Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on Machine Learning-Based Forecasting Models for SSE Indexes-Analysis From The Perspective of Quantitative Time-Timing

Tingyue Wang *

Business School, The Hong Kong University of Science and Technology, Hong Kong, China

* Corresponding Author Email: twangcp@connect.ust.hk

Abstract. Quantitative investment is a hot topic in recent years, through machine learning algorithms to deal with a large amount of data, saving investors' time and energy, to provide investors with reliable investment prediction results, the use of machine learning algorithms in quantitative investment has a broad prospect. This paper is based on the machine learning model to predict the rise and fall of the SSE index, using evaluation indicators to measure the model performance, select the best model for further optimisation; and in the backtest to construct a quantitative time-testing trading strategy, to verify that the model is good profitability in the real market, which is of great significance; and finally to analyse the importance of the features, to increase the interpretability of the model. In this paper, the decision tree, random forest and XGBoost models are used to predict the rise and fall of the index respectively, and the sample interval of the historical data of the SSE index is from 19 December 1990 to 6 December 2022, and the training set and test set are divided according to 7 to 3. After a multi-model comparison of the prediction ability by machine learning evaluation metrics such as Accuracy, Precision, Recall, F1 Score and AUC, it is concluded that the XGBoost model has the best performance in upward and downward prediction. Secondly, the paper successfully re-optimises the prediction accuracy of the XGBoost model in the upward and downward prediction of the Shanghai Stock Exchange (SSE) index using an innovative Bayesian optimisation method. Furthermore, the quantitative strategies constructed based on the optimised XGBoost model predictions are backtested and found to have good profitability. Finally, in order to enhance the interpretability of the model, the SHAP interpretable method is used to analyse the important variables that predict the rise and fall of the index.

Keywords: Machine Learning, Quantitative Investment, Shanghai Stock Exchange.

1. Introduction

Since the reform and opening up, with the deepening of the market economy and the tilting of policies, China's various industries have flourished, especially the stock market, which has expanded rapidly since 1990 and has become increasingly well regulated. Diversified investment needs have driven the emergence of new investment strategies such as statistical arbitrage and hedge trading. In recent years, the widespread application of computer technology in the financial field has fuelled the rise of quantitative investment. Quantitative investment reduces reliance on personal experience, analyses historical data through algorithms and models, and is able to process large-scale information, greatly enhancing decision-making efficiency. This approach avoids subjective errors and helps maintain rationality during market volatility. Europe and the United States have been applying quantitative investment techniques for a long time, especially the success of machine learning in predicting the stock market. China still has much room for development in the field of quantitative investment. The continuous innovation of various strategies has enriched the market application, which is expected to further popularise and optimise quantitative models to enhance market effectiveness and investment returns [1-2].

As an emerging technology, quantitative investment is of great significance and vast opportunities, and should be the focus of academic and industry attention. Today's quantitative investment no longer relies on personal experience but builds trading strategies with the help of big data and machine

learning, which is especially suitable for dealing with the complex high-noise and non-linear characteristics of the stock market. Previous studies have shown that the application of machine learning in quantitative investing is feasible, especially in predicting the prices and ups and downs of financial assets.

This paper focuses on the practical application of machine learning techniques in the quantitative field, aiming to build models that can predict the future ups and downs of the Shanghai Stock Exchange Index, and help investors obtain stable returns. The study compares the prediction effects of multiple machine learning models, selects the optimal model and carries out further optimisation. Subsequently, quantitative strategies are designed and backtested, and model outputs are analysed using interpretable methods to enhance the credibility and interpretability of the models [3]. Finally, the effectiveness of machine learning in quantitative investment is summarised and its future prospects are outlined.

The research idea of this paper is divided into three core issues: firstly, to explore the effectiveness of machine learning methods in predicting the future rise and fall of the SSE index and the performance differences of different models. By evaluating multiple metrics such as Accuracy, Precision, Recall, F1 Score and AUC, the optimal model is selected, and its prediction ability is enhanced through Bayesian optimisation. Secondly, quantitative strategies are built based on the prediction results and their return performance in simulated trading is evaluated to verify the practical significance of machine learning model prediction. Finally, the SHAP visualisation method is used to analyse the interpretability of the model and identify the important variables affecting the prediction results, so as to enhance the interpretability and application effect of the model [4-5].

Methodologically, this paper combines theoretical and empirical research to learn in-depth about machine learning algorithms and their applications in quantitative investment, with special attention to the literature on explainability. The SSE index is empirically predicted by a variety of machine learning models, and the model with the best effect is selected to construct a quantitative strategy, with a view to achieving a high return target. Meanwhile, XGBoost, Random Forest, Decision Tree and XGBoost model optimised by Bayesian optimisation are used for quantitative analysis, and combined with qualitative analysis, to combine financial theories with practical problems, to comprehensively explore the effect of the application of machine learning in quantitative investment and the future development trend.

This paper innovatively employs Bayesian optimisation to tune the selected optimal model with hyperparameters in order to improve the prediction accuracy. Most of the current quantitative investment research is limited to the prediction and testing of different models, and Bayesian optimisation is less frequently used to optimise the models [6-7].

In the field of machine learning, common hyperparameter optimisation methods include Grid Search, Stochastic Grid Search and Halving Grid Search, but they have limitations in terms of computational efficiency and accuracy. In contrast, Bayesian optimisation methods are considered to be the state-of-the-art optimisation framework that can strike a good balance between efficiency and quality of results. Many modern efficient and superior hyperparametric optimisation methods are derived from the basic concepts of Bayesian optimisation.

Nonetheless, the study in this paper still has some shortcomings, especially in feature selection, which relies too much on stock index level features and does not cover a sufficiently rich feature type. Future research can consider introducing more macroeconomic features to further improve the model prediction effect.

2. Machine Learning Models and Optimisation Methods

2.1. Decision Tree Models

Decision tree is a commonly used machine learning algorithm that can be applied to classification and regression tasks and is suitable for both discrete and continuous variables. The CART algorithm was proposed by Breiman et al. in 1984 and is a commonly used algorithm for generating decision trees. The full name of CART is Classification and Regression Tree, i.e. "Classification and Regression Tree". The algorithm mainly consists of two parts: decision tree generation and pruning [8] and adopts the "greedy strategy" for feature selection.

The CART algorithm adopts the Gini index minimisation criterion as the feature selection method, and after feature selection, binary tree division and other processing, the tree-based method is used to construct classifiers and regression prediction models, which can be used for the processing of discrete variables and continuous variables.

2.2. Random forest model

Random Forest is an integrated learning algorithm that belongs to the Bagging method. It constructs a "forest" by integrating multiple decision trees and is used to predict the final outcome. When constructing a Random Forest, Bootstrap sampling is used to extract training data to form multiple training sets, and then a decision tree is trained for each training set. These decision trees are independent of each other, the selection of data and features are random, and a feature is also randomly selected at each node for division to avoid model overfitting.

In the prediction phase, new samples are fed into each decision tree and each tree gives a classification result. Eventually, the final classification result is determined by voting, i.e., the classification with the most votes is chosen as the prediction result.

The construction and prediction process of the random forest model can be divided into several steps: firstly, a portion of samples from all training samples is randomly and with put-back selected to constitute the training set; secondly, a portion of features among all features is randomly selected to constitute the feature subset, and a decision tree is constructed based on these features and the training subset; lastly, a classification prediction is made for each decision tree for the new test samples, and finally, a final classification result is determined by means of voting to determine the prediction results.

2.3. XGBoost model

In recent years, with the rapid development of machine learning, Gradient Boosting algorithm has become one of the important research directions in the field of natural language processing. Among them, XGBoost is an optimised distributed Gradient Boosting library that can solve multiple data science problems quickly and accurately. The algorithm is implemented under the Gradient Boosting framework, which is efficient, flexible and portable.

2.4. Bayesian Optimisation

Bayesian Optimization (BOP) is a method of black-box function optimisation using agent models and Bayesian inference. In the conventional grid search or random search and other tuning methods, it is necessary to design a specific search space for each parameter, while Bayesian Optimisation does not need to specify the distribution of the parameters, and automatically adjusts the parameter values in one iteration and gradually learns the distribution of the optimal set of parameters, thus reducing the complexity of tuning. The core idea of Bayesian optimisation lies in modelling the objective function, based on the posterior probability, calculating the prior distribution probability of the objective function by continuously updating the use of the Bayesian formula, obtaining a new test point and evaluating its function value, and then calculating the posterior distribution probability of the point using the Bayesian method, so as to iteratively optimise[9-10].

3. Empirical studies and interpretable analyses

In this paper, three machine learning models (Decision Tree, Random Forest and XGBoost) are used respectively to predict the rise and fall of the SSE index in the coming day. All model construction in the paper is done in Python.

The training set and test set are divided in a ratio of 7 to 3. The training set is the first 5467 samples, and the sample interval is from 19 December 1990 to 21 April 2013. Meanwhile, the random seed random_state is specified to ensure that the experimental results can be reproduced. The metrics used in this paper to evaluate the model performance are Accuracy, Precision, Recall, F1 Score and AUC, and the optimal model is selected comprehensively for subsequent model optimisation and prediction work.

Table 1 Selection of model parameters

model	Parameter Selection
Decision Trees	random_state=2022, max_depth=6
Random Forest	random_state=2022, max_depth=6, n_estimators=50, class_weight={0:1.5,1:1}
XGBoost	max_depth=6, n_estimators=100, learning_rate=0.5

Note: All other parameters not mentioned are default values.

In the decision tree model, the maximum depth of the tree is 6, which is used to control the structure of the tree and prevent overfitting. The value of the random seed for control data assignment is 2022, and specifying the same random seed ensures the same result for each assignment.

In the random forest model, the maximum depth of the tree is 6, the number of trees in the forest is 50, and the random seed value assigned to the control data is 2022.The category weights used to deal with the category imbalance problem are used here with customised {0:1.5,1:1} weights, which means that samples labelled 0 are set to a weight of 1.5, and samples labelled 1 are set to a weight of 1.

In the XGBoost model, the maximum depth of the tree is 6. This parameter is used to control the structure of the tree to prevent overfitting. In the forest, the number of trees is 100. the learning rate is 0.5, which is used to control the speed of error correction for each tree during the iteration process.

A test will be conducted to compare three different machine learning models to predict the future rise and fall of the SSE index, and after optimising the best of the three, a quantitative timing strategy will be constructed based on it.

3.1. Model Prediction and Optimisation

The test set is a posterior 2343 samples, the sample interval is from 22 April 2013 to 6 December 2022, the test set data plays an important role in measuring the generalisation ability of the model. Models with good generalisation ability are able to adapt to different datasets and handle situations such as noise, missing values and outliers well without overfitting or underfitting. The accuracy on the test set reflects how well the model performs on unknown data. At the same time, the test set can also help us to check how well the model performs in practical applications.

Based on the evaluation metrics presented above, the prediction effects of the three models used, Decision Tree, Random Forest and XGBoost, are compared and evaluated.

The model confusion matrix is shown in Figure 1 to Figure 3.

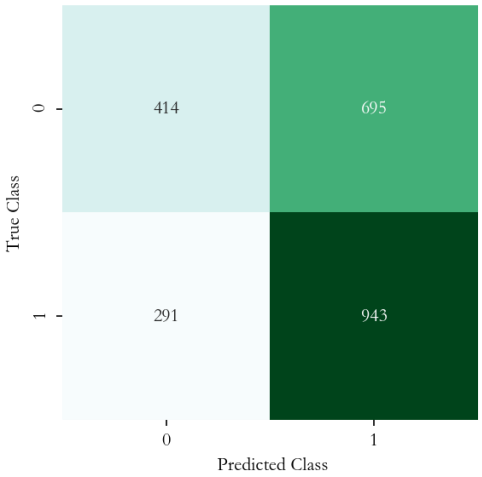

Figure 1 Decision tree model confusion matrix

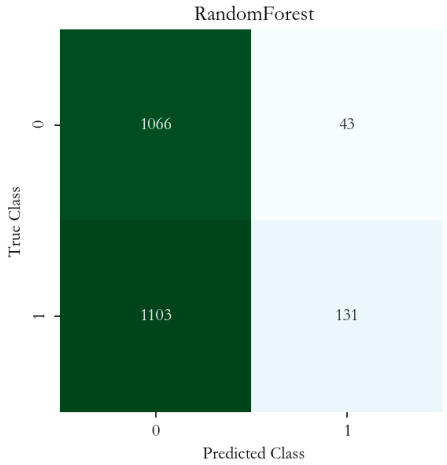

Figure 2 Confusion matrix for random forest model

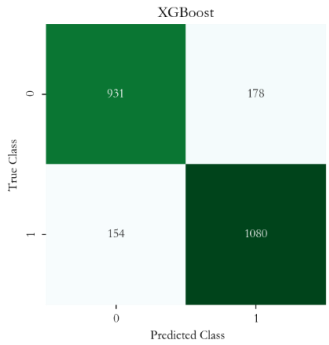

Figure 3 XGBoost model confusion matrix

The performance of the five-evaluation metrics Accuracy, Precision, Recall, F1 Score & AUC used in this paper in the three-model test set data is shown in Table 2.

Table 2 Performance of model evaluation indicators

modelling	Accuracy	Precision	Recall	F1 Score	AUC
Decision Tree	0.5792	0.5757	0.7642	0.6567	0.6137
Random Forest	0.5109	0.7529	0.1062	0.1861	0.6782
XGBoost	0.8583	0.8585	0.8752	0.8668	0.9361

It can be seen from the performance of the three prediction models on the metrics in Table 2:

Firstly, the decision tree model and the random forest model have their own strengths in the values of the metrics, and the decision tree model is better than the random forest in terms of accuracy, recall and F1 Score. With a largely balanced sample, the decision tree model has a slightly better miss rate. The random forest model has better accuracy and AUC than the decision tree, with fewer misclassifications.

Secondly, the XGBoost model outperforms the decision tree and random forest on all indicators, representing that XGBoost is the best model for prediction in the prediction problem of upward and downward movement of the SSE index.

Through the above analysis, it can be concluded:

When using historical data as a sample to predict the rise and fall of the SSE index, Decision Tree and Random Forest have some out-of-sample prediction ability, but there is still room for improvement, while XGBoost's out-of-sample prediction ability is higher than the other two models. Among these three models, XGBoost is the best-performing model, which will be further tuned for optimisation and re-prediction.

In this paper, it is hypothesised that XGBoost's better prediction results are due to its ability to handle high-dimensional sparse data and to automatically learn complex relationships between features, thus providing a great advantage in the task of financial market index prediction. In addition, XGBoost uses a range of feature engineering techniques, such as weight adjustment and missing value filling, to further improve the model's performance. In contrast, while decision trees and random forests can also solve classification problems well, they do not place as much emphasis on feature selection and optimisation as XGBoost. Decision trees may overfit the training data, which leads to poor performance on the test data; while random forests, although they can reduce the risk of overfitting, may ignore some important features. In addition, XGBoost is highly scalable and flexible, capable of handling large-scale data and highly concurrent requests. This makes it popular in the financial industry and adopted by many organisations and enterprises.

Therefore, it is very important to choose appropriate machine learning algorithms in financial real-world applications, taking into account the specific tasks and data situations.

The Bayesian optimal tuning method has a wide range of applications in optimising the hyperparameters of XGBoost classifiers. The method achieves an efficient parameter optimisation process by building an agent model on the objective function, quickly finding the optimal parameter combinations, and updating the agent model in real time in continuous iterations. The objective function of Bayesian optimisation consists of the hyperparameters of the XGBoost classifier, such as the learning rate, the depth of the tree, the proportion of subsamples, etc., and for each set of hyperparameters, we use cross-validation to obtain the performance evaluation metrics of the model, which are used as the return value of the objective function. Selection of model parameters are shown in table 3.

Table 3 Selection of model parameters

Model	Model
Bayesian optimised	max_depth=5, n_estimators=300, learning_rate=0.3
XGBoost	verbosity=0, n_jobs=-1, random_state=2022

Note: All other parameters not mentioned are default values.

In the optimised XGBoost model, the maximum depth of the tree is 5. When it is set to a smaller value, it prevents overfitting but can also lead to underfitting. The number of trees is specified as 300, a larger value usually improves model performance but may also lead to overfitting problems. A

learning rate of 0.3 per iteration is used to control how quickly the model learns at each iteration step. A smaller learning rate increases the stability of the model but may also lead to a slower training process or a tendency to fall into local optima.

In addition to this, there are three default parameters that are also assigned specific values. When the level of detail of debugging information (verbosity) is set to 0, the least amount of information is output. The number of parallel operations (n_jobs) was set to -1, when all available CPUs were used. the random seed was set to 2022, which ensures that the same sequence of random numbers is generated every time the code is run, thus making the experimental results reproducible.

The Bayesian-optimised XGBoost model outperforms the original XGBoost model in all metrics.

In order to show more visually the gap between the performance of the models (as seen by the RUC-AUC), Figure 4 plots the ROC curves of the decision tree, random forest, XGBoost, and Bayesian-optimised XGBoost models in a single graph.

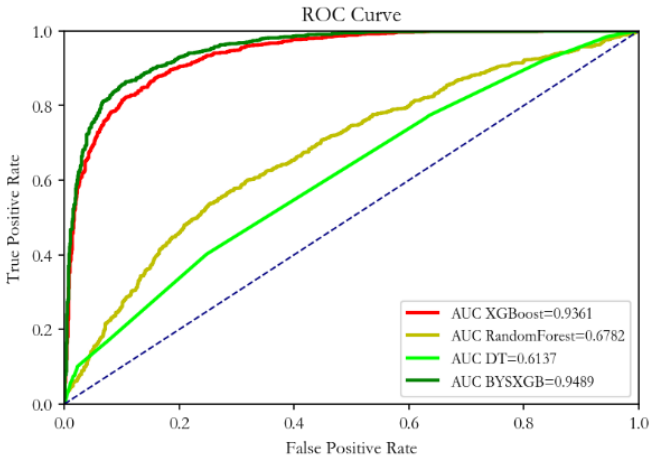

Figure 4 Decision tree, random forest, XGBoost and optimised XGBoost model ROC curves

Overall, the decision tree and random forest have similar ability to predict the ups and downs of the SSE index, with AUC values between 0.6 and 0.7. XGBoost is significantly different from these two models with an AUC value of 0.9361, and the Bayesian-optimised XGBoost is again slightly better than the original model with an AUC value of 0.9489. The above study on performance evaluation using cross-validation methodology demonstrates the better predictive ability of the XGBoost classifier compared to decision trees and random forests, as well as the significant benefits of Bayesian optimisation methods in machine learning. The application of Bayesian optimisation methods in XGBoost can effectively improve the prediction accuracy and stability of the model.

Therefore, it can be concluded:

When predicting the ups and downs of the SSE index, the performance of the XGBoost model can be significantly improved by using the Bayesian optimization method to adjust the parameter and provide more accurate and reliable prediction results for practical applications. The XGBoost model and the Bayesian optimization method play an important role in machine learning, which can provide a strong support for the training and adjustment of the model and make the machine learning technology better applied to the real-world scenarios. in real-world scenarios.

3.2. Backtesting experiment and result analysis

Model backtesting is a method used to test the effectiveness of financial models. Model backtesting is usually done by simulating real trades in the financial market during the test period using the model's predictions and measuring the effectiveness of the model based on the returns obtained. This type of test is more convincing to investors when it is based on real trading.

3.2.1. Trading strategies and evaluation indicators

Pre-conditions: In order to avoid the impact of tracking error and liquidity constraints to the maximum extent possible, "buying or selling SSE index" in the subsequent trading strategy means that all the constituent stocks in the SSE index are considered as a whole portfolio in accordance with the proportion and weighting of the index rules, and at the same time, the liquidity problem of buying and selling certain stocks is not taken into account. It is equivalent to buying and selling an ETF fund that closely tracks the SSE Index on its own.

Trading strategy: (1) Input the pre-processed daily frequency data of the SSE index into the Bayesian optimised XGBoost model; (2) The model outputs the prediction results of "up" and "down" of the SSE index on the next day, with the following values (2) The model outputs the "up" and "down" prediction results of the SSE index on the next day, which is indicated by "1" and "0"; (3) If the output is "up", the whole position will be bought in the SSE index, and if the output is "down", the whole position will be sold in the SSE index. (4) Repeat the above process for each trading day and adjust the position.

3.2.2. Backtest results and analysis

Backtest evaluation index: the profitability of a quantitative model can be evaluated by selecting the cumulative return index as the evaluation index of quantitative trading in the following section.

Backtesting is performed over the time period corresponding to the test set, corresponding to the latter 2343 samples, dated from 22 April 2013 to 6 December 2022. The cumulative returns achieved by the strategy during the period are shown in Figure 5.

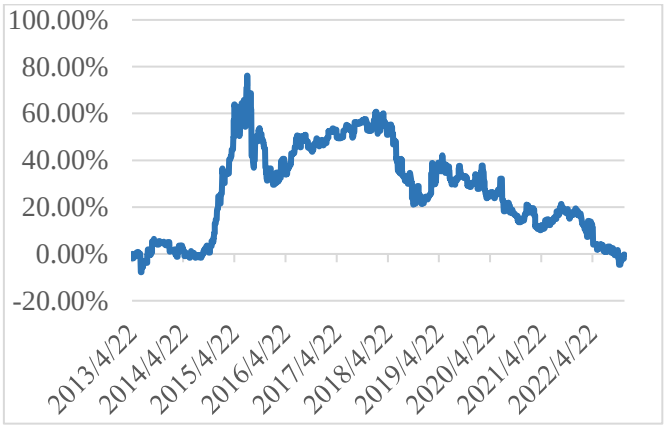

Figure 5 Strategy Cumulative Returns

In China's stock market, the common situation is "bear market", "bull market" and "downturn shock", the stock market in these states alternately back and forth. The market performance of the intrinsic law and the quantitative model varies in different market situations.

According to Figure 5, the performance of the model varies under different market conditions. In this paper, the testing period is divided into four phases based on market conditions.

The period from April 2013 to December 2014 belongs to the depressed and oscillating phase, where the market has a small volatility level and trading volume, investor confidence is low, and the index data is oscillating. During this phase, the cumulative return of the model was somewhat affected by the overall low level of the general market.

The period from January 2015 to December 2015 was a period of intense change in the stock market, with the market experiencing a transition between a "bull market" and a "bear market", and also the widely known "5000 point mark" period. It is also widely known as the "5000 point mark" period. The SSE index experienced a "bull market" close to 5000 points, and experienced a sharp decline back to 3000 points, completed the "bull" and "bear" market conversion. During this period, the strategy showed good returns, with a large increase in cumulative returns, reaching a maximum of

76.16%. In this phase, there were both up and down-market performances, and the model showed good returns, indicating the effectiveness of the quantitative strategy based on the predicted results of the XGBoost model after Bayesian optimisation.

January 2016 to June 2018 was the stabilisation phase where the stock market, which had experienced a crash, levelled off. The cumulative return also remained above and below 55% during this time period, which is a good performance.

The period from July 2016 to December 2022 will be a combination of "bull" and "bear" markets, including short-term oscillations, and generally a period of intense change. 2016 saw the "619 crash", which triggered a potentially volatile market. After the "619 crash" in 2016, the market was greatly hurt by the potential risk exposure, and it was not until 2019 that a new wave of upward market was ushered in. The bullish bubble burst after the Chinese New Year in 2021, ushering in a modest decline. After a short-term shock, from February 2022, the stock market fell and oscillated until 6 December 2022 (6 December 2022 is the date of this writing). The quantitative strategies underperformed during this period. It is inferred that the possible reasons for this are that, in addition to the effect of the market decline, this time period is also too far away from the training period, coupled with the fact that the intrinsic nature of the market changes at each stage, and that there is a large discrepancy between the data from the training samples that are too early and the current market's operating pattern, leading to deviations from the model's predictions. In the future, new data can be added to train the model to further improve the performance of the strategy.

3.3. SHAP interpretability analysis

In the above empirical analysis, the model prediction effect is measured using the machine learning evaluation metrics Accuracy, Precision, Recall, F1 Score and AUC, and the Bayesian optimised XGBoost performs optimally on all metrics, with an F1 Score of 0.8827 and an AUC value of as high as 0.9489. Further on, the construction of the trading strategy, it is found that the cumulative return of the strategy reaches up to 76.19% without considering the condition of transaction costs, which can bring sufficient returns for investors.

However, in reality, it is not enough to construct a model with good prediction effect, and there are high requirements for interpretability in actual operation. When recommending a stock, the investment consultant must explain to the client the reason for recommending the stock. If it is not possible to explain, even if the model has a good prediction effect, it is difficult to convince the client. In addition, the operation process of machine learning is a "black box" state, and the complex model structure is accompanied by a very low interpretability, and there is an urgent need to introduce some indicators to enhance the interpretability of the model. In this paper, the SHAP method is chosen to explain the final Bayesian optimised XGBoost model for prediction.

SHAP (SHapley Additive exPlanation) is a feature importance assessment method for machine learning algorithms. SHAP utilises the concept of Shapley values, considers the contribution of each feature in the prediction, and transforms the global model interpretation into a local-based model interpretation problem. In the context of machine learning, SHAP is able to assess the extent to which each feature contributes to the prediction by constructing an explanatory model for each input sample. The core of SHAP is the SHAP value, which represents the average contribution of each feature to the final prediction. SHAP uses an additive function to assign the contribution of each feature to the prediction to each feature. Applications of SHAP include model SHAP can also be used to interpret various types of models (including neural networks) and therefore has a wide range of applications in practice.

3.4. Feature Importance and Feature Effect

In machine learning, sample data usually consists of one or more features (also known as "variables"), each of which has an effect on model prediction. Therefore, it makes sense to understand the importance of each feature to the model prediction.

Feature importance refers to the extent to which features contribute to model predictions. Specifically, it can tell us how the model predictions change when a feature changes. By revealing feature importance, we can better understand the behaviour of the model.

In the following section, the Summary Plot and Beeswarm Plot in the SHAP method will be plotted to visualise the importance ranking of each feature, see Figure 6 to Figure 7.

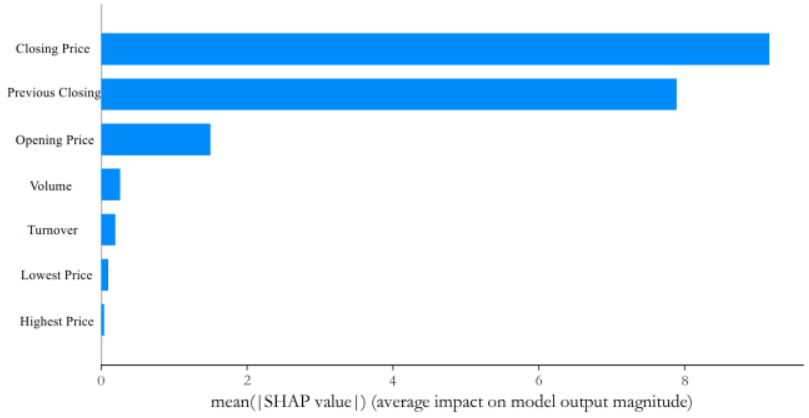

Figure 6 Summary Plot

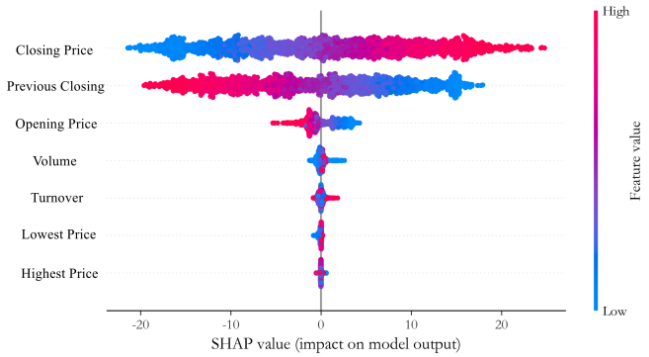

Figure 7 Feature density scatter plot: Beeswarm Plot

The Summary Plot ranks feature regardless of their values. The Summary Plot averages the absolute values of the SHAP values for each feature to show the distribution of feature importance, blurring positive and negative influences compared to the Beeswarm Plot. The vertical axis of the bar graph indicates the feature name, and the horizontal axis indicates the range of SHAP values.

As seen in Figure 6, the two variables that have the greatest impact on the model output are the closing price and the previous closing. The SHAP values of these two variables are calculated to be +9.16 & +7.89 respectively.

The Beeswarm Plot (Feature Density Scatter Plot) distinguishes between feature values to rank features and shows the distribution of SHAP values for each feature. Each scatter represents a sample, the horizontal axis represents the value of the feature, and the vertical axis represents the SHAP value; a higher SHAP value (larger vertical scale) indicates that the feature contributes more to the model output, while a lower SHAP value (smaller vertical scale) indicates that the feature contributes less or negatively to the model output.

Density scatter plots allow you to compare the magnitude and direction of influence between multiple features by displaying their SHAP values one at a time. Each point represents a sample, and the distribution of points also presents the magnitude and extent of the feature's influence on the model output at different values. To the left and right of the points, there are usually two sets of scattered points with higher density, which indicates that the corresponding feature values may have positive and negative impacts on the model predictions.

As can be summarised from Figure 7, the closing price is the variable that contributes the most to the output, with high closing prices positively affecting the forecasts and low closing prices negatively affecting the forecasts, and the previous closing is the variable that contributes the second most, with high previous closing values negatively affecting the forecasts and low previous closing values positively affecting the forecasts.

4. Conclusion

This paper uses three machine learning models (decision tree, random forest and XGBoost) in order to predict the rise and fall of the SSE index in the coming day and draws the following main conclusions.

Firstly, the XGBoost model has a high accuracy of upward and downward prediction. According to the evaluation index of machine learning models, the accuracy rate of XGBoost model is calculated to be 48.24% higher than that of decision tree and 68.02% higher than that of random forest; the F1 Score is 32.03% higher than that of decision tree and 364.06% higher than that of random forest; and the AUC value is 52.45% higher than that of decision tree and 37.97% higher than that of random forest. This indicates that the XGBoost model has better performance in the problem of predicting the rise and fall of indices in the market.

Secondly, the prediction accuracy of the XGBoost model is further improved after adjusting the parameters by Bayesian optimisation. The prediction accuracy of the optimised XGBoost model is higher than that of the original model by 1.83%, the F1 Score is higher than that of the original model by 1.83%, and the AUC value is higher than that of the original model by 1.36%. It can be found that the Bayesian optimisation method is effective in improving the performance of the XGBoost model, resulting in better prediction.

Thirdly, the quantitative strategy constructed based on the output predictions from the Bayesian optimised XGBoost model, which has the best performance in the evaluation metrics, achieves significant gains. When investing in the entire portfolio and not considering transaction costs, the cumulative return over the test period 2013-2022 reached a maximum of 76.16%. In addition, the significant returns obtained when forecasting using only historical data as sample data also indicate that the Chinese stock market has not reached weak efficient.

Fourth, in the interpretability analysis, this paper conducts a feature importance analysis to effectively enhance the interpretability of the XGBoost model, and initially explores the "thinking process" of the machine learning model in the "black box". The results show that the most important influencing variables are closing price and previous closing, high closing price has a positive influence on the prediction results, and high previous closing value has a negative influence on the prediction results.

In recent years, with the continuous progress of technology, machine learning has developed rapidly and become one of the popular areas. This paper has chosen to study only a few representative machines learning classic models, due to time and energy constraints, may not be able to cover some of the other possible prediction of better machine learning models, the future can be expanded in this area. The index predicted in this paper is the Shanghai Stock Exchange (SSE) index, and in the future, this paper can also introduce the prediction of other indexes to test the model robustness and prediction effect. At the same time, the SHAP interpretation method can also be extended to other models and indices to help "unravel" the logic of machine learning algorithms.

In conclusion, although there is still a lot of room for development in China's research on machine learning prediction of the rise and fall of financial targets, with the strengthening of people's understanding and the improvement of technology, a more perfect and accurate prediction system and quantitative strategy will be formed in the future. In addition to the pursuit of prediction results, it is urgent to improve the inherent interpretability of machine learning models due to the relevant norms and ethical requirements. Understanding the decision-making process of the model is beneficial for improving the trust in the prediction results. At present, the number of studies exploring

the internal logic of machine learning is still relatively small, and in future research, this paper hope to unveil the process of machine learning models to a greater extent, so that machine learning algorithms can benefit society more and more.

References

[1] Dong Liang. Quantitative trading's best underlying range is CSI 300 constituent stocks[N]. Beijing Business News, 2024-06-11 (006).

[2] Wei Zhaoyu. Orient Red Asset Management's Xu Xijia: Digging deep into long-term effective factors to make "understandable" quantitative products[N]. China Securities Journal, 2024-06-03 (J08).

[3] Ge Yao. Guolian Fund's Chen Qianyu: Empowering index enhancement with AI quantification[N]. China Securities Journal, 2024-05-20 (J02).

[4] Song ZHH. Dilemmas and Cracking Strategies of Quantification of Collective Operational Property Revenue Right [J]. Finance and Economics Law, 2024, (03): 35-50.

[5] Zhu Yan. Western Leader's Chen Yuanhua: Adopting low valuation strategy to explore dividend value opportunities[N]. Shanghai Securities News, 2024-05-12 (003).

[6] WU Qi, WANG Xiaoqian. Microcap strategy is being abandoned as quantitative public equity firms embrace fundamental stock picking[N]. Securities Times, 2024-04-22 (A07).

[7] Dong Peng. Regulatory inquiries lead to A-share dividend "game": inquiry letters, abstentions and the difficulty of quantifying dividend capacity[N]. 21st Century Business Herald, 2024-04-16 (010).

[8] Liu WJ. Nanhua Fund Huang Zhigang: Subjective stock selection and quantitative investment[N]. China Securities Journal, 2024-04-08 (J08).

[9] Wei, Zhaoyu. Lianbo fund Zhu Liang: Quantitative fundamentals escort to dig for gold with value strategy[N]. China Securities Journal, 2024-03-25 (J05).

[10] Tang YF, Wei YT. The feast falls apart. Quantitative industry at crossroads[N]. Shanghai Securities News, 2024-03-19 (003).