Feature Importance Analysis for Customer Management of Insurance Products

Misbah Sohail, Pedro Peres, Yuhua Li
School of Computer Science & Informatics
Cardiff University, UK
{sohailm, GanhaoPeresP, LiY180}@cardiff.ac.uk

***Abstract*— Optimizing customer contact strategies is important to improving customer experience, increasing sales and improving business profitability. This study focuses on finding an optimal time to contact customers, using demographic features provided by a private insurance broker and area characteristics from national census data. We train machine learning models and interpret the results using SHAP to analyze how each feature explains customer contactability. Among all the interesting results, we find that for older people, the best time to contact is during late evenings and nights.**

Keywords—Machine Learning, LightGBM, Optuna, SHAP Analysis, Insurance

I. INTRODUCTION

Contacting a person over landline telephone or mobile phone for conducting surveys, interviews and customer service has become a tedious, resource draining and inefficient task. Frequent call backs are usually required to successfully reach the intended person. However, keeping in mind the limitations in time, money, and workforce, it is practically impossible to perform frequent call backs to all customers. Finding an effective solution which could accurately schedule calls such that the call becomes successful on the first trial has become vital.

Early research of call scheduling focused on finding the optimal call times and cost-effective techniques for conducting surveys and interviews via telephone. Telephonic interviews were considered to be the cheapest and most efficient form of communication when compared to personal or mail-based surveys [1] [2]. Different methods which include parametric and nonparametric empirical methods were applied to detect productive calling times [3]. The results all showed strong consistency in regard to timing. However, in today's era, most people have switched to a mode of wireless communication, and mainly due to its nature of mobility, the findings obtained from the studies related to the scheduling for telephone samples may no longer be applicable.

We hypothesize that previous call trials, a person's demographics and area characteristics altogether affect predicting the best time to conduct calls. Effective models which consider all the above-mentioned factors together are yet to be developed. Hence, there is a need for exploiting the rich customer data of telephone and mobile phone users to develop an effective call scheduling strategy.

In this paper, we made the following novel contributions. First, we generate more meaningful features out of the raw customer data by performing feature engineering. Second, we employ state-of-the-art machine learning model, LightGBM [4], to model complex customer data and use a relatively new framework, Optuna [5], to optimize model hyperparameters. Third, we apply SHAP (SHapley Additive exPlanations) [6] to interpret model prediction, identifying and presenting the most important features that influence the probability of successful calls.

The remainder of the paper is structured as follows: Section II reviews related research work on customer contactability; Section III explains the proposed methodology and framework; Section IV presents experiment details and results analysis; Section V concludes the paper with a summary of findings and a discussion of further study for improvements.

II. RELATED WORK

Back during the early 1980s, Weeks, et al. [7]performed an empirical study to schedule the best time to contact people for conducting surveys in person, followed by Vigderhous's work [8], whose goal was to optimize landline calling-based interviews. They both, however, produced similar results, concluding that weekday evenings are the best time to contact people.

Weeks, et al. [9] later researched on how the results of previous trial calls could be utilized to improve the chances of successful calls for landlines. The results were consistent with the above studies found for the first trial.

Massey, et al.[10], followed by Triplett [11], applied probabilistic approach and ordinary least squares regression respectively, to analyze different demographic features determining productivity of random digit dial (RDD) telephone surveys. The results were similar, with people's age and gender directly correlated with how they respond to surveys.

Moving away from probabilistic methods, Stokes & Greenberg [12] and Brick, et al. [13] built simple machine learning models, using linear regression, to predict probability of a successful call to an individual at a particular time. The

common independent variables for both the works included features associated with the calling history and the timings of current call. As well as the call related variables, Brick, et al. [13] also added the feature of area characteristics identified by the telephone numbers, such as median year of education and logs of median home value. For both studies all factors were found to be significant predictors for the outcome. Calling history did not emerge as a good predictor for Brick, et al.

Given the prevalent increase in the usage of wireless phones during the last two decades [14], the literature significantly shifted its focus to finding out the feasibility and effectiveness of conducting surveys using this new mode of communication [15] [16] [17] [18]. The studies suggested that the response rate is lower when contacted over mobile phone due to several different factors involved, such as the user being busy or driving. However, the study of Link, et al. [19]contradicted the above findings deducing that a general reluctance is observed in responding to surveys, regardless of the gadget being used by the receiver.

In evaluating the optimal time to contact on mobile phone, Zuwallack 's study [20] put forward the result that any time after 5 p.m. during weekdays is considered to be most productive while the second most is in the afternoons on weekends. Carley-Baxter, et al. [21] found that all times (of day) and days (of week) carry equal weight in terms of contactability. Reimer, et al. [22], along with devising the best times, performed logistic regression to analyze the effect of time of call, lag between calls and previous number of trials towards the outcome of existing calls. Reimer, et al. [22] devised that the best time to call are weekday afternoons. They also advised to wait longer to call back.

Similar to [10] and [11], Vicente & Lopes [23] aimed to find differences in contacting different groups of people in respect of age, gender, education and location (in house or outside the house). It was concluded that young men were harder to contact. The results in terms of gender even after more than a decade remained coherent.

Shino & McCarty [3] analyzed the effect of time over the respondents' answers of surveys. Their analysis was based on data of the Consumer Sentiment Index (CSI) survey from year 2010 to 2017 in Florida. Contact times (year, time of the day, day of the week) were used as independent variables, and dependent variable was the log of completed interview divided by the total number of dials at that time. Unlike the previous literature, where logistic regression had been the most common technique used, the random forest was used in their study to predict the dependent variable. The findings contradicted the results found more than a decade ago in terms of the most productive hours and weekdays. They showed that afternoon hours are as productive as evening and in regard to day of the week, Monday is the most productive day with the weekend being the least productive.

III. METHODOLOGY

The overall workflow of the proposed methodology to achieve the novel contributions is illustrated in Figure 1. It is composed of data retrieval / collection, feature engineering, implementation and evaluation of machine learning models, model interpretation and integration with the existing customer management system in the partner company. The integrated system has been in operation in the Company recently, which has seen improved customer contact experience and sales opportunities.

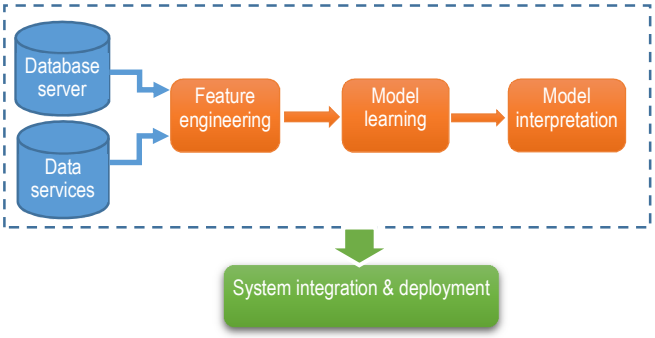

Figure 1. Architecture of the proposed system.

The data used in this project were collected from the Company's database server and public data services including national census data.

During the initial steps, data were cleansed, new insightful features were created, and all the values were converted to numerical form. All the features were then normalized and divided into training and test sets.

After the hyperparameter optimization, using Optuna, the models based on LightGBM with the best set of hyperparameters were trained. SHAP analysis was then performed to obtain the most influential features in finding the probability of a call being successful.

Other components of the proposed methodology are explained in the following sub-sections.

A. Model Optimization Using Optuna

A key task in model optimization is to determine the appropriate set of hyperparameters for machine learning models such that the models' performance is optimised on the data set available. Hyperparameter optimization is a non-trivial task, its search in the hyperparameter space is a time-consuming process. We employ one of the latest hyperparameter optimization algorithms, Optuna [5], for our model hyperparameter tuning.

Optuna [5] is a software framework, scripted in python, to perform this very task. It utilizes the Bayesian optimization method to find the right set of hyperparameters. Unlike other methods such as random search or grid search using trial and error until they show the best results, Bayesian optimization considers the result of previous trials to choose and evaluate the set of hyperparameters for the next trial. This framework comes with better cost effectiveness and is scalable and versatile to deploy.

Optuna comprises of three main concepts [5] [24]:

- Objective Function: It returns the numerical value of the metric which is used to evaluate the performance of trials. The function is defined by the user.

This project is funded by Innovate UK and Walsh Government. (KTP11894).

- Trial: A single execution of the objective function. Each trial receives a different set of hyperparameters from the ranges defined by the user (in objective function).
- Study: An optimization session, which is a set of trials. This sets the direction of optimization (maximize or minimize).

We use Optuna to optimise hyperparameters of models in this study due to its effectiveness and efficiency [5].

B. Customer Contact Modelling Using LightGBM

After a pilot investigation of some popular machine learning methods, we chose to use one of the most advanced frameworks of gradient boosting, LightGBM, to build machine learning models on the data sets. This decision tree-based algorithm has an advantage over deep learning methods on certain aspects: they are easy to interpret, perform better with imbalanced data and take less time to train. Gradient boosting decision tree (GBDT) is one of the gradient boosting algorithms which combines multiple weak learners (decision trees) to create a better performing model. Due to its state-of-the-art performance in machine learning tasks, such as multi-class classification, click prediction, ranking problems in terms of efficiency, accuracy and interpretability, it has become one of the most widely used algorithms [25]. LightGBM is a framework of gradient boosting and trains a GBDT. It comes with a histogram-based method which groups features into bins and performs splitting on them instead of conducting it naively. This reduces time as well as computational complexity.

For the abovementioned reasons, LightGBM is known for its good performance with large datasets, in terms of accuracy and scalability. Thus, LightGBM was chosen and its hyperparameters optimized using Optuna in this study.

C. Model Interpretability and Feature Importance Using SHAP

After the model has been created, it is important to find an effective way to make it more interpretable, understandable and obtain better insights from it. Simple models such as linear models are easier to understand but with the growing amount of data, models have tended to become more complex, for example deep neural network, making them harder to interpret [6].

We adopt a relatively new approach, SHAP [6], to evaluate the feature importance in predicting the output level. The concept of Shapley values has been inspired from cooperative game theory where each participating player is equivalent to a feature, the reward of game is the prediction given by the model for an observation, and the aim is to find the contribution of each feature towards the final outcome. SHAP is known as the state-of-the-art model for evaluating features and interpreting machine learning models. SHAP analysis can provide global as well as local interpretability. Local interpretability is specific to analyzing feature importance for a particular given observation. Global interpretability shows the contribution direction of each feature to the model prediction for all observations. This characteristic of interpretability allows us to identify the key factors that affect customer contactability in this study.

IV. EXPERIMENTAL EVALUATION

A. Dataset

The data provided by the company comprises three datasets: Life, Income and Health, each associated with the insurance product the company deals with.

Apart from the person's demographics, area characteristics and calling history data, the datasets also contain all the information necessary to decide upon the best insurance policies. Each insurance type depends on a different set of factors, so the datasets have varying features. For example, Life insurance is more related to the number of dependents and family members, while Health insurance is more focused towards the health and wellbeing of an individual. Income Protection takes into account factors such as a person's occupation and employment.

The timings of the trial calls are automatically recorded when the calls are performed by the company's sales advisors. The number of calls to each customer ranged from one to dozens, with extreme cases over one hundred. Only ten calling records are kept for analysis such that if the customer ever picks up the call, the data of last ten calls are kept and if the customer never picks up the call, the trials after the first ten calls are removed. All the trial occurring after the customer picks up the call are also removed from the analysis [26].

The datasets include all the variables relating to calls which include the time of the latest trial call, the time of previous trial calls, the time difference between the latest and the one before it, number of trials etc. The data about the individual's demographics and other information related to insurance is provided by the customer when they complete the initial quote on the company's website. Demographics associated with the postal code, more commonly called as area characteristics, e.g., happiness, anxiety, life satisfaction, worthwhile scores, mean/median salary, proportion of people based on different age ranges, as done in the previous literature [27] was readily available in the dataset as it had already been collected for another analysis using the census data. The output variable is dichotomized to just two classes:

- 0 – if the call is unsuccessful (customer never picks up the call, or the dialled number is busy)
- 1 – if the call was picked up (even if the interview is not completed or the customer asks to call again sometime later, the call is considered as successful) [26]

If the customer picks up the call at the n-th trial, the null values of all the variables featuring the rest of trials are revalued as -1.

The datasets, covering the period from 1st January 2017 to 8th August 2020, are fetched from the company's database using Microsoft SQL server by performing relevant SQL queries. The calls from 1st January 2017 to 15th September 2019 are used for model development, and the calls from 15th September 2019 to 8th August 2020 are used as the testing data.

B. Data preparation

Data pre-processing is carried out with an aim to achieve best machine learning modelling performance. The datasets have

three different type of variables (DateTime, Categorical, Numerical).

1) DateTime Variables

A number of DateTime type variables are included in the datasets. The most common ones are as follows:

1. *Date_created* (time when the customer first created the profile on the website)
2. Recent *Call_time* (time when the customer was most recently called)
3. Call times of previous 9 trials (-1 if no value)

Time elements were extracted from those variables which are:

1. Hour of the day
2. Day of the week
3. Month
4. Quarter
5. Year
6. For the previous trials, Hour and Day of the week were extracted and *'_Ti'* was added as the suffix, where *i* is the trial number.

Some of the calculated variables are:

1. 'Vacation' (a binary value to indicate if the date lies during the vacation period):
   - December 20 to January 5
   - April 1 to April 15
   - July 20 to August 1
2. 'Diff_call_created_day' – number of days lapsed between the time of call and when the profile was created.
3. 'difference_from_prev_trial_in_DAY' – number of days between the previous call and the day of the most recent call.

2) Categorical Variables

The categorical variables were dealt with in the following two ways:

1. Textual categorical variables involving many different values were processed, including converting all the words to lower case, removing the stop words and punctuations, stemming and lemmatizing using *sklearn.nltk* library. The least common values for each categorical variable were stored as '*others*'. One-hot encoding was performed to convert them to a numerical form.
2. To obtain better insight of the categorical variables and to make them more useful for the model, call success rates associated with the categorical values were calculated.

3) Data Balancing

All three datasets from insurance products of Life, Income and Health, respectively, have an imbalance ratio of roughly 9:1 between class 0 and class 1. With the available resources, under sampling method is performed to balance the datasets.

Table 1 provides a summary of the obtained datasets used in the following experiments.

Table 1. Summary of Datasets.

Product	Train (imbalanced)	Train (balanced)	Test	Number of variables
Income	829223	159416 (-80%)	308899	326
Health	384070	95432 (-74%)	265452	214
Life	434945	78936 (-80%)	80654	213

C. Model Implementation and Evaluation

The first group of experiments were performed for the balanced datasets as well as imbalanced datasets to compare how the model performed on both of them. First, 100 trials were run to find the best set of hyperparameters for LightGBM model using Optuna, and the ones which resulted in the highest F1 score were finalized. The performance of best models on imbalanced and balanced datasets were then compared.

Table 2 shows the set of values of hyperparameters which gives out the maximum value of objective function, *F1 score* for each model.

Table 2. Values of optimized hyperparameters.

	Balanced			**Imbalanced**		
	Income	Health	Life	Income	Health	Life
'learning_rate'	0.206	0.592	0.248	0.242	0.239	0.385
'n_estimators'	7	31	18	43	46	30
'num_leaves'	25	73	63	86	53	37
'reg_lambda'	216	670	6.194	597	1201	1509
'scale_pos_weight'	-	-	-	2.223	4.539	5.673

For all models, the maximum score of objective function reached to the highest of 35%.

D. Machine Learning Model

The values of hyperparameters given in Table 2 were finally used to train the models. We use classification accuracy, precision, recall, F1 score and AUC of ROC to assess the performance of trained models. Results are listed in Table 3.

Table 3. Experimental results of different insurance products.

	Balanced			**Imbalanced**		
	Income	Health	Life	Income	Health	Life
Accuracy	58.24	61.01	59.37	71.24	73.09	75.49
Precision	19.49	21.43	17.66	23.98	25.54	22.59
Recall	75.76	69.17	77.77	60.50	50.66	54.49
F1 Score	31.00	32.72	28.78	34.25	34.05	31.94
ROC_AUC	70.53	69.51	72.99	72.32	70.45	74.10

Table 3 shows the performance of LightGBM models on balanced and imbalanced datasets. It can be clearly observed that the performance is much better with the imbalanced datasets in terms of all the above-mentioned metrics except recall. The reason for the high recall in the balanced sets is that the model predicted most of the cases as class 1, see Figure 2 - Figure 4 for reference.

Figure 2 to Figure 4 comprise of different histogram graphs which illustrate the distribution of predicted probabilities (by the models) for the test sets, separated by the actual classes (0 – red, 1 – blue). A large number of false positives (with the threshold

being set to default value of 0.5) can be observed for the bbalanced datasets for all products while an improvement in models' performance, in regards to precision, can be seen with the imbalanced datasets.

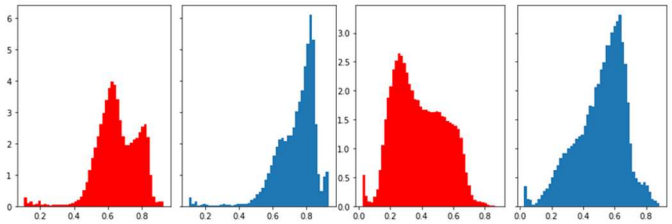

Figure 2. Income (Balanced VS Imbalanced)

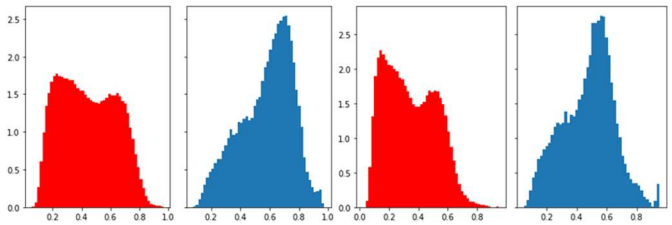

Figure 3. Health (Balanced VS Imbalanced)

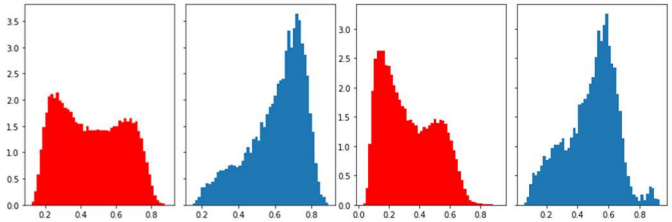

Figure 4. Life (Balanced VS Imbalanced)

- Probability distribution test cases in class 0
- Probability distribution test cases in class 1

A large number of false positives can lead to unnecessary call backs, so in order to save the resources, the datasets which caused the models to perform better in terms of precision were prioritized.

It can be observed from Table 3 that models trained on balanced datasets do not improve models' performance for other metrics all as well over original imbalanced datasets, so we will focus result analysis on the models obtained from original imbalanced datasets.

E. Results Analysis

1) Feature Importance

For each product, the trained model is interpreted using SHAP to explain feature importance (in terms of SHAP value in log-odds).

Figure 5 - Figure 7 show the top 20 impacting features on the model output for all three products, respectively. These plots also show how the values of each feature influences (negatively or positively) towards the SHAP values. They actually aggregate SHAP values of all the customer cases (rows) for the top 20 features to see the influence on global level for SHAP values.

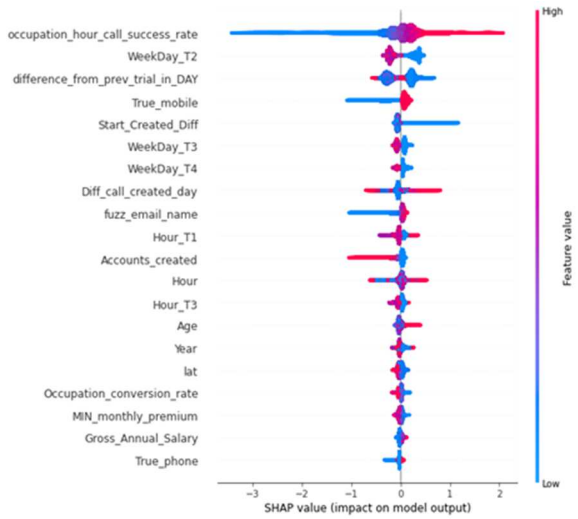

Figure 5. Summary SHAP plot Income

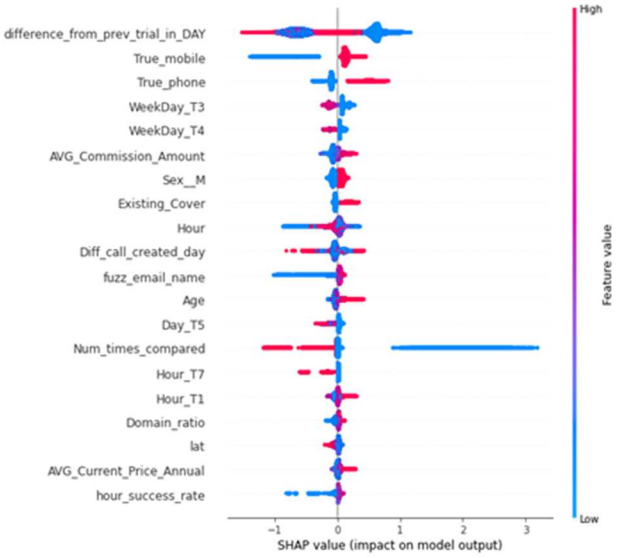

Figure 6. Summary SHAP plot Health

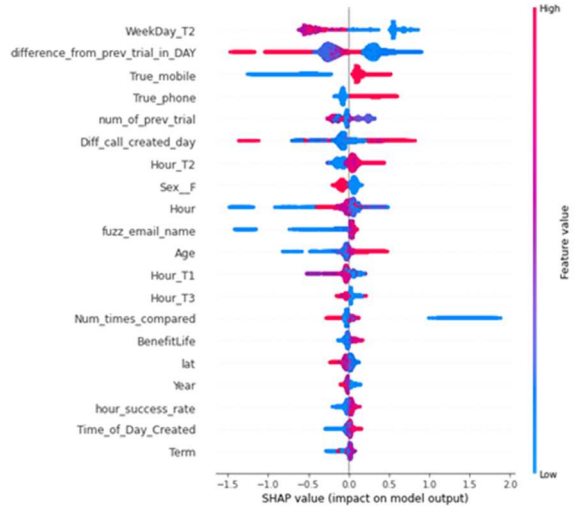

Figure 7. Summary SHAP plot Life

For Income as shown in Figure 5, amongst the top variables is 'occupation_hour_call_success_rate'; a higher value of this variable gives a high SHAP value, impacting positively towards the overall SHAP value. Weekday's for previous trials ('WeekDay_T2', 'WeekDay_T3', 'WeekDay_T4'), with a negative relationship, can be observed to be appearing in the list for all three of the products. This means that the increase in weekday values of trials (which ranges from 1 to 7 and -1) , decreases the total SHAP value for the particular case and hence the chances of attending the call decreases. The impact is however not that large.

Moreover, it can be seen in all figures that the high value of the feature, 'difference_from_prev_trial_DAY' results in a lower SHAP value and vice versa. This relation is not that linearly strong, as for some cases, smaller value of 'difference_from_prev_trial_DAY' gives a smaller corresponding SHAP value, which is opposite to the general negative relationship observed. 'True_mobile' and 'True_phone' always influence positively on the SHAP values.

Age is also in the list of top 20 influencing features for all three figures. The positive influence is highest for Life, followed by Health and very little for Income.

It was also found that for all three of the models, 'Age' had the strongest interaction with 'True_mobile'. Figure 8 to Figure 10 show how the SHAP value of 'Age' interacts with 'True_mobile'.

It can be observed from Figure 8 that an increase in the value of 'Age' generally increases the SHAP value for 'Age'. The relationship between 'Age' and its SHAP value is found to be nonlinear. The cause of nonlinearity is due to the interaction of another variable, 'True_mobile' (1 if the gadget being called on is mobile phone, and 0 for vice versa). 'True_mobile' goes in contrast with 'True_phone'. If the value of 'True_mobile' is 0, for 'True_phone' it is 1. This high interaction explains that customer contactability depends on their age and the type of phones they use. People over 50 are more contactable and are more likely contacted over landline telephone rather than mobile phone, vice versa for people of younger age range (between 20 and 50). We also analyzed interactions between other pairs of features to observe any significant relationship between them. In particular, the results show that there is a slight relationship between the time of day of contacting and the medium contacted on. SHAP values are higher towards the second half of the day when the call medium is a telephone ('True_mobile' = 0), hence positively impacting towards the final calculated SHAP value. Comparatively, using the call medium of mobile in the first half of the day will likely result in a higher success rate.

1) Discussion

SHAP analysis is used to observe if demographic and area characteristic features influence the probability of making a successful call. Experimental results show that features relating to gender and age are in the top 20 features for all three of the products, but the impact is small. However, no feature corresponding to any area characteristic are found to be significant in predicting call success rate, unlike [13] [27]. The reason might be that since Durrant, et al.[27]'s focus was more on finding an optimal time to do face to face surveys, hence these features might be suitable at predicting best times to contact in person rather than over the phone.

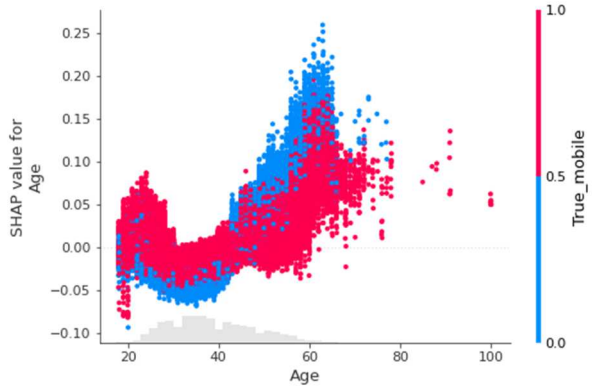

Figure 8. Effect of 'Age' and its interaction with 'True_mobile' on the SHAP value for 'Age' for Income

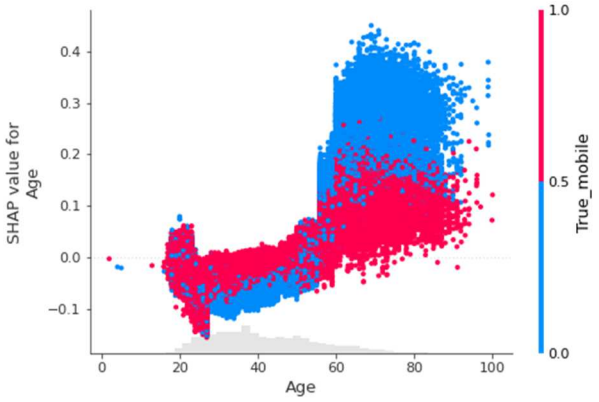

Figure 9. Effect of 'Age' and its interaction with 'True_mobile' on the SHAP value for 'Age' for Health

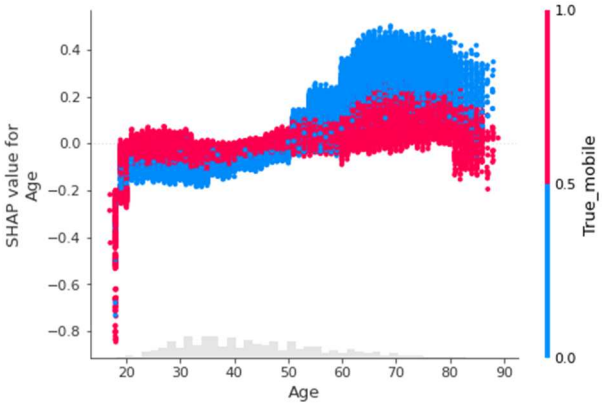

Figure 10. Effect of 'Age' and its interaction with 'True_mobile' on the SHAP value for 'Age' for Life

Although occupation of the recipient or time of the current call (in terms of hour) does not appear to be the most influencing feature, interestingly, the combined variable of these two, 'occupation_hour_call_success_rate', which featured the call success rate for each hour based on the person's occupation, proved to have a great positive impact for Income model. This implies that the call success rate can vary at different times of

day for people with different occupations and it can be taken into consideration while making a call.

The results also show that the variables associated with the previous trials can bring an impact towards the probability of calls being successful. The presence of variables such as 'Weekday_T2' and 'Weekday_T3' amongst the top features for all three of the products and the negative relationship with its corresponding SHAP value, either means that if the day of the week for the nth call resides at the end of the week, it may reduce the chances of the call being picked up on the n+1 th trial. It also explains that the calls conducted at the earliest trials tend to be most successful, since the null values for no calls being done on later trials after the call was picked were replaced by -1. This second opinion is also favored by the impact of *num_of_prev_trials* variable, which appears among the top five for Life, and has a negative relationship with the SHAP values. It can be deduced that with the increase in number of trials, the call success rate decreases. This was also supported by Reimer, et al. [22].

Moreover, the SHAP results with some of the individual features depicts that the values of features 'Age', 'True_mobile' and 'Hour' strongly interact with each other to produce varying impact overall SHAP value. The outcome for the three different models had very similar results; the chances of making a successful call increases if the group of older people are contacted over telephone rather than mobile phones. The findings also imply that the calling medium affects the probability of success over different hours of the day. Hence, it can be concluded that the likelihood of contact increases for telephone medium during late afternoons and evenings, whereas for mobile phones, the better time to contact is during the morning hours.

V. CONCLUSIONS

This paper presents a practical application of the latest advances of machine learning modelling and explainable artificial intelligence to the development of call scheduling models for the insurance industry. Experimental results demonstrate that the proposed approach not only predict the probability of a person answering the call based on the given data, but also explains how individual factors can affect prediction outcomes. The developed models avoid unnecessary call-backs which help in saving cost and time. These models were recently integrated with the existing customer management system of our partner Company, which has seen improved customer contact experience and sales opportunities.

In terms of timing, based on the results, it is advised to wait longer to attempt to call a particular customer again, if the previous call remained unsuccessful. Moreover, the results indicate that the timings of the preceding calls, in terms of the weekday, significantly affects the end result for the calls that follow.

None of the area-related features are among the top influencing variables. The variable presenting the hourly call success rate for different occupations had a great influence over foreseeing the probability of success rate for Income customers. The variable 'occupation' combining with other variables can be used to suggest customer call time for Income Insurance. This variable, however, does not seem to be useful for devising plans for Health or Life insurance, although adding this variable to the application can be useful for predicting suitable call times to the customer. It is also advised to contact older people on telephone rather than on mobile phone, with the best time to contact them being evenings or late afternoon. In short, it is better to schedule the calls associated with the telephone number during the above-mentioned timings.

The cases for the two classes of contactability are very much intermingled, and to separate them can be a very difficult task, requiring further data and feature engineering. Other data balancing methods in combination with different machine learning models is therefore worthy of further investigation.

Moreover, this analysis does not consider the company's resources in terms of availability of staff and the number of calls they could possibly make in a given time. Some people wait longer to see if the customer picks up the call, while others hang up quickly. It is practically impossible to schedule calls in a same slot, so it is necessary to divide them evenly. This work can be improved by incorporating these factors to make it more practical.

REFERENCES

[1] J. R. Hochstim, "A critical comparison of three strategies of collecting data from households," Journal of the American Statistical Association, vol. 62, pp. 976-989, 1967.

[2] J. Moore, K. P. Uhl and B. Schoner, "Marketing research: Information systems and decision making," Journal of Marketing, vol. 35, no. 1, pp. 109, 1971.

[3] E. Shino and C. McCarty, "Telephone survey calling patterns, productivity, survey responses, and their effect on measuring public opinion," Field Methods, vol. 32, no. 3, pp. 291-308, 2020.

[4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, Curran Associates Inc., 2017, pp. 3149–3157.

[5] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, Association for Computing Machinery, 2019, pp. 2623-2631.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model," in Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, Curran Associates Inc., 2017, pp. 4768–4777.

[7] M. F. Weeks, B. L. Jones, R. E. Folsom, J. and C. H. Benrud, "Optimal times to contact sample households," Public Opinion Quarterly, vol. 44, pp. 101-114, 1980.

[8] G. Vigderhous, "Scheduling telephone interviews: A study of seasonal patterns," Public Opinion Quarterly, vol. 45, no. 2, pp. 250-259, 1981.

[9] M. F. Weeks, R. . A. Kulka and S. A. Pierson, "Optimal call scheduling for a telephone survey," Public Opinion Quarterly, vol. 51, no. 1, pp. 540-549, 1987.

[10] J. T. Massey, P. R. Barker and S. Hsiung, "An investigation of response in a telephone survey," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 1981, pp. 426-431.

[11] T. Triplett, "What is gained from additional call attempts and refusal conversion and what are the cost implications?," Urban Institute, Washington, DC, 2002.

[12] S. L. Stokes and B. S. Greenberg, "A priority system to improve callback success in telephone surveys," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 1990.

[13] J. M. Brick, B. Allen and P. Cunningham, "Outcomes of a calling protocol in a telephone survey," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 1996.

[14] C. Tucker, J. M. Brick and B. Meekins, "Household telephone service and usage patterns in the United States in 2004: Implications for telephone samples," Public Opinion Quarterly, vol. 71, pp. 3-22, 2007.

[15] L. Piekarski, "Cellular phones: Challenges and opportunities," Survey Research, vol. 34, no. 2, 2003.

[16] A. Y. Yuan, B. Allen, J. M. Brick, S. Dipko, S. Presser, C. Tucker, D. Han, L. Burns and M. Galesic, "Surveying households on cell phones—results and lessons," in Proceedings of the Survey Research Methods Section, American Statistical Association, 2005.

[17] J. M. Brick, P. D. Brick, S. Dipko, S. Presser, C. Tucker and Y. Yuan, "Cell phone survey feasibility in the U.S.: Sampling and calling cell numbers versus landline numbers," Public Opinion Quarterly, vol. 71, no. 1, pp. 23-39, 2007.

[18] C. Steeh, "A new era for telephone surveys," in Annual Conference of the American Association for Public Opinion Research, Phoenix, 2004.

[19] M. W. Link, M. P. Battaglia, M. R. Frankel, L. Osborn and A. H. Mokdad, "Reaching the U.S. cell phone generation comparison of cell phone survey results with an ongoing landline telephone survey," Public Opinion Quarterly, vol. 71, no. 5, pp. 814-839, 2007.

[20] R. Zuwallack, "Piloting data collection via cell phones: Results, experiences, and lessons learned," Field Methods, vol. 21, no. 4, pp. 388-406, 2009.

[21] L. R. Carley-Baxter, A. Peytchev and M. C. Black, "Comparison of cell phone and landline surveys: A design perspective," Field Methods, vol. 22, no. 1, pp. 3-15, 2009.

[22] B. Reimer, V. Roth and R. Montgome, "Optimizing call patterns for landline and cell phone surveys," Proceedings. American Statistical Association. Annual Meeting, vol. 2012, pp. 4648–4660, 2012.

[23] P. Vicente and I. Lopes, "When should I call you? An analysis of differences in demographics and responses according to respondents' location in a mobile CATI survey," Social Science Computer Review, vol. 33, no. 6, pp. 766-778, 2015.

[24] M. Corporation, "LightGBM," 2020. [Online]. Available: https://lightgbm.readthedocs.io/en/latest/. [Accessed 20 October 2020].

[25] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[26] J. Wagner, "Adaptive contact strategies in telephone and face-to-face surveys," Survey Research Methods, vol. 7, no. 1, pp. 45-55, 2013.

[27] G. B. Durrant, J. D'Arrigo and F. Steele, "Using paradata to predict best times of contact, conditioning on household and interviewer influences," Journal of the Royal Statistical Society. Series A, vol. 174, no. 4, pp. 1029-1049, 2011.