The Impact of Transaction Costs on Forecast-Based Trading Strategy Performance

James Lewis-Cheetham
School of Computer Science and Informatics
Cardiff University, United Kingdom
lewis-cheethamjw@cardiff.ac.uk

Yuhua Li
School of Computer Science and Informatics
Cardiff University, United Kingdom
liy180@cardiff.ac.uk

Federico Liberatore
School of Computer Science and Informatics
Cardiff University, United Kingdom
liberatoref@cardiff.ac.uk

Qingwei Wang
Cardiff Business School
Cardiff University, United Kingdom
wangq30@cardiff.ac.uk

***Abstract*—Active investing strategies have poor historical long-term performance compared to passive strategies. Furthermore, many active strategies use forecasting of various signals even though the efficient market hypothesis posits that stock prices rapidly incorporate information. Despite this, recent studies continue to explore the use of forecast-based strategies and report positive evaluation results. We closely investigate several inter-day active management strategies based on forecasting models via a test set and a U.S. market backtest. We approximated transaction costs via bid and ask prices and percentage costs.**

Test metrics indicated that models could not forecast over week-long horizons but demonstrated some forecasting ability over monthly and quarterly horizons. Backtest results indicate a null correlation between test metrics and strategy performance. For low transaction cost rates, several strategies achieved higher Sharpe ratios than a benchmark passive strategy. At 25 basis points a single strategy had comparable performance to the benchmark. It is unclear to what extent this observation of superior strategy performance was due to the quantity of strategies we evaluated. Our findings emphasise the importance of evaluating strategies under realistic trading conditions and suggest further investigation into the null correlation between forecast accuracy and strategy performance.

***Index Terms*—Machine Learning, Algorithmic Trading, Financial Forecasting, Backtesting, Efficient Market Hypothesis, Temporal Arbitrage.**

I. Introduction

Active strategies are trading strategies that frequently trade based on predictions of returns determined by analysis or forecasting. They aim to produce positive alpha, referring to risk-adjusted returns greater than those of a benchmark or market index. In contrast, passive strategies limit trading activity after initial capital investment, investing in assets to track market or sector performance. Analysis of historical data shows that active trading strategies rarely outperform passive alternatives, in part due to the additional costs these strategies incur [1], [2], [3]. Passive strategies are increasingly adopted by investors, but active strategies remain commonly used [4].

Many active strategies rely on forecasting future returns or price movements. Forecasting aims to identify market inefficiencies, a situation where an asset's market price does not accurately reflect all available information. Market inefficiencies present opportunities for temporal arbitrage - the practice of capitalizing on the price differences for the same asset at different times.

The Efficient Market Hypothesis (EMH) [5], [6] argues that information is factored into prices almost immediately in well-developed markets. Thus, the EMH argues that forecast-based trading strategies cannot consistently generate positive alpha. Furthermore, the relationship between forecast accuracy and strategy performance is unclear [7]. Despite these arguments, many studies demonstrate the effectiveness of active management strategies with evaluations using historical data [8].

The performance of these strategies may not translate to real-world market performance due to the simplification of evaluations. Many studies evaluate methods without consideration for transaction costs which can have a large impact on strategy performance. To determine if strategies truly manage to identify and exploit opportunities for temporal arbitrage, thorough evaluations must be conducted.

Given this research problem, the objective of our study is to further investigate active strategy performance. Specifically, this study investigates to what extent various models that are commonly used in time-series forecasting, and those recently published in the literature [9], outperform a buy-and-hold strategy. We conducted a trading backtest, which included approximations of transaction costs, on historical data.

We consider the following null hypothesis:

- The active trading strategies included in this evaluation do not demonstrate superior performance compared to a buy-and-hold benchmark when transaction costs approximating those of real-world markets are considered.

If the active strategies outperform a passive benchmark, this suggests that forecasting models can identify previously unknown opportunities for temporal arbitrage. This has implications for financial researchers as it contradicts the EMH, the cornerstone of modern portfolio theory. It also has implications for financial traders.

II. Literature Review

A. Technical Analysis

In our work, we use *technical analysis* and derived *technical indicators* for forecasting. Technical analysis refers to analysis which uses historical price information, volume, and derived features such as moving averages and identification of trading patterns. Proponents of technical analysis argue that it can give an insight into market trends and investor psychology. Yet, the utility of technical analysis is disputed [2]. As part of our work, we investigate whether technical analysis can be successfully used with the forecasting models under examination.

B. The Efficient Market Hypothesis

The EMH argues that prices reflect all available information. The term 'efficient' specifically refers to informational efficiency. Thus, the EMH argues that as all information is already factored into prices, technical analysis will confer no benefits. The EMH is a theoretical model but still has considerable utility as the conditions of major U.S. stock exchanges provide reasonable approximations of the assumptions made.

Large price changes for liquid assets in response to news events have been generally found to occur on the scale of minutes or faster [10], [11], [12]. Thus, forecasting over an inter-day period may not be able to identify worthwhile opportunities for temporal arbitrage. Minor opportunities for temporal arbitrage may be identifiable but persist because the benefits do not outweigh the costs. They could also persist due to limits to arbitrage, such as bans on short selling.

C. Evaluation of Trading Strategies

Using forecasting metrics alone to determine if an algorithm can be used for a real-world trading strategy can be misleading. Real-world trading strategies have to account for constraints such as transaction costs, order execution time, and information acquisition. A more thorough analysis of trading strategy performance can be done by performing a *backtest*, which involves testing the performance of a trading strategy on historical data. Many studies do not perform backtesting, despite the additional validity it confers to model and strategy evaluations [9].

Models can also appear unreasonably effective on historical data as a result of data mining [13], [14]. Data mining refers to the practice of evaluating a range of different models on a task and selecting those which achieved the best performance. The models that demonstrated good performance may simply have achieved these results due to chance.

D. Machine Learning in Financial Markets

Numerous publications have investigated the forecasting of stock returns or prices, which generally takes the form of a time-series forecasting problem. This task is commonly approached via statistical or machine learning methods. More recently, researchers have explored the use of convolutional neural networks (CNNs), recurrent neural networks (RNNs), CNN-RNN hybrid architectures, and transformer-based architectures [8], [15].

To our knowledge, a commonly used benchmark dataset for financial time-series forecasting does not exist. However, the results of the most recent M-series competition (M-6) [7] provide a useful comparison of different methodologies. Their analysis found that forecast accuracy only had a limited connection to investment performance and that the EMH held strongly. They also found that participants faced great difficulty with outperforming the market, even though no transaction costs were applied. In the competition results both traditional and machine learning models were equally represented among the top forecast accuracy rankings.

E. Re-imagining Price Trends

We include in our evaluations a model presented in the recently published paper *(Re-)Imag(in)ing Price Trends* [9]. The authors conducted return forecasts using price visualisations and supervised learning with a convolutional neural network model. They found that their image-based strategies generally outperformed pre-existing technical analysis-based strategies. We included this model in our analysis as it demonstrated excellent performance that bears investigating further.

F. The Importance of Transaction Costs

Transaction costs are the expenses that result from trading an asset. Transaction costs are the culmination of various fees (brokerage, exchange, regulatory, clearing and settlement), and trade execution costs (slippage and bid-ask spread induced). Further costs may also be considered, such as costs associated with data acquisition, infrastructure, and taxation.

Accurate transaction costs are difficult to determine using observable market information. Calculating costs requires detailed trade and market information and may also require information that is private to trade participants.

A small number of studies have approximated transaction costs and evaluated trading strategies with them taken into consideration [16], [17]. References [16] and [18] present 25 basis points as a typical one-way transaction cost for a value-weighted strategy. Reference [16] states that this is often two to three times higher for equal-weighted strategies but how these approximations were produced is unclear.

G. Theoretical and Practical Implications

The deficiencies in the literature have led us to the following conclusions:

- Accurate forecasts cannot necessarily be used to construct a successful trading strategy. The relationship between forecast accuracy and strategy performance is unclear [7].
- Evaluations must be conducted more thoroughly to accurately estimate real-world performance. We assert that transaction costs must be approximated when evaluating real-world strategy feasibility.
- Strategies must be critically examined and it must be justified why they succeed. It must be evident why inefficiencies persist for exploitation by the strategy, what form they take, and why no other market participants have identified and exploited them already.

III. Methodology

A. Data Acquisition

Our study used historical price data consisting of the open, high, low, close, bid, ask, and volume data for all available stocks in the New York Stock Exchange, American Stock Exchange, and Nasdaq Exchange. Data was obtained at a daily frequency for all trading dates from 1993-1-1 to 2023-12-31 from the Center for Research in Security Prices *Daily Stock* database from Wharton Research Data Services[19] (WRDS). Price adjustment factors, stock delisting event notifications and stock delisting prices were also obtained from this database. Strategy performance analysis used *WRDS Stock File Indices* [20] and Federal Reserve Economic Data *3-Month Treasury Bill Secondary Market Rate, Discount Basis* data [21].

Due to the large size of the complete dataset, we selected a subset of 5000 randomly selected stocks (out of approximately 27000) for training and evaluation.

B. Model Architectures

TABLE I: A summary of the models used in this study. Model acronyms are as follows: LR (logistic regression), XGB (XGBoost), MLP (multi-layer perceptron), TF (transformer), E-CNN (ensemble convolutional neural network), E-CNN-V (ensemble convolutional neural network using visualisations), SMAC (simple moving-average crossover), and SMACI (simple moving-average crossover inverse). P_t is the adjusted close price at time t, and MA_t is the moving average of the adjusted close price at time t. L represents the context window length in business days. h and w are the height and width of the visualisation image. All models produced a single output value per sample. For probabilistic models, this was the predicted probability of a positive return ranging from zero to one. The expression $P(R_{t+H} = 1 \mid X_t)$ denotes the probability of a positive return label at time t plus horizon length H, given window X. The SMAC and SMACI models produced a signed price difference, where d_t represents the price difference at time t.

Model	Input	Input Dimension	Output
LR	Window	$(L \times 7)$	$P(R_{t+H} = 1 \mid X_t)$
XGB	Window	$(L \times 7)$	$P(R_{t+H} = 1 \mid X_t)$
MLP	Window	$(L \times 7)$	$P(R_{t+H} = 1 \mid X_t)$
TF	Window	$(L, 7)$	$P(R_{t+H} = 1 \mid X_t)$
E-CNN	Window	$(1, L, 7)$	$P(R_{t+H} = 1 \mid X_t)$
E-CNN-V	Visualisation	$(1, h, w)$	$P(R_{t+H} = 1 \mid X_t)$
SMAC	P_t, MA_t	$(L, 2)$	$d_t = P_t - MA_t$
SMACI	P_t, MA_t	$(L, 2)$	$d_t = MA_t - P_t$

We evaluated the performance of various model architectures for this forecasting task. A summary of the models is given in Table I. For each model type, combinations of context length L and forecasting horizon H were used, as it was unclear which would produce the most performant strategies.

We used a range of models of varying complexity. The justification for the choice of models is as follows:

Logistic regression models are a simple benchmark. XGBoost and multi-layer perceptron models are computationally simple to train, yet often demonstrate good performance for predictive modelling. Transformer models are state-of-the-art for various problems and although we used a relatively simple architecture it can indicate whether transformer models are worth investigating further for this task.

The ensemble-CNN model using visualisations, presented by [9], is included here as a point of comparison to a recent well-performing method from the literature. Each ensemble model consists of five CNN models, with the average of the five model outputs being used as the ensemble output. We opted to use the smaller visualisations developed for this method as [9] reported that the smaller visualisations produced better performance, and the larger visualisations were computationally intensive. We also evaluated this CNN architecture without the use of visualisations to briefly investigate if the visualisation approach conferred a forecasting benefit.

Simple moving average crossover and simple moving average crossover inverse models were used primarily as benchmarks, and to represent simple strategies based on trend and mean reversal.

Reasonable values were chosen for the architecture sizes and hyperparameters. To reduce overfitting, regularization was implemented through the use of L1 (lasso) and L2 (ridge) losses and dropout layers.

C. Data Pre-processing and Model Input

Various pre-processing steps were required. Prices and outstanding shares were adjusted to account for distribution events using adjustment factors provided by WRDS. Small-scale occurrences of missing values in the data were forward-filled. To improve the statistical properties of the data, prices were converted to log returns, and the volume column was differenced.

The dataset was divided into three parts, with an approximate split of 58%, 16%, and 26%:

- **Training Set:** January 1, 1993 – December 31, 2010
- **Validation Set:** January 1, 2011 – December 31, 2015
- **Test Set:** January 1, 2016 – December 31, 2023

This provides a suitably large test period to account for general shifts in market conditions, and the inclusion of the 2020 COVID-19 recession tests the robustness of the strategies to extreme market conditions. Min-max scaling (-1 to 1) was applied after splitting the data.

Due to the differences in model context lengths and forecast horizons, several different datasets were created. For each dataset, samples were produced by running a sliding window over each stock time series.

All models used a fixed number of prior days as input, the length of which was determined by the context length of the model/strategy, L. For a model with context length L forecasting at date t, the window contained features for $t - L - 1$ to t business days.

$$\text{Window} = \{\mathbf{x}_{t-L-1}, \mathbf{x}_{t-L}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t\}$$

where

$$\mathbf{x}_i = \{\text{Close log return}_i, \text{Open log return}_i, \text{High log return}_i, \text{Low log return}_i, \text{Bid log return}_i, \text{Ask log return}_i, \text{Volume difference}_i\}$$

represents the 7 normalized features for each business day i. Log return and difference correspond to that day and the previous business day.

Due to the computational time required to produce visualisations, and the over-abundance of weekly horizon data, for some datasets a proportion of samples were skipped. Each dataset contained approximately 100,000 training samples.

These features, with the addition of a normalised moving average of length equal to the window length, were used to produce visualisations for the ensemble CNN. See [9] for details of their visualisation method.

The simple moving average models used the most recent adjusted close price and a moving average of the close price with a length equal to the window length.

All models used an identical method for producing prediction labels. For each forecast horizon, binary categorical labels were generated by calculating the log returns over the forecast horizon. A label of zero indicates a negative log return over the horizon. A label of one indicates a positive log return over the horizon.

Context windows and labels were only generated for the earliest business day of each period. We assumed that temporal locality to the start of a week/month/quarter could affect price patterns.

D. Model Training and Evaluation

To reduce the amount of time required for training, and given that we observed that metrics stabilised after processing only a small proportion of the total dataset, we did not use a fixed number of training epochs. Instead, models were trained until metric stabilisation. Validation and testing used the same fixed size every time.

We explored the hyperparameters most likely to affect trading strategy performance: context length and forecast horizon. Further hyperparameter tuning was not performed, given the time and computing constraints for this study. Reasonable hyperparameter values were chosen for the models. More thorough hyperparameter tuning could potentially improve model performance, though whether any improvement would be notable is unclear.

E. Trading Backtest

We further evaluate model performance using a backtest; a trading simulation on historical data of the 5000 randomly sampled stocks. We evaluated every combination of model, context length, forecast horizon and transaction cost, 360 strategies (+ 1 benchmark) in total.

Traders were modelled as agents, each with a unique combination of strategy parameters and transaction costs. Forecasting models formed the basis for trading strategies. Forecasts were used to select which stocks to include in the trader's portfolio. Traders placed trades as instructed by their strategy frequency. For all strategies, the rebalance frequency was equal to the forecast horizon. The backtest used historical bid-ask prices for trading prices.

Portfolios were created by producing return predictions for every tradeable stock and ranking these predictions from largest to smallest. The top 10% of predictions were used to construct the portfolio (predictions for negative returns were excluded).

Probabilistic models ranked stocks based on the probability of positive returns. Class prediction used a threshold of 0.5. The simple moving-average crossover models ranked stocks based on the price difference observed (see Table I. Class prediction used a threshold of 0.

All strategies used value-weighting to construct portfolios, as this reduces the effect of transaction costs and better reflects real-world portfolio construction. For benchmarking purposes, the backtest included a trader utilising a value-weighted buy-and-hold strategy with yearly rebalancing.

We evaluate strategies with a range of one-way transaction costs from zero to 25 basis points. This transaction cost rate approximates all of the various costs associated with trading shares, except for the bid-ask spread. The bid-ask spread is inherently factored in by using bid and ask prices for trading. Traders were assigned a fixed transaction cost, that applied to all of their trades as a percentage cost.

In the dataset, bid-ask data is only available for the end of a trading day, after market close. Therefore, we assume that traders in our backtest perform trades shortly before market close and that bid-ask prices do not change much between this trading time and market close. We considered alternative approaches but concluded that making this assumption was the most reasonable method.

Note the simplifications of our simulation. The exchange order book was not modelled, liquidity was not considered, and trades did not induce price changes. We also assumed that traders received information immediately and could place orders immediately. Each trader also acted independently of all other traders.

No strategies using shorting were implemented, as this would only marginally add to our evaluation while increasing simulation complexity.

IV. Results

A. Test Results

We evaluated models on the test sets using AUROC (Area Under the Receiver Operating Characteristic) and F_1 score. Monthly and quarterly datasets were weighted towards the positive class. Hence, we present F_1 instead of accuracy. Test sets consisted of pre-generated sample-label pairs from the date range 2016-01-01 to 2023-12-31. These test results relate only to forecast performance, not investment decision-making. Table II shows the models with the context length that achieved the highest AUROC for each strategy-forecast horizon combination.

TABLE II: A summary of model performance metrics on the test dataset from 2016-01-01 to 2023-12-31. The results include the F_1 score and AUROC (Area Under the Receiver Operating Characteristic), rounded to two decimal places for clarity. Model acronyms are: LR (logistic regression), XGB (XGBoost), MLP (multi-layer perceptron), TF (transformer), E-CNN (ensemble convolutional neural network), E-CNN-V (ensemble CNN using visualisations), SMAC (simple moving-average crossover), and SMACI (simple moving-average crossover inverse). H represents the forecasting horizon in business days, and L denotes the context length for each model. For each metric and forecast horizon, the largest value is highlighted in bold.

Model	H	L	AUROC	F_1 Score
LR	Week	5	**0.51**	**0.70**
XGB		60	**0.51**	0.61
MLP		5	**0.51**	**0.70**
TF		5	**0.51**	**0.70**
E-CNN		60	0.50	**0.70**
E-CNN-V		5	**0.51**	0.64
SMAC		60	0.50	0.51
SMACI		5	**0.51**	0.49
LR	Month	5	0.48	**0.86**
XGB		60	0.53	**0.86**
MLP		5	0.47	**0.86**
TF		5	**0.54**	**0.86**
E-CNN		5	0.48	**0.86**
E-CNN-V		5	0.51	**0.86**
SMAC		60	**0.54**	0.59
SMACI		5	0.49	0.55
LR	Quarter	20	0.47	**0.99**
XGB		60	**0.84**	**0.99**
MLP		20	0.47	**0.99**
TF		5	0.66	**0.99**
E-CNN		20	0.48	**0.99**
E-CNN-V		5	0.53	**0.99**
SMAC		20	0.61	0.61
SMACI		60	0.43	0.63

The relationship between the context length, forecasting horizon, and test metrics is unclear. For most models, varying the context length only had a negligible effect on metrics.

All of the models struggled to achieve good performance at a weekly forecasting horizon. Although some models appeared to achieve reasonably good F_1 scores (of 0.70), in practice these scores were achieved by predicting the positive class for most samples. The AUROC scores of approximately 0.5 indicate that the models achieve performance equivalent to random guessing as to the probability of positive returns.

For a monthly horizon, model performance was more variable, with the XGB, TF, and SMAC models achieving a minor improvement in AUROC. All other models achieved similar or worse AUROC scores. The identical F_1 scores may indicate that the models are correctly classifying sample classes, even if predicted probabilities are incorrect.

At a quarterly horizon, XGB, TF, and SMAC models once again achieved the highest AUROC. XGB achieved an AUROC notably higher than all other models. Again, F_1 scores were largely similar between models.

SMAC and SMACI models had a notably different performance from other models. This result is not surprising, given that they are fixed, non-learning models. These models also used unique variations of the raw data, and rather than using a probability threshold, used a price difference threshold to produce categorical predictions.

The visualisation method used by [9] provided a minor improvement to model performance. However, even when using visualisations, the ensemble CNN model was consistently outperformed by other models, so the increased computational complexity was not justified.

B. Backtest Results

Table III shows the backtest results for the top 4 performing models, ranked by mean Sharpe ratio. The benchmark results for the yearly buy-and-hold strategy are also included. Full results are available in the supplementary material [22].

The mean Sharpe ratio is used to show model performance across the range of transaction costs, penalising strategies for which performance decreases as transaction costs increase. We chose not to rank strategies by alpha, as it heavily favoured strategies that had a negative correlation to the market index, even if they produced large negative returns. Although such a strategy could be useful for diversification, strategies were intended to produce positive returns.

Several of the strategies outperformed the benchmark buy-and-hold strategy, achieving greater alphas and Sharpe ratios, despite notably higher turnovers. The top strategies did not include any model with a weekly forecast horizon, and in general, these strategies performed poorly. All three variations in window length are among the top-performing strategies.

The trading performance of the strategies deviates from what would be expected, given the results in Table II. Despite the SMACI model performing exceptionally poorly in test set evaluations, the strategy performed well in the backtest. The presence of logistic regression and ensemble CNN models in the top-performing models is also unexpected.

We conducted simple statistical tests to determine if there was a significant correlation between test metrics and backtest performance. Spearman's rank correlation tests at the 5% confidence level, performed between AUROC and mean SR, and F_1 and mean SR, for each forecast horizon, found no significant correlations.

The quarterly XGBoost model achieved a relatively high SR, despite a negative alpha. The standard deviation of excess returns, and maximum drawdown, for this strategy were low in comparison to the benchmark and alternative strategies.

Strategy performance degraded as transaction costs increased. The effect on the benchmark strategy was negligible due to its low turnover. Performance for the top strategies was still reasonable at 25 basis points, but at this cost rate, most models achieved lower Sharpe ratios than the benchmark. The quarterly logistic regression model is an exception, in that it achieved an equal Sharpe ratio and higher alpha.

Sharpe ratios are lower than generally expected for all models, which is caused by the effect of the COVID-19

TABLE III: Backtest results from 2016-01-01 to 2023-12-31. Model acronyms are: B&H (buy-and-hold), SMACI (simple moving-average crossover inverse), LR (logistic regression), XGB (XGBoost), and E-CNN (ensemble convolutional neural network). Here, H represents the model forecast horizon or rebalancing frequency, and L is the model context length. $\overline{\text{SR}}$ denotes the mean Sharpe Ratio across all transaction cost rates. BETC indicates the Break-Even Transaction Cost, the transaction cost rate at which alpha is zero. This was approximated using a linear regression of transaction rate against alpha. Transaction costs are one-way and expressed in basis points, with one basis point equal to 0.01%. α represents excess return relative to an index composed of the New York Stock Exchange, American Stock Exchange, and Nasdaq Exchange. *StdDev* denotes the standard deviation of the portfolio's excess returns. *SR* is the Sharpe Ratio, a measure of risk-adjusted performance compared to a risk-free asset. *CAGR*(%) represents the compound annual growth rate. *TO* (%) is portfolio turnover. *MDD* (%) indicates the maximum drawdown, representing the downside risk of the portfolio. For each metric, the best value corresponding to a given transaction cost rate is highlighted in bold. Values are rounded to two significant figures for clarity.

Strategy	**H**	**L**	$\overline{\text{SR}}$	**BETC**	**Metrics**	**One-Way Transaction Cost Rate (bps)**					
						0	**5**	**10**	**15**	**20**	**25**
B&H	Year	N/A	0.43	221	*SR*	0.43	0.43	0.43	0.43	0.43	**0.43**
					α	0.01	0.01	0.01	0.01	0.01	0.01
					StdDev	0.17	0.17	0.17	0.17	0.17	0.17
					CAGR	8.6	8.5	8.5	8.5	8.5	8.5
					TO	**0.23**	**0.23**	**0.23**	**0.23**	**0.23**	**0.23**
					MDD	-31	-31	-31	-31	-31	-31
SMACI	Month	5	0.52	57	*SR*	**0.65**	**0.60**	**0.55**	**0.50**	**0.45**	0.40
					α	0.12	0.11	0.10	0.09	0.08	0.07
					StdDev	0.22	0.21	0.21	0.21	0.21	0.20
					CAGR	**15**	**13**	**12**	11	10	8.9
					TO	80	79	79	80	80	80
					MDD	-36	-36	-36	-36	-36	-36
LR	Quarter	20	0.46	231	*SR*	0.49	0.48	0.47	0.45	0.44	**0.43**
					α	**0.20**	**0.19**	**0.19**	**0.18**	**0.18**	**0.17**
					StdDev	0.33	0.32	0.32	0.32	0.32	0.32
					CAGR	13	12	12	11	11	11
					TO	29	29	29	29	29	29
					MDD	-48	-48	-48	-49	-49	-49
XGB	Quarter	60	0.41	-73	*SR*	0.52	0.48	0.44	0.39	0.35	0.31
					α	-0.04	-0.04	-0.05	-0.05	-0.05	-0.05
					StdDev	**0.08**	**0.08**	**0.08**	**0.08**	**0.08**	**0.08**
					CAGR	6.2	5.8	5.5	5.1	4.8	4.5
					TO	25	24	24	24	24	24
					MDD	**-19**	**-19**	**-19**	**-19**	**-19**	**-19**
E-CNN	Month	60	0.41	161	*SR*	0.45	0.44	0.42	0.40	0.39	0.37
					α	0.36	0.35	0.34	0.33	0.32	0.31
					StdDev	0.50	0.49	0.48	0.48	0.48	0.47
					CAGR	15	14	13	**12**	**11**	**10**
					TO	59	58	58	58	58	59
					MDD	-55	-55	-56	-57	-57	-58

recession on our test dataset. The dataset is not long enough to smooth out the effect on return volatility induced by this recession.

V. Discussion

A. Interpretation

Our results indicate that technical analysis can not be used to predict short-term returns, however, long-term forecasting is potentially possible, as indicated by the AUROC scores in Table III.

The inability of models to forecast weekly horizons could be due to high volatility making forecasting challenging over this horizon. The improvement in forecasting performance for long-term horizons can be explained by the reduced volatility of long-term forecasts.

Our findings do not indicate a clear relationship between window length, forecast horizon, and model performance. This result is not in agreement with existing literature [23]. This could be due to differences in model architectures and datasets but could bear further investigation.

Reference [16] states that transaction costs are approximately 50 basis points (for an equally-weighted strategy). This is inclusive of the bid-ask spread. Thus, we interpret our results with 25 basis points as an approximate transaction cost.

Assuming a real-world transaction cost rate of 25 basis points, our results tentatively reject the null hypothesis. They indicate that a simple logistic regression can potentially outperform a benchmark passive strategy. We must acknowledge the tentative nature of this result. Given the uncertainty in approximating transaction cost rates and the lack of explanation

for why a logistic regression model outperformed alternatives, we cannot assert this result with great confidence.

At lower transaction cost rates, several strategies outperformed the benchmark, but this rate may be unable to be achieved in real-world trading. The failure of most strategies to outcompete a passive benchmark at 25 basis points aligns with the EMH. The strategies may only be identifying strategies that persist because transaction costs make them detrimental to trade on.

The SMACI strategy which was among the top-performing strategies may have been exploiting the short-term reversal effect. The SMAC and SMACI models also considered the magnitude of predicted price changes, unlike the other models. This may have had a beneficial effect on strategy performance. However, further investigation would be required to confirm these possibilities and to ascertain why model performance was not reflected well in the test metrics.

The null correlation between test performance and backtest performance is a notable result. A similar result was found by [7]. Several potential explanations could explain this finding. Technical analysis may possess no forecasting power, as asserted by the EMH. An alternative explanation is that the current separation of forecasting and strategy execution is not conducive to a successful investment strategy. It may be beneficial to train models using backtest performance. This could enable the model to consider additional aspects such as excess return and the transaction costs incurred. Considering the magnitude of returns, rather than just positive or negative returns, may also produce performance improvements.

As is common knowledge, our backtest results show that buy-and-hold strategies are resilient to transaction costs.

Good strategy performance may be due to data mining. We evaluated many different models and may have identified some reasonably well-performing models due to chance. Statistical analysis of backtest results could indicate if they are significant.

B. Broader Implications

Our results indicate that researchers should be thorough in their analysis of financial forecasting models for trading. Our findings indicate that forecasting performance does not necessarily correspond to the performance of a trading strategy using those forecasts.

The inability of most models to notably outperform B&H strategies in an approximation of real-world conditions suggests that technical indicators have already been factored into prices to a great extent. A forecasting model that does not consider transaction costs is likely to identify already well-studied effects like momentum and short-term reversal. These are effects that do not survive transaction costs [2, Chapter 6].

Despite the excellent performance of the logistic regression model, most strategies underperformed the benchmark. This aligns with the EMH and the observations presented in [7]. Forecast-based trading strategies do not inherently account for the problem that prices rapidly factor in new information. This problem is unaddressed by all the approaches we evaluated.

It remains uncertain whether strategies processing and acting on data at an inter-day frequency, such as those used in this study, can be consistently successful. Given that information is rapidly factored into prices, we are motivated to explore the feasibility of high-frequency strategies. The theoretical basis of high-frequency strategies aligns with a practical consideration of the EMH, which acknowledges that market prices will have a small response time to new information.

Researchers should consider including transaction costs in both trading strategy development and evaluation. Our results contribute to the body of evidence that transaction costs cannot be overlooked as they can drastically reduce strategy performance. This aligns with the findings in [16]. Although realistic transaction costs are hard to approximate, using no approximation will lead to over-optimistic strategy results.

C. Limitations

We have not analysed why the top-performing strategies achieved that performance. As discussed, good strategy performance could be due to data mining effects. We also only briefly explore the relationship between test metrics and strategy performance.

As previously discussed, the uncertainty around determining realistic transaction cost rates limits our ability to draw confident conclusions.

Due to the stochastic nature of neural network weight initialisation, differences in performance could be evident when evaluating multiple versions of the same model. Training and evaluating multiple instances of the same model would enable us to perform statistical analysis of model variation.

As noted in Section III-E, our backtest contained several intentional simplifications. Although it was expected that these simplifications would have a negligible impact on results, we do not quantify this impact. In particular, the validity of the assumption that trading could take place using end-of-day prices could be contested.

All strategies we evaluated used the same approach to use predictions to construct a portfolio. Using the same predictions but different trading strategies could yield different results. Strategies could be formulated specifically to account for additional factors such as transaction costs [16], [24].

D. Future Research

Models that focus specifically on maximising risk-adjusted returns and limiting costs, may be able to achieve better trading performance. This may enable models to optimise with consideration for factors that might be otherwise obfuscated.

Additional features could be incorporated into model training data, such as fundamental indicators. However, this does not solve the problem of prices factoring in information before the model can exploit temporal arbitrage opportunities.

A higher data frequency rate would enable intra-day and high-frequency trading strategies to be explored. This could potentially enable temporal arbitrage opportunities to be identified and acted upon before they are factored into price.

But, this would likely also require order book simulation and estimations of both strategy and trade execution times.

Future research could focus on forecasting based on primary data sources. Technical indicators contain a lag in that price changes only occur after an event that induced the price change. Using primary data sources would remove the reliance on the actions of other traders and improve the time available to identify and act on temporal arbitrage opportunities.

Further statistical analysis of our results could identify data mining effects and quantify to what extent model and strategy performance is due to chance.

Given the large uncertainty around what constitutes a realistic transaction cost rate, studies that investigate more accurate approximations for transaction costs would improve the accuracy of strategy evaluations.

VI. Conclusion

We evaluated the performance of a range of forecast-based active management trading strategies. The success of such strategies would contradict the efficient market hypothesis, which implies that such strategies will not perform favourably as prices quickly factor in all available information. We investigated whether the performance of these strategies persisted in a trading backtest with transaction costs.

The results of our study tentatively reject the null hypothesis as we developed a strategy that outperformed the passive benchmark. We cannot express this result with great confidence due to the uncertainty in approximating a realistic transaction cost rate, and the possibility that this model only outperformed the benchmark due to chance. Several strategies outperformed the benchmark when transaction costs were absent. However, transaction costs degraded the performance of these strategies. We found no correlation between model test metrics and strategy performance, which bears investigating further.

Our study contributes further evidence to the applicability of the EMH, the robustness of forecast-based trading strategies, and the relationship between forecasts and trading strategy performance. Future work could involve investigating this discrepancy or investigating high-frequency trading strategies.

Acknowledgment

We thank Jiang, Kelly, and Xiu [9] for providing their study code. Due to licensing restrictions, we regret that we are unable to share the raw data used in our research. The study code and supplementary material are available at [22].

References

[1] B. G. Malkiel, "Reflections on the Efficient Market Hypothesis: 30 Years Later," *Financial Review*, vol. 40, no. 1, pp. 1–9, 2005.

[2] ——, *A Random Walk Down Wall Street: The Best Investment Guide That Money Can Buy (13th Edition)*. New York, New York, United States: W. W. Norton & Company, Jan. 2023.

[3] B. Armour. (2023, Feb.) Active Funds Continue to Fall Short of Their Passive Peers. https://www.morningstar.com/etfs/active-funds-continue-fall-short-their-passive-peers.

[4] Y. Millo, C. Spence, and J. J. Valentine, "Active fund managers and the rise of passive investing: Epistemic opportunism in financial markets," *Economy and Society*, vol. 52, no. 2, pp. 227–249, Apr. 2023.

[5] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, no. 2, p. 383, May 1970.

[6] ——, "Efficient Capital Markets: II," *The Journal of Finance*, vol. 46, no. 5, pp. 1575–1617, 1991.

[7] S. Makridakis, E. Spiliotis, R. Hollyman, F. Petropoulos, N. Swanson, and A. Gaba, "The M6 forecasting competition: Bridging the gap between forecasting and investment decisions," Oct. 2023, unpublished.

[8] T. J. Strader, J. J. Rozycki, and T. H. Root, "Machine Learning Stock Market Prediction Studies: Review and Research Directions," *Journal of International Technology and Information Management*, vol. 28, no. 4, Article 3, 2017.

[9] J. Jiang, B. Kelly, and D. Xiu, "(Re-)Imag(in)ing Price Trends," *The Journal of Finance*, vol. 78, no. 6, pp. 3193–3249, 2023.

[10] L. H. Ederington and J. H. Lee, "The Short-Run Dynamics of the Price Adjustment to New Information," *The Journal of Financial and Quantitative Analysis*, vol. 30, no. 1, p. 117, Mar. 1995.

[11] M. Bank and R. H. Baumann, "Price formation, market quality and the effects of reduced latency in the very short run," *Research in International Business and Finance*, vol. 37, pp. 629–645, May 2016.

[12] R. M. Brooks, A. Patel, and T. Su, "How the Equity Market Responds to Unanticipated Events*," *The Journal of Business*, vol. 76, no. 1, pp. 109–133, Jan. 2003.

[13] H. White, "A Reality Check for Data Snooping," *Econometrica*, vol. 68, no. 5, pp. 1097–1126, 2000.

[14] P. R. Hansen, "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, vol. 23, no. 4, pp. 365–380, 2005.

[15] C. Wang, Y. Chen, S. Zhang, and Q. Zhang, "Stock market index prediction using deep Transformer model," *Expert Systems with Applications*, vol. 208, p. 118128, Dec. 2022.

[16] R. Novy-Marx and M. Velikov, "A Taxonomy of Anomalies and Their Trading Costs," *The Review of Financial Studies*, vol. 29, no. 1, pp. 104–147, Jan. 2016.

[17] A. Detzel, R. Novy-Marx, and M. Velikov, "Model Comparison with Transaction Costs," *The Journal of Finance*, vol. 78, no. 3, pp. 1743–1775, 2023.

[18] A. W. Lynch and P. Balduzzi, "Predictability and Transaction Costs: The Impact on Rebalancing Rules and Behavior," *The Journal of Finance*, vol. 55, no. 5, pp. 2285–2309, 2000.

[19] W. R. D. Services. (2024, Jan.) Daily Stock File. https://wrds-www.wharton.upenn.edu/pages/get-data/center-research-security-prices-crsp/annual-update/stock-security-files/daily-stock-file/.

[20] ——. (2024, Jan.) Crsp stock file indexes. https://wrds-www.wharton.upenn.edu/pages/get-data/center-research-security-prices-crsp/annual-update/index-stock-file-indexes/market-cap-daily/.

[21] Board of Governors of the Federal Reserve System (US). (2024, Jan.) 3-Month Treasury Bill Secondary Market Rate, Discount Basis. https://fred.stlouisfed.org/series/TB3MS.

[22] J. Lewis-Cheetham, "The Impact of Transaction Costs · GitLab," https://gitlab.com/JamesLewisCheetham/the-impact-of-transaction-costs.

[23] Y. Shynkevich, T. M. McGinnity, S. A. Coleman, A. Belatreche, and Y. Li, "Forecasting price movements using technical indicators: Investigating the impact of varying input window length," *Neurocomputing*, vol. 264, pp. 71–88, Nov. 2017.

[24] T. I. Jensen, B. T. Kelly, S. Malamud, and L. H. Pedersen, "Machine Learning and the Implementable Efficient Frontier," *Swiss Finance Institute*, Aug. 2022.