores, in which, the proposed GA-GBR model has the highest predictive accuracy.

(2) During the genetic search process, all the 32 GA generations had a higher predictive accuracy (R^2) over the traditional GBR model, with an advantage of predictive accuracy at 1.3% on the Chinese dataset, 3.57% on the American dataset, and 2.4% on the UK dataset. This proves the proposed GA-GBR model not only has achieved a decent predictive accuracy, but also has a high generalization capability for property valuation.

7.2.3 Validation on the divided datasets representing different perspectives

The relationships between the GA-GBR model and the input features will be further explored in this section. The three big datasets were divided into different groups according to eight different perspectives, with 1000 house transaction data in each of them. For instance, the Chinese dataset was divided into 22 groups by different building categories (3 groups), building structures (3 groups), renovation conditions (3 groups), and districts (13 groups). The American dataset was divided into 23 groups by different cities (20 groups) and different types of garages (3 groups). The UK dataset was divided into 6 groups by different property types (4 groups) and different types of Price Paid transactions (2 groups).

The predicted price by the GBR and GA-GBR models was compared with the actual price using different datasets from different perspectives. From Figure 7-8 to Figure 7-15, the bar chart showed the predicted values of house price by the two models and the actual price in the three groups, while the line chart on the top explained the price difference of the two models by using the regression metrics MAPE, which is often calculated as a percentage:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{14}$$

where A_t is the actual value and F_t is the predicted value.

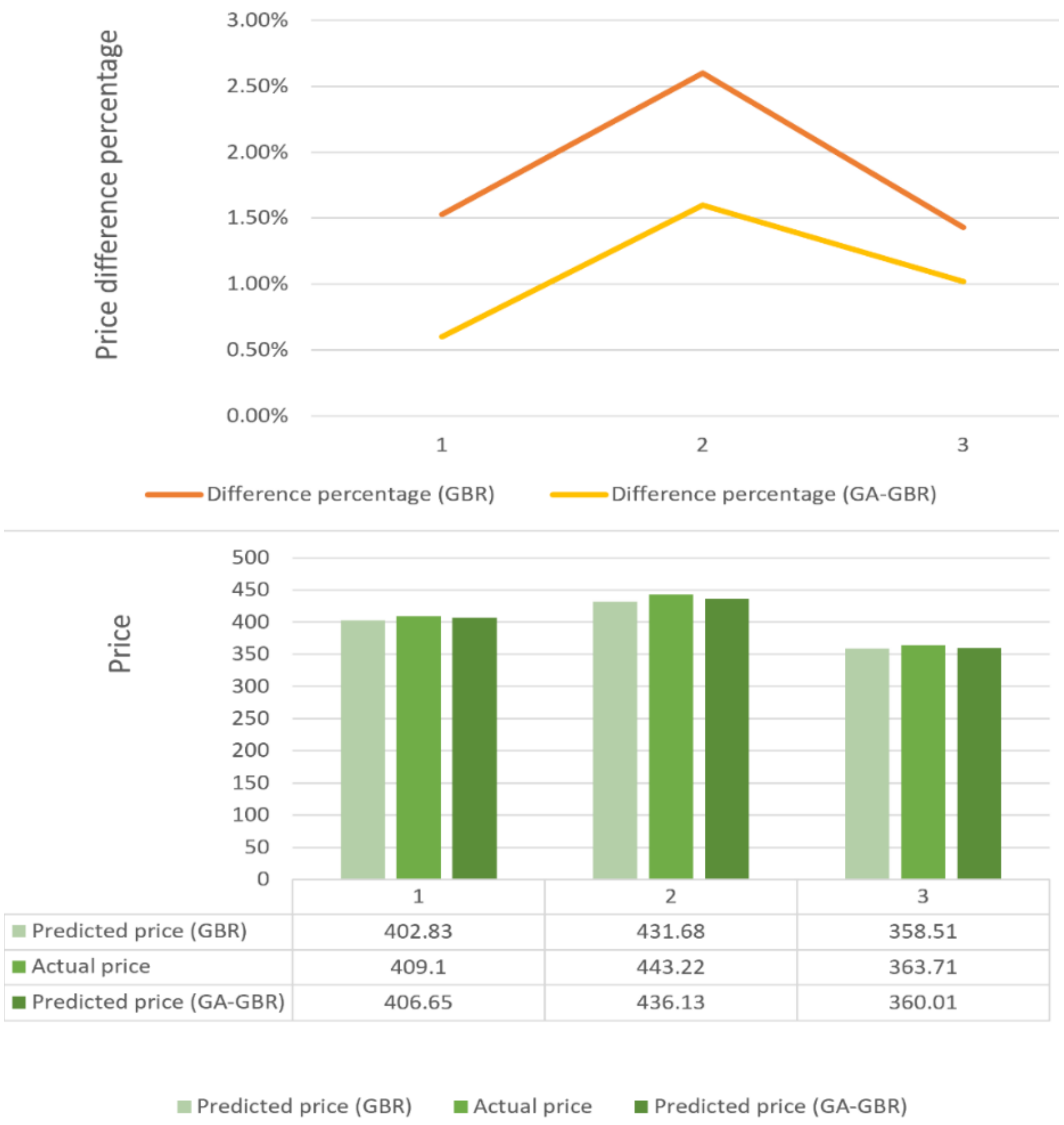
1) Testing on the Chinese dataset

The Chinese dataset was tested from four different perspectives including different building categories, building structures, districts, and renovation conditions.

- *From the perspective of different building categories*

The predicted price by the GBR and GA-GBR models was firstly compared with the actual price using the three different building-category-related datasets: (1) the tower group, (2) the combination of plate and tower group, and (3) the plate group. From Figure 7-8, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE at 0.6%, 1.6% and 1.02% in the three different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.

To understand the predictive performance improvement of the GA-GBR model, Table 7-10 listed the top three important features to the two models in the three groups, which were calculated using the embedded feature importance function in the decision tree-based models. In group 1 (the tower group), the second important feature – *CommunityAverage* (the community average price) in the GBR model was replaced by the feature – *Elevator* in the GA-GBR model, with other two features weights changed: 21.11% to 14.58% for the *Square* feature and 11.66% to 13.47% for the Trade time



	1	2	3
🟩 Predicted price (GBR)	402.83	431.68	358.51
🟩 Actual price	409.1	443.22	363.71
🟩 Predicted price (GA-GBR)	406.65	436.13	360.01

Figure 7- 8: Comparison of the actual price and predicted price by the two models in the three building-category-related datasets

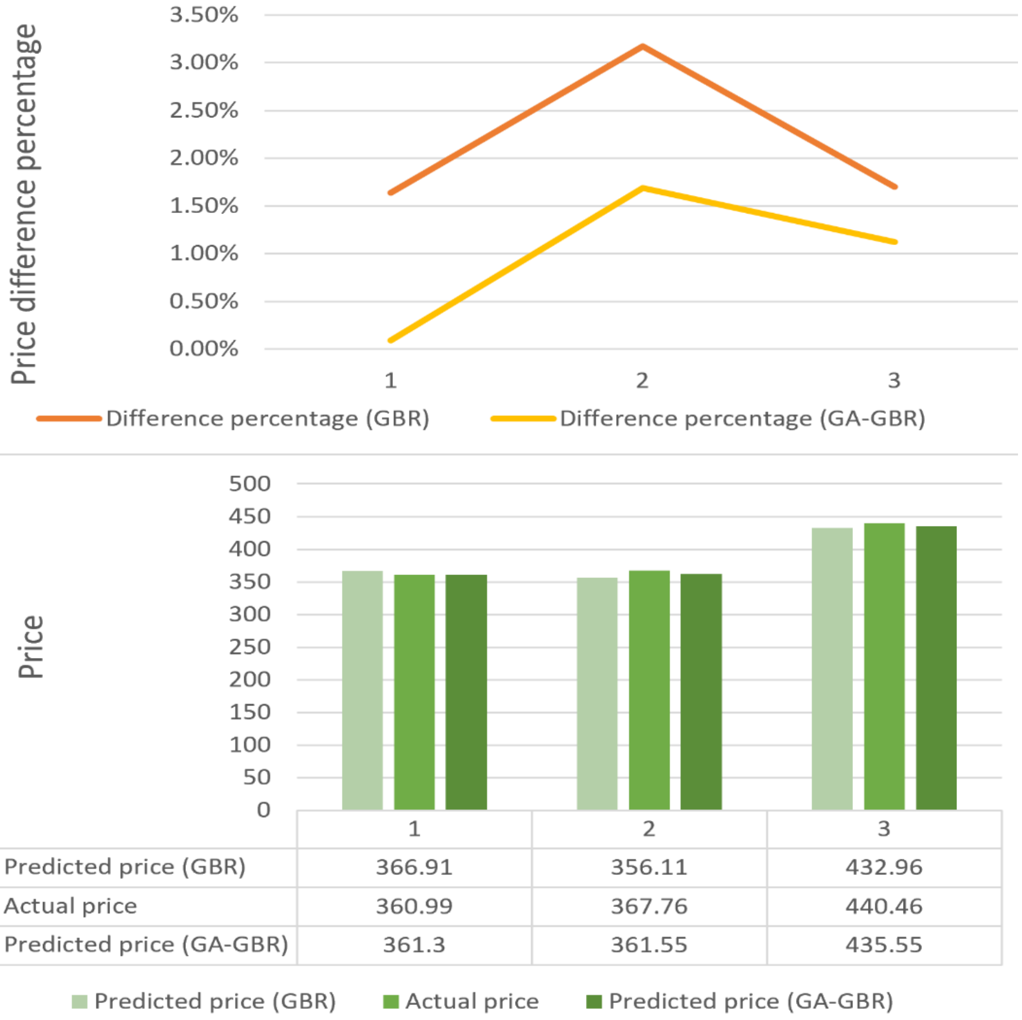
(*TradeTime)* feature. In group 2 (the combination of plate and tower group), the *Square* and *CommunityAverage* features in the GBR model were replaced by the *DOM* (active days on market) *and Floor* (the height of houses) in the GA-GBR model. In group 3 (the plate group), the *Square* and *CommunityAverage* features in the GBR model were replaced by the *DOM* (active days on market) *and ConstructionTime* (the time of construction) in the GA-GBR model. It was discovered that the Trade Time feature was important to both the GA-GBR and GBR in all the three groups. For the tower building type, the Elevator feature was considered more important to the GA-GBR model than the GBR model. For the plate building type, the DOM (active days on market) and Construction Time features were considered more important to the GA-GBR than the GBR model.

Table 7- 10: Feature importance ranking (top 3) by GBR and GA-GBR in the three building-category-related datasets

Tower		**Plate and tower**		**Plate**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
Square	21.11%	*Square*	19.74%	*CommunityAverage*	18.82%
CommunityAverage	13.67%	*CommunityAverage*	16.69%	*TradeTime*	15.07%
TradeTime	11.66%	*TradeTime*	14.76%	*Square*	13.36%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
Square	14.58%	*DOM*	26.85%	*ConstructionTime*	25.52%
Elevator	13.94%	*TradeTime*	25.81%	*DOM*	16.86%
TradeTime	13.47%	*Floor*	14.67%	*TradeTime*	12.72%

- *From the perspective of different building structures*

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the three different building-structure-related datasets: (1) the mixed group, (2) the brick and concrete group, and (3) the steel-concrete composite group. In Figure 7-9, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE at 0.09%, 1.69% and 1.12% in the three different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.



	1	2	3
🟩 Predicted price (GBR)	366.91	356.11	432.96
🟩 Actual price	360.99	367.76	440.46
🟩 Predicted price (GA-GBR)	361.3	361.55	435.55

Figure 7- 9: Comparison of the actual price and predicted price by the two models in the three building-structure-related datasets

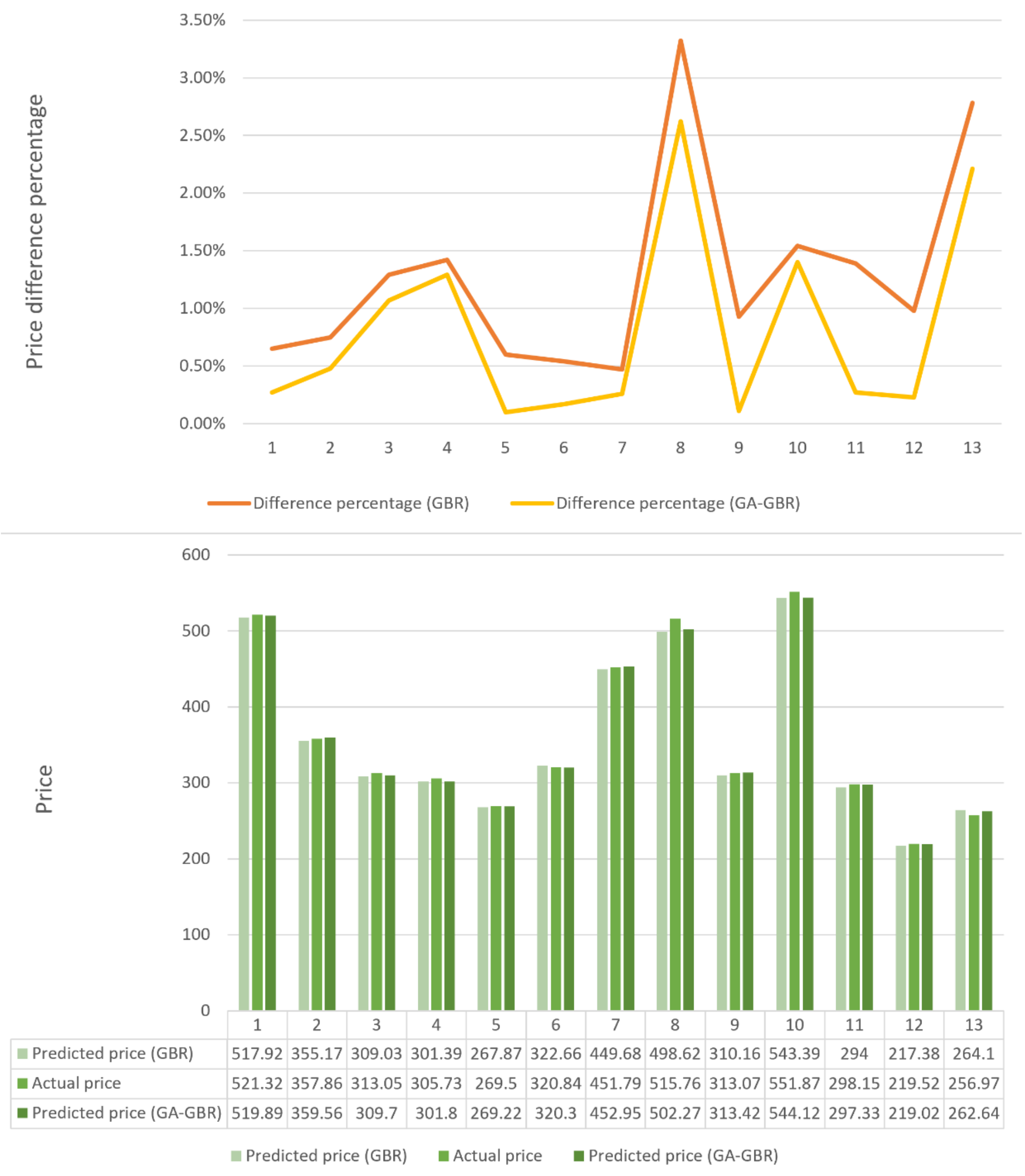
To understand the predictive performance improvement of the GA-GBR model, Table 7-11 listed the top three important features to the two models in the three groups, which were calculated using the embedded feature importance function in the decision tree-based models. In group 1 (the mixed group), the second important feature – *CommunityAverage* (the community average price) in the GBR model was replaced by the feature – *Lng* (Longitude) in the GA-GBR model, with other two features weights changed: 8.6% to 13.15% for the *DOM* (active days on market) feature and 27.39% to 27.12% for the *TradeTime* feature. In group 2 (the brick and concrete group), the *Square* and *CommunityAverage* features in the GBR model were replaced by the *DOM* (active days on market) *and ConstructionTime* (the time of construction) in the GA-GBR model. In group 3 (the steel-concrete composite group), the *Square* and *CommunityAverage* features in the GBR model were replaced by the *Followers and LadderRatio* in the GA-GBR model.

It was discovered that the Trade Time feature was important to both the GA-GBR and GBR in all the three groups. For the brick and concrete group, the Construction Time feature was considered more important to the GA-GBR model than the GBR. For the steel-concrete composite group, the Followers and LadderRatio features were considered more important to the GA-GBR than the GBR.

Table 7- 11: Feature importance ranking (top 3) by GBR and GA-GBR in the three building-structure-related datasets

Mixed		**Brick and concrete**		**Steel-concrete composite**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
TradeTime	27.39%	*TradeTime*	23.69%	*CommunityAverage*	20.04%
CommunityAverage	22.92%	*CommunityAverage*	23.68%	*Square*	14.31%
DOM	8.6%	*Square*	8.02%	*TradeTime*	11.5%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
TradeTime	27.12%	*ConstructionTime*	22.63%	*Followers*	26.02%
Lng	14.31%	*DOM*	12.54%	*TradeTime*	21.16%
DOM	13.15%	*TradeTime*	10.83%	*LadderRatio*	19.33%

- *From the perspective of different districts*



	1	2	3	4	5	6	7	8	9	10	11	12	13
Predicted price (GBR)	517.92	355.17	309.03	301.39	267.87	322.66	449.68	498.62	310.16	543.39	294	217.38	264.1
Actual price	521.32	357.86	313.05	305.73	269.5	320.84	451.79	515.76	313.07	551.87	298.15	219.52	256.97
Predicted price (GA-GBR)	519.89	359.56	309.7	301.8	269.22	320.3	452.95	502.27	313.42	544.12	297.33	219.02	262.64

Figure 7- 10: Comparison of the actual price and predicted price by the two models in the 13 district-related datasets

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the 13 different district-related datasets: (1) the DongCheng district, (2) the FengTai district, (3) the DaXing district, (4) the FaXing district, (5) the FangShang district, (6) the ChangPing district, (7) the ChaoYang district, (8) the HaiDian district, (9) the ShiJingShan district, (10) the XiCheng district, (11) the TongZhou district, (12) the ShunYi district, and (13) the MenTouGou district. In Figure 7-10 above, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE in all the 13 different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.

Table 7- 12: Feature importance ranking (top 3) by GBR and GA-GBR in the four district-related datasets

XiCheng district		**FengTai district**		**ChaoYang district**		**ShunYi district**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
square	23.82%	*square*	28.29%	*square*	23.11%	*tradeTime*	28.46%
tradeTime	21.44%	*tradeTime*	16.10%	*community Average*	12.98%	*square*	18.22%
DOM	10.30%	*DOM*	11.32%	*tradeTime*	12.07%	*DOM*	11.38%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
Lng	33.72%	*Lat*	50.32%	*followers*	27.37%	*Lat*	25.76%
Lat	20.91%	*ladderRatio*	11.35%	*floor*	16.44%	*followers*	19.05%
square	9.32%	*Lng*	9.03%	*Lat*	14.03%	*tradeTime*	19.05%

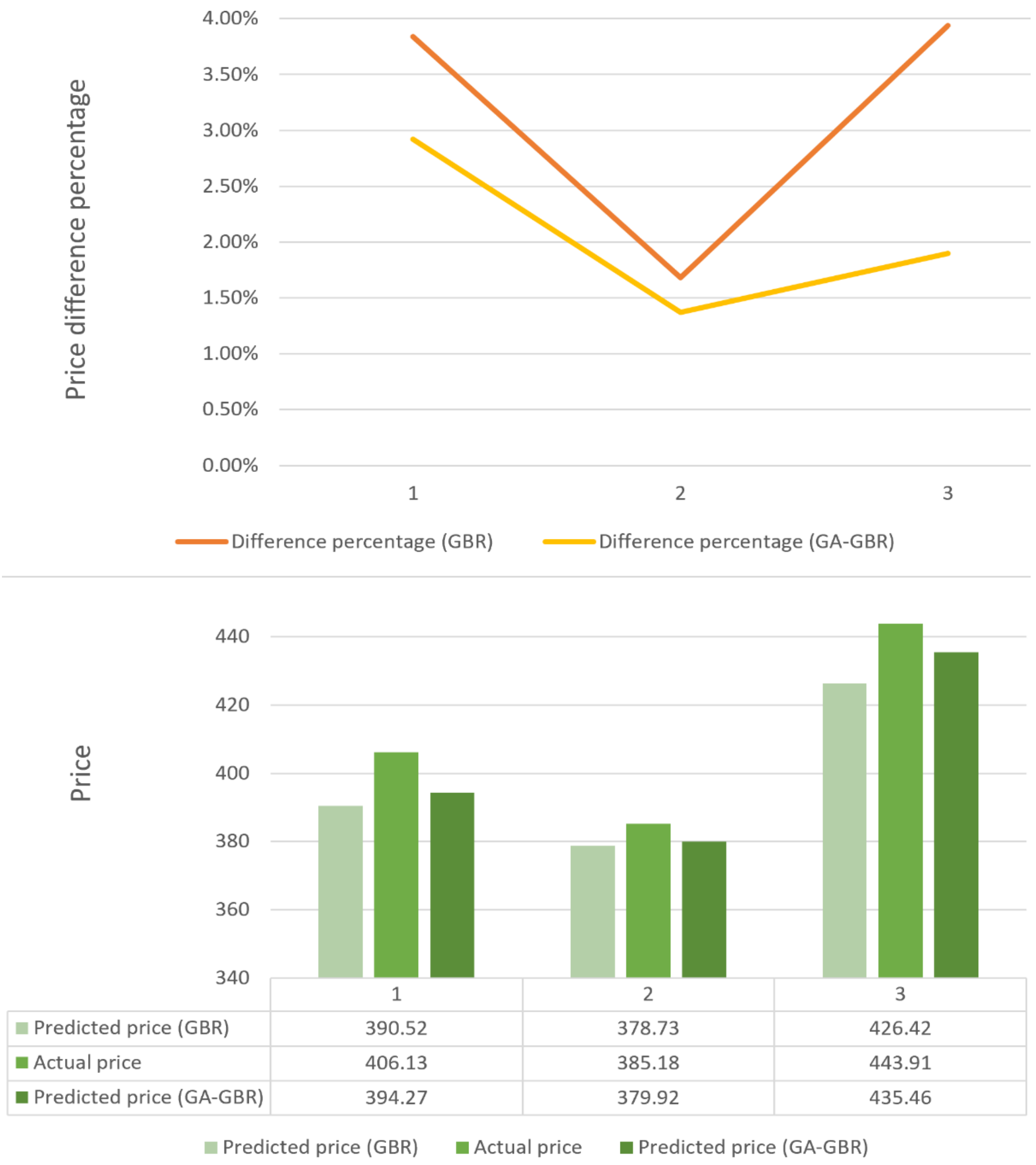
To understand the predictive performance improvement of the GA-GBR model, Table 7-12 listed the top three important features to the two models in the four groups including the XiCheng district (group 10), the FengTai district (group 2), the ChaoYang district (group7), and the ShunYi district (group 12), in which the XiCheng district is the closest one to the centre of Beijing, the ShunYi district is the farthest one to the centre of Beijing, and the other two are in the middle. In the XiCheng district, the *tradeTime* and *DOM* (active days on market) features in the GBR model were replaced by the *Lng and Lat* features in the GA-GBR model. In the ShunYi district, the *square* and *DOM* features in the GBR model were replaced by the *followers and Lat* features in the GA-GBR model. In the FengTai district, the *square, tradeTime* and *DOM* features in the GBR model were replaced by the *Lng, ladderRatio and Lat* features in the GA-GBR model. In the ChaoYang district, the *square, tradeTime* and *communityAverage* features in the GBR model were replaced by the *followers, floor and Lat* features in the GA-GBR model.

It was discovered that in the XiCheng and FengTai districts, the Longitude *(Lng)* and Latitude *(Lat)* features (the location related features) were considered more important to the GA-GBR model than the GBR. In the ShunYi district, the Followers and TradeTime features were considered more important to the GA-GBR model than the GBR.

- *From the perspective of different renovation conditions*

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the three different renovation-related datasets: (1) the rough group, (2) the simplicity group, and (3) the hardcover group. In Figure 7-11, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE at 2.92%, 0.14% and 1.37% in the three different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.

To understand the predictive performance improvement of the GA-GBR model, Table 7-13 listed the top three important features to the two models in the three groups, which were calculated using the embedded feature importance function in the decision tree-based models. In group 1 (the rough group), the most important feature – *square* in the GBR model was replaced by the feature – *bathroom* in the GA-GBR model. In group 2 (the simplicity group), the *Square* and *CommunityAverage* features in the GBR model were replaced by the *ladderRatio and followers* features in the GA-GBR model. In group 3 (the hardcover group), the *Square* and *CommunityAverage* features in the GBR model were replaced by the *bathRoom and Lat* in the GA-GBR model.



	1	2	3
🟩 Predicted price (GBR)	390.52	378.73	426.42
🟩 Actual price	406.13	385.18	443.91
🟩 Predicted price (GA-GBR)	394.27	379.92	435.46

Figure 7- 11: Comparison of the actual price and predicted price by the two models in the three renovation-related datasets

It was discovered that in the rough and hardcover group, the Bathroom feature was considered more important to the GA-GBR model than the GBR. In the simplicity group, the Followers and LadderRatio features were considered more important to the GA-GBR than the GBR.

Table 7- 13: Feature importance ranking (top 3) by GBR and GA-GBR in the three renovation-related datasets

Rough		**Simplicity**		**Hardcover**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
square	16.84%	*communityAverage*	15.69%	*square*	19.57%
communityAverage	15.64%	*square*	15.66%	*communityAverage*	14.35%
tradeTime	13.83%	*tradeTime*	12.32%	*tradeTime*	11.91%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
bathRoom	24.73%	*ladderRatio*	22.81%	*tradeTime*	35.13%
Lng	23.84%	*followers*	18.11%	*bathRoom*	21.01%
Lat	13.31%	*tradeTime*	17.94%	*Lat*	10.33%

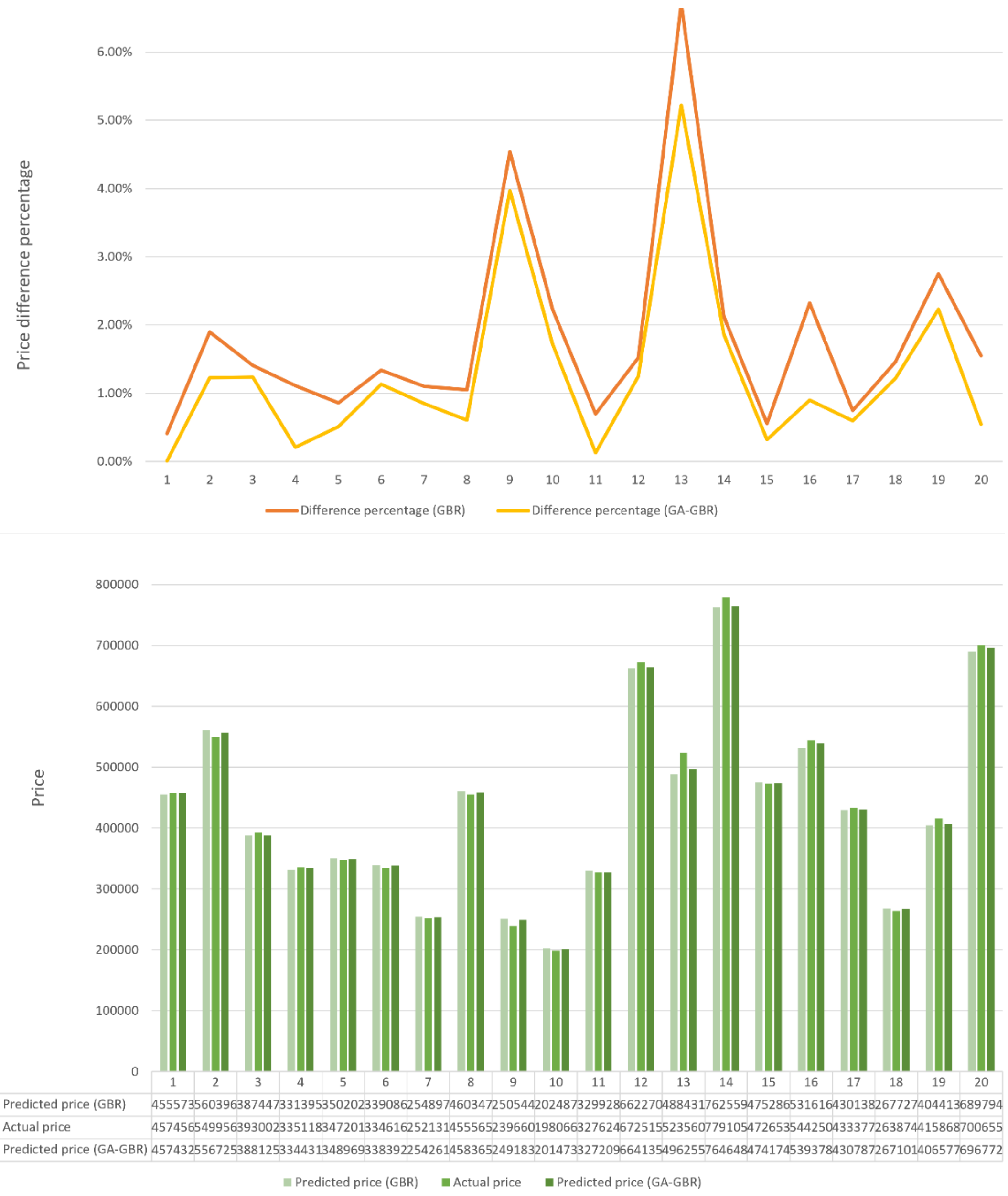
2) Testing on the American dataset

The American dataset was tested from two different perspectives including different cities and garages.

- *From the perspective of different cities*

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the 20 different location-related datasets: (1) the Chadstad, (2) the Coletown, (3) the Davidfort, (4) the Amychester, (5) the East_Lucas, (6) the Hallfort, (7) the Jeffreyhaven, (8) the Joshuafurt, (9) the Lake_Carolyn, (10) the Lake_Christina, (11) the Lake_Dariusborough, (12) the Lake_Jack, (13) the Lewishaven, (14) the Morris_port, (15) the North_Erinville, (16) the Port_Andrealand, (17) the

Port_Jonathanborough, (18) the Scottberg, (19) the South_Anthony, and (20) the West_Ann. In Figure 7-12 above, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE in all the 20 different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Predicted price (GBR)	455573	560396	387447	331395	350202	339086	254897	460347	250544	202487	329928	662270	488431	762559	475286	531616	430138	267727	404413	689794
Actual price	457456	549956	393002	335118	347201	334616	252131	455565	239660	198066	327624	672515	523560	779105	472653	544250	433377	263874	415868	700655
Predicted price (GA-GBR)	457432	556725	388125	334431	348969	338392	254261	458365	249183	201473	327209	664135	496255	764648	474174	539378	430787	267101	406577	696772

Figure 7- 12: Comparison of the actual price and predicted price by the two models in the 20 city-related datasets

To understand the predictive performance improvement of the GA-GBR model, Table 7-13 listed the top three important features to the two models in the four groups including the Morris_port (group 14), the Amychester (group 4), the Lake_Carolyn (group 9), and the Lake_Christina (group 10), in which the Morris_port (group 14) is in the New York city in the east of America with the highest price among the 20 groups, the Lake_Christina (group 10) is in Tampa in the south of America with the lowest price among the 20 groups, the Amychester (group 4) is in Los Angeles in the west and the Lake_Carolyn (group 9) is in Dallas in the middle of America with price in between. In the Morris_port group, the most important feature – *livable_sqft* in the GBR was replaced by the feature – *stories* in the GA-GBR. In the Lake_Christina group, the most important feature – *livable_sqft* in the GBR was replaced by the feature – *num_bedrooms* in the GA-GBR. In the Amychester group), the *total_sqft* and *has_pool* features in the GBR model were replaced by the *full_bathrooms and year_built* features in the GA-GBR model. In the Lake_Carolyn group, all the top three features - *livable_sqft, total_sqft* and *year_built* were replaced by the *stories, full_bathrooms* and *num_bedrooms* in the GA-GBR model.

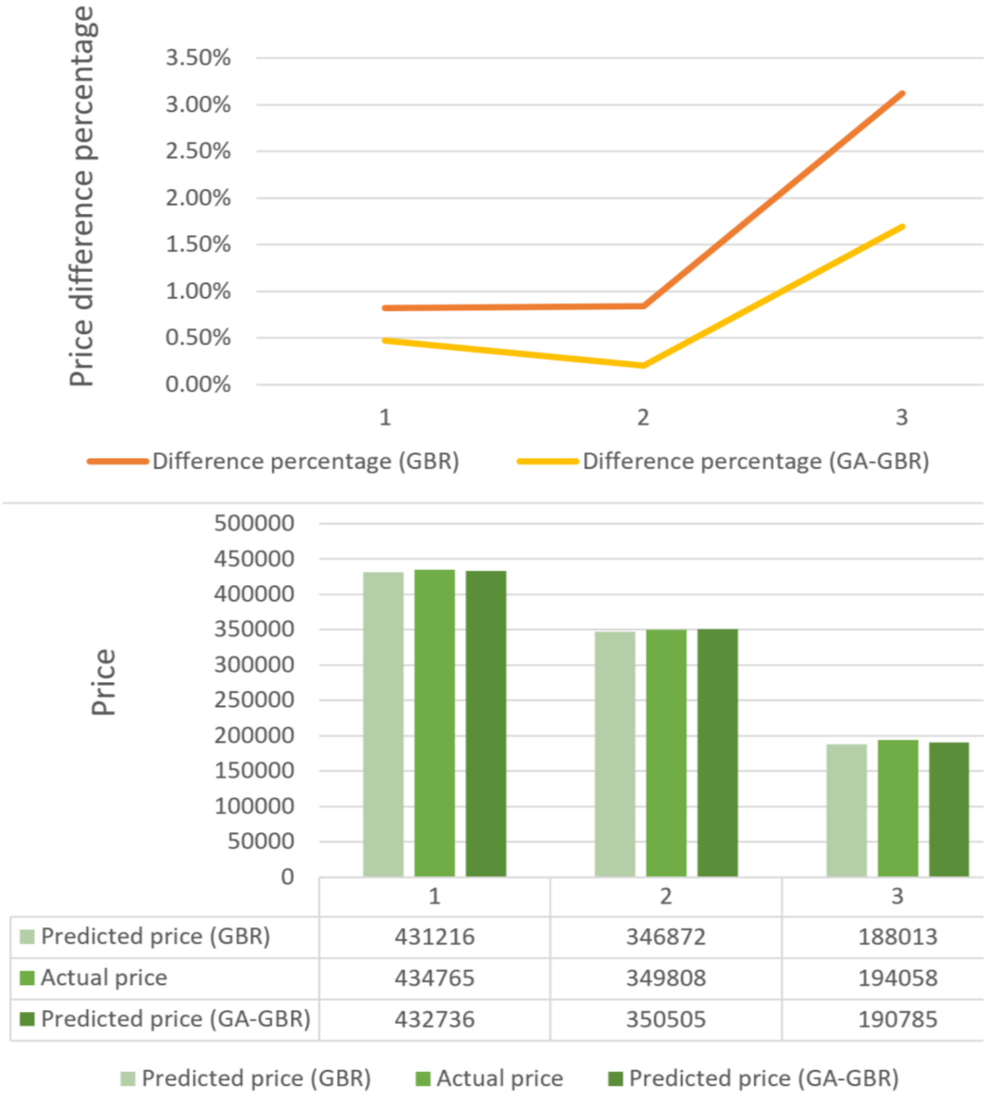
From the experiment results in Table 7-14, it was discovered that that the *livable_sqft* and *total_sqft* features were considered as the top two important to the GBR model in all the four groups. These two common features were replaced by other features such as *num_bedrooms, stories, full_bathrooms, half_bathrooms,* and *year_built* in the GA-GBR model. The abovementioned features are generally considered as important features in the house transaction market, however, the features selected by GA-GBR are more independent than those selected by the GBR. For instance, the *livable_sqft* and *total_sqft* features selected by GBR are highly connected to each other. As mentioned earlier, one of the important rules to generate an ensemble model is 'as independent as possible'. This might be the reason that the GA-GBR models had achieved an improved predictive accuracy.

Table 7- 14: Feature importance ranking (top 3) by GBR and GA-GBR in the four city-related datasets

Morris_port		**Amychester**		**Lake_Carolyn**		**Lake_Christina**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
livable_sqft	39.91%	*livable_sqft*	22.08%	*livable_sqft*	35.55%	*livable_ sqft*	29.30%
total_ sqft	23.38%	*total_sqft*	21.94%	*total_sqft*	21.22%	*total_sqft*	19.07%
has_ pool	8.95%	*has_pool*	14.80%	*year_built*	11.76%	*year_built*	13.00%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
stories	42.51%	*year_built*	24.70%	*stories*	42.99%	*num_ bedrooms*	39.96%
num_ bedrooms	26.74%	*livable_sqft*	22.71%	*full_ bathrooms*	21.05%	*full_ bathrooms*	21.50%
half_ bathrooms	17.04%	*full_ bathrooms*	20.26%	*num_ bedrooms*	19.38%	*stories*	12.93%

- *From the perspective of different garage types*

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the three different garage-related datasets: (1) the attached garage group, (2) the detached garage group, and (3) the none garage group. In Figure 7-13, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE at 0.47%, 0.2% and 1.69% in the three different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.



	1	2	3
Predicted price (GBR)	431216	346872	188013
Actual price	434765	349808	194058
Predicted price (GA-GBR)	432736	350505	190785

Figure 7- 13: Comparison of the actual price and predicted price by the two models in the three garage-related datasets

To understand the predictive performance improvement of the GA-GBR model, Table 7-15 listed the top three important features to the two models in the three groups, which were calculated using the embedded feature importance function in the decision tree-based models. In the attached group, the *livable_sqft* and *has_pool* features in the GBR model were replaced by the Full Bathroom *(full_bathrrom)* and Carport Size (*carport_sqft)* features in the GA-GBR model. In group 2 (the detached group), the *total_sqft* and *has_fireplace* features in the GBR model were replaced by the Number of Bedrooms (num_bedrooms) and Stories in the GA-GBR model. In group 3 (the none garage group), the most important feature – *carport_sqft* in the GBR model was replaced by the Full bathroom feature in the GA-GBR model.

From the experiment results, it was discovered that in the attached garage group, the Carport size and Full bathroom features were considered more important to the GA-GBR than the GBR. In the detached garage group, the Number of Bedrooms and Stories features were considered more important to the GA-GBR than the GBR. In the none garage group, the Full bathroom feature was considered more important to the GA-GBR than the GBR. Compare the attached garage group with the none garage group, the Carport size feature was considered more important by the garage owners, while it was surprising to see that the none garage owners paid more attention to the Full bathroom feature.

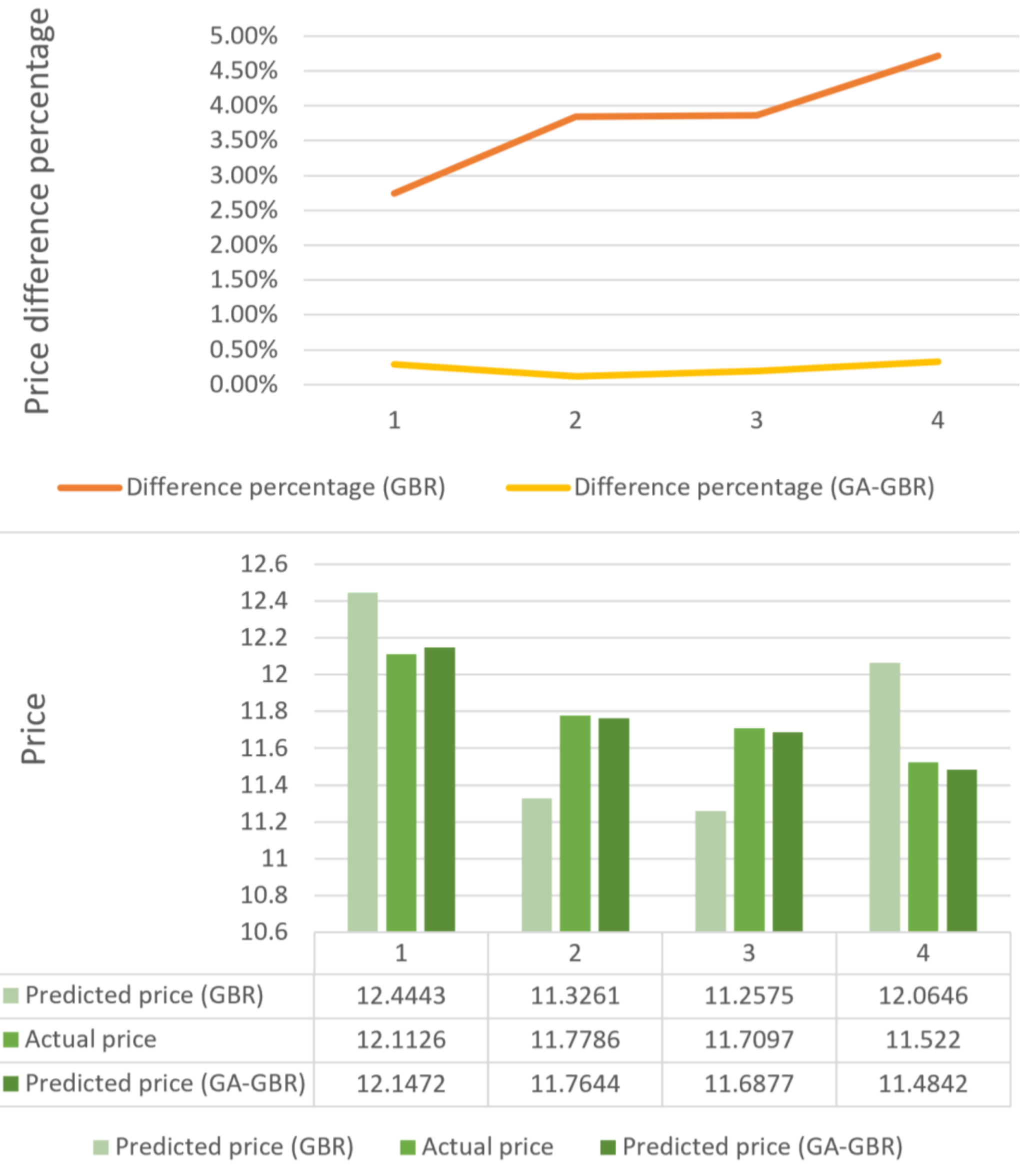
Table 7- 15: Feature importance ranking (top 3) by GBR and GA-GBR in the three garage-related datasets

Attached		**Detached**		**None**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
livable_sqft	18.85%	*livable_sqft*	17.16%	*carport_sqft*	14.75%
total_sqft	16.62%	*total_sqft*	13.58%	*livable_sqft*	12.79%
has_pool	9.30%	*has_fireplace*	10.86%	*year_built*	9.29%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
full_bathrooms	52.80%	*livable_sqft*	19.93%	*full_bathrooms*	17.25%
total_sqft	12.09%	*num_bedrooms*	17.38%	*livable_sqft*	13.98%
carport_sqft	5.69%	*stories*	12.04%	*year_built*	11.24%

3) Testing on the UK dataset

The UK dataset was tested from two different perspectives including different property types and PPD (Price Paid Transaction) types.

- *From the perspective of different property types*



	1	2	3	4
Predicted price (GBR)	12.4443	11.3261	11.2575	12.0646
Actual price	12.1126	11.7786	11.7097	11.522
Predicted price (GA-GBR)	12.1472	11.7644	11.6877	11.4842

Figure 7- 14: Comparison of the actual price and predicted price by the two models in the four property-type-related datasets

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the four different property type-related datasets: (1) the detached group, (2) the flats group, (3) the semi-detached group, and (4) the terraced group. In Figure 7-14, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE at 0.29%, 0.12%, 0.19% and 0.33% in the three different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.

Table 7- 16: Feature importance ranking (top 3) by GBR and GA-GBR in the four property-type-related datasets

Detached (1)		**Flats (2)**		**Semi-detached (3)**		**Terraced (4)**	
GBR feature	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**	**GBR feature**	**Rank**
Town/City	78.81%	*Town/City*	27.74%	*year*	60.22%	*year*	52.83%
index	19.78%	*County*	25.75%	*County*	14.69%	*County*	17.72%
District	0.65%	*District*	21.80%	*Town/City*	7.22%	*Town/City*	8.63%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**	**GA-GBR feature**	**Rank**
District	52.95%	*year*	46.01%	*Town/City*	48.43%	*District*	34.38%
Town/City	15.38%	*Town/City*	15.83%	*index*	33.20%	*index*	27.99%
index	10.59%	*County*	13.67%	*District*	17.66%	*Town/City*	22.72%

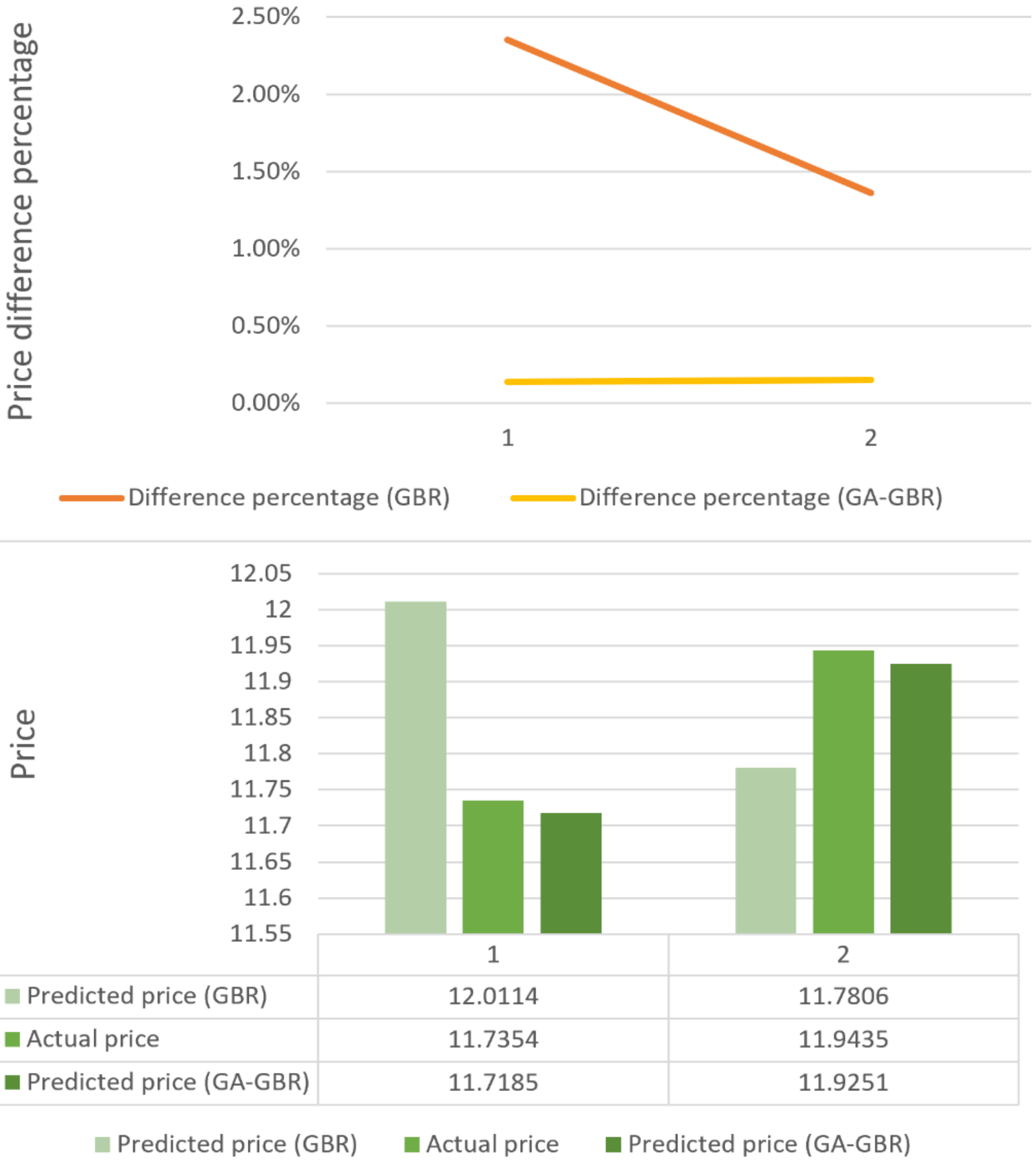
To understand the predictive performance improvement of the GA-GBR model, Table 7-16 listed the top three important features to the two models in the four groups, which were calculated using the embedded feature importance function in the decision tree-based models. In group 1 (the detached), the *Town/City, index* and *District* features were selected by both the GBR and GA-GBR models, with different weights of importance. In group 2 (the flats), the *District* feature in the GBR model were replaced by the *year* feature in the GA-GBR model, with the other two features' weights changed. In group 3 (the semi-detached) and group 4 (the terraced), the *year* and *county* features in the GBR model were replaced by the *index and District* features in the GA-GBR model.

The *Town/City* feature was considered as important to both GBR and GA-GBR models in all the four groups, which can be concluded as a general feature in different building types. For the detached and semi-detached groups, the *Town/City, index* and *District* features were considered more important to GA-GBR models than that to GBR models. For the flats group, the *year* feature was selected as the most important one to the GA-GBR, which indicated that this type of building might have a close connection to the time-related features.

- *From the perspective of different PPD Categories (Price Paid transaction)*

In this subsection, the predicted price by the GBR and GA-GBR models was compared with the actual price using the two different PPD-related datasets: (A) the Standard Price Paid group, (B) the Additional Price Paid group. In Figure 7-15, it was observed the predicted price by the GA-GBR model was closer to the actual price than that predicted by the GBR model, with a smaller MAPE at 0.14% and 0.15% in the two different groups respectively. This proved the advantage of the proposed GA-GBR model with a higher prediction accuracy than the traditional GBR model.

To understand the predictive performance improvement of the GA-GBR model, Table 7-17 listed the top three important features to the two models in the two groups, which were calculated using the embedded feature importance function in the decision tree-based models. In group A (the Standard Price Paid), the *year, county* and *Property_Type_is__D* features selected by the GBR were replaced by the *Town/City, index* and *District* features in the GA-GBR models. In group B (the Additional Price Paid), the *county, Town/City* and *Property_Type_is__D* features selected by the GBR were replaced by the *month, index* and *District* features in the GA-GBR models.



	1	2
🟩 Predicted price (GBR)	12.0114	11.7806
🟩 Actual price	11.7354	11.9435
🟩 Predicted price (GA-GBR)	11.7185	11.9251

Figure 7- 15: Comparison of the actual price and predicted price by the two models in the two PPD-related datasets

The abovementioned features suggest that the location-related features such as *County, District* and *Town/City* were important to the PPD related house transactions, in which the *District* feature was selected as the most important one to GA-GBR in both groups. As for time-related features, while the *year* feature was considered as the most important one to the GBR in the Standard Price Paid group, the *month* feature was considered as important to the GA-GBR in the Additional Price Paid group.

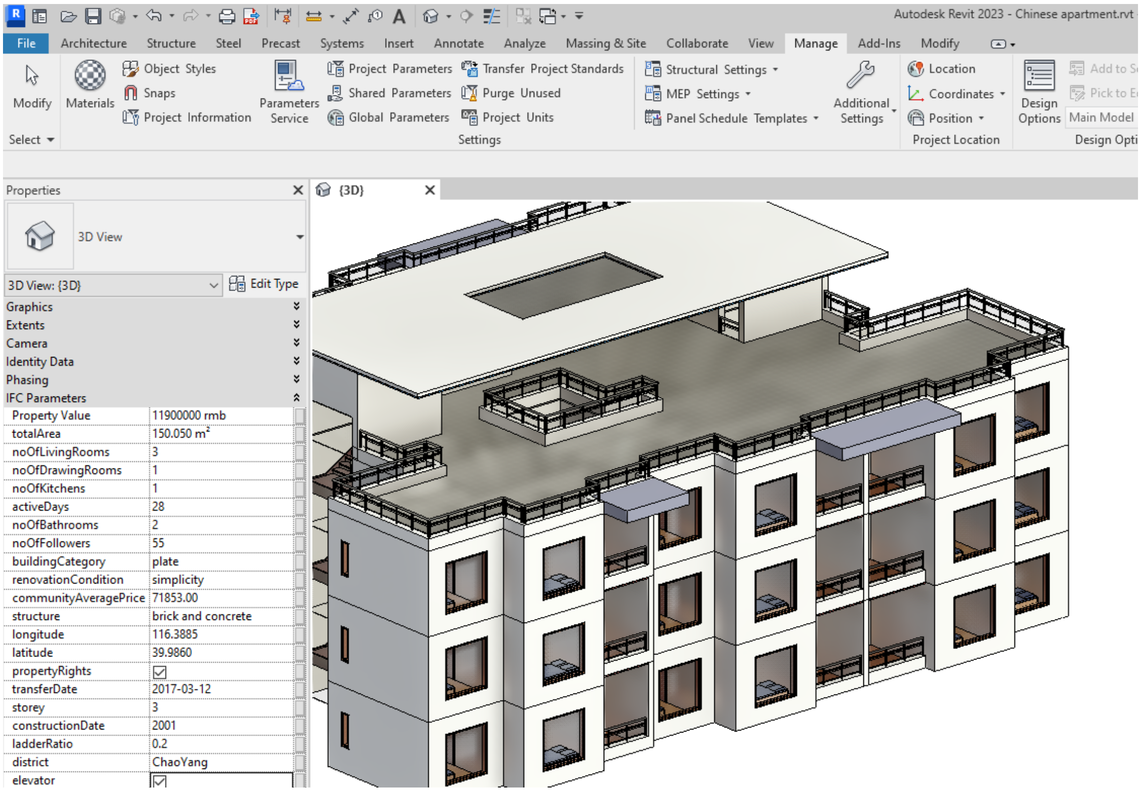
Table 7- 17: Feature importance ranking (top 3) by GBR and GA-GBR in the two PPD-related datasets

Standard Price Paid		**Additional Price Paid**	
GBR feature	**Rank**	**GBR feature**	**Rank**
year	46.69%	*County*	24.26%
County	13.19%	*Town/City*	23.77%
Property_Type_is__D	9.04%	*Property_Type_is__D*	11.70%
GA-GBR feature	**Rank**	**GA-GBR feature**	**Rank**
District	52.95%	*District*	33.00%
Town/City	15.38%	*index*	20.69%
index	10.59%	*month*	18.02%

7.3 Validation of the IFC-based Information Extraction as Required

In this section, three Revit models were tested for the IFC-based information extraction for property valuation. Due to limited time and resources, the value-related information for the three BIM models was achieved from three well-known real estate brokerage companies, namely the *Lianjia* company from China, the *Zillow* company from US, and the *Zoopla* company from the UK. After that, the required value-related information for property valuation was added into the spaces and zones defined in the Revit models, based on the proposed property sets and properties in the extended IFC schema and the input features in the three testing datasets. The syntactic and semantic validation of the IFC models were performed on *Solibri Model Checker* referring to the ISO 10303-11 (ISO 2014), with no missing mandatory entities or incorrect data structure. Lastly, the required value-relevant information was extracted using the developed IFC-based information extraction algorithm automatically.

1) IFC-based information extraction from the Chinese BIM model as required



IFC Parameters	
Property Value	11900000 rmb
totalArea	150.050 m²
noOfLivingRooms	3
noOfDrawingRooms	1
noOfKitchens	1
activeDays	28
noOfBathrooms	2
noOfFollowers	55
buildingCategory	plate
renovationCondition	simplicity
communityAveragePrice	71853.00
structure	brick and concrete
longitude	116.3885
latitude	39.9860
propertyRights	☑
transferDate	2017-03-12
storey	3
constructionDate	2001
ladderRatio	0.2
district	ChaoYang
elevator	☑

Figure 7- 16: An IFC-based BIM model of the duplex house with required value-related information added according to the 22 input features in the Chinese dataset

Referring to the 22 input features in the Chinese dataset and the extended IFC schema, value-related information (collected from the *Lianjia* company) for property valuation in terms of property sets, properties, and the nominal value of the properties were added into the BIM model through the shared parameters under the Manage tab setting panel. The added properties and their nominal value were displayed on the left sidebar in Figure 7-16, for instance, the IfcLabel ‘brick and concrete’ was added into the Structure property under the Pset_PV_Building property set, the IfcInteger ‘50’ was added into the *activeDays* property under the Pset_PV_Transaction property set, and the IfcReal

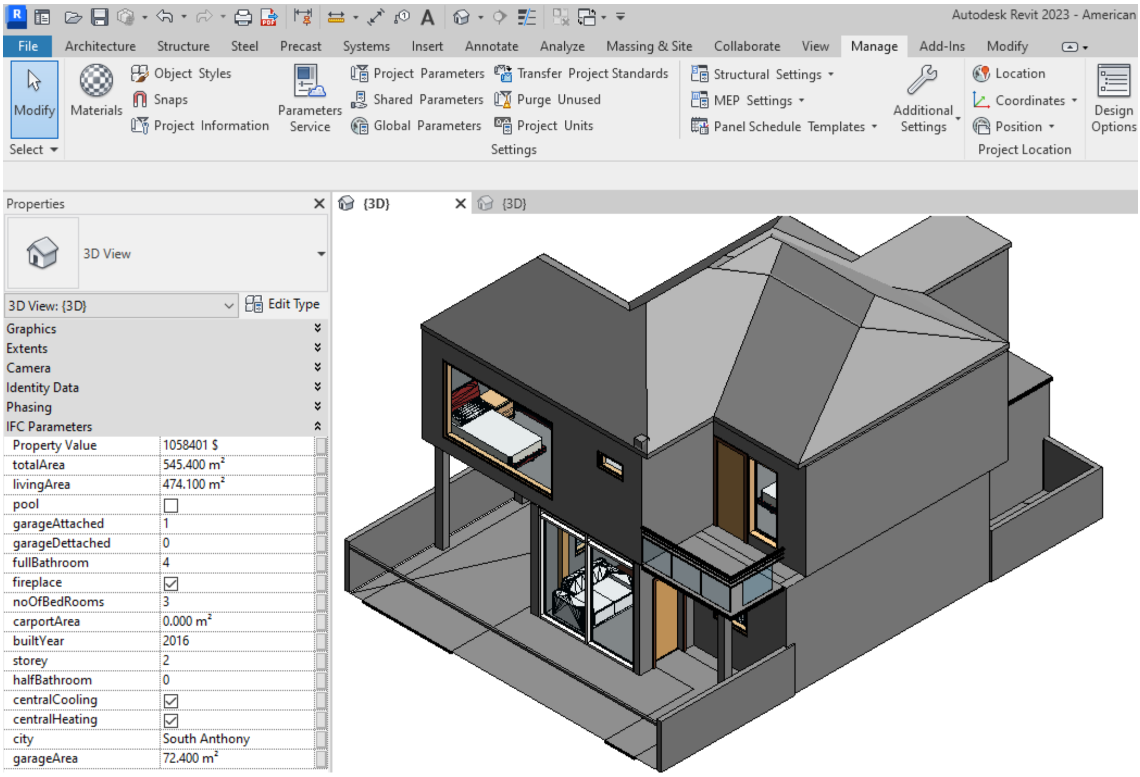
‘11900000’ was added into the *Property value* property under the Pset_PV_Valuation property set.

Table 7- 18: The extracted value-related information from the Chinese BIM model

Property set name	**Property name**	**Data type**	**Adapted nominal value**
Pset_PV_Building	totalArea	IfcAreaMeasure (150.05)	150.05 ㎡
Pset_PV_Building	noOflivingRooms	IfcInteger (3)	3
Pset_PV_Building	noOfDrawingRooms	IfcInteger (1)	1
Pset_PV_Building	noOfKitchens	IfcInteger (1)	1
Pset_PV_Transaction	transferDate	IfcDateTime (2017-03-12)	2017-03-12
Pset_PV_Transaction	activeDays	IfcInteger (28)	28
Pset_PV_Building	noOfBathrooms	IfcInteger (2)	2
Pset_PV_Transaction	noOfFollowers	IfcInteger (55)	55
Pset_PV_Building	buildingCategory	IfcLabel (‘plate’)	plate
Pset_PV_Building	renovationCondition	IfcLabel (‘simplicity’)	simplicity
Pset_PV_Transaction	communityAveragePrice	IfcReal (71853)	71853 RMB
Pset_PV_Building	structure	IfcLabel (‘brick and concrete’)	brick and concrete
Pset_PV_Building	elevator	IfcBoolean(.F.)	False
Pset_PV_Parcel	longitude	IfcLabel (‘116.3885’)	116.3885
Pset_PV_Parcel	latitude	IfcLabel (’39.9860’)	39.9860’
Pset_PV_Transaction	propertyRights	IfcBoolean(.T.)	True
Pset_PV_Building	storey	IfcInteger (3)	3
Pset_PV_Building	constructionDate	IfcDateTime (2001)	2001
Pset_PV_Building	ladderRatio	IfcReal (0.2)	0.2
Pset_PV_Parcel	district	IfcLabel (‘ChaoYang’)	ChaoYang
Pset_PV_Valuation	Property Value	IfcReal (11900000)	11900000 RMB

The value-related information regarding the 22 input features in the Chinese dataset was extracted using the developed information extraction algorithm, which is displayed in Table 7-18.

2) FC-based information extraction from the American BIM model as required



3D View: {3D}

Property Value	1058401 $
totalArea	545.400 m^2
livingArea	474.100 m^2
pool	□
garageAttached	1
garageDettached	0
fullBathroom	4
fireplace	☑
noOfBedRooms	3
carportArea	0.000 m^2
builtYear	2016
storey	2
halfBathroom	0
centralCooling	☑
centralHeating	☑
city	South Anthony
garageArea	72.400 m^2

Figure 7- 17: An IFC-based BIM model of the duplex house with required value-related information added according to the 15 input features in the American dataset

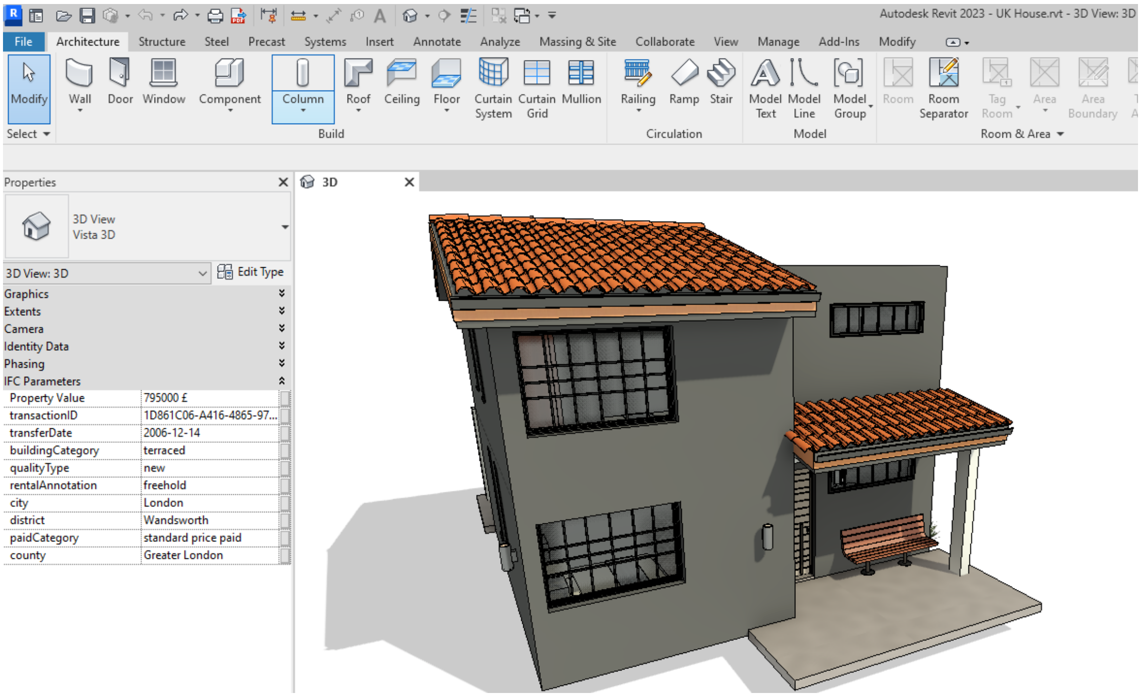
Referring to the 15 input features in the American dataset and the extended IFC schema, required information (collected from the *Zillow* company) for property valuation in terms of property sets, properties, and the nominal value of the properties were added into the BIM model through the shared parameters under the Manage tab setting panel. The added properties and their nominal value were displayed on the left sidebar in Figure 7-17, for instance, the IfcAreaMeasure '5454 ㎡' was added into the *totalArea* property under the Pset_PV_Building property set, the IfcLabel 'South Anthony' was added into the *city* property under the Pset_PV_Parcel property set, and the IfcReal '1058401' was added into the *Property value* property under the Pset_PV_Valuation property set.

The value-related information regarding the 15 input features in the American dataset was extracted using the developed information extraction algorithm, which is displayed in Table 7-19.

Table 7- 19: The extracted value-related information from the American BIM model

Property set name	Property name	Data type	Adapted nominal value
Pset_PV_Building	totalArea	IfcAreaMeasure (545.4)	545.4 ㎡
Pset_PV_Building	livingArea	IfcAreaMeasure (474.1)	474.1 ㎡
Pset_PV_Building	pool	IfcBoolean(.F.)	False
Pset_PV_Building	garageAttached	IfcInteger (1)	1
Pset_PV_Building	garageDettached	IfcInteger (0)	0
Pset_PV_Building	fullBathroom	IfcInteger (4)	4
Pset_PV_Building	fireplace	IfcBoolean(.T.)	True
Pset_PV_Building	noOfBedRooms	IfcInteger (4)	4
Pset_PV_Building	carportArea	IfcAreaMeasure (0)	0
Pset_PV_Building	builtYear	IfcInteger (2016)	2016
Pset_PV_Building	storey	IfcInteger (2)	2
Pset_PV_Building	halfbathroom	IfcInteger (0)	0
Pset_PV_Building	centralCooling	IfcBoolean(.T.)	True
Pset_PV_Building	centralHeating	IfcBoolean(.T.)	True
Pset_PV_Parcel	city	IfcLabel ('South Anthony')	South Anthony
Pset_PV_Building	garageArea	IfcAreaMeasure (72.4)	72.4 ㎡
Pset_PV_Valuation	Property Value	IfcReal ('1058401')	1058401 $

3) IFC-based information extraction from the UK BIM model as required



Properties

3D View
Vista 3D

3D View: 3D — Edit Type

Graphics
Extents
Camera
Identity Data
Phasing
IFC Parameters

Property Value	795000 £
transactionID	1D861C06-A416-4865-97...
transferDate	2006-12-14
buildingCategory	terraced
qualityType	new
rentalAnnotation	freehold
city	London
district	Wandsworth
paidCategory	standard price paid
county	Greater London

Figure 7- 18: An IFC-based BIM model of the duplex house with required value-related information added according to the 9 input features in the UK dataset

Referring to the 9 input features in the UK dataset and the extended IFC schema, required information (collected from the *Zoopla* company) for property valuation in terms of property sets, properties, and the nominal value of the properties were added into the BIM model through the shared parameters under the Manage tab setting panel. The added properties and their nominal value were displayed on the left sidebar in Figure 7-18, for instance, the IfcLabel ‘detached house' was added into the *buildingCategory* property under the Pset_PV_Building property set, the IfcLabel ‘Standard Price Paid’ was added into the *paidCategory* property under the Pset_PV_Transaction property set.

The value-related information regarding the 9 input features in the UK dataset was extracted using the developed information extraction algorithm, which is displayed in Table 7-20.

Table 7- 20: The extracted value-related information from the UK BIM model

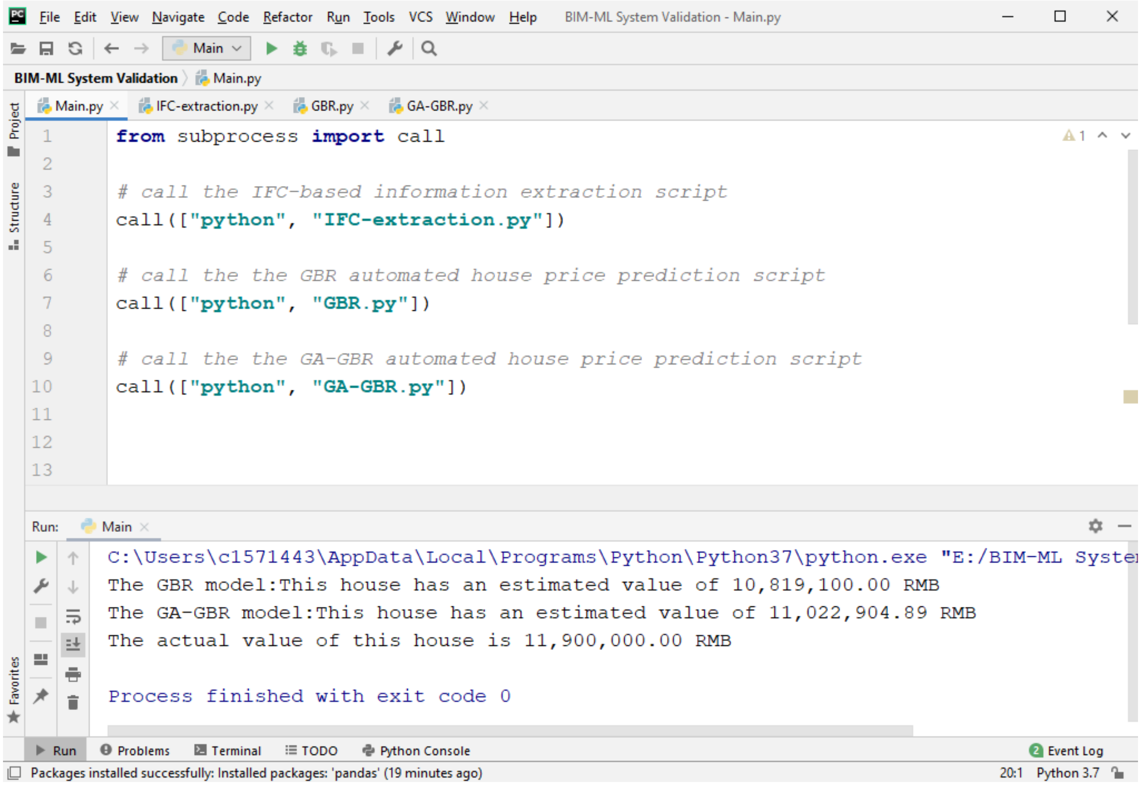
Property set name	**Property name**	**Data type**	**Adapted nominal value**
Pset_PV_Transaction	transactionID	IfcLabel (‘1D861C06-A416-4865-973C-4956DB12CD12’)	1D861C06-A416-4865-973C-4956DB12CD12
Pset_PV_Transaction	transferDate	IfcDateTime (2006-12-14)	2006-12-14
Pset_PV_Building	buildingCategory	IfcLabel (‘terraced’)	terraced
Pset_PV_Building	qualityType	IfcLabel (‘new’)	new
Pset_PV_Transaction	rentalAnnotation	IfcLabel (‘freehold’)	freehold
Pset_PV_Parcel	city	IfcLabel (‘London’)	London
Pset_PV_Parcel	district	IfcLabel (‘Wandsworth’)	Wandsworth
Pset_PV_Parcel	county	IfcLabel (‘Greater London’)	Greater London
Pset_PV_Transaction	paidCategory	IfcLabel (‘standard price paid’)	standard price paid
Pset_PV_Valuation	Property Value	IfcReal (‘795000’)	795000 £

From the testing results on the three IFC-based BIM models, it was summarized by the author as follows:

The proposed properties and property sets in the extended IFC schema are functional for property transaction case management, and the required value-related information for property valuation can be extracted from the IFC instance model automatically. This can ease the use of information in the AEC projects for the property valuation professionals who are lack of BIM knowledge and skills. However, during the testing process, there are some mismatched issues regarding the input feature names in the three testing datasets, which in turn increases the requirement for the standardized data format among the property valuation processes. For instance, the building category feature in the Chinese dataset is named as the property type in the UK dataset, the built year feature in the American dataset has the same meaning with the construction time feature in the Chinese dataset.

7.4 Validation of the Proposed BIM-ML Framework as a Complete Artifact

The validation of the whole BIM-ML system was conducted through a main python script on the PyCharm platform, calling the IFC-based information extraction algorithm and the GA-GBR model. Figure 7-19 displayed the validation of the complete BIM-ML system using the Chinese BIM model introduced in Figure 7-16. Based on the extracted information from the Chinese BIM model, the predicted house value using the GBR and the GA-GBR models are 10819100 RMB and 11022905 RMB respectively. With the actual value at 11900000 RMB, the value predicted by the GA-GBR model is more accurate than the GBR mode.


File Edit View Navigate Code Refactor Run Tools VCS Window Help
BIM-ML System Validation - Main.py
Main
BIM-ML System Validation
Main.py
IFC-extraction.py
GBR.py
GA-GBR.py
Project
Structure
Favorites
Run:
Main
Run
Problems
Terminal
TODO
Python Console
Event Log
Packages installed successfully: Installed packages: 'pandas' (19 minutes ago)
20:1
Python 3.7

```python
from subprocess import call

# call the IFC-based information extraction script
call(["python", "IFC-extraction.py"])

# call the the GBR automated house price prediction script
call(["python", "GBR.py"])

# call the the GA-GBR automated house price prediction script
call(["python", "GA-GBR.py"])
```

```
C:\Users\c1571443\AppData\Local\Programs\Python\Python37\python.exe "E:/BIM-ML Syste
The GBR model:This house has an estimated value of 10,819,100.00 RMB
The GA-GBR model:This house has an estimated value of 11,022,904.89 RMB
The actual value of this house is 11,900,000.00 RMB

Process finished with exit code 0
```

Figure 7- 19: Validation of the complete BIM-ML system

For the American BIM model, using the same approach, the predicted house value using the GBR and the GA-GBR models are 1152242 $ and 1087472 $ respectively. With the actual value at 1058401 $, the value predicted by the GA-GBR model is more accurate than the GBR mode.

For the UK BIM model, using the same approach, the predicted house value using the GBR and the GA-GBR models are 128300 £ and 129700 £ respectively. With the actual value at 135860 £, the value predicted by the GA-GBR model is more accurate than the GBR mode.

As introduced in Section 2.1.1, the international acceptable margin error of the predictive accuracy in the appraisal domain is between ± 0 and 10% (Brown et al. 1998; Abidoye and Chan 2018). The predictive error of the complete BIM-ML system tested using the three BIM models from China, US and the UK are 7.4%, 2.7% and 4.5% respectively, which complies with the international valuation standard.

The impact of the complete artifact in practical use:

The proposed BIM-ML system was implemented for five months in a commercial real estate appraisal and advisory company (HXZH) in China in 2021. The traditional valuation process applied in this commercial company can be classified into five main steps:

1) Task initialization - identifying the objectives and problems
2) Data collection through building survey onsite and online
3) Preliminary data analysis and provide an initial report of the indicated value
4) Calculation of the market value of the subject asset using the three traditional approaches
5) Deliver the final valuation report including the building information and the selected method to the client.

Comparing the traditional valuation process in the commercial real estate company and the developed BIM-ML system framework (illustrated in Figure 5-1), the main differences lie in the data collection and information exchange process (step 2 above) and the valuation method (step 4 above).

1) Data collection and information exchange

The data in the commercial real estate company is normally collected through building survey onsite and online, which normally takes at least several hours or several days. The collected data from onsite will be stored in a word file, with descriptions of this evaluation task, all the data for the evaluation, and attached pictures. It is manual based, time-consuming, and error-prone. In contrast, in this research partial data of the building to be evaluated is collected through BIM models, using the defined IFC property valuation extension and the information extraction algorithm. The information exchange process is based on the standard data format, which is automatic, efficient and less information missing or misunderstanding. Following the same process, different types of data such as subjective factors and green factors can be defined in the IFC extension and information extraction. In the long term, this IFC-based data definition and information exchange can save time and human costs for the real estate industry.

2) Valuation method

The sales comparison method, one of the most popular traditional property valuation methods, contains three main steps: (1) select the appropriate traded house in the same real estate transaction market in recent years; (2) compare the feature differences of the selected traded house as a reference with the subject house to be evaluated; and (3) make the value adjustments based on the feature differences according to the sales comparison grid and the comparison factor correction coefficient grid. There are a large amount of human subjective judgements on the value adjustments for different features and how many weights to be assigned for individual features. These human subjective judgements existed are more likely to produce an evaluated house price with a range, which makes it hard to compare the predictive accuracy with the predicted price by the GA-GBR model in this research. However, comparing the decision making in the two methods, in this research the value adjustment based on the feature differences and different weights allocated to individual features is implemented by the GA-GBR models, finding the implicit patterns from the three big data sets, which is more objective and less errors from human bias.

In addition, the feedback documents of the performance comparison of the developed BIM-ML system and traditional valuation method were provided in Figure E-1 in the Appendix E, which is summarized as follows:

- Compared with the traditional valuation process, the BIM-ML framework can produce accurate prediction results when dealing with residential buildings.
- Compared with the traditional valuation process, the BIM-ML framework can produce more objective results, which has the potential to reduce prediction error caused by human bias on selected variables and individual judgements on value adjustments.
- Compared with the traditional valuation process which normally take several hours gathering building information from building survey on site and generating the final results, the BIM-ML framework is much faster that can produce prediction results based on information acquired from BIM models in several minutes. The BIM-ML framework can be extremely helpful when a company are facing a large amount of prediction tasks.

7.5 Conclusion

This chapter presented the three validations of the developed BIM-ML system, including the validation of the developed AVM (the GA-GBR model), the validation of the IFC-based information extraction, and the validation of the complete artifact. It is concluded by the author that the developed BIM-ML system is reliable in terms of three aspects: (1) the developed IFC property valuation extension provides a standard way to store and exchange information in the real estate industry; (2) the IFC-based information extraction can save time and human costs on the data collection and information exchange in the commercial real estate companies; and (3) the enhanced AVM (the GA-GBR model) improves the predictive accuracy of house price prediction and helps understand the reasons behind different market factors and property transaction activities. The BIM-ML system has facilitated the automation of the data collection and information exchange and improved the objectiveness and predictive accuracy of property valuation.

Chapter 8. Discussion

Reflecting on the observations and findings from previous sections, this chapter provides a summary discussion of this research by revisiting the hypothesis and research questions. Subsequently, the limitations of the conducted research and recommended future work are presented.

8.1 Answer the Research Hypothesis

The research hypothesis tested for this research was: *A BIM and Machine Learning integration framework that allows the interpretation of value-relevant design information, information retrieval from BIM models automatically and an AI enhanced automated property valuation by leveraging existing BIM data and comprehensive property value determinants to enhance the decision-making processes about property valuation.*

The proposed research hypothesis has been broken down into five research questions (**Q1 – Q5**) which determine the contents of each chapter in this thesis. The five research questions and their answers will be provided as follows.

Q1: What is the current BIM and Machine Learning implementation on property valuation and What are the opportunities and challenges concerning automated property valuation and information exchange between AEC projects and property valuation?

The answer to the first research question was aimed at summarizing the research findings from literature review and identifying the research gaps. The current implementation of BIM and Machine Learning on property valuation is summarized as:

- Since traditional valuation approaches are questioned as inaccurate, inefficient and unreliable, in the last two decades, there has been a move towards the advanced valuation approaches due to the increasing complexity of property transaction and many advantages of the AI-enhanced AVMs. The advantages of Machine Learning for property valuation include: (1) efficiently assess information from big data; (2) identify non-linear relationships between house characters, market factors and property price; and (3) make decisions in the property valuation process with less human bias.

- While ANNs has attracted more attention than other algorithms, it is often recognized as ‘black box’ and shows limits on explaining the relationship between the input variables and the target price. Compared to neural networks, decision tree-based ensemble learning has advantages in terms of model interpretability and flexibility, which is more suitable for knowledge mining and system development. Since genetic algorithm (GA) optimized neural networks have achieved good prediction accuracy for property valuation, the integration of GA and ensemble learning has the potential to achieve good predictive performance for property valuation as well as good model interpretability.
- While research on BIM for property valuation is still at the early stage, the benefits of using BIM for property valuation have gained researchers and professionals’ attention. The value-relevant design information existed in AEC projects has not been widely utilized for property valuation, and there is a need to improve information exchange between AEC projects and property valuation. As the volume of data in BIM is rising exponentially, data analytics concepts and tools integrated BIM might bring added value and produce revolutionary influence on the construction industry, however, there is no framework establishment research related to the integration of BIM and Machine Learning for property valuation.

Q2: How innovative information technologies such as BIM and Machine Learning (ML) can improve the current valuation process and what are the information requirements for property valuation?

The answer to the second research question was aimed at providing a specific review of the current valuation process and future possible changes that BIM and ML will bring, whilst considering requirements for the envisaged proposed system. This is summarized as:

- The influence of AI enhanced AVMs on the current valuation process was summarized as: (1) data collection and exchange: to be more automated, and improved accuracy and efficiency with data standardization; (2) the valuation method: from traditional methods to AVMs; and (3) the role of valuers: from performing traditional property inspection and data analysis to interpreting the outcome of AMVs.
- The influence of BIM on information flows in the current valuation process was summarized as improved data exchange, less data input and sharing errors, data standardization, linked data with other information sources, saved time and costs, and improved productivity.
- Among 95 variables reviewed in the literature, 62 of them were identified as relevant to this research. The identified variables that have potential to be associated with BIM related concepts were classified as six main types and 28 subtypes of information related to property valuation, which were further used for the IFC extension development at the system development stage.

Q3: What kind of automated valuation models (AVMs) might have a better prediction performance for property valuation and how to improve the current AVMs?

The answer to the third research question addresses one of the three main components of the proposed BIM-ML system – the optimized AVM, which is the main focus of this research. It involves dealing with the suitability of the selected machine learning model (gradient boosting ensemble machine) for automated property valuation, and the optimized structure design for the proposed GA-GBR.

The findings from literature suggested that ensemble learning based AVMs are emerging, and the integration of GA and ensemble learning has the potential to achieve good predictive performance for property valuation as well as good model interpretability. To test this theory, it is necessary to conduct a comparison experiment with eleven AI-enhanced AVMs and compare their model performances. The eleven AI-enhanced AVMs are linear regression, ridge regression, lasso regression, elastic net regression, KNN, SVM, ANN, CART, AdaBoost, Random forest, and gradient boosting ensemble (GBR). The experiment was tested with the UCI Machine learning repository Boston housing dataset, which included 506 entries represent aggregated data with 14 variables for house price prediction in Boston in 1978. Model performances of the eleven AVMs were measured using the MAE, MAPE, MSE, and RMSE (measuring prediction errors) and the R-squared (measuring prediction accuracy).

From the comparison of the experiment results, it indicated that the classic linear models showed the poorest model performance and predictive accuracy, and the decision tree-based models including AdaBoost, Random Forest and GBR showed the highest model performance and predictive accuracy. The KNN and SVM models showed advantages over the linear models but showed disadvantages over the ANN and CART. The AdaBoost, Random Forest and GBR models showed advantages over the ANN and CART. It is worth to mention that the GBR model has the highest model prediction accuracy of the eleven different types of AVMs, with the MAPE at 10.4%, the MSE at 7.6, the RMSE at 2.7, and the mean R^2 at 90.3%. This complies with the findings from the literature and validate the logic of choosing the GBR model for the proposed system.

To solve the conflicts between the accuracy of individual weak learners and the diversity among them, whilst considering the exploration of the relationship between the input features and the target price, this research presented a study on an AVM based on genetic algorithm optimized gradient descent regression ensemble (GA-GBR) for property valuation. The genetic algorithm (GA) in the GA-GBR works as an evolutionary feature selection engine to search the near optimal feature subset which is further used to train a good boosting ensemble.

There are two major advantages of the proposed GA-GBR model:

1) For data with a big number of input features, input feature manipulation method often gives a good result. The manipulation of input features using GA increases the diversity of individual base learners. The evolutionary feature selection engine eliminates the redundant and irrelevant features without affecting the prediction accuracy, which avoids the overfitting of traditional GBR machines. As a good ensemble depends on the individual base learners being as accurate, and as diverse as possible, the increased diversity of individual base learners and the reduced

data dimensionality ensure the GA-GBR model with good prediction accuracy and model generalization capability.

2) The GA searches the suitable number of input features and updates the weights of them, which enables the GA-GBR to explore the relationship between the input features and the target price, and therefore gives an improved model interpretability over traditional boosting ensemble machines.

After explaining the structure of the proposed GA-GBR model, an initial test of the GA-GBR with the UCI Machine Learning repository - Boston dataset was performed. Compared to a similar house price prediction study using random forest machine learning with the same Boston housing dataset (Adetunji et al. 2022), in terms of R^2, showing 90.0% for the random forest, the proposed GA-GBR showed slight superiority of 0.5%. Considering that there are only 506 entries data in the Boston housing dataset, the performance of the proposed GA-GBR could be improved when using dataset with a bigger number of house transaction cases.

Q4: How to implement the BIM-ML integration framework and how to develop the three main components accordingly?

The answer to the fourth research question was aimed at developing the three main components of the BIM-ML system.

Firstly, an IFC extension for property valuation and an IFC-based partial information extraction were developed. The IFC extension definition came from 95 variables reviewed in the literature, of which 62 variables were identified as relevant to this research. Based on the identified 62 influential variables and the Valuation Information Model (LADM_VM), the required property sets and their properties are proposed to add to the IfcSpace and IfcZone entities. Seven property sets are proposed to add to the IfcSpace entity, including Pset_PV_Transaction, Pset_PV_Parcel, Pset_PV_Building, Pset_PV_CondominiumUnit, Pset_PV_Valuation, Pset_PV_MassValuation, and Pset_PV_Annex. Each of these property sets covers a number of properties that related to real estate transactions, the subject buildings, the valuation methods, and other special considerations from stakeholders. In total, there are 104 properties and 7 property sets proposed for the IFC Property Valuation extension (Table B-1 in the Appendix B).

After the analysis of information elements and their relationships between *IfcObject* and *IfcProperty*, the IFC-based information extraction algorithm was designed and developed on Python 3.7 using *IfcOpenshell-python* module on *Pycharm* software. The extraction algorithm was developed to extract a partial model for property valuation based on the internal data structure and internal relationships of the IFC schema, with the aim to ease the use of the value-relevant design information for property valuation professionals who were lack of BIM related knowledge and skills.

Secondly, before fitting data to the proposed AVM, the corelation relationship between the input features and the target price was explored, with the aims at discovering the implicit patterns in the data set. The exploratory data analysis concluded that the relationship between the input features and the target price are complex and significantly non-linear, and the Total Area variable shows the strongest positive correlation with the house price in both the Chinese and American data sets. It is indicated that the complex relationships between the input features and the target price are difficult to be directly estimated using simple linear machine learning models such as Linear Regression and KNN, whereas complex models such as ANN and decision tree-based machine learning models have the potential to fit well with the data.

After that, the optimal input feature subsets were explored using two typical feature selection techniques, namely the wrapper method and the embedded method. In the Chinese data set, the wrapper method – RFE selected 9 features including Total Area, Height, Trade Time, Active Days, Construction Time, Community Average Price, Followers, Latitude, and Longitude. The top 9 features ranked by the embedded methods were Community Average Price, Total Area, Trade Time, Bathroom, Active days, Living Room, Latitude, District, and Longitude. Compared with the top 9 features by the two methods, the Bathroom, Living room and District features selected by the embedded method were replaced by the Height, Construction time and Followers features in the wrapper method, while the other 6 features remain the same. In total, there are 12 important features selected by the two methods, and 6 common features including Community Average Price, Total Area, Trade Time, Active Days, Latitude, and

Longitude. These 6 features are given more attention when trying to improve the model performance of the GA-GBR model in the experiment stage. In the American data set, the wrapper method – RFE selected 6 features including Total Area, Living Area, Garage Area, Built Year, Stories, and Pool. The top 6 features ranked by the embedded methods were Total Area, Living Area, Garage Area, Built Year, Full Bathroom, and Pool. Compared with the top 6 features by the two methods, the Full Bathroom feature selected by the embedded method was replaced by the Stories feature in the RFE method, whereas the other 5 features remain the same. In total, there are 7 important features selected by the two methods, and 5 common features including Total Area, Living Area, Garage Area, Built Year, and Pool. These 5 features are given more attention when trying to improve the model performance of the GA-GBR model in the experiment stage.

Q5: How reliable is the proposed BIM-ML integration framework that can facilitate information exchange and support automated property valuation?

The answer to the last research question aimed at validating the proposed BIM-ML system, which was divided into three steps: (1) validate the trained GA-GBR model, (2) validate the IFC-based information extraction as required, and (3) validate the proposed BIM-ML system as a complete piece.

1) Validation of the trained GA-GBR model

The validation of the proposed GA-GBR model has been performed on the three different datasets, divided and undivided. In the next, the divided datasets will be clearly mentioned, whereas it means the datasets are undivided if there are no additional statements.

Basically, there are two goals of the proposed GA-GBR model, one is to improve the predictive accuracy of current AVMs, and the other one is to improve the model interpretation capability.

- ***Model predictive accuracy***

From the validation of the trained GA-GBR model with three datasets from different countries including China, US, and the UK, it was surprising that the proposed GA-GBR model achieved the highest predictive accuracy of the twelve different AVMs in terms of all the five typical regression accuracy metrics in all the three datasets. Moreover, on the divided datasets, it was observed the predicted price by the GA-GBR model was closer to the actual price than the predicted price by the GBR model, with a smaller MAPE in all the divided groups. This proves the proposed GA-GBR model not only has achieved a decent predictive accuracy, but also has a high generalization capability for property valuation.

In terms of coefficient of determination (R^2), the model accuracy of the GA-GBR had an advantage of 1.3% over the traditional GBR model in the Chinese dataset, an advantage of 3.57% in the American dataset, and an advantage of 2.4% in the UK dataset. The difference of the improved predictive accuracy on the three datasets probably because of the GBR models' predictive accuracy score baseline and the number of input variables in the GBR model after one-hot encoding. The GA-GBR model had a small predictive accuracy improvement (1.3%) on the Chinese dataset, because the traditional GBR model' predictive accuracy score baseline is already as high as 93.9%. This means that the decision-tree based structure had already fit well with the Chinese dataset, and the evolutionary feature selection function of the proposed GA-GBR model has limited positive effects on improving the predictive accuracy. In contrast, there was a big predictive accuracy improvement (3.57%) on the American dataset, because the traditional GBR model' predictive accuracy score baseline is only at 78.92%. This means that the decision-tree based structure had not fit so well with the American dataset, and therefore the evolutionary feature selection function of the proposed GA-GBR model has major positive effects on improving the predictive accuracy. However, the GA-GBR model had a medium predictive accuracy improvement (2.4%) on the UK dataset with the traditional GBR model' predictive accuracy score baseline at 73.2%. This is probably because of the data dimension reduction effect, which caused by the evolutionary feature selection function of the proposed GA-GBR model, requires a certain number of input features to improve the predictive accuracy. However, there are only 17 input features in the UK dataset after one-hot encoding, in contrast that there are 56 and 61 input features in the Chinese and American datasets.

From the comparison with the predictive accuracy of the twelve different AVMs, in general, linear regression models did not fit well on all the three datasets, which complies with the discovery from the exploratory data analysis before training the AVMs that the correlation relationship between the input features and the target price are complex and significantly non-linear. The KNN and SVM preformed worse than the linear regression models on the Chinese and American datasets but showed better performance than them on the UK dataset. The ANN showed advantages over the linear regression models on the American and UK datasets but shows disadvantages on the Chinese dataset. In general, the decision-tree based models had better predictive performances than all other models with high predictive accuracies. Surprisingly, the proposed GA-GBR model had the best predictive performance on all the three datasets. While a number of researchers in the literature indicated that ANN had achieved good predictive accuracy on property valuation, in this research the proposed GA-GBR model performed better predictive results than the ANN. The possible reason behind is that the ANN is good at dealing with unstructured input data, and the feature selection operation in the ANN is performed automatically by the embedded functions. In contrast, the proposed GA-GBR model, which has the decision tree-based structure, is good at dealing with structured input data (which is the case in this research), and the genetic algorithm or the evolutionary function engine in the proposed GA-GBR model has searched the optimal feature subsets which are more suitable for property valuation.

Compared with a similar research using the same Chinese dataset conducted by Quang et al. (2020), the extreme gradient boosting (XGBoost), which is the industrial application of the GBR model, had the predictive error at 0.1660 on the test set in terms of RMSLE. Whereas, in this research the RMSLE of the proposed GA-GBR model achieved smaller predictive error at 0.1626 on the test set, which is significantly better than the results of XGBoost in that paper. The Root Mean Squared Logarithmic Error (RMSLE) evaluation function used is calculated as follows:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\log(A_t + 1) - \log(F_t + 1))^2} \tag{16}$$

where A_t is the actual value and F_t is the predicted value.

- ***Model interpretability***

From the experiment results of the feature importance rankings calculated by the GBR and GA-GBR in the three datasets, it was discovered that the evolutionary feature selection engine in the proposed GA-GBR model had changed the weights of the input features, which make it more suitable for generating a good machine learning model. In the Chinese dataset, the total feature importance of the top 9 features selected by the GBR and GA-GBR account for 87.32% and 91.02% of all the 56 input features respectively. Compared with the six common important features selected by the two feature selection methods in Section 6.3.3, the five generic important features in the Chinese datasets are concluded by the author including the Total area, Trade time, Active days, Latitude, and Longitude features. Compare the top 9 features selected by the GBR and GA-GBR models, the experiment results indicated the Construction time, Trade time, and Follower features are more important to the GA-GBR model, which generated better predictive performance that the features selected by the traditional GBR model.

In the American dataset, the total feature importance of the top 6 features selected by the GBR and GA-GBR account for 70.68% and 73.72% of all the 61 input features respectively. Compared with the five common important features selected by the two feature selection methods in Section 6.3.3, there are two generic important features in the American dataset concluded by the author including the Total area and Living area features. Compare the top 6 features selected by the GBR and GA-GBR models, the experiment results indicated that the Number of bedrooms feature is more important to generate a good machine learning model than that in the traditional GBR model.

In the UK dataset, the total feature importance of the top 7 features selected by GBR and GA-GBR account for 94.03% and 98.9% of all the 17 input features respectively. Compared with the 6 common important features selected by the two feature selection methods in Section 7.2.1, there are 5 generic important features in the UK dataset concluded by the author including Year, Town/City, County, Property type, and District. Compare the top 7 features selected by the GBR and GA-GBR models, the experiment results indicated that the Month feature is more important to the GA-GBR model, which generated a higher model predictive accuracy than that in the traditional GBR model.

From the testing results of the feature importance rankings in the three datasets, it was concluded by the author as follows:

1. The top N features contributed more than 70% feature importance to the proposed GA-GBR model ($N \leq 10$).
2. There are five generic features in the Chinese dataset: Total area, Trade time, Active days, Latitude, and Longitude features; two generic features in the American dataset: Total area and Living area features; five generic features in the UK dataset: Year, Town/City, County, Property type, and District.
3. From the Comparison of the top N features selected by the GBR and GA-GBR models in the three datasets, it indicated that real estate transactions in Beijing happen more frequently, since the Construction time, Trade time, and Follower features are more important to generate a higher predictive performance model. In the American dataset customers seem to focus more on the number of bedrooms and garage-related features, whereas in the UK dataset time-related factors such as the Month feature is more important.

The relationships between the input features and the GA-GBR model were further explored in the divided datasets representing different perspectives. The Chinese dataset was divided into 22 groups by different building categories (3 groups), building structures (3 groups), renovation conditions (3 groups), and districts (13 groups). The American dataset was divided into 23 groups by different cities (20 groups) and different types of garages (3 groups). The UK dataset was divided into 6 groups by different property types (4 groups) and different PPD Categories (2 groups).

In the divided Chinese datasets, from the perspective of different building categories, it was discovered that the Trade time feature was important to both the GA-GBR and GBR in all the three groups. For the tower building type, the Elevator feature was considered more important to the GA-GBR model than the GBR model, which was more likely to represent the correlation relationship between the market factors and the house price in the real life. For the plate building type, the DOM (active days on market) and Construction time features were considered more important to the GA-GBR than the GBR.

From the perspective of different building structures, it was discovered that the Trade Time feature was important to both the GA-GBR and GBR in all the three groups. For the brick and concrete group, the Construction time feature was considered more important to the GA-GBR model than the GBR. For the steel-concrete composite group, the Followers and LadderRatio features were considered more important to the GA-GBR than the GBR. In the Chinese house transaction market, the steel-concrete composite building structure are normally used in a building with high stories and this type of buildings often have a higher price over other types of building structure. Therefore, the followers and LadderRatio features might contribute more to the model construction in the steel-concrete composite group.

From the perspective of different districts, it was discovered that in the XiCheng and FengTai districts, the Longitude *(Lng)* and Latitude *(Lat)* features (the location related features) were considered more important to the GA-GBR model than the GBR. In the ShunYi district, the Followers and TradeTime features were considered more important to the GA-GBR model than the GBR. In the Chinese house transaction market, the location related feature has always been one of the most important features to the house price. This might be the reason that the GA-GBR model considered the Longitude and Latitude features more important to the price of houses in the XiCheng district, which was the closest district to the city centre and had the highest price in all the 13 groups. The ShunYi district, which had the longest distance to the city centre and the lowest price in the 13 groups, the Followers and TradeTime features selected by the GA-GBR model indicated that houses in this district were traded more frequently.

From the perspective of different renovation conditions, it was discovered that in the rough and hardcover group, the Bathroom feature was considered more important to the GA-GBR model than the GBR. In the Chinese house transaction market, the bathroom might be renovated frequently, and houses with simplicity renovation conditions are the most frequently traded type in the three groups. That might be the reason that the GA-GBR model considered the Bathroom feature more important in the rough and hardcover group, and considered the Followers feature more important in the simplicity group.

In the divided American datasets, from the perspective of different cities, it was discovered that that the *livable_sqft* and *total_sqft* features were considered as the top two important to the GBR model in all the four groups. These two common features were replaced by other features such as *num_bedrooms, stories, full_bathrooms, half_bathrooms,* and *year_built* in the GA-GBR model. The abovementioned features are generally considered as important features in the house transaction market, however, the features selected by GA-GBR are more independent than those selected by the GBR. For instance, the *livable_sqft* and *total_sqft* features selected by GBR are highly connected to each other. As mentioned earlier, one of the important rules to generate an ensemble model is 'as independent as possible'. This might be the reason that the GA-GBR models had achieved an improved predictive accuracy.

From the perspective of different garage types, it was discovered that in the attached garage group, the Carport size and Full bathroom features were considered more important to the GA-GBR than the GBR. In the detached garage group, the Number of Bedrooms and Stories features were considered more important to the GA-GBR than the GBR. In the none garage group, the Full bathroom feature was considered more important to the GA-GBR than the GBR. Compare the attached garage group with the none garage group, the Carport size feature was considered more important by the garage owners, while it was surprising to see that the none garage owners paid more attention to the Full bathroom feature.

In the divided UK datasets, from the perspective of different property types, it was discovered that the *Town/City* feature was considered as important to both the GBR and GA-GBR models in all the four groups, which can be concluded as a general feature in different building types. For the detached and semi-detached groups, the *Town/City, index* and *District* features were considered more important to the GA-GBR model than that to the GBR model. For the flats group, the *year* feature was selected as the most important one to the GA-GBR, which indicated that this type of building might have a close connection to the time-related factors.

From the perspective of different PPD Categories, it was discovered that the location-related features such as *County, District and Town/City* were important to the PPD related house transactions, in which the *District* feature was selected as the most important one to the GA-GBR in both groups. As for time-related features, while the *year* feature was considered as the most important one to the GBR in the Standard Price Paid group, the *month* feature was considered as important to the GA-GBR in the Additional Price Paid group.

From the testing results of the feature importance rankings in the divided datasets representing different perspectives, it was concluded by the author as follows:

1. The top three important features selected by the GA-GBR model are closely connected to the corresponding represented perspectives. For instance, in the divided Chinese datasets, from the perspective of different building categories, the Elevator feature selected by the GA-GBR model is closely connected to the tower building category. From the perspective of different building structures, the LadderRatio feature selected by the GA-GBR model is closely connected to the steel-concrete composite building structure. From the perspective of different districts, the Longitude and Latitude features (the location related features) selected by the GA-GBR model is closely connected to the Xicheng district, which was the closest district to the city center and had the highest price in all the 13 groups. From the perspective of different renovation conditions, the Bathroom and Followers features selected by the GA-GBR model are closely connected to the hardcover renovation condition and the simplicity renovation condition respectively. In the divided American datasets, from the perspective of different garage types, the Carport size feature was considered more important by the garage owners.
2. The evolutionary feature selection engine in the GA-GBR model had changed the weights of the input features and increased the independence of individual features. For instance, in the divided American datasets, from the perspective of different cities, the Living area and Total area features selected by GBR are highly connected to each other. This correlated relationship has been minished in the GA-GBR model, which increased the model diversity and therefore improved the predictive accuracy.
3. In addition, it was discovered that time- related features are more important in the divided UK datasets, while it was surprising to see that the none garage owners paid more attention to the Full bathroom feature from the perspective of different garage types in the divided American datasets.

To sum up, the proposed GA-GBR model has not only improved the predictive accuracy of the traditional GBR, but also achieved improved model interpretation capability. The intended purpose of the proposed GA-GBR model in Section 5.3 has been achieved and the superiority of the GA-GBR model has been proved.

The housing prices are influenced by many different variables, some of them are objective factors such as number of rooms, building structure, size, age, stories and garages; the others are subjective factors that related to the economic, social and political indicators such as unemployment rate, Consumer Price Index (CPI), population, and gross domestic product (GDP). The objective factors are stationary, while features related to subjective factors are rather unstable or fluctuated. For instance, the DOM (active days on market) feature in the Chinese data set is influenced by the market demand that might go up or down based on the confidence of consumers. The national average house price fluctuates seasonally, but generally increases in the long-term period. As a time-series regression task, there are two strategies dealing with the subjective factors on machine learning based house price prediction. First, these subjective factors or time-dependent variables can be addressed by selecting representative data sets over a long period that the AVM can learn the patterns including the influence of the subjective factors which are unstable or fluctuated and further predict the house price today using the ‘learned knowledge’. That is one of the reasons why the collected data sets require to be explored and pre-processed before the machine learning model training, making sure that the data sets are representative for this particular task. In this research, the time period of the Chinese data set ranges from 2010 to 2018, the American data set ranges from 1889 to 2017, and the UK data set ranges from 1995 to 2017.

Second, for the data sets with a short time period, the percentage change of housing price caused by the time dependent parameters can be tackled with some time series models (i.e., VAR, ARIMA) or calculations for time difference error correction. For instance, Yan et al. (2007) used commercial housing sales price index as a basis for comparison, the time difference correlation analysis predicts that national commercial housing price would rise 6.88% in Q4-2006 and 6.64% in Q1-2007. Wang et al. (2019) combined the ARIMA (a well-known time series model) with deep learning for house price prediction, the experimental results showed that the proposed approach showed an advantage over a

SVR method and the predicted house price trend was basically consistent with the real data when dealing with short-term prediction.

2) Validation of the IFC-based information exchange as required

The proposed property sets (e.g. Pset_PV_Transaction, Pset_PV_Parcel, Pset_PV_Building and Pset_PV_Valuation) and properties (e.g. propertyRights, renovationCondition and communityAveragePrice) in the extended IFC schema were tested using three IFC-based Revit models from China, US and the UK. After which, the required value-relevant information was automatically extracted using the developed IFC-based information extraction algorithm, which indicated only the necessary information was extracted. This demonstrated the IFC-based information extraction was functional and reliable. As mentioned in Chapter 4, more information technologies are expected to be involved in the data collection, data exchange, and data processing processes. The proposed property sets and properties in the extended IFC schema not only fill a knowledge gap that considering design-related information for property valuation, but also could be used in 3D in an accurate and efficient manner. In the long term, the extended IFC schema can be further developed with more entities and property sets, which can be a valuable information enrichment for property valuation.

The literature indicated the professionals are exploring the use of BIM for property valuation, but it is challenging since there are a large number of various types of information in the IFC-based BIM models. The proposed IFC-based information extraction algorithm helps property valuation professionals who lack of BIM knowledge and digital skills to acquire value-specific information from AEC projects automatically. In practise, the proposed IFC-based information exchange for property valuation enables effective and efficient human decision-making in selecting the design alternatives with the highest value to different stakeholders. Besides, this method can be easily adapted to support other automation tasks in the AEC industry, for instance, automated energy prediction, automated clash detection, and automated compliance checking.

During the testing process, there are also some mismatched issues regarding the input feature names in the three testing datasets, which in turn increases the importance of using the standardized data format such as the IFC format among the property valuation processes. For instance, the building category feature in the Chinese dataset is named as the property type in the UK dataset, the built year feature in the American dataset has the same meaning with the construction time feature in the Chinese dataset. According to Ventolo (2015), the data collection in the traditional building survey can come from more than 40 data sources: regional government officials, property managers, professional journals, financial institutions, building architects, contactors, engineers and so on. All market actors in property markets can create their own sets of raw data in the building lifecycle, or they can collect and process information from other information source suppliers. Different market actors use different descriptive ways to interpret information in different data formats, which means information exchange issues will inevitably happen. The IFC-based data interpretation and information extraction provided a standardized method for managing value-related information on trading cases.

3) Validation of the proposed BIM-ML framework as a complete system

The validation of the whole BIM-ML system was conducted through a main python script on the PyCharm platform, putting the IFC-based information extraction algorithm and the GA-GBR model together. The predictive error of the complete BIM-ML system tested using the three BIM models from China, US and the UK are 7.4%, 2.7% and 4.5% respectively, which complies with the international acceptable margin error of the predictive accuracy in the appraisal domain is between ± 0 and 10%.

The literature indicated that the use of innovative information technology BIM and machine learning could be a revolution for data-driven applications in construction industry, but there is a significant gap in property valuation domain. In this research, this gap was filled by the development and validation of the proposed BIM-ML system. In practice, the real-time valuation results from the proposed BIM-ML system can be treated as constraints to optimize design, construction and operation strategies, which can be further developed as a decision-making tool for construction companies or property investors.

Based on the system development and the three validation tests, it can be concluded that the hypothesis is true, the IFC-based data interpretation and information extraction provided a standardized method for managing value-related information, the use of value-related information existed in AEC projects was promoted for the valuation professionals, and the predictive accuracy of automated valuation model was significantly improved.

8.2 Research Limitations

There are some deficiencies concerning the data, the methodology and the system used in implementation and testing in this research. The value-related information was firstly achieved from the valuation reports of three real estate transaction companies and then translated into the three IFC-based BIM models. Ideally, the IFC model should be collected from a real estate company who has both the trading cases information and the related BIM models. In addition, the proposed IFC Property Valuation extension has not been verified the experts in the valuation domain, due to the limited time and resources. While the proposed GA-GBR has showed improved predictive accuracy and model interpretation capability, the evolutionary feature selection process has heavy computing cost.

8.3 Future Work

Although this research has achieved its main aims through the development and validation of the proposed BIM-ML system, a series of suggestions for future work are summarised as follows.

1) The framework explored a way of using BIM as an information source for automated property valuation, in the future, more information systems could be integrated in this framework. For instance, the geographic information system (GIS) is a valuable information source for external environmental information.

2) Technology infusion of BIM, Machine Learning and other emerging digital technologies (IoT, digital twin systems, block chain and cloud computing) is worth exploring for property valuation and construction industry.
3) To support a more comprehensive property valuation, the entities, property sets and properties in the IFC Property Valuation can be further developed.
4) The proposed GA-GBR model used the integration of genetic algorithm and gradient boosting decision trees, in the future, more hybrid methods can be explored such as the integration of machine learning and deep learning models. While the GA-GBR model is tested with improved predictive accuracy using the evolutionary feature selection engine for tree-based models, the other optimization strategies such as decreasing the computing cost should be investigated.
5) This research focused on improving the predictive accuracy and automation process of property valuation. As mentioned in the literature review, sustainable property valuation is another popular research trend, it is worth exploring how the green parameters contribute to the proposed GA-GBR model in the future.

Chapter 9. Conclusion

To facilitate information exchange between AEC projects and property valuation and support automated property valuation, this thesis presented a BIM and Machine Learning integration framework for property valuation, which contains three main components: (1) an IFC extension for property valuation; (2) an IFC-based information extraction; and (3) an advanced automated valuation model (GA-GBR). Along with the developments of the three components in the proposed BIM-ML system, this research contributes as follows.

Firstly, this research contributes to the knowledge development of an extended IFC schema for property valuation. Among 95 variables reviewed in the literature, 62 of them are identified as relevant to this research, which is further used for the definition of the IFC extension. The extended IFC schema includes 7 new property sets and 104 new properties. The seven proposed property sets are Pset_PV_Transaction, Pset_PV_Parcel, Pset_PV_Building, Pset_PV_CondominiumUnit, Pset_PV_Valuation, Pset_PV_MassValuation, and Pset_PV_Annex. Each of these property sets contains a number of specific properties which can be used for property valuation. The proposed property sets and properties in the extended IFC schema can serve as a valuable information source for property valuation in the future. After that, the required value-specific design information is extracted automatically from an IFC-based BIM instance model using the developed information extraction algorithm.

Secondly, a genetic algorithm optimized gradient boosting ensemble model (GA-GBR) is firstly applied to automated property valuation, with the aim at improving the predictive accuracy of current AVMs and exploring the implicit patterns between the input features and the target price. Based on the findings from literature, the innovative GA-GBR model design starts with testing 11 different types of AVMs including linear regression models, SVM, ANN and decision tree-based models, and discovered the traditional GBR model has the highest model prediction accuracy. After that, to make up the deficiency and improve the predictive accuracy of the traditional GBR model, whilst considering the exploration of the relationship between the input features and the target price, the genetic approach for optimizing boosting ensemble is proposed. The proposed GA-GBR model was firstly validated on three different datasets from China, US, and the UK, comparing predictive accuracy with the other 11 AVMs using five typical regression accuracy metrics (MAE, MAPE, MSE, RMSE, and R^2), during which the GA-GBR model showed the highest predictive accuracy. Moreover, the proposed GA-GBR model was validated on 51 divided datasets representing 8 different perspectives including different building categories, building structures, renovation conditions, districts, cities, garages, property types, and PPD categories, during which the model interpretability was explored to understand the optimal input feature subsets, the individual feature importance to the GA-GBR model, and the patterns representing property transaction activities in the real life.

Lastly, as the volume of data in BIM is rising exponentially, data mining concepts and tools integrated BIM are expected to bring added value and produce revolutionary influence on industrial practices, but it exists a significant gap in the property valuation field. In this research a BIM-ML integration system was designed and implemented in property valuation, which filled that gap.

Several main findings have been identified in this research, which are highlighted as follows.

This research not only improved the predictive accuracy of current AVMs, but also achieved a high generalization capability for property valuation.

- From the validation results of the proposed GA-GBR model on three different datasets from China, US, and the UK, it was surprising that the GA-GBR model showed the highest predictive accuracy of the 12 AVMs in terms of all the five typical regression accuracy metrics (MAE, MAPE, MSE, RMSE, and R^2) on all the three test datasets.
- In terms of coefficient of determination (R^2), the model accuracy of the GA-GBR had an advantage of 1.3% over the traditional GBR model in the Chinese dataset, an advantage of 3.57% in the American dataset, and an advantage of 2.4% in the UK dataset.
- Moreover, on the divided datasets, it was discovered that the predicted price by the GA-GBR model was closer to the actual price than the predicted price by the GBR model, with a smaller MAPE in all the divided groups.

- Compared with a similar research using the same Chinese dataset conducted by Quang et al. (2020), the extreme gradient boosting (XGBoost), which is the industrial application of the GBR model, had the predictive error at 0.1660 on the test set in terms of RMSLE. Whereas, in this research the proposed GA-GBR model achieved smaller predictive error at 0.1626 on the test set (2% improvement), which is significantly better than the results of the XGBoost in that paper.

This research contributes to the knowledge mining from datasets and recognizing the non-linear relationships between the input features and the target price, which offers insights into the reasons for different market factors and property transaction activities in the real life.

- Real estate transactions in Beijing happen more frequently, since the Construction time, Trade time, and Follower features are more important to generate a higher predictive performance model. In the American dataset customers seem to focus more on the number of bedrooms, whereas in the UK dataset time-related factors such as the Month feature is more important. It was surprising to see that the none garage owners paid more attention to the Full bathroom feature from the perspective of different garage types in the divided American datasets.
- The top N features contributed more than 70% feature importance to the decision tree – based structure of Machine Learning ($N \leq 10$).
- The top three important features selected by the GA-GBR model are closely connected to the corresponding represented perspectives.
- The evolutionary feature selection engine in the GA-GBR model had changed the weights of the input features and increased the independence of individual features, which increased the model diversity and therefore improved the predictive accuracy.

The proposed BIM-ML system, provided a valuable information source for property valuation, eased the use of BIM knowledge and skills for the valuation professionals, enhanced the automated valuation process, and helped understand the implicit patterns behind property valuation.

Bibliography

Abdullah, L. et al. 2016. A Conceptual Framework of Green Certification Impact on Property Price. In: *MATEC Web of Conferences*. EDP Sciences. doi: 10.1051/matecconf/20166600033.

Abidoye, R.B. and Chan, A.P.C. 2017a. Artificial neural network in property valuation: application framework and research trend. *Property Management* 35(5), pp. 554–571. doi: 10.1108/PM-06-2016-0027.

Abidoye, R.B. and Chan, A.P.C. 2017b. Critical review of hedonic pricing model application in property price appraisal: A case of Nigeria. *International Journal of Sustainable Built Environment* 6(1), pp. 250–259. doi: 10.1016/j.ijsbe.2017.02.007.

Abidoye, R.B. and Chan, A.P.C. 2018. Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal* 24(1), pp. 71–83. Available at: https://doi.org/10.1080/14445921.2018.1436306.

Adair, A.S. et al. 1996. Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research* 13(1), pp. 67–83. doi: 10.1080/095999196368899.

Adetunji, A.B. et al. 2022. House Price Prediction using Random Forest Machine Learning Technique. In: *Procedia Computer Science*. Elsevier, pp. 806–813. doi: 10.1016/J.PROCS.2022.01.100.

Ahn, J.J. et al. 2012. Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications* 39(9), pp. 8369–8379. Available at: http://dx.doi.org/10.1016/j.eswa.2012.01.183.

Aladwan, Z. and Ahamad, M.S.S. 2019. Hedonic Pricing Model for Real Property Valuation via GIS - A Review. *Civil and Environmental Engineering Reports* 29(3), pp. 34–47. doi: 10.2478/ceer-2019-0022.

Antwi-Afari, M.F. et al. 2018. Critical success factors for implementing building information modelling (BIM): A longitudinal review. *Automation in Construction* 91(3), pp. 100–110. doi: 10.1016/j.autcon.2018.03.010.

Arcuri, N. et al. 2020. Automated valuation methods through the cost approach in a BIM and GIS integration framework for smart city appraisals. *Sustainability* 12(18), p. 7546. doi: 10.3390/su12187546.

Artus, M. and Koch, C. 2021. *Modeling Physical Damages Using the Industry Foundation Classes – A Software Evaluation*. Springer International Publishing. Available at: http://dx.doi.org/10.1007/978-3-030-51295-8_36.

Bajpai, P. and Kumar, M. 2010. Genetic Algorithm-an Approach to Solve Global Optimization Problems. *Indian Journal of Computer Science and Engineering* 1(3), pp. 199–206.

Bazjanac, V. 2004. Virtual Building Environments - Applying Information Modeling to Buildings. In: *European Conference on Product and Process Modeling in the Building and Construction Industry (ECPPM)*.

Benedetto, M. et al. 2015. Using Genetic Algorithms in the Housing Market Analysis. In: *Computational Science and Its Applications -- ICCSA 2015.*, pp. 36–45. doi: https://link.springer.com/chapter/10.1007/978-3-319-21470-2_3.

Bienert, S. et al. 2009. Integration of energy efficiency and LCC into property valuation practise-Transforming green features into values. Available at: https://epub.uni-regensburg.de/16209 [Accessed: 18 February 2020].

Bilal, M. et al. 2016. Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics* 30(3), pp. 500–521. Available at: http://dx.doi.org/10.1016/j.aei.2016.07.001.

BIM industry working group 2011. *A report for the government construction client group Building Information Modelling (BIM) working party strategy paper*.

Birje, M.N. et al. 2017. Cloud computing review: Concepts, technology, challenges and security. *International Journal of Cloud Computing* 6(1), pp. 32–57. doi: 10.1504/IJCC.2017.083905.

Bonci, A. et al. 2019. A cyber-physical system approach for building efficiency monitoring. *Automation in Construction* 102(6), pp. 68–85. doi: 10.1016/j.autcon.2019.02.010.

Borrmann, A. et al. 2018. *Building Information Modeling: Technology Foundations and Industry Practice*. 1st ed. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-92862-3.

Borrmann, A. et al. 2019. The IFC-Bridge project – Extending the IFC standard to enable high-quality exchange of bridge information models. *Proceedings of the 2019 European Conference on Computing in Construction* 1, pp. 377–386. doi: 10.35490/ec3.2019.193.

Bosteels, T. and Sweatman, P. 2016. Sustainable real estate investment: Implementing the Paris climate agreement - An action framework. Available at: http://www.unepfi.org/work-streams/property/sustainablerei/.

Boton, C. and Forgues, D. 2020. Construction 4.0: The next revolution in the construction industry. Available at: https://www.canbim.com/articles/construction-4-0 [Accessed: 9 March 2021].

Boyes, H. 2017. *A security framework for cyber-physical systems*. Coventry: University of Warwick. Available at: https://www.coursehero.com/file/86339403/A-Security-Framework-for-Cyber-Physical-Systemspdf/.

Boyes, H. et al. 2018. The industrial internet of things (IIoT): An analysis framework. *Computers in Industry* 101(10), pp. 1–12. doi: 10.1016/j.compind.2018.04.015.

Brown, G.R. et al. 1998. Valuation uncertainty and the Mallinson Report. *Journal of Property Research* 15(1), pp. 1–13. Available at: https://www.tandfonline.com/doi/abs/10.1080/095999198368473 [Accessed: 18 April 2022].

BSI 2014. BS 1192-4 : 2014 Collaborative production of information Part 4 : Fulfilling employer' s information exchange requirements using COBie - Code of practice. *British Standards Institution (BSI)* , p. 58. Available at: http://shop.bsigroup.com/forms/BS-1192-4/.

buildingSMART 2013. IFC Release Notes - buildingSMART Technical. Available at: https://technical.buildingsmart.org/standards/ifc/ifc-schema-specifications/ifc-release-notes/ [Accessed: 7 May 2021].

buildingSMART 2017a. IFC4 Addendum 2 Technical Corrigendum 1. Available at: https://technical.buildingsmart.org/standards/ifc/ifc-schema-specifications/.

buildingSMART 2017b. Information Delivery Manuals. Available at: https://www.buildingsmart.org/standards/bsi-standards/information-delivery-manual/.

buildingSMART International User Group 2012. An integrated process for delivering ifc based data exchange.

Cannon, S.E. and Cole, R.A. 2011. How Accurate Are Commercial Real Estate Appraisals? Evidence from 25 Years of NCREIF Sales Data . *The Journal of Portfolio Management* 37(5), pp. 68–88. Available at: https://jpm.pm-research.com/content/37/5/68 [Accessed: 10 May 2021].

Celik Simsek, N. and Uzun, B. 2021. Building Information Modelling (BIM) for property valuation: A new approach for Turkish Condominium Ownership. *Survey Review* . Available at: https://doi.org/10.1080/00396265.2021.1905251 [Accessed: 8 May 2021].

Chau, K.W. and Chin, T.L. 2003. A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications* 27(2), pp. 145–165.

Chen, K. and Xue, F. 2020. The renaissance of augmented reality in construction: history, present status and future directions. *Smart and Sustainable Built Environment* . doi: 10.1108/SASBE-08-2020-0124.

Chen, X. and Jeong, J. 2007. Enhanced recursive feature elimination-Web of Science Core Collection. In: *Sixth International Conference on Machine Learning and Applications (ICMLA 2007).*, pp. 429–435. Available at: 10.1109/ICMLA.2007.35.

Chen, Y.C. et al. 2016. Attention-Based User Interface Design for a Tele-Operated Crane. *Journal of Computing in Civil Engineering* 30(3). doi: 10.1061/(asce)cp.1943-5487.0000489.

Cheng, Y. et al. 2015. GA-based multi-level association rule mining approach for defect analysis in the construction industry. *Automation in Construction* 51(5), pp. 78–91. Available at: http://dx.doi.org/10.1016/j.autcon.2014.12.016.

Chicco, D. et al. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science* 7, p. e623. Available at: https://peerj.com/articles/cs-623 [Accessed: 11 March 2022].

CIC BIM 2050 group 2014. *Built environment 2050: a report on our digital future*.

Colwell, P.F. et al. 2009. Expert Testimony: Regression Analysis and Other Systematic Methodologies. *Appraisal Journal* 77(3), pp. 253–262.

Colwell, P.F. and Dilmore, G. 1999. Who was first? An examination of an early hedonic study. *Land Economics* 75(4), pp. 620–626. doi: 10.2307/3147070.

Couto, P. et al. 2021. Real-Estate Valuation Based on BIM Methodology. In: *Advances in Science,* *Technology and Innovation*. Springer Nature, pp. 15–19. Available at: https://link.springer.com/chapter/10.1007/978-3-030-35533-3_3 [Accessed: 8 May 2021].

Craft 2022. Zoopla Company Profile - Office Locations, Competitors, Financials, Employees, Key People, News | Craft.co. Available at: https://craft.co/zoopla-property-group [Accessed: 13 September 2022].

Craveiro, F. et al. 2019. Additive manufacturing as an enabling technology for digital construction: A perspective on Construction 4.0. *Automation in Construction* 103, pp. 251–267. doi: 10.1016/j.autcon.2019.03.011.

D'Amato, M. and Kauko, T. 2017. *Advances in automated valuation modeling:AVM After the Non-Agency Mortgage Crisis*. Springer International Publishing.

Deng, X. et al. 2020. Generic language for partial model extraction from an IFC model based on selection set. *Applied Sciences* 10(6), p. 1968. doi: 10.3390/app10061968.

Diaz, G.I. et al. 2017. An effective algorithm for hyperparameter optimization of neural networks. *IBM Journal of Research and Development* 61(4/5), pp. 9:1-9:11. doi: https://ieeexplore.ieee.org/abstract/document/8030298.

Dietterich, T.G. 2000. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*. Springer, Berlin, Heidelberg, pp. 1–15. doi: https://doi.org/10.1007/3-540-45014-9_1.

Dimopoulos, T. and Bakas, N. 2019. An artificial intelligence algorithm analyzing 30 years of research in mass appraisals. *RELAND: International Journal of Real Estate & Land Planning* 2, pp. 10–27. doi: https://doi.org/10.26262/reland.v2i0.6749.

Dorchester, J. 2011. Market value, fair value, and duress. *Journal of Property Investment and Finance* 29(4), pp. 428–447. doi: 10.1108/14635781111150321.

Eastman, C. et al. 2008. *BIM Handbook Paul Teicholz Rafael Sacks*. Hoboken, NJ: John Wiley and Sons.

Eastman, C. et al. 2011. *BIM handbook: a guide to building information modeling for owners, managers, designers, engineers and contractors.* Hoboken: Wiley.

Eastman, C.M. 1975. The Use of Computers Instead of Drawings in Building Design. *AIA Journal* 63(3), pp. 46–50.

El-gohary, N. 2010. Model-Based Automated Value Analysis of Building Projects. *Analysis* , pp. 16–18.

Forcael, E. et al. 2020. Construction 4.0: A Literature Review. *Sustainability* 12(22). doi: 10.3390/su12229755.

Ghaffarianhoseini, A. et al. 2017. Building Information Modelling (BIM) uptake: Clear benefits, understanding its implementation, risks and challenges. *Renewable and Sustainable Energy Reviews* 75(8), pp. 1046–1053. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1364032116308413.

Giudice, V. Del et al. 2017. Using genetic algorithms for real estate appraisals. *Buildings* 7(2), pp. 1–12. doi: 10.3390/buildings7020031.

Glumac, B. and Des Rosiers, F. 2020. Practice briefing – Automated valuation models (AVMs): their role, their advantages and their limitations. *Journal of Property Investment and Finance* 39(5), pp. 481–491. doi: 10.1108/JPIF-07-2020-0086.

Graczyk, M. et al. 2010. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5991 LNAI(PART 2), pp. 340–350. doi: 10.1007/978-3-642-12101-2_35.

Grover, R. 2016. Mass valuations. *Journal of Property Investment and Finance* 34(2), pp. 191–204. doi: 10.1108/JPIF-01-2016-0001.

GSA 2003. *3D-4D Building Information Modeling | GSA*. Available at: https://www.gsa.gov/real-estate/design-construction/3d4d-building-information-modeling [Accessed: 6 May 2021].

GSA 2022. BIM guides. Available at: https://www.gsa.gov/real-estate/design-and-construction/3d4d-building-information-modeling/bim-guides [Accessed: 19 April 2022].

Harrison, D. and Rubinfeld, D.. 1978. The Boston Housing Dataset. Available at: https://www.kaggle.com/datasets/schirmerchad/bostonhoustingmlnd [Accessed: 6 May 2022].

Hassanat, A. et al. 2019. Choosing mutation and crossover ratios for genetic algorithms-a review with a new dynamic approach. *Information (Switzerland)* 10(12). doi: 10.3390/info10120390.

Hastie, T. et al. 2008. *The elements of statistical learning: data mining, inference, and prediction*.

2nd ed. California: Springer.

Heaton, J. 2017. An Empirical Analysis of Feature Engineering for Predictive Modeling. *arXiv* , pp. 0–5.

Herr, C.M. and Fischer, T. 2019. BIM adoption across the Chinese AEC industries: An extended BIM adoption model. *Journal of Computational Design and Engineering* 6(2), pp. 173–178. Available at: https://doi.org/10.1016/j.jcde.2018.06.001.

HM Land Registry 2017. UK Housing Prices Paid | Kaggle. Available at: https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid [Accessed: 6 May 2022].

Hofmann, E. and Rüsch, M. 2017. Industry 4.0 and the current status as well as future prospects on logistics. *Computers in Industry* 89, pp. 23–34. doi: 10.1016/j.compind.2017.04.002.

Holst, E. and Thyregod, P. 1999. A statistical test for the mean squared error. *Journal of Statistical Computation and Simulation* 63(4), pp. 321–347. doi: 10.1080/00949659908811960.

IAAO 2003. *Standard on Automated Valuation Models (AVMs)*.

IfcOpenShell 2018. IfcOpenShell-python: A python module based on IfcOpenShell. Available at: http://ifcopenshell.org/python.

IPF 2014. *A Vision for Real Estate Finance in the UK Recommendations for reducing the risk of damage SPONSORED BY :*

Isakson, H. 2002. The Linear Algebra of the Sales Comparison Approach. *Journal of Real Estate Research* 24(2), pp. 117–128. Available at: https://www.tandfonline.com/action/journalInformation?journalCode=rjer20 [Accessed: 13 May 2021].

Isikdag, U. et al. 2015. Utilizing 3D Building and 3D Cadastre Geometries for Better valuation of existing real estate., pp. 17–21.

ISO 2014. Industrial automation systems and integration — Product data representation and exchange — Part 11: Description methods: The EXPRESS language reference manual.

ISO 2017. ISO 29481-1 : 2017 BSI Standards Publication Building information models –

Information delivery manual.

ISO 2018. ISO 16739-1:2018 - Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries — Part 1: Data schema. Available at: https://www.iso.org/standard/70303.html [Accessed: 19 March 2021].

IVS 2016. IVS 105: Valuation approaches and methods. *International Valuation Standards Council*

IVS 2019. *International Valuation Standards*. London: IVS.

IVSC 2016. IVS 104: Bases of value. *International Valuation Standards Council* (April), pp. 1–23. Available at: https://www.ivsc.org/files/file/view/id/646.

Jabbar, H.K. and Khan, R.Z. 2014. Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study). In: *Computer Science, Communication and Instrumentation Devices*. Kochi. doi: 10.3850/978-981-09-5247-1_017.

Jaly-zada, A. et al. 2015. Ifc extension for design change management. *Proc. of the 32nd CIB W78 Conference 2015, 27th-29th October 2015, Eindhoven, The Netherlands* , pp. 327–335.

De Jong, K.A. and Spears, W.M. 1991. An analysis of the interacting roles of population size and crossover in genetic algorithms. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Berlin, Heidelberg, pp. 38–47. Available at: https://link.springer.com/chapter/10.1007/BFb0029729 [Accessed: 1 May 2022].

Kanan, H.R. et al. 2007. Feature Selection Using Ant Colony Optimization (ACO): A New Method and Comparative Study in the Application of Face Recognition System. In: *Industrial conference on data mining*. Springer, Berlin, Heidelberg, pp. 63–76. Available at: https://link.springer.com/chapter/10.1007/978-3-540-73435-2_6 [Accessed: 30 April 2022].

Kara, A. et al. 2018. Supporting fiscal aspect of land administration through a LADM-based valuation information model. In: *Land Governance in an interconnected World: 19th Annual World Bank Conference on Land and Poverty*., pp. 1–34.

Katepalli, P. 2017. ml_house_data_set.csv. Available at: https://github.com/pavankat/flask-ml/blob/master/ml_house_data_set.csv [Accessed: 6 May 2022].

Kaul, A. et al. 2017. Autolearn - automated feature generation and selection. *Proceedings - IEEE International Conference on Data Mining, ICDM* 11, pp. 217–226. doi: 10.1109/ICDM.2017.31.

Kettani, O. and Oral, M. 2015. Designing and implementing a real estate appraisal system: The case of Québec Province, Canada. *Socio-Economic Planning Sciences* 49, pp. 1–9. doi: 10.1016/j.seps.2014.12.003.

Kok, N. et al. 2017. Big data in real estate? from manual appraisal to automated valuation. *Journal of Portfolio Management* 43(6), pp. 202–211. Available at: https://jpm.pm-research.com/content/43/6/202 [Accessed: 10 May 2021].

Kontrimas, V. and Verikas, A. 2011. The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing Journal* 11(1), pp. 443–448. doi: 10.1016/j.asoc.2009.12.003.

Krishnan, S. and Padmavathi, S. 2017. Feature ranking procedure for automatic feature extraction. *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings* , pp. 1613–1617. doi: 10.1109/SCOPES.2016.7955713.

Kumar, G. et al. 2020. Mathematics for Machine Learning. *Journal of Mathematical Sciences & Computational Mathematics* 1(2), pp. 229–238. doi: 10.15864/jmscm.1208.

Kutasi, D. and Badics, M.C. 2016. Valuation methods for the housing market: Evidence from Budapest. *Acta Oeconomica* 66(3), pp. 527–546. doi: 10.1556/032.2016.66.3.8.

L-TSV, N. 2018. *Draft Approach to Determine Long-Term Sustainable Value (L-TSV)*. Available at: http://ltsv.info/ltsv/ltsv.

Laakso, M. and Kiviniemi, A. 2012. The IFC standard: A review of History, development, and standardization, Information Technology - University of Salford Institutional Repository. *ITcon* 17(9), pp. 134–161. Available at: http://usir.salford.ac.uk/id/eprint/28373/ [Accessed: 7 May 2021].

Lambourne, T. 2021. Valuing sustainability in real estate: a case study of the United Arab Emirates. *Journal of Property Investment and Finance* . doi: 10.1108/JPIF-04-2020-0040.

Łaszek, J. et al. 2018. *Recent trends and its analysis in the real estate market*. 1st ed. SGH Warsaw School of Economics.

Lewis, O.M. et al. 1997. A novel neural network technique for the valuation of residential property. *Neural Computing & Applications* 5(4), pp. 224–229. Available at: http://link.springer.com/10.1007/BF01424227.

Lindblad, H. 2013. Study of the implementation process of BIM in construction projects: Analysis of the barriers limiting BIM adopotion in the AEC industry. *MSc Thesis* (263), p. 64. Available at: http://kth.diva-portal.org/smash/get/diva2:633132/FULLTEXT01.

Lipowski, A. and Lipowska, D. 2012. Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications* 391(6), pp. 2193–2196. doi: 10.1016/J.PHYSA.2011.12.004.

Lipscomb, J. and Gray, B. 1990. An Empirical Investigation of Four Market-Derived Adjustment Methods. *Journal of Real Estate Research* 5(1), pp. 53–66. Available at: https://www.tandfonline.com/doi/abs/10.1080/10835547.1990.12090602 [Accessed: 12 May 2021].

Lisi, G. 2019. Sales comparison approach, multiple regression analysis and the implicit prices of housing. *Journal of Property Research* 36(3), pp. 272–290. Available at: https://www.tandfonline.com/action/journalInformation?journalCode=rjpr20 [Accessed: 12 May 2021].

Liu, X.S. et al. 2011. Real estate appraisal system based on GIS and BP neural network. *Transactions of Nonferrous Metals Society of China* 21(Supplement 3), pp. s626–s630. Available at: http://dx.doi.org/10.1016/S1003-6326(12)61652-5.

Liu, Z. et al. 2019. A review and scientometric analysis of Global Building Information Modeling (BIM) Research in the Architecture, Engineering and Construction (AEC) industry. *Buildings* 9(10), p. 210. Available at: www.mdpi.com/journal/buildings [Accessed: 5 May 2021].

Lorentzon, J. 2011. *Att värdera tillgångar: Verkligt värde inom skogs- och fastighetsbranschen*. University of Gothenburg.

Lorenz, D. and Lützkendorf, T. 2008. Sustainability in property valuation: Theory and practice. *Journal of Property Investment and Finance* 26(6), pp. 482–521. doi: 10.1108/14635780810908361.

Lorenz, D. and Lützkendorf, T. 2011. Sustainability and property valuation: Systematisation of existing approaches and recommendations for future action. *Journal of Property Investment and Finance* 29(6), pp. 644–676. doi: 10.1108/14635781111171797.

Lorenz, D.P. et al. 2007. Exploring the relationship between the sustainability of construction and market value: Theoretical basics and initial empirical results from the residential property sector. *Property Management* 25(2), pp. 119–149. doi: 10.1108/02637470710741506.

Lützkendorf, T. and Lorenz, D. 2007. Integrating sustainability into property risk assessments for market transformation. *Building Research and Information* 35(6), pp. 644–661. doi: 10.1080/09613210701446374.

Lützkendorf, T. and Lorenz, D. 2011. Capturing sustainability-related information for property valuation. *Building Research & Information* 39(3), pp. 256–273. Available at: http://www.tandfonline.com/doi/abs/10.1080/09613218.2011.563929 [Accessed: 18 February 2020].

Mahamadu, A.M. et al. 2014. Determinants of Building Information Modelling (BIM) acceptance for supplier integration: A conceptual model. In: *In Proceedings 30th Annual ARCOM Conference*. Portsmouth

Mahdjoubi, L. et al. 2013. Providing real-estate services through the integration of 3D laser scanning and building information modelling. *Computers in Industry* 64(9), pp. 1272–1281. doi: 10.1016/j.compind.2013.09.003.

Mard, M.J. and Todd, J. 2010. The fair market value of land and buidling creates fair value error. *Real Estate Finance* 27(1), pp. 7–11.

Mark, J. and Goldberg, M.A. 1988. Multiple Regression Analysis and Mass Assessment: A Review of the Issues. *Appraisal Journal* 56(1), p. 89.

Marmo, R. et al. 2020. Building performance and maintenance information model based on IFC schema. *Automation in Construction* 118. doi: 10.1016/j.autcon.2020.103275.

Matthews, J. et al. 2015. Real time progress management: Re-engineering processes for cloud-based BIM in construction. *Automation in Construction* 58, pp. 38–47. doi: 10.1016/j.autcon.2015.07.004.

Mazairac, W. and Beetz, J. 2013. BIMQL - An open query language for building information models. *Advanced Engineering Informatics* 27(4), pp. 444–456. doi: 10.1016/j.aei.2013.06.001.

McCluskey, W.J. et al. 2013. Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research* 30(4), pp. 239–265. doi: 10.1080/09599916.2013.781204.

McGraw Hill Construction 2012. *The Business Value of BIM in North America*.

Miller, N. et al. 2008. Does green pay off? *Journal of Real Estate Portfolio Management* 14(4), pp. 385–399.

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill.

Montgomery, D.C. et al. 2015. *Introduction to linear regression analysis*. 5th ed. Hoboken, NJ: Wiley.

Mora-Esperanza, J.G. 2004. Artificial intelligence applied to real estate valuation: An example for the appraisal of Madrid. Catastro. *Catastro* 4(1), pp. 255–265.

Morano, P. et al. 2015. Artificial intelligence in property valuations An application of artificial neural networks to housing appraisal PIERLUIGI. *Advances in Environmental Science and Energy Planning* , pp. 23–29.

Munir, M. et al. 2019. BIM business value for asset owners through effective asset information management. *Facilities* 38(3–4), pp. 181–200. doi: 10.1108/F-03-2019-0036.

Myttenaere, A. de et al. 2016. Mean Absolute Percentage Error for regression models. *Neurocomputing* 192(6), pp. 38–48. doi: 10.1016/J.NEUCOM.2015.12.114.

Natekin, A. and Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* 7(DEC). doi: 10.3389/fnbot.2013.00021.

NBS 2020. *10th Annual UK's National Building Specification Report 2020*. Available at: https://www.thenbs.com/knowledge/national-bim-report-2020.

Nesta 2022. Looking at how AI could reduce the carbon footprint of your dinner. Available at: https://www.nesta.org.uk/project-updates/looking-at-how-ai-could-reduce-the-carbon-footprint-of-your-dinner/ [Accessed: 13 September 2022].

NIST 2018. IFC File Analyzer | NIST. Available at: https://www.nist.gov/services-resources/software/ifc-file-analyzer [Accessed: 13 April 2021].

Noran, O. et al. 2020. Exploring the Path Towards Construction 4.0: Collaborative Networks and Enterprise Architecture Views. In: *IFIP Advances in Information and Communication Technology*. Springer, pp. 547–556. Available at: https://doi.org/10.1007/978-3-030-57997-5_63 [Accessed: 5 March 2021].

Pagourtzi, E. et al. 2003. Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance* 21(4), pp. 383–401. Available at: http://www.emeraldinsight.com/doi/10.1108/14635780310483656.

Park, B. and Bae, J.K. 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications* 42(6), pp. 2928–2934. Available at: http://dx.doi.org/10.1016/j.eswa.2014.11.040.

Peffers, K. et al. 2007. A design science research methodology for information systems research. *Journal of Management Information Systems* 24(3), pp. 45–77. doi: 10.2753/MIS0742-1222240302.

Qiu, Q. 2018. Housing price in Beijing | Kaggle. Available at: https://www.kaggle.com/datasets/ruiqurm/lianjia [Accessed: 6 May 2022].

Qiu, Y. et al. 2017. Soak up the sun: Impact of solar energy systems on residential home values in Arizona. *Energy Economics* 66, pp. 328–336. doi: 10.1016/j.eneco.2017.07.001.

Quang, T. et al. 2020. Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science* 174, pp. 433–442. doi: 10.1016/J.PROCS.2020.06.111.

Rafiei, M.H. and Adeli, H. 2016. A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units. *Journal of Construction Engineering and Management* 142(2), p. 04015066. doi: 10.1061/(asce)co.1943-7862.0001047.

Rastogi, S. 2017. Construction 4.0: THE 4 th generation revolution. In: *Indian lean construction conference–ILCC 2017*. Available at: https://www.researchgate.net/publication/331131645_CONSTRUCTION_40_THE_4_th_GENERATION_REVOLUTION [Accessed: 5 March 2021].

Ratajczak, J. et al. 2019. BIM-based and AR Application Combined with Location-Based Management System for the Improvement of the Construction Performance. *Buildings* 9(5), p. 118. Available at: https://www.mdpi.com/2075-5309/9/5/118 [Accessed: 6 March 2021].

Renaud, O. and Victoria-Feser, M.P. 2010. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference* 140(7), pp. 1852–1862. doi: 10.1016/j.jspi.2010.01.008.

RICS 2013. *Sustainability and commercial property valuation: RICS Professional Guidance, Global*.

RICS 2016. BIM and commercial property: opportunities for property professionals. Available at: https://www.isurv.com/info/390/features/8870/bim_and_commercial_property_opportunities_for_property_professionals [Accessed: 8 May 2021].

RICS 2017a. *RICS valuation - global standards 2017 : incorporating the IVSC international valuation standards*.

RICS 2017b. *The Future of Valuations*.

RICS 2019. *RICS Valuation - Global Standards*. London: RICS.

Sacks, R. et al. 2018. SeeBridge as next generation bridge inspection: Overview, Information Delivery Manual and Model View Definition. *Automation in Construction* 90(2), pp. 134–145. Available at: https://doi.org/10.1016/j.autcon.2018.02.033.

Sattler, L. et al. 2019. Interoperability aims in building information modeling exchanges: A literature review. *IFAC-PapersOnLine* 52(13), pp. 271–276. doi: 10.1016/j.ifacol.2019.11.180.

Saunders, M. et al. 2015. *Research Methods for Business Students*. 7th ed. Harlow: Pearspm Education Limited.

Sayce, S. et al. 2006. *Real Estate Appraisal From Value to Worth*. Blackwell Publishing Ltd.

Schittenkopf, C. et al. 1997. Two strategies to avoid overfitting in feedforward networks. *Neural Networks* 10(3), pp. 505–516.

Schutt, R.K. 2011. *Investigating the Social World: The Process and Practice of Research*. Pine Forge Press.

Selim, H. 2009. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications* 36(2), pp. 2843–2852. doi: 10.1016/j.eswa.2008.01.044.

Sevinç, E. and Coşar, A. 2011. An evolutionary genetic algorithm for optimization of distributed database queries. *Computer Journal* 54(5), pp. 717–725. doi: 10.1093/comjnl/bxp130.

Shehzad, H.M.F. et al. 2020. Recent developments of BIM adoption based on categorization, identification and factors: a systematic literature review. *International Journal of Construction Management* . Available at: https://doi.org/10.1080/15623599.2020.1837719.

Sherwin, R. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of political economy* 82(1), pp. 34–55. doi: https://doi.org/10.1086/260169.

Smith, D.K. and Tardif, M. 2009. *Building Information Modeling: A Strategic Implementation Guide for Architects, Engineers, Constructors, and Real Estate Asset Managers | Wiley*. 1st ed. N: John Wiley and Sons.

Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45(4), pp. 427–437. doi: 10.1016/j.ipm.2009.03.002.

Spears, W.M. and Anand, V. 1991. A study of crossover operators in genetic programming. In: *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 409–418. Available at: https://link.springer.com/chapter/10.1007/3-540-54563-8_104 [Accessed: 1 May 2022].

Succar, B. and Kassem, M. 2015. Macro-BIM adoption: Conceptual structures. *Automation in Construction* 57, pp. 64–79. doi: 10.1016/j.autcon.2015.04.018.

Sumonja, N. et al. 2019. Automated feature engineering improves prediction of protein–protein interactions. *Amino Acids* 51(8), pp. 1187–1200. Available at: https://doi.org/10.1007/s00726-019-02756-9.

Sun, Y. 2019. Real estate evaluation model based on genetic algorithm optimized neural network. *Data Science Journal* 18(1), pp. 1–9. doi: 10.5334/dsj-2019-036.

Sylvain, A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4, pp. 40–79. doi: https://doi.org/10.1214/09-SS054.

Taffese, W.Z. 2007. Case-based reasoning and neural networks for real estate valuation. *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2007* , pp. 84–89.

Tang, S. et al. 2020. BIM assisted Building Automation System information exchange using BACnet and IFC. *Automation in Construction* 110, p. 103049. doi: 10.1016/j.autcon.2019.103049.

Tay, D.P.H. and Ho, D.K.H. 2004. Artificial Intelligence and the Mass Appraisal of Residential Apartments. *Journal of Property Valuation and Investment* 10(2), pp. 525–540. Available at: http://dx.doi.org/10.1108/14635789210031181.

Thant, Z.O. 2014. *Critical Success Factors for Application of BIM for Singapore Architectural Firms*.

Ullah, K. et al. 2019. An overview of BIM adoption in the construction industry: Benefits and barriers. *Emerald Reach Proceedings Series* 2, pp. 297–303. doi: 10.1108/S2516-285320190000002052.

Valier, A. 2020. Who performs better? AVMs vs hedonic models. *Journal of Property Investment and Finance* 38(3), pp. 213–225. doi: 10.1108/JPIF-12-2019-0157.

Vanlande, R. et al. 2008. IFC and building lifecycle management. *Automation in Construction* 18(1), pp. 70–78. doi: 10.1016/j.autcon.2008.05.001.

Ventolo, W.L. 2015. *Fundamentals of real estate appraisal*. 12th ed. La Crosse: DF Institure, Inc.

Villani, V. et al. 2018. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* 55, pp. 248–266. doi: 10.1016/j.mechatronics.2018.02.009.

Volk, R. et al. 2014. Building Information Modeling (BIM) for existing buildings - Literature review and future needs. *Automation in Construction* 38, pp. 109–127. Available at: http://dx.doi.org/10.1016/j.autcon.2013.10.023.

Volker, T. 2011. *Industry Foundation Classes (IFC) - BIM Interoperability through a vendor-independet file format*.

Wang, F. et al. 2019. House Price Prediction Approach based on Deep Learning and ARIMA Model. In: *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*. Institute of Electrical and Electronics Engineers Inc., pp. 303–307. doi: 10.1109/ICCSNT47585.2019.8962443.

Weimer, M. et al. 1972. *Real estate*. 6th ed. New York: The Ronald Press Company.

Weise, M. et al. 2003. Generalised model subset definition schema. *Cib Report* 284, p. 440. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Generalised+model+subset+definition+schema#0.

Wilkinson, S.J. and Jupp, J.R. 2016. Exploring the value of BIM for corporate real estate. *Journal of Corporate Real Estate* 18(4), pp. 254–269. doi: 10.1108/JCRE-11-2015-0040.

Witten, I.H. et al. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Won, J. et al. 2013. No-Schema Algorithm for Extracting a Partial Model from an IFC Instance Model. *Journal of Computing in Civil Engineering* 27(6), pp. 585–592. doi: 10.1061/(asce)cp.1943-5487.0000320.

Wong, F. and Tan, C. 1994. *Hybrid neural, genetic and fuzzy systems*. New York: Wiley.

World Economic Forum 2016. *Environmental Sustainability Principles for the Real Estate Industry*. Available at: www.weforum.org [Accessed: 15 May 2021].

Wu, J.Y. 2017. Housing Price prediction Using Support Vector Regression. Available at: https://scholarworks.sjsu.edu/etd_projects/540/.

El Yamani, S. et al. 2019. BIM potential for an enhanced real estate valuation approach based on the hedonic method. *WIT Transactions on the Built Environment* 192. Available at: https://orbi.uliege.be/handle/2268/250013 [Accessed: 8 May 2021].

El Yamani, S. et al. 2021. 3d variables requirements for property valuation modeling based on the integration of bim and cim. *Sustainability* 13(5), pp. 1–22. doi: 10.3390/su13052814.

Yan, H. et al. 2020. Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction* 119(11), p. 103331. Available at: https://doi.org/10.1016/j.autcon.2020.103331.

Yan, Y. et al. 2007. Method for Housing Price Forecasting based on TEI@I Methodology. *System engineering* 27(7), pp. 1–9. doi: https://doi.org/10.1016/S1874-8651(08)60047-2.

Yang, z et al. 2020. Rethinking Bias-Variance Trade-off for Generalization of Neural Networks. In: *Proceedings of the 37th International Conference on Machine Learning,PMLR*. Available at:

https://arxiv.org/abs/2002.11328.

Yang, L. and Shami, A. 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* 415, pp. 295–316. doi: 10.1016/j.neucom.2020.07.061.

Yang, Y. 2016. *Temporal data mining via unsupervised ensemble learning*. Elsevier.

Ying, L. et al. 2020. BIM based cyber-physical systems for intelligent disaster prevention. *Journal of Industrial Information Integration* 20, p. 100171. doi: 10.1016/j.jii.2020.100171.

Yu, H. and Liu, Y. 2016. Integrating Geographic Information System and Building Information Model for Real Estate Valuation. In: *FIG Working Week 2016 : Recovery from Disaster*. Christchurch: OICRF.

Yu, W. and Lin, H.W. 2006. A VaFALCON neuro-fuzzy system for mining of incomplete construction databases. *Automation in Construction* 15(1), pp. 20–32. doi: 10.1016/j.autcon.2005.01.006.

Zhao, J. et al. 2018. *Data-driven prediction for industrial processes and their applications*. Springer. Available at: https://doi.org/10.1007/978-3-319-94051-9.

Zhao, X. 2017. A scientometric review of global BIM research: Analysis and visualization. *Automation in Construction* 80, pp. 37–47. doi: 10.1016/j.autcon.2017.04.002.

Zheng, A. and Casari, A. 2018. *Feature engineering for machine learning*. O'Reilly Media, Inc.

Zhiliang, M. et al. 2011. Application and extension of the IFC standard in construction cost estimating for tendering in China. *Automation in Construction* 20(2), pp. 196–204. Available at: http://dx.doi.org/10.1016/j.autcon.2010.09.017.

Zhou, Z.H. 2012a. Ensemble Learning. Available at: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/springerEBR09.pdf.

Zhou, Z.H. 2012b. *Ensemble methods: foundations and algorithms*. Boca Raton: CRC Press.

Appendix A: Information Requirements for Property Valuation

Table A- 1: Holistic data interpretation of value-relevant attributes for property valuation

Type of Information	Subtype	Performance indicator and attribute	A	B	C	D
Environmental Quality	Local Environmental Impact	Climate Change	■		■	
	Pollution	Noise from transport service and building service equipment, water pollution, land contamination, electromagnetic pollution	■	■	■	■
	Land Use	Soil Characteristics	■	■	■	■
		Layout, size, inclination, topography	■		■	
	Sustainable Resource	Rainwater use		■		
		Green area	■	■	■	■
		Sunlight/Solar potential		■	■	
	Waste Water Volume	Waste water disposal	■	■		■
Social and Economic Quality	Commercial Viability	Policy and economic situation	■			
		Demographic structure and development	■			
		Purchasing power, letting prospects, expected rates of return	■			
		Rental growth potential, inflation expectations, rental payments, other payments	■			

		Payments for construction, acquisition, disposal, payments for operating costs, marketing / letting fee, payments for revitalization	🟨	🟩	🟥	🟥
		Number of tenants, Duration and structure of rental contracts	🟨			
		Vacancy rate, tenant fluctuation	🟨			
	Safety and Security	Location regrading natural hazards (risk of floods, landslides, collapse)	🟨	🟩	🟥	🟥
	Lifecycle Cost	Water demand and price, energy demand and price	🟨	🟩	🟥	🟥
Functional Quality	Indoor Air Quality	Sufficient natural air flow, low emitting material	🟨	🟩	🟥	🟥
	Acoustic Comfort	Noise reduction	🟨	🟩	🟥	🟥
	Visual Comfort	Good scene view, sufficient natural light	🟨	🟩	🟥	🟥
	Thermal Comfort	Hygrothermal rating		🟩	🟥	
	Flexibility and Adaptability	Flexibility of use (residential, office, medical practice), adaptability to users	🟨	🟩	🟥	🟥
		Wheelchair accessibility	🟨	🟩	🟥	🟥
		Wheelchair accessible washrooms		🟩	🟥	
		Usability of outside space	🟨	🟩	🟥	🟥
		Elevators (for all stories or not)	🟨	🟩	🟥	🟥
		Wide doors and wide halls	🟨	🟩	🟥	🟥
		Floor plan, storey height	🟨	🟩	🟥	🟥

	Brand Value	Green certification		🟩		
		Famous designer	🟨		🟥	
	Design/Aesthetic Quality	Architectural quality, Holistic monument	🟨	🟩	🟥	▨
Process Quality	Sustainability Aspects in Tender Phase	Ecological or recycled construction materials, risks and impacts for the local environment and residence		🟩	🟥	
	Documentation for Sustainable Management	Documented maintenance and servicing activities	🟨	🟩	🟥	▨
	Urban Planning and Design Procedure	Public accessibility, quality of layout	🟨	🟩	🟥	▨
	Construction Process/Site	Quality control during construction (air-tightness, thermography, sound insulation)		🟩	🟥	
	FM-compliant Planning	Maintenance management		🟩	🟥	
	Basic Information	Structure, age, size, construction type, main construction materials	🟨		🟥	
		Availability of green roofs/green facades	🟨	🟩	🟥	▨
		Renovation condition, construction quality	🟨		🟥	
		Building equipment and appliances	🟨	🟩	🟥	▨
	Sound Insulation	Noise Protection Techniques and Components	🟨	🟩	🟥	▨
	Quality of the Building Envelope	Heat insulation	🟨	🟩	🟥	▨
		Moisture proofing of the thermal building envelope	🟨	🟩	🟥	▨

Technical Quality	Ease of Cleaning Building Components	Ease of conducing cleaning, building services and maintenance works	●	●	●	●
	Recyclability and Energy efficiency	Ease of recovery and recycling, efficiency of heating ventilation, solar radiation, rainwater use		●	●	
	Immission Control	External and internal accessibility	●	●	●	●
	Infrastructure	Fitness	●	●	●	●
	Quality of Indoor and Outdoor Spaces	Balcony, storage space	●	●	●	●
	Safety and Security	Clear arrange routes for escape	●	●	●	●
		Protection against burglary	●	●		●
		Fire Protection	●	●	●	●
		Quality of sanitary and electronic fixtures	●	●	●	●
		Structural Safety		●	●	
		Durability of building components		●	●	
Site Quality	Local Environment and Policy	Visual context, building permission and planning regulations	●	●		●
	Transport Access	Public transport, parking	●	●	●	●
	Amenities	Area and distance to facilities (shopping, social and medical)	●	●	●	●

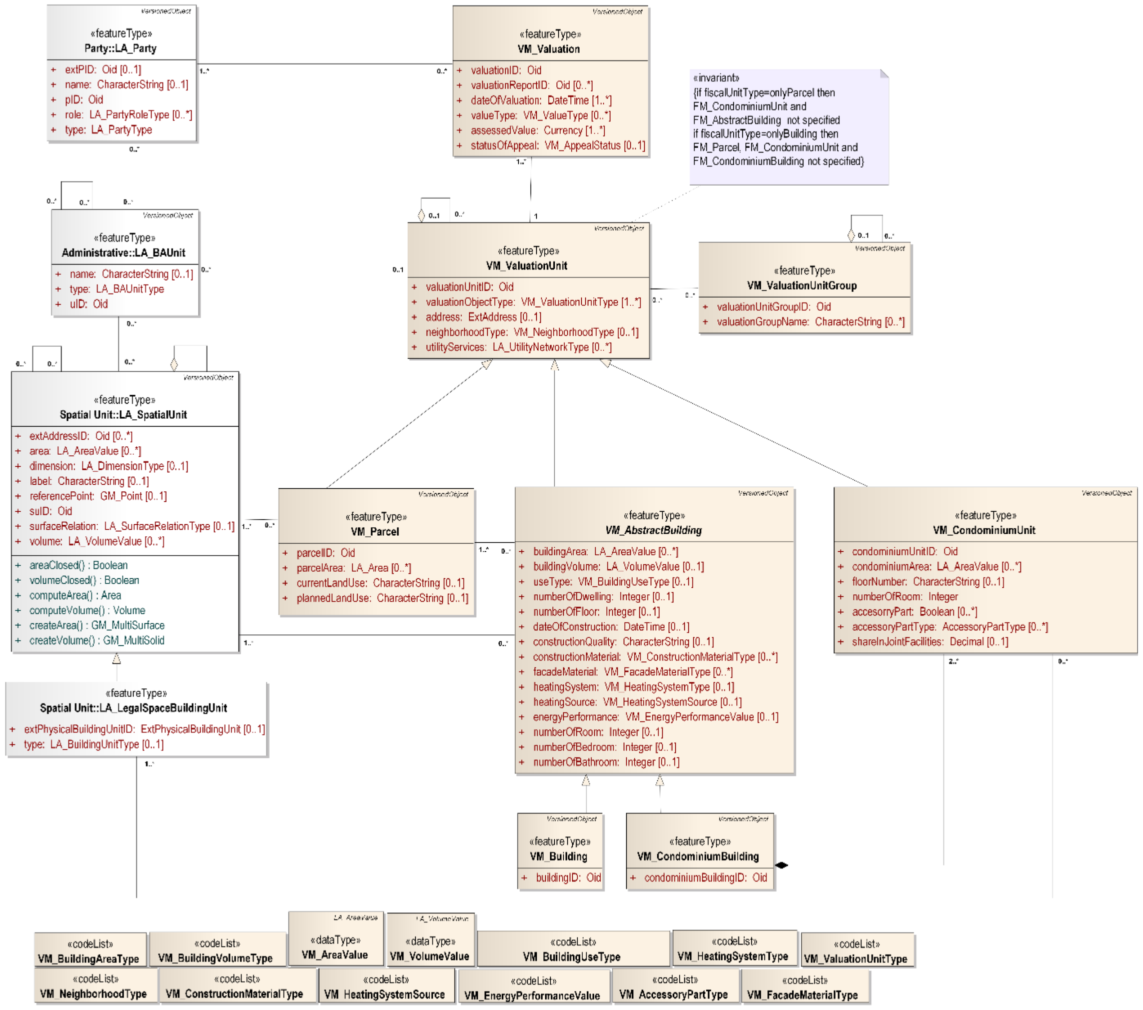

VersionedObject
«featureType»
Party::LA_Party
+ extPID: Oid [0..1]
+ name: CharacterString [0..1]
+ pID: Oid
+ role: LA_PartyRoleType [0..*]
+ type: LA_PartyType
1..*
0..*
0..*
VersionedObject
«featureType»
VM_Valuation
+ valuationID: Oid
+ valuationReportID: Oid [0..*]
+ dateOfValuation: DateTime [1..*]
+ valueType: VM_ValueType [0..*]
+ assessedValue: Currency [1..*]
+ statusOfAppeal: VM_AppealStatus [0..1]
1..*
1
«invariant»
{if fiscalUnitType=onlyParcel then
FM_CondominiumUnit and
FM_AbstractBuilding not specified
if fiscalUnitType=onlyBuilding then
FM_Parcel, FM_CondominiumUnit and
FM_CondominiumBuilding not specified}
0..*
0..*
0..*
VersionedObject
«featureType»
Administrative::LA_BAUnit
+ name: CharacterString [0..1]
+ type: LA_BAUnitType
+ uID: Oid
0..*
0..*
0..1
0..*
0..1
VersionedObject
«featureType»
VM_ValuationUnit
+ valuationUnitID: Oid
+ valuationObjectType: VM_ValuationUnitType [1..*]
+ address: ExtAddress [0..1]
+ neighborhoodType: VM_NeighborhoodType [0..1]
+ utilityServices: LA_UtilityNetworkType [0..*]
0..*
0..*
0..1
0..*
VersionedObject
«featureType»
VM_ValuationUnitGroup
+ valuationUnitGroupID: Oid
+ valuationGroupName: CharacterString [0..*]
0..*
0..*
0..*
VersionedObject
«featureType»
Spatial Unit::LA_SpatialUnit
+ extAddressID: Oid [0..*]
+ area: LA_AreaValue [0..*]
+ dimension: LA_DimensionType [0..1]
+ label: CharacterString [0..1]
+ referencePoint: GM_Point [0..1]
+ suID: Oid
+ surfaceRelation: LA_SurfaceRelationType [0..1]
+ volume: LA_VolumeValue [0..*]
+ areaClosed() : Boolean
+ volumeClosed() : Boolean
+ computeArea() : Area
+ computeVolume() : Volume
+ createArea() : GM_MultiSurface
+ createVolume() : GM_MultiSolid
1..*
0..*
VersionedObject
«featureType»
VM_Parcel
+ parcelID: Oid
+ parcelArea: LA_Area [0..*]
+ currentLandUse: CharacterString [0..1]
+ plannedLandUse: CharacterString [0..1]
1..*
0..*
VersionedObject
«featureType»
VM_AbstractBuilding
+ buildingArea: LA_AreaValue [0..*]
+ buildingVolume: LA_VolumeValue [0..1]
+ useType: VM_BuildingUseType [0..1]
+ numberOfDwelling: Integer [0..1]
+ numberOfFloor: Integer [0..1]
+ dateOfConstruction: DateTime [0..1]
+ constructionQuality: CharacterString [0..1]
+ constructionMaterial: VM_ConstructionMaterialType [0..*]
+ facadeMaterial: VM_FacadeMaterialType [0..*]
+ heatingSystem: VM_HeatingSystemType [0..1]
+ heatingSource: VM_HeatingSystemSource [0..1]
+ energyPerformance: VM_EnergyPerformanceValue [0..1]
+ numberOfRoom: Integer [0..1]
+ numberOfBedroom: Integer [0..1]
+ numberOfBathroom: Integer [0..1]
VersionedObject
«featureType»
VM_CondominiumUnit
+ condominiumUnitID: Oid
+ condominiumArea: LA_AreaValue [0..*]
+ floorNumber: CharacterString [0..1]
+ numberOfRoom: Integer
+ accesoryPart: Boolean [0..*]
+ accessoryPartType: AccessoryPartType [0..*]
+ shareInJointFacilities: Decimal [0..1]
2..*
0..*
1..*
0..*
«featureType»
Spatial Unit::LA_LegalSpaceBuildingUnit
+ extPhysicalBuildingUnitID: ExtPhysicalBuildingUnit [0..1]
+ type: LA_BuildingUnitType [0..1]
1..*
VersionedObject
«featureType»
VM_Building
+ buildingID: Oid
VersionedObject
«featureType»
VM_CondominiumBuilding
+ condominiumBuildingID: Oid
«codeList»
VM_BuildingAreaType
«codeList»
VM_BuildingVolumeType
LA_AreaValue
«dataType»
VM_AreaValue
LA_VolumeValue
«dataType»
VM_VolumeValue
«codeList»
VM_BuildingUseType
«codeList»
VM_HeatingSystemType
«codeList»
VM_ValuationUnitType
«codeList»
VM_NeighborhoodType
«codeList»
VM_ConstructionMaterialType
«codeList»
VM_HeatingSystemSource
«codeList»
VM_EnergyPerformanceValue
«codeList»
VM_AccessoryPartType
«codeList»
VM_FacadeMaterialType

Figure A- 1: The VM_ValuationUnit and related classes in the Valuation Information Model (LADM_VM) from (Kara et al. 2018)

Appendix B: IFC Extension and IFC-based Information Extraction

Table B- 1: The proposed property sets and properties to be added in the IfcSpace and IfcZone entities for property valuation

Property Set	Property Name	Property Type	Data Type
Pset_PV_Transaction	transactionID	IfcPropertySingleValue	IfcIdentifier
	registrationDate	IfcPropertySingleValue	IfcDateTime
	activeDays	IfcPropertySingleValue	IfcInteger
	transferDate	IfcPropertySingleValue	IfcDateTime
	paidCategory	IfcPropertyEnumerated Value	IfcLable
	easement	IfcPropertyEnumerated Value	IfcLable
	mortgage	IfcPropertySingleValue	IfcBoolean
	rentalAnnotation	IfcPropertySingleValue	IfcBoolean
	noOfFollowers	IfcPropertySingleValue	IfcInteger
	communityAverag ePrice	IfcPropertySingleValue	IfcReal
	propertyRights	IfcPropertyEnumerated Value	IfcLable
Pset_PV_Parcel	propertyNumber	IfcPropertySingleValue	IfcInteger
	parcelNumber	IfcPropertySingleValue	IfcInteger
	area	IfcPropertySingleValue	IfcAreaMeasure
	ID	IfcPropertySingleValue	IfcIdentifier
	parcelUseType	IfcPropertyEnumerated Value	IfcLable
	parcelGeometry	IfcPropertySingleValue	IfcBoolean

Property Set	Property Name	Property Type	Data Type
	parcelFrontage	IfcPropertySingleValue	IfcReal
	parcelLocation	IfcPropertyEnumerated Value	IfcLable
	city	IfcPropertyEnumerated Value	IfcLable
	district	IfcPropertyEnumerated Value	IfcLable
	town	IfcPropertyEnumerated Value	IfcLable
	county	IfcPropertyEnumerated Value	IfcLable
	longitude	IfcPropertyEnumerated Value	IfcLable
	latitude	IfcPropertyEnumerated Value	IfcLable
Pset_PV_Building	buildingID	IfcPropertySingleValue	IfcIdentifier
	totalArea	IfcPropertySingleValue	IfcAreaMeasure
	livingArea	IfcPropertySingleValue	IfcAreaMeasure
	garageArea	IfcPropertySingleValue	IfcAreaMeasure
	carportArea	IfcPropertySingleValue	IfcAreaMeasure
	builtDate	IfcPropertySingleValue	IfcDateTime
	noOfBedrooms	IfcPropertySingleValue	IfcInteger
	noOfDrawingRooms	IfcPropertySingleValue	IfcInteger
	noOfGarages	IfcPropertySingleValue	IfcInteger
	noOfBathrooms	IfcPropertySingleValue	IfcInteger
	noOfKitchens	IfcPropertySingleValue	IfcInteger
	bathroomType	IfcPropertyEnumerated Value	IfcLable

Property Set	Property Name	Property Type	Data Type
	garageType	IfcPropertyEnumerated Value	IfcLable
	buildingCategory	IfcPropertyEnumerated Value	IfcLable
	storey	IfcPropertySingleValue	IfcInteger
	height	IfcPropertySingleValue	IfcReal
	volume	IfcPropertySingleValue	IfcVolumeMeasure
	noOfFloors	IfcPropertySingleValue	IfcInteger
	constructionType	IfcPropertyEnumerated Value	IfcLable
	qualityType	IfcPropertyEnumerated Value	IfcLable
	renovationCondition	IfcPropertyEnumerated Value	IfcLable
	constructionDate	IfcPropertySingleValue	IfcDateTime
	heatingCooling	IfcPropertySingleValue	IfcBoolean
	centralCooling	IfcPropertySingleValue	IfcBoolean
	centralHeating	IfcPropertySingleValue	IfcBoolean
	structure	IfcPropertyEnumerated Value	IfcLable
	elevator	IfcPropertySingleValue	IfcInteger
	ladderRatio	IfcPropertySingleValue	IfcReal
	pool	IfcPropertySingleValue	IfcBoolean
	fireplace	IfcPropertySingleValue	IfcBoolean
	balcony	IfcPropertySingleValue	IfcBoolean
	energyEfficiency	IfcPropertyEnumerated Value	IfcLable

Property Set	Property Name	Property Type	Data Type
	indoorSoundLevel	IfcPropertyEnumerated Value	IfcLable
	indoorDaylight	IfcPropertySingleValue	IfcRatioMeasure
Pset_PV_Condominium Unit	condominiumUnitID	IfcPropertySingleValue	IfcIdentifier
	totalArea	IfcPropertySingleValue	IfcAreaMeasure
	livingArea	IfcPropertySingleValue	IfcAreaMeasure
	garageArea	IfcPropertySingleValue	IfcAreaMeasure
	carportArea	IfcPropertySingleValue	IfcAreaMeasure
	builtDate	IfcPropertySingleValue	IfcDateTime
	noOfBedrooms	IfcPropertySingleValue	IfcInteger
	noOfDrawingRooms	IfcPropertySingleValue	IfcInteger
	noOfGarages	IfcPropertySingleValue	IfcInteger
	noOfBathrooms	IfcPropertySingleValue	IfcInteger
	noOfKitchens	IfcPropertySingleValue	IfcInteger
	bathroomType	IfcPropertyEnumerated Value	IfcLable
	garageType	IfcPropertyEnumerated Value	IfcLable
	condominiumUnit Category	IfcPropertyEnumerated Value	IfcLable
	storey	IfcPropertySingleValue	IfcInteger
	height	IfcPropertySingleValue	IfcReal
	volume	IfcPropertySingleValue	IfcVolumeMeasure
	noOfFloors	IfcPropertySingleValue	IfcInteger

Property Set	Property Name	Property Type	Data Type
	constructionType	IfcPropertyEnumerated Value	IfcLable
	qualityType	IfcPropertyEnumerated Value	IfcLable
	renovationCondition	IfcPropertyEnumerated Value	IfcLable
	constructionDate	IfcPropertySingleValue	IfcDateTime
	heatingCooling	IfcPropertySingleValue	IfcBoolean
	centralCooling	IfcPropertySingleValue	IfcBoolean
	centralHeating	IfcPropertySingleValue	IfcBoolean
	structure	IfcPropertyEnumerated Value	IfcLable
	elevator	IfcPropertySingleValue	IfcInteger
	ladderRatio	IfcPropertySingleValue	IfcReal
	pool	IfcPropertySingleValue	IfcBoolean
	fireplace	IfcPropertySingleValue	IfcBoolean
	balcony	IfcPropertySingleValue	IfcBoolean
	energyEfficiency	IfcPropertyEnumerated Value	IfcLable
	indoorSoundLevel	IfcPropertyEnumerated Value	IfcLable
	indoorDaylight	IfcPropertySingleValue	IfcRatioMeasure
Pset_PV_Valuation	valuationID	IfcPropertySingleValue	IfcIdentifier
	valuationPurpose	IfcPropertyEnumerated Value	IfcLable
	valuationDate	IfcPropertySingleValue	IfcDateTime
	valuationMethod	IfcPropertyEnumerated Value	IfcLable

Property Set	Property Name	Property Type	Data Type
	Property value	IfcPropertySingleValue	IfcReal
Pset_PV_MassValuation	valuationID	IfcPropertySingleValue	IfcIdentifier
	valuationPurpose	IfcPropertyEnumeratedValue	IfcLable
	valuationDate	IfcPropertySingleValue	IfcDateTime
	algorithm	IfcPropertyEnumeratedValue	IfcLable
	Property value	IfcPropertySingleValue	IfcReal
Pset_PV_Annex	type	IfcPropertyEnumeratedValue	IfcLable
	sID	IfcPropertySingleValue	IfcIdentifier

Before developing an IFC-based information extraction algorithm, it is necessary to understand the types of information elements and their relationships between *IfcObject* and *IfcProperty* in an IFC-based instance model. Referring to IFC4-ADD2 schema, *IfcObject* and *IfcProperty* are linked directly and indirectly. The relationships between them are displayed in Figure B-1 and Figure B-2 (ISO 2018). On the one hand, they are directly linked through *IfcRelDefinesByProperties* (Figure B-1) – an objectified relationship that defines the relation between objects and property sets. In an IFC instance model, an instance of a property set (*IfcPropertySetDefinition*) is directly linked to an instance of a building object (*IfcObjectDefinition*) by *IfcRelDefinesByProperties.* For example, a specific instance of *IfcWindow* can be associated with a specific instance of *IfcPropertySet* through *IfcRelDefinesByProperties.*

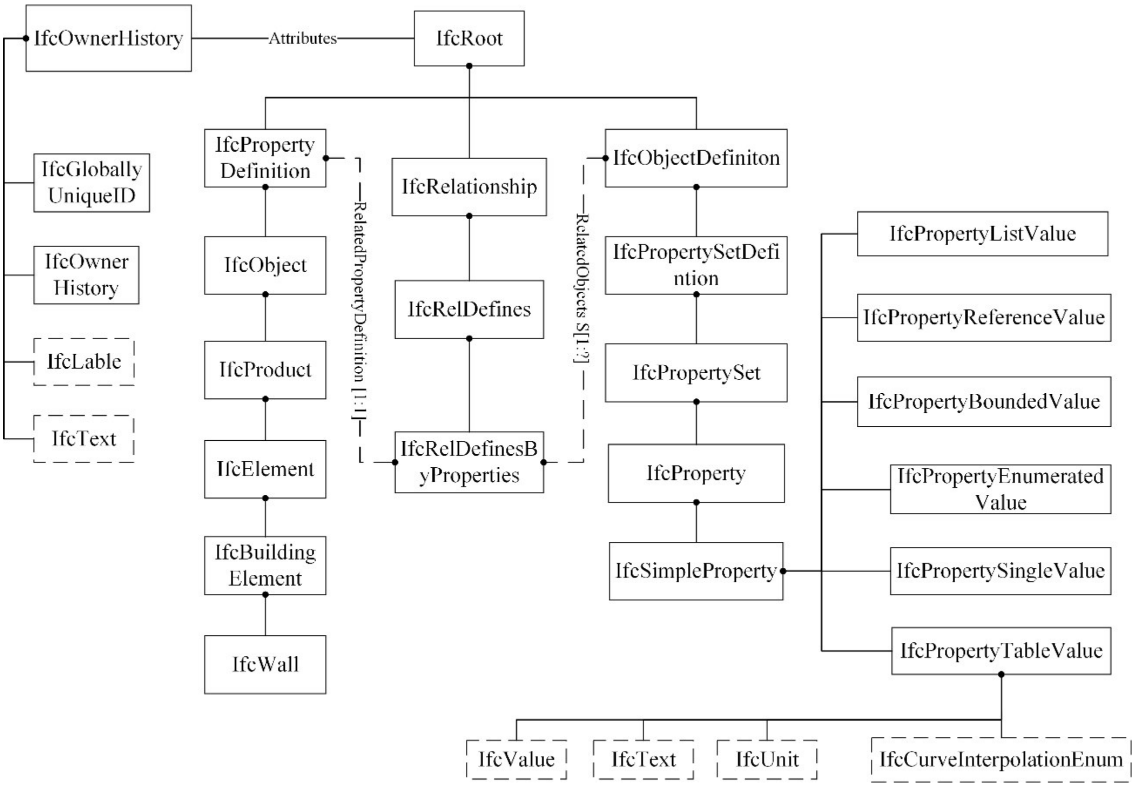


Figure B- 1: Direct relationships between *IfcObject* and *IfcProperty* in IFC4-ADD2

On the other hand, building objects and their properties are indirectly linked through *IfcRelDefinesByType* (Figure B-2). *IfcRelDefinesByType* defines the relationship between an object type and object occurrences which can leverage a one-to-N relationship. The one-to-N relationship can link one or more objects to one object type, which means these objects are sharing the same object type and related property sets (ISO 2018). For instance, multiple instances of *IfcSlab* can be related to an instance of *IfcSlabType* through *IfcRelDefinesByType,* and all these instances of *IfcSlab* share the same property sets assigned to the specific instance of *IfcSlabType*.

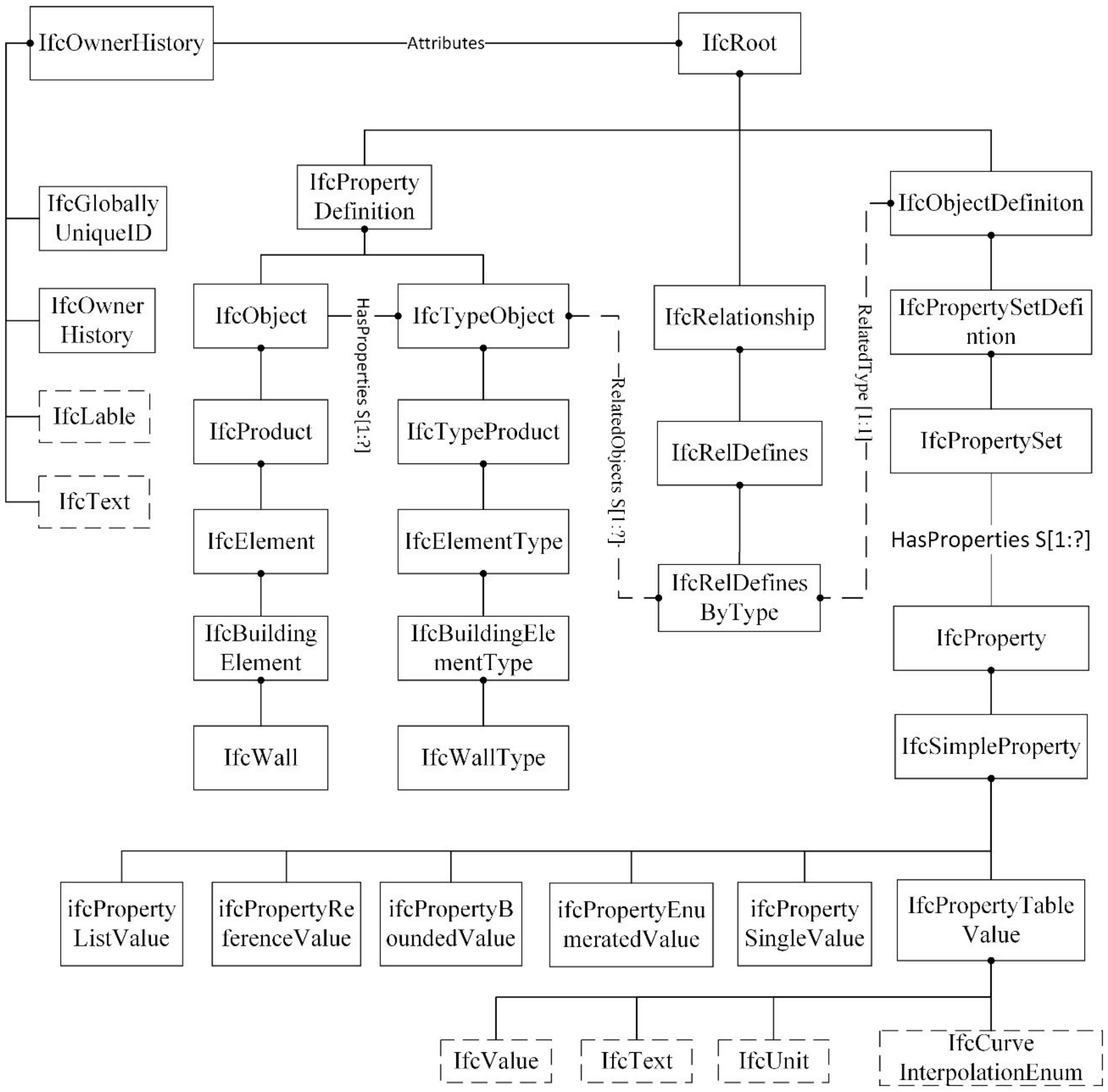


Figure B- 2: Indirect relationships between *IfcObject* and *IfcProperty* in IFC4-ADD2

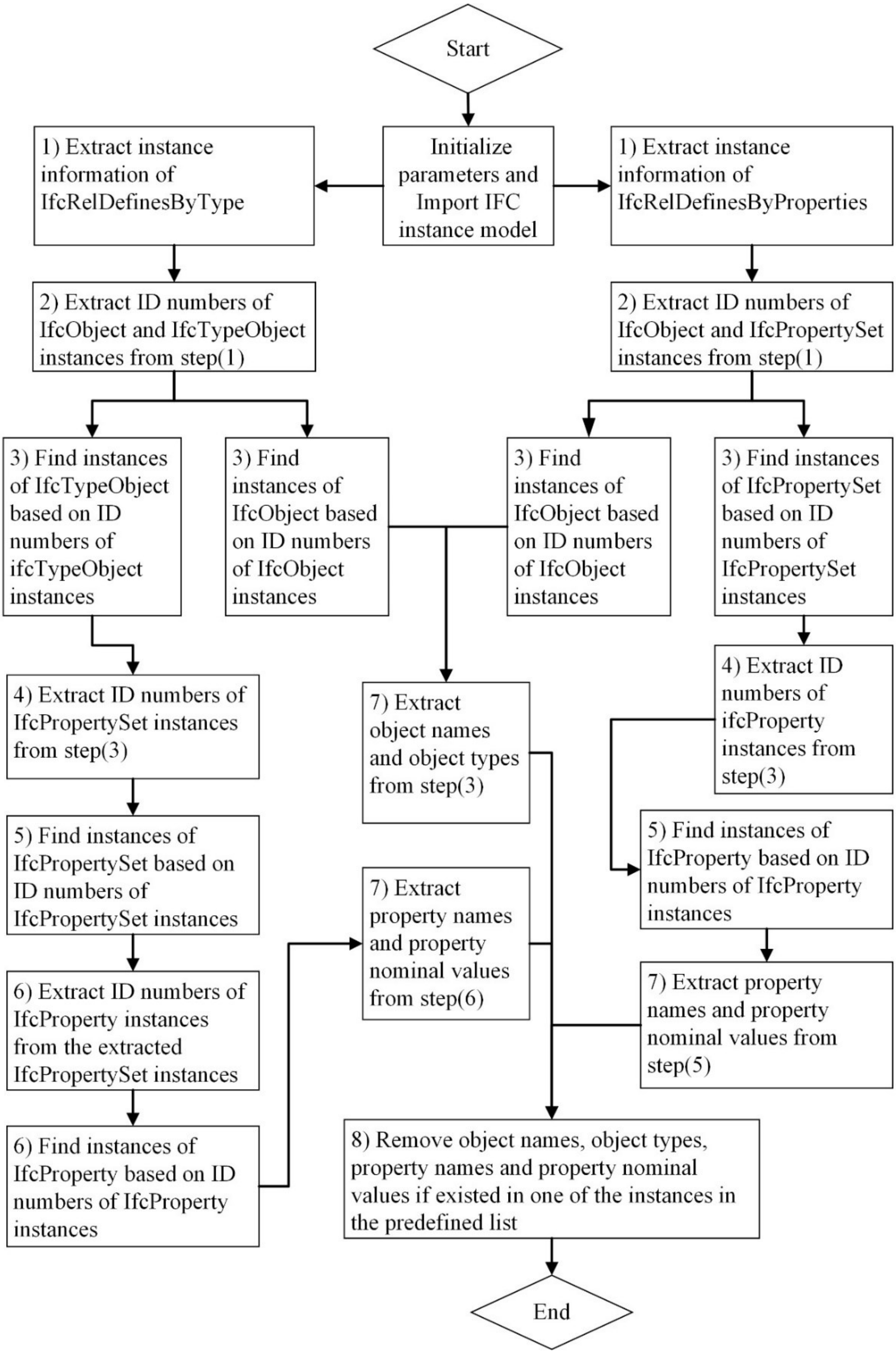

Start
Initialize parameters and Import IFC instance model
1) Extract instance information of IfcRelDefinesByType
1) Extract instance information of IfcRelDefinesByProperties
2) Extract ID numbers of IfcObject and IfcTypeObject instances from step(1)
2) Extract ID numbers of IfcObject and IfcPropertySet instances from step(1)
3) Find instances of IfcTypeObject based on ID numbers of ifcTypeObject instances
3) Find instances of IfcObject based on ID numbers of IfcObject instances
3) Find instances of IfcObject based on ID numbers of IfcObject instances
3) Find instances of IfcPropertySet based on ID numbers of IfcPropertySet instances
4) Extract ID numbers of IfcPropertySet instances from step(3)
7) Extract object names and object types from step(3)
4) Extract ID numbers of ifcProperty instances from step(3)
5) Find instances of IfcPropertySet based on ID numbers of IfcPropertySet instances
7) Extract property names and property nominal values from step(6)
5) Find instances of IfcProperty based on ID numbers of IfcProperty instances
6) Extract ID numbers of IfcProperty instances from the extracted IfcPropertySet instances
7) Extract property names and property nominal values from step(5)
6) Find instances of IfcProperty based on ID numbers of IfcProperty instances
8) Remove object names, object types, property names and property nominal values if existed in one of the instances in the predefined list
End

Figure B- 3: Flowchart of the developed IFC-based information extraction algorithm

The extraction based on direct relationships was displayed in Figure B-4 (a).

First, an *IfcRelDefinesByProperties* instance with the ID number of #563 was extracted. Subsequently, the ID numbers of an *IfcWall* instance (#215) and an *IfcPropertySet* instance (#558) were extracted from *IfcRelDefinesByProperties* instance (#563).

Second, an *IfcWall* instance (#215) and an *IfcPropertySet* instance (#558) were found by the algorithm. Subsequently, the ID number of the *IfcProperty* instance (#557) was extracted from the *IfcPropertySet* instance (#558).

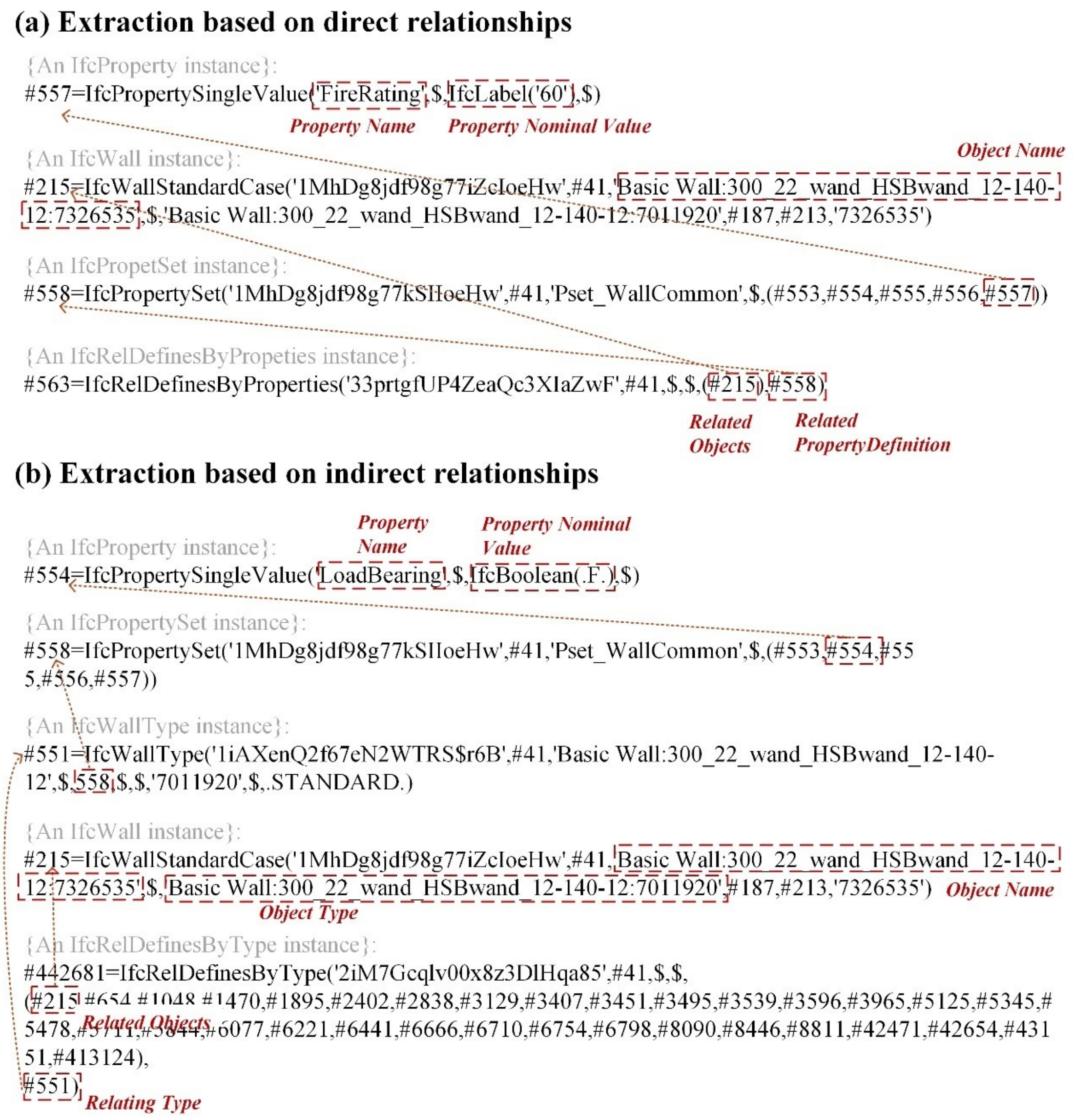
Third, an *IfcProperty* instance (#557) was found by the algorithm. After that, the object name (Basic Wall:300_22_wand_HSBwand_12-140-12:7326535), property name (*FireRating*) and property nominal value (*IfcLabel('60'))* were extracted with removed duplicated data from the *IfcWall* instance (#215) and the *IfcProperty* instance (#557).

The extraction based on indirect relationships is displayed in Figure B-3 (b).

First, an *IfcRelDefinesByType* instance with the ID number of #442681 was extracted. Subsequently, the ID numbers of an *IfcWall* instance (#215) and an *IfcWallType* instance (#551) were extracted from instance (#442681).

Second, an *IfcWall* instance (#215) and an *IfcWallType* instance (#551) were found by the algorithm. Subsequently, the ID number of the *IfcPropertySet* instance (#558) was extracted from the *IfcPropertySet* instance (#551).

Third, an *IfcPropertySet* instance (#558) was found by the algorithm. After that, the object name (Basic Wall:300_22_wand_HSBwand_12-140-12:7326535), object type (Basic Wall:300_22_wand_HSBwand_12-140-12:7011920), property name (*LoadBearing*) and property nominal value (*IfcBoolean(.F.))* were extracted with removed duplicated data from the *IfcWall* instance (#215) and the *IfcProperty* instance (#554).



(a) Extraction based on direct relationships

```
{An IfcProperty instance}:
#557=IfcPropertySingleValue('FireRating',$,IfcLabel('60'),$)

{An IfcWall instance}:
#215=IfcWallStandardCase('1MhDg8jdf98g77iZcIoeHw',#41,'Basic Wall:300_22_wand_HSBwand_12-140-
12:7326535',$,'Basic Wall:300_22_wand_HSBwand_12-140-12:7011920',#187,#213,'7326535')

{An IfcPropetSet instance}:
#558=IfcPropertySet('1MhDg8jdf98g77kSIIoeHw',#41,'Pset_WallCommon',$,(#553,#554,#555,#556,#557))

{An IfcRelDefinesByPropeties instance}:
#563=IfcRelDefinesByProperties('33prtgfUP4ZeaQc3XIaZwF',#41,$,$,(#215),#558)
```

(b) Extraction based on indirect relationships

```
{An IfcProperty instance}:
#554=IfcPropertySingleValue('LoadBearing',$,IfcBoolean(.F.),$)

{An IfcPropertySet instance}:
#558=IfcPropertySet('1MhDg8jdf98g77kSIIoeHw',#41,'Pset_WallCommon',$,(#553,#554,#55
5,#556,#557))

{An IfcWallType instance}:
#551=IfcWallType('1iAXenQ2f67eN2WTRS$r6B',#41,'Basic Wall:300_22_wand_HSBwand_12-140-
12',$,(#558),$,$,'7011920',$,.STANDARD.)

{An IfcWall instance}:
#215=IfcWallStandardCase('1MhDg8jdf98g77iZcIoeHw',#41,'Basic Wall:300_22_wand_HSBwand_12-140-
12:7326535',$,'Basic Wall:300_22_wand_HSBwand_12-140-12:7011920',#187,#213,'7326535')

{An IfcRelDefinesByType instance}:
#442681=IfcRelDefinesByType('2iM7Gcqlv00x8z3DlHqa85',#41,$,$,
(#215,#654,#1048,#1470,#1895,#2402,#2838,#3129,#3407,#3451,#3495,#3539,#3596,#3965,#5125,#5345,#
5478,#5711,#5844,#6077,#6221,#6441,#6666,#6710,#6754,#6798,#8090,#8446,#8811,#42471,#42654,#431
51,#413124),
#551)
```

Figure B- 4: An example for using the IFC-based information extraction algorithm

Appendix C: Feature Importance Ranking Calculated by Decision Tree – based Embedded Methods

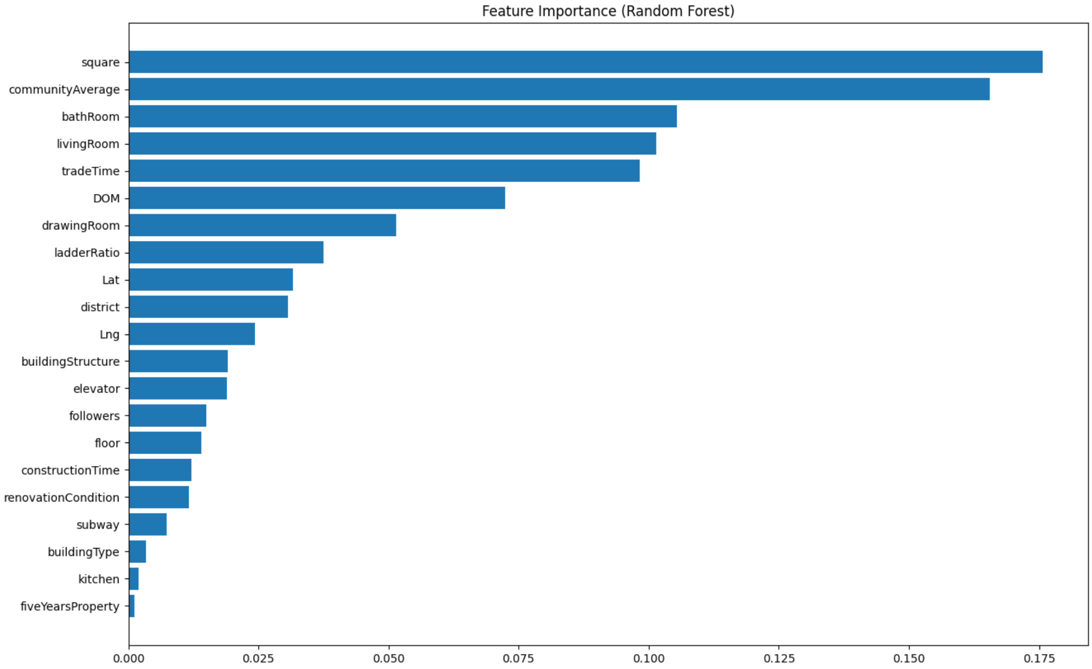


Figure C- 1: Feature importance ranking calculated by GBDT in the Chinese dataset

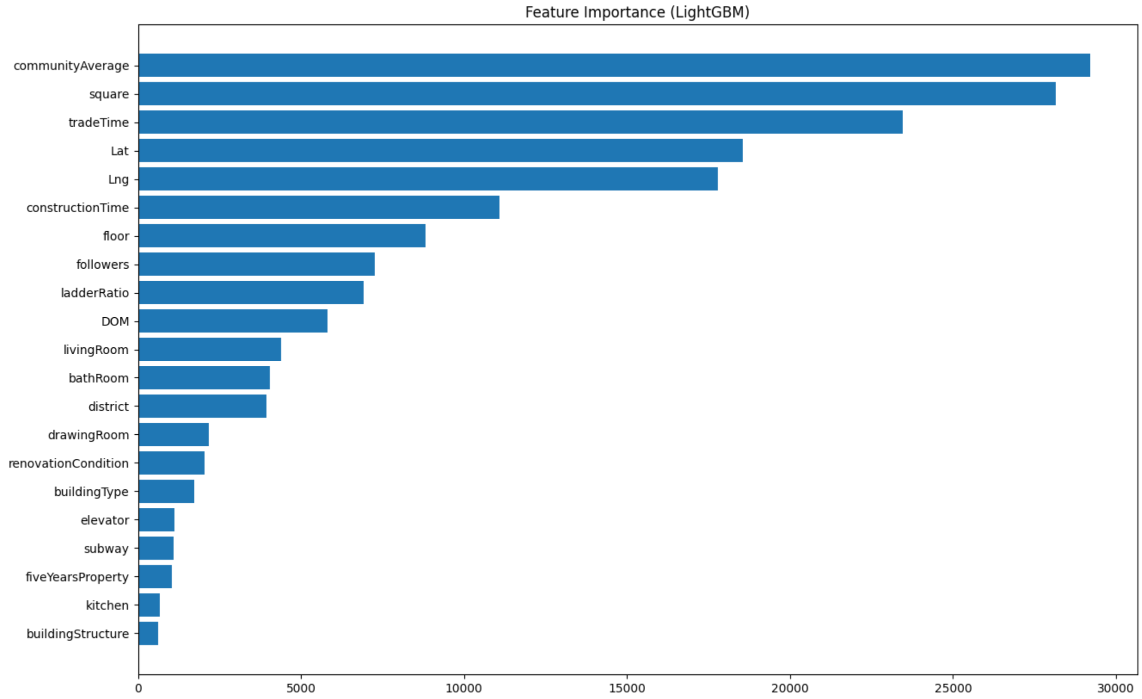


Figure C- 2: Feature importance ranking calculated by LightGBM in the Chinese dataset

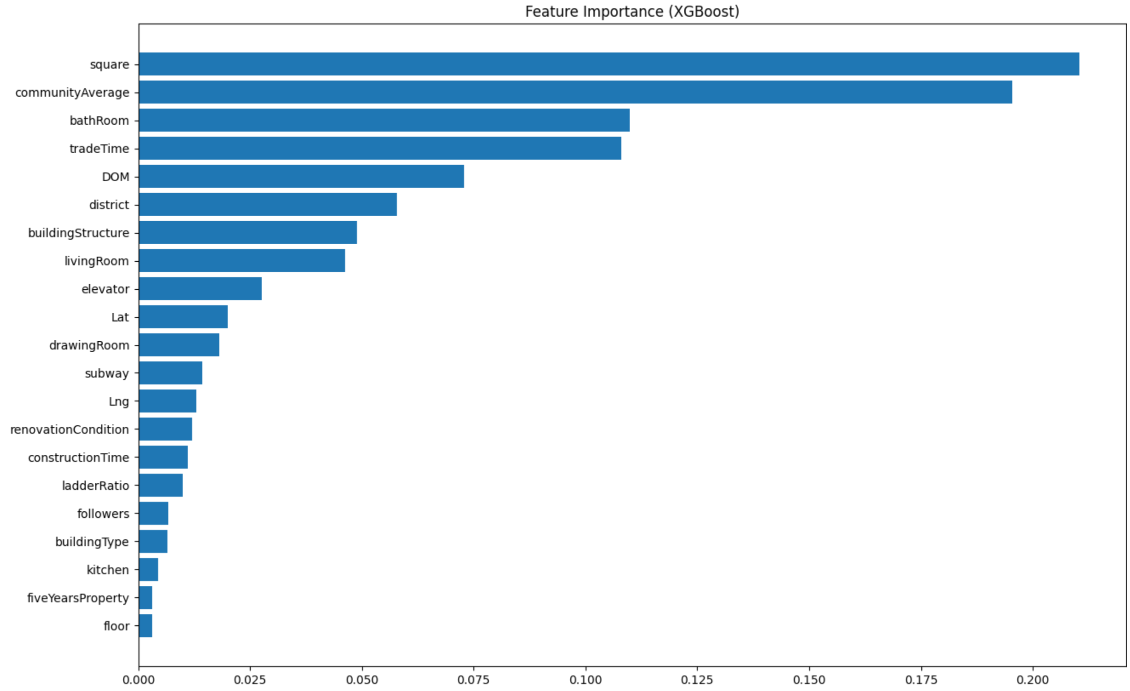


Figure C- 3: Feature importance ranking calculated by XGBoost in the Chinese dataset

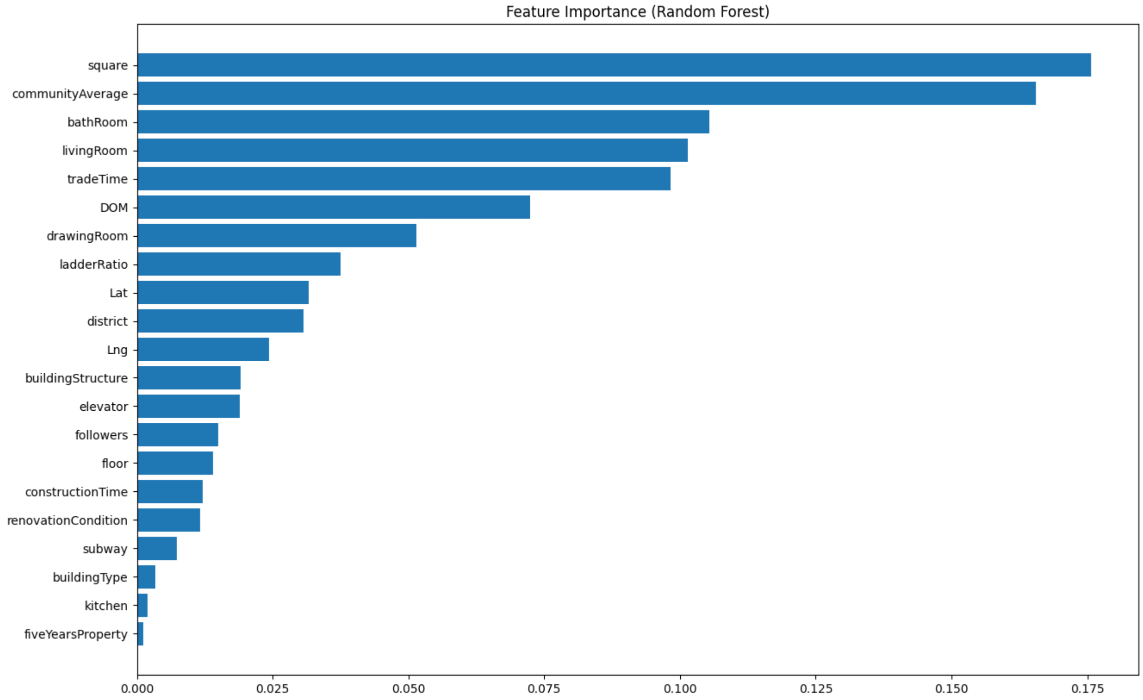


Figure C- 4: Feature importance ranking calculated by Random Forest in the Chinese dataset

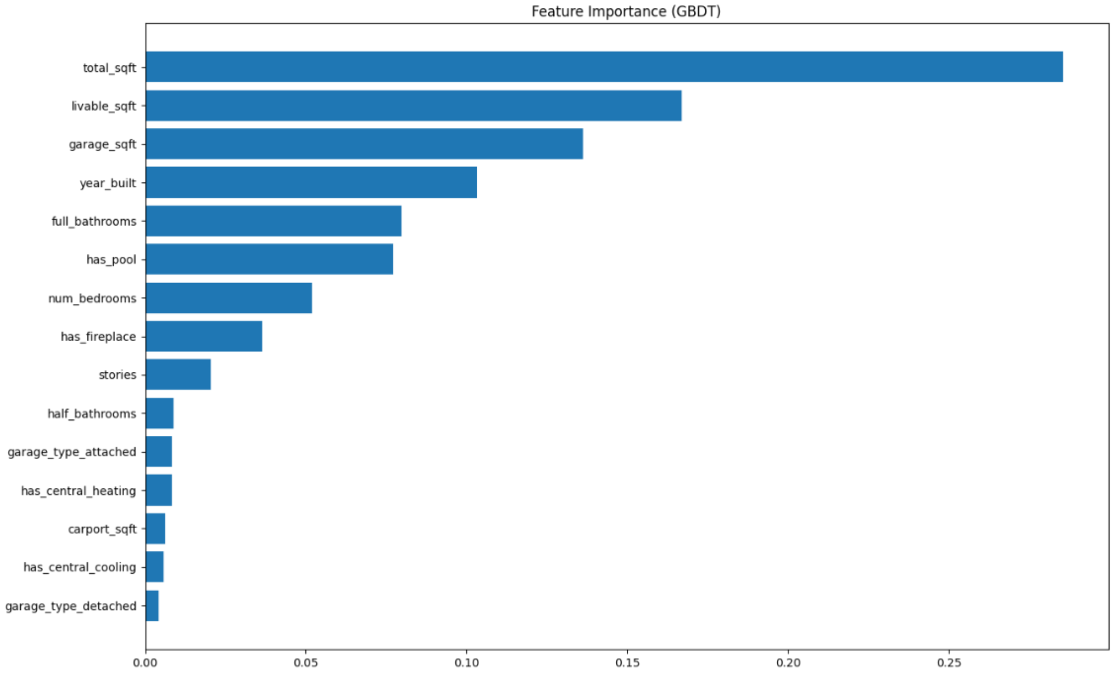


Figure C- 5: Feature importance ranking calculated by GBDT in the American dataset

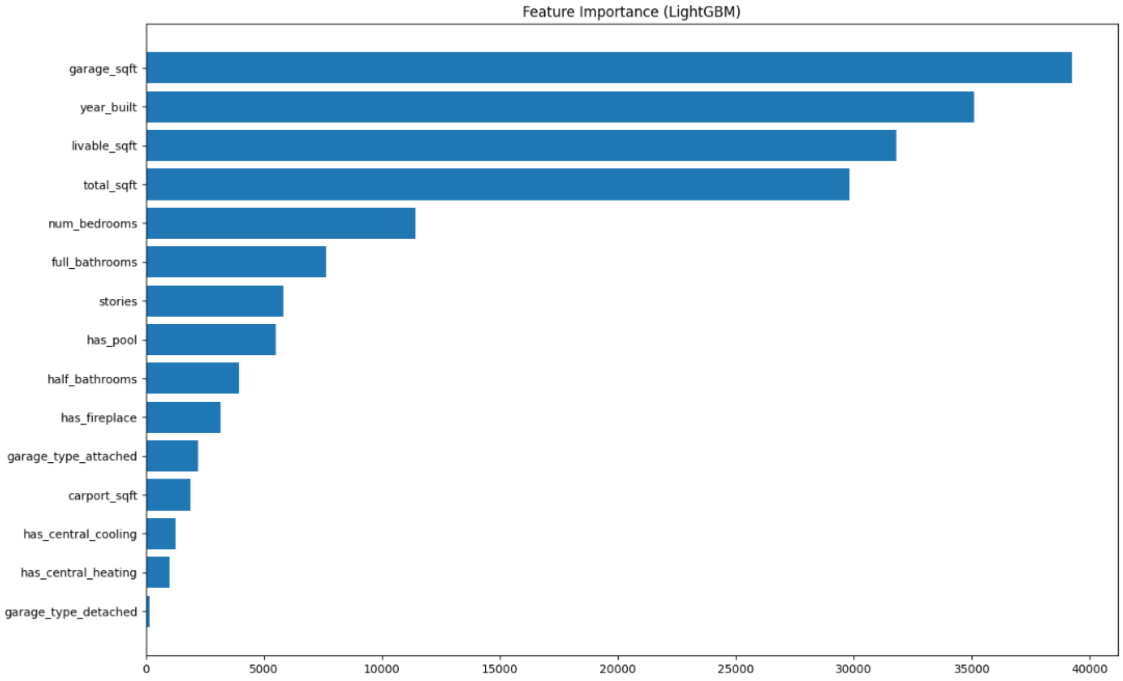


Figure C- 6: Feature importance ranking calculated by LightGBM in the American dataset

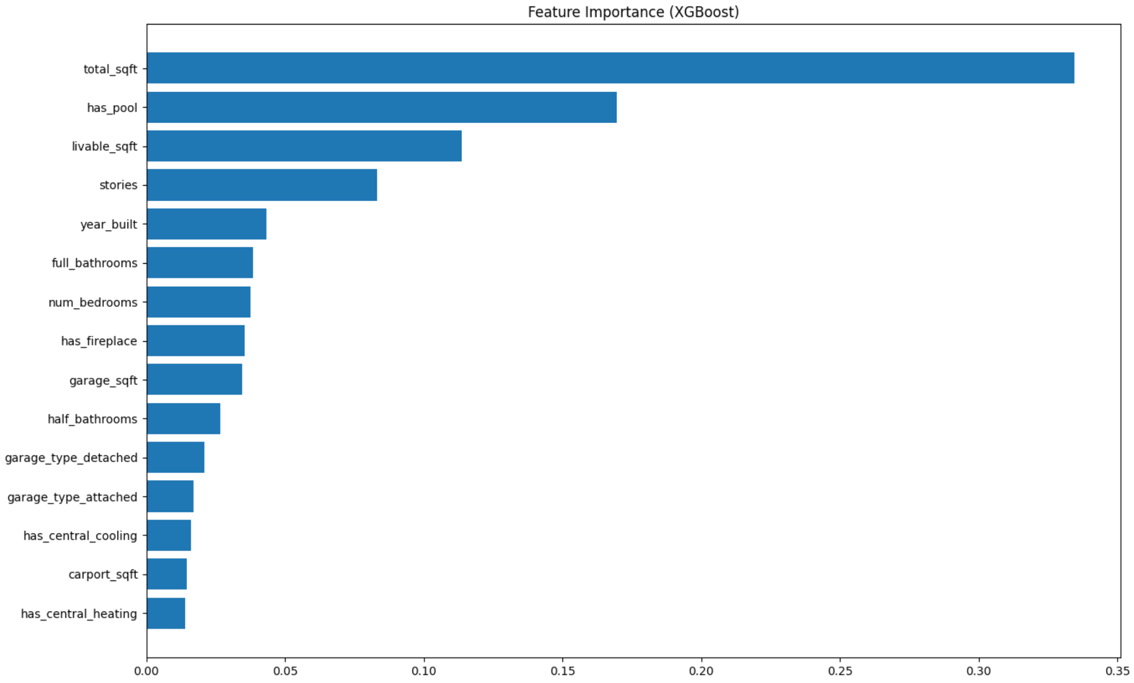


Figure C- 7: Feature importance ranking calculated by XGBoost in the American dataset

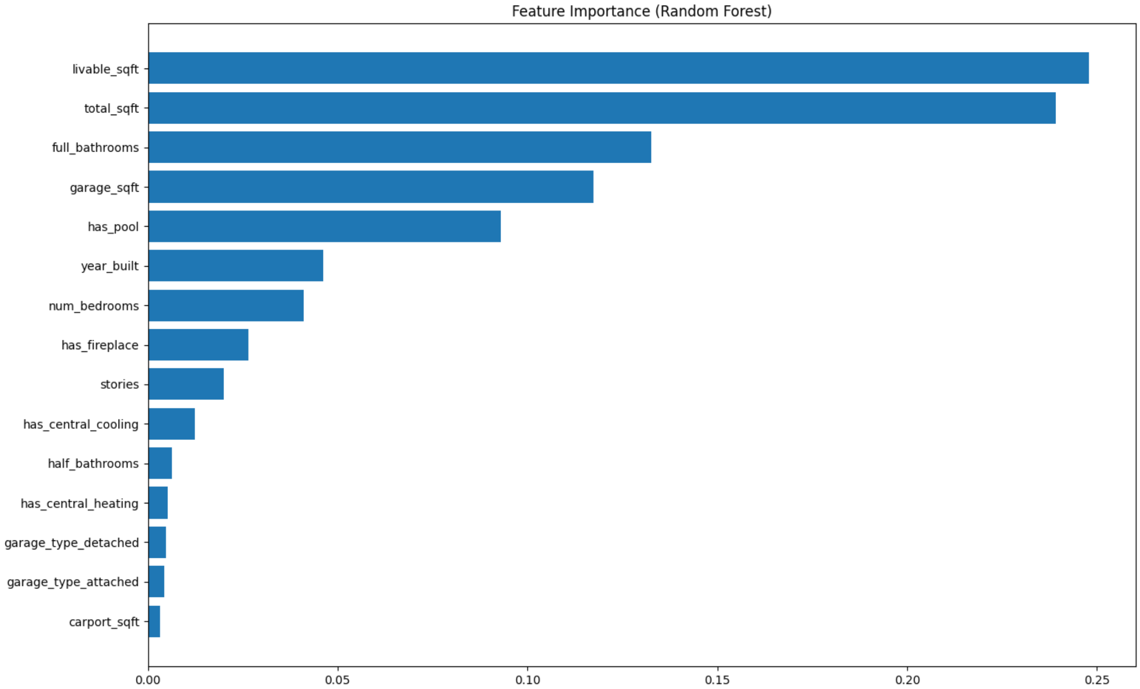


Figure C- 8: Feature importance ranking calculated by Random Forest in the American dataset

Appendix D: Feature Importance Ranking Calculated by GBR and GA-GBR Model

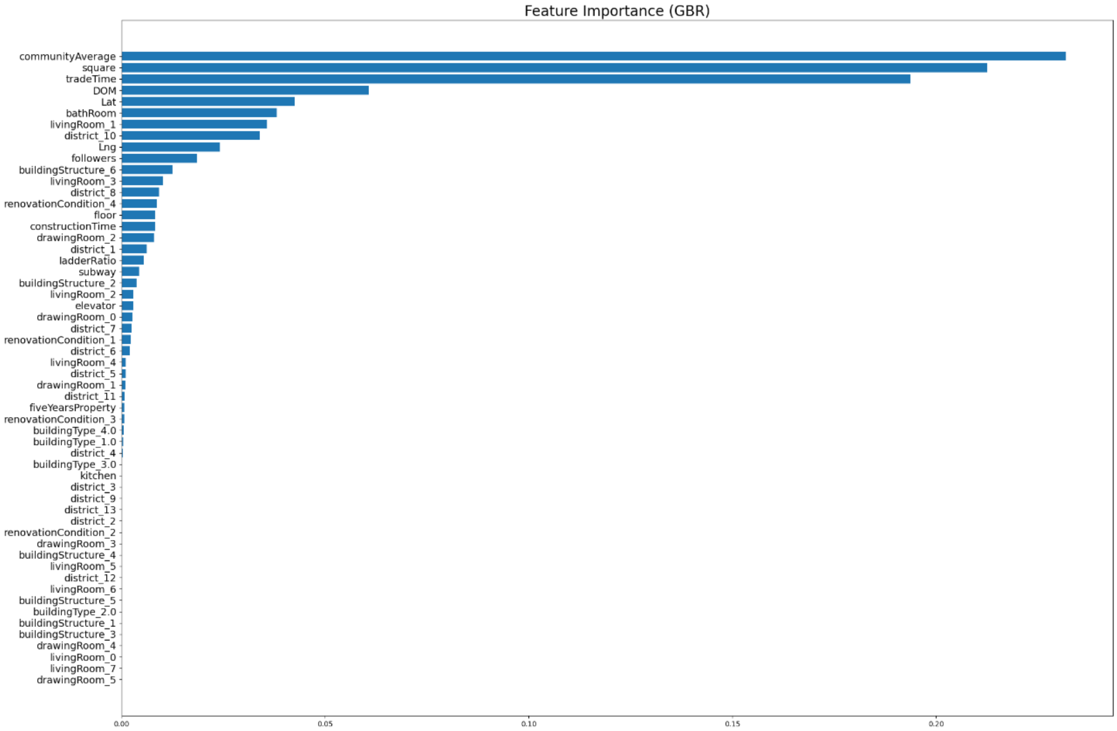


Figure D- 1: Feature importance ranking calculated by GBR model in the Chinese dataset

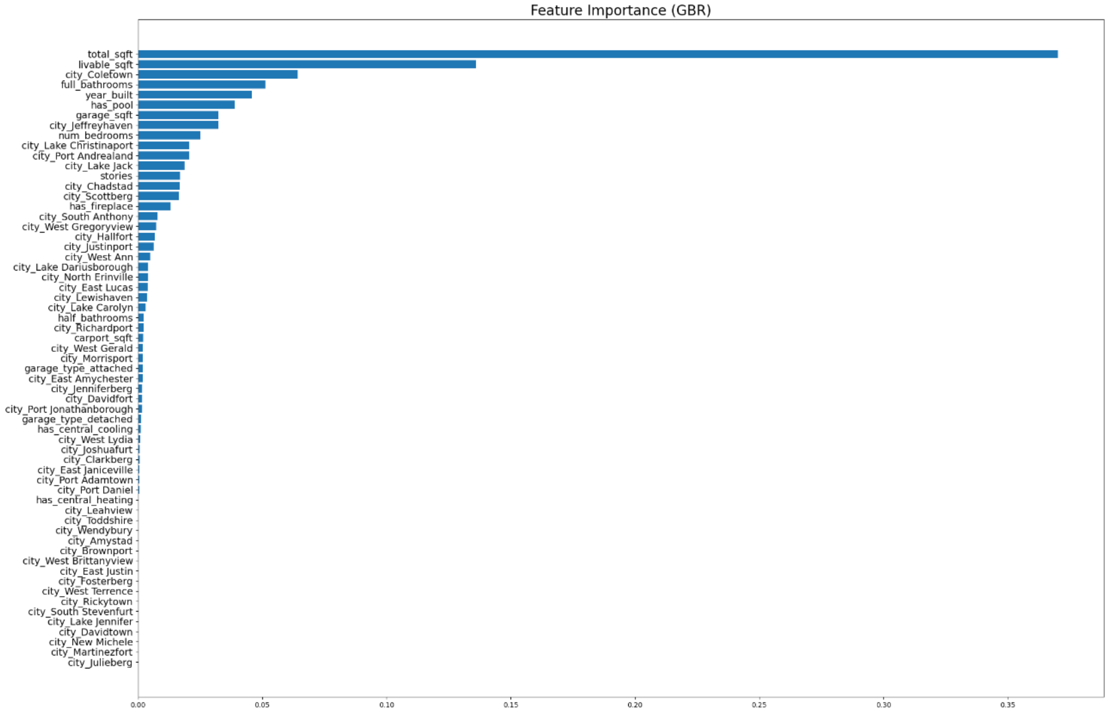


Figure D- 2: Feature importance ranking calculated by GBR model in the American dataset

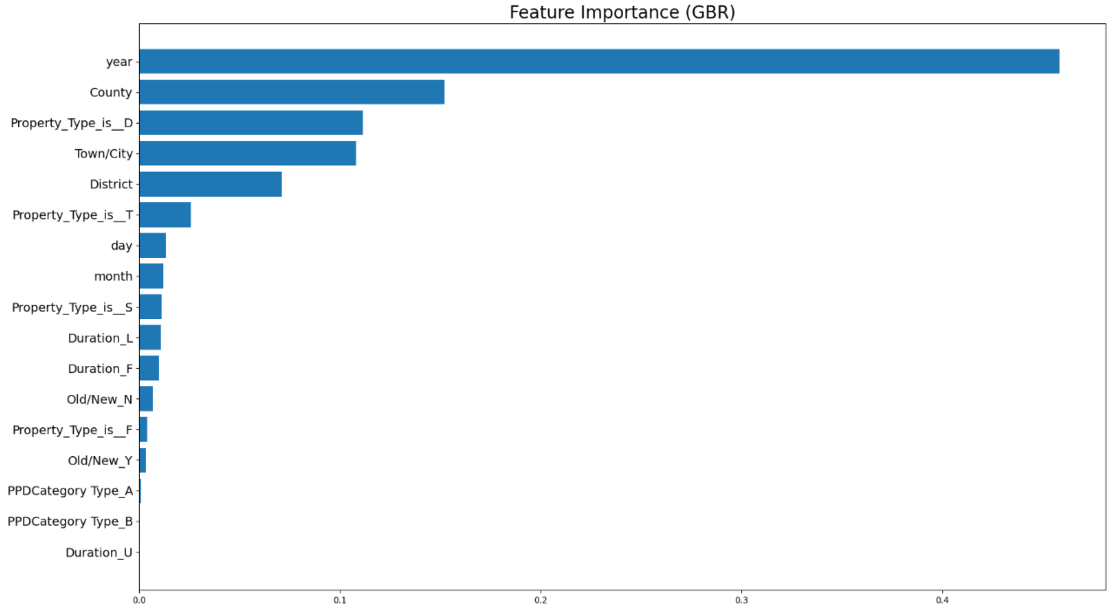


Figure D- 3: Feature importance ranking calculated by GBR model in UK dataset

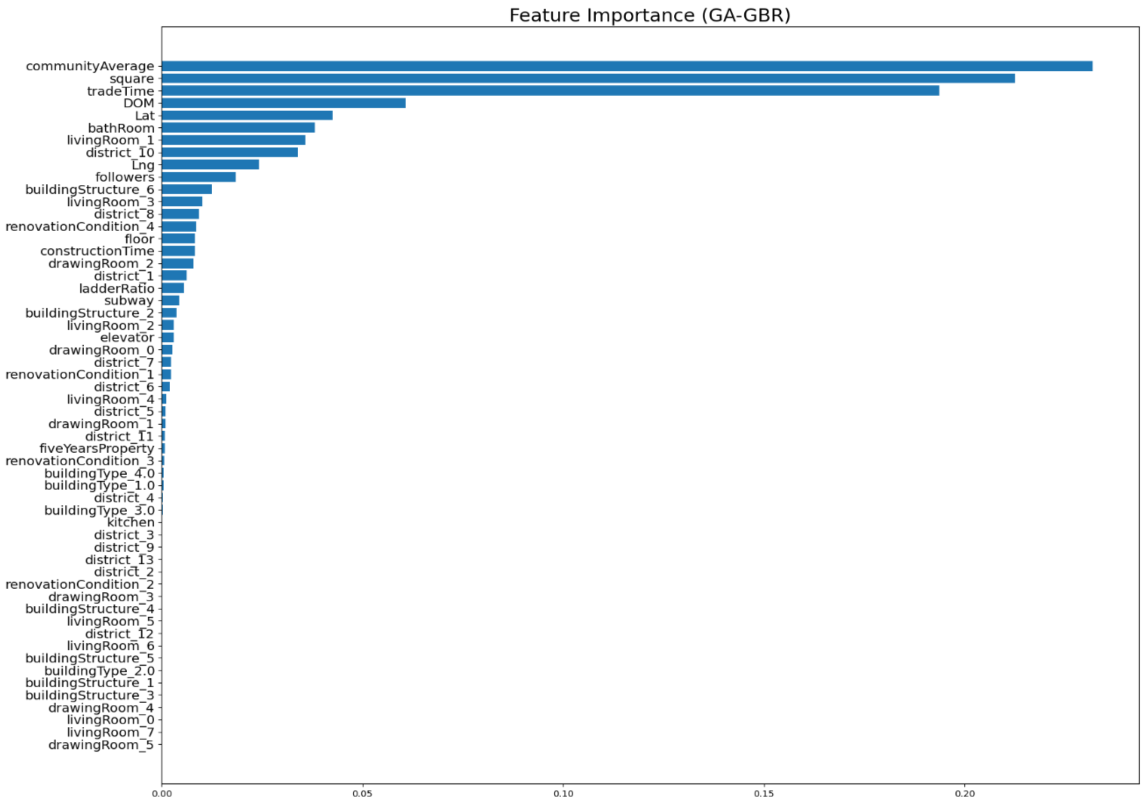


Figure D- 4: Feature importance ranking calculated by GA-GBR model in the Chinese dataset

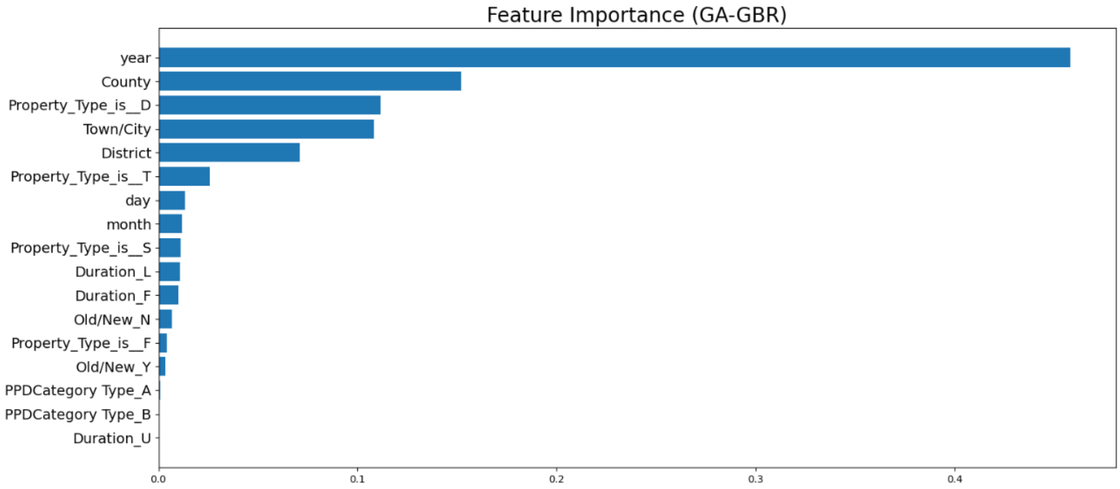


Figure D- 5: Feature importance ranking calculated by GA-GBR model in UK dataset

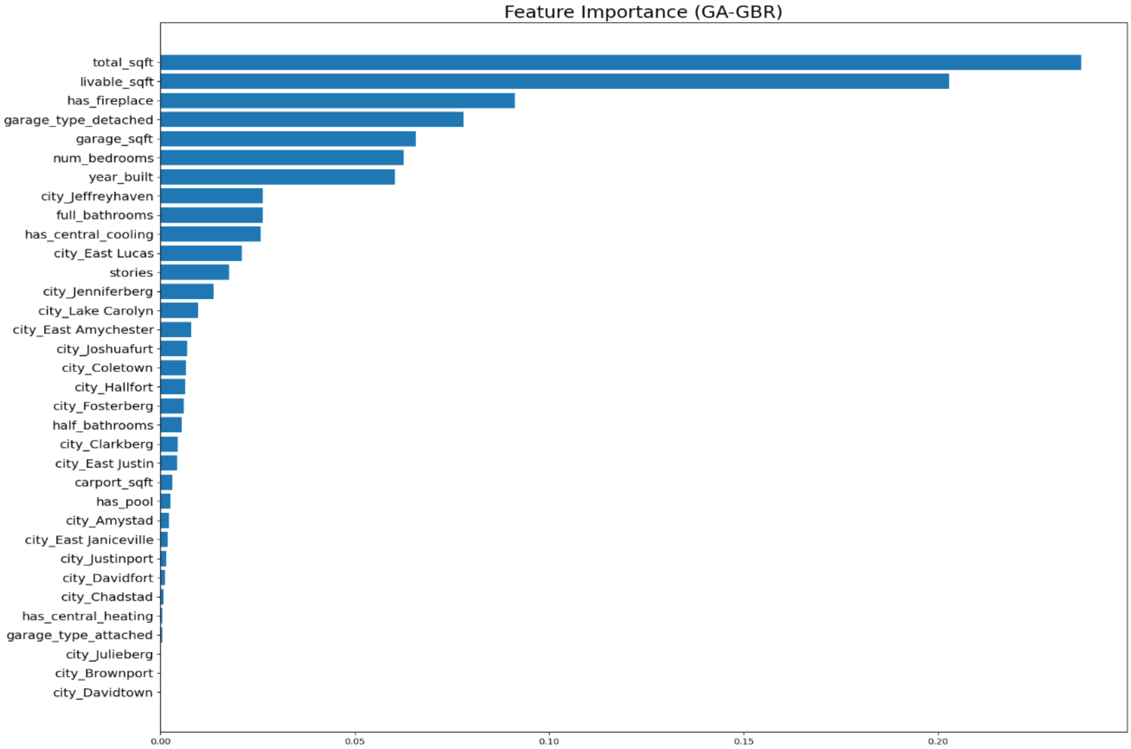
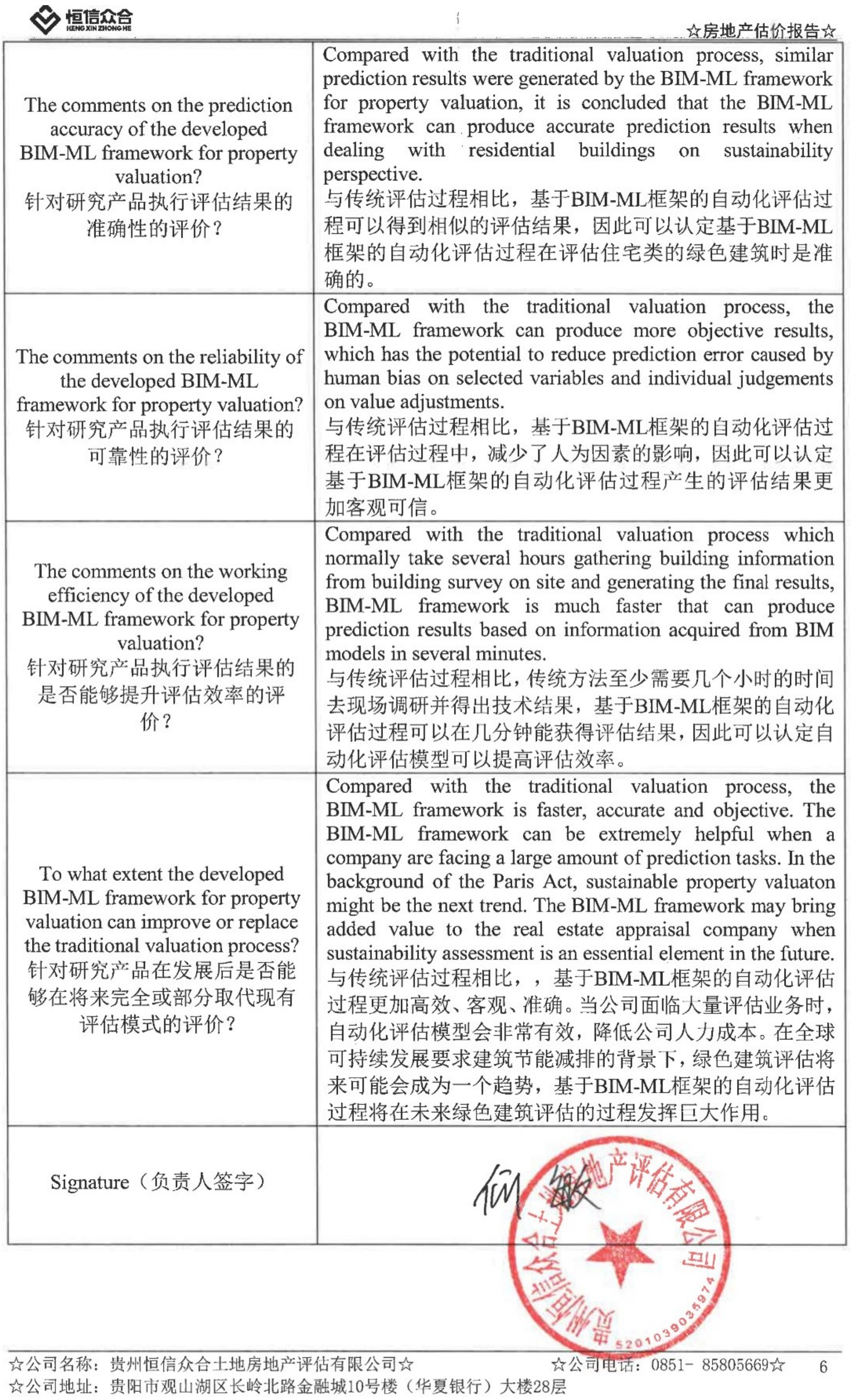


Figure D- 6: Feature importance ranking calculated by GA-GBR model in the American dataset.

Appendix E: Delivered Feedback Form from Industry - Validation of the Comprehensive BIM-ML Integration Framework



恒信众合
HENG XIN ZHONGHE

☆房地产估价报告☆

The comments on the prediction accuracy of the developed BIM-ML framework for property valuation? 针对研究产品执行评估结果的准确性的评价？	Compared with the traditional valuation process, similar prediction results were generated by the BIM-ML framework for property valuation, it is concluded that the BIM-ML framework can produce accurate prediction results when dealing with residential buildings on sustainability perspective. 与传统评估过程相比，基于BIM-ML框架的自动化评估过程可以得到相似的评估结果，因此可以认定基于BIM-ML框架的自动化评估过程在评估住宅类的绿色建筑时是准确的。
The comments on the reliability of the developed BIM-ML framework for property valuation? 针对研究产品执行评估结果的可靠性的评价？	Compared with the traditional valuation process, the BIM-ML framework can produce more objective results, which has the potential to reduce prediction error caused by human bias on selected variables and individual judgements on value adjustments. 与传统评估过程相比，基于BIM-ML框架的自动化评估过程在评估过程中，减少了人为因素的影响，因此可以认定基于BIM-ML框架的自动化评估过程产生的评估结果更加客观可信。
The comments on the working efficiency of the developed BIM-ML framework for property valuation? 针对研究产品执行评估结果的是否能够提升评估效率的评价？	Compared with the traditional valuation process which normally take several hours gathering building information from building survey on site and generating the final results, BIM-ML framework is much faster that can produce prediction results based on information acquired from BIM models in several minutes. 与传统评估过程相比，传统方法至少需要几个小时的时间去现场调研并得出技术结果，基于BIM-ML框架的自动化评估过程可以在几分钟能获得评估结果，因此可以认定自动化评估模型可以提高评估效率。
To what extent the developed BIM-ML framework for property valuation can improve or replace the traditional valuation process? 针对研究产品在发展后是否能够在将来完全或部分取代现有评估模式的评价？	Compared with the traditional valuation process, the BIM-ML framework is faster, accurate and objective. The BIM-ML framework can be extremely helpful when a company are facing a large amount of prediction tasks. In the background of the Paris Act, sustainable property valuaton might be the next trend. The BIM-ML framework may bring added value to the real estate appraisal company when sustainability assessment is an essential element in the future. 与传统评估过程相比，，基于BIM-ML框架的自动化评估过程更加高效、客观、准确。当公司面临大量评估业务时，自动化评估模型会非常有效，降低公司人力成本。在全球可持续发展要求建筑节能减排的背景下，绿色建筑评估将来可能会成为一个趋势，基于BIM-ML框架的自动化评估过程将在未来绿色建筑评估的过程发挥巨大作用。
Signature（负责人签字）	何敏 贵州恒信众合土地房地产评估有限公司 5201039035974

☆公司名称：贵州恒信众合土地房地产评估有限公司☆ ☆公司电话：0851- 85805669☆ 6
☆公司地址：贵阳市观山湖区长岭北路金融城10号楼（华夏银行）大楼28层

Figure E- 1: The feedback document of the comprehensive BIM-ML framework from the HXZH commercial real estate appraisal Co., Ltd.