Dean&Francis
ISSN 2959-6130

Machine Learning Methods for Predicting the Price of Exchange-Traded Funds

Zhibo Feng

College of Finance, Nanjing Agricultural University, Nanjing, Jiangsu, 210014, China
fzhibo@stu.njau.edu.cn

Abstract:

This study applies machine learning algorithms to the field of quantitative finance. By employing both Random Forest and Extreme Gradient Boosting (XGBoost) models to predict price movements of nine different Exchange-traded funds (ETFs) from the US, it assesses the practical performance of machine learning in ETFs' price forecasting, thereby assisting investors and institutions in better evaluating future ETF trends. The ETFs' price data used in this research are sourced from the US ETF Prices datasets on Kaggle. Technical indicators such as Bollinger Bands, Relative Strength Index (RSI), and Moving Average (MA) were incorporated into the models through feature engineering. The performance of both models was evaluated across different time windows and ETF products. Comparative analysis revealed that both Random Forest and XGBoost perform well within the 5 to 200-day forecasting horizon. The results indicate that larger sample sizes positively impact the goodness-of-fit of the Random Forest model, while excessively large samples may lead to degraded performance in XGBoost. In conclusion, while machine learning algorithms show strong promise in predicting ETF price movements, practitioners should still integrate market experience, sentiment analysis, and multi-factor evaluation to comprehensively assess ETF performance.

Keywords: Exchange-traded Funds; XGBoost Model; Random Forest model; Quantitative investment

1 Introduction

Nowadays, high-frequency trading of Exchange Traded Funds (ETFs) has enhanced market liquidity, provided investors with more flexible investment strategy combinations, and promoted the development of financial markets. ETFs are a type of open-end fund that tracks a specific investment target by forming a portfolio comprising a range of financial products, including stocks, bonds, and cash instruments. Compared to traditional stocks, ETFs offer greater liquidity, lower transaction costs, and a broader array of investment strategy options for investors. Benefiting from high liquidity and relatively lower

ISSN 2959-6130

risk than stocks, ETFs are increasingly favored by institutional and individual investors and are widely traded in the secondary market.

Predicting ETFs' prices has long been a key focus for numerous experts in the field of quantitative finance. The Autoregressive Integrated Moving Average (ARIMA) model, as an efficient econometric statistical tool for time series analysis, has been widely used for forecasting prices of individual ETFs. Banhi successfully applied the ARIMA model to predict the price movements of gold ETFs in India and yielded a coefficient of determination (R^2) as high as 0.993 [1]. His research revealed that although the predictive assumptions of the ARIMA model are based on a strictly linear pattern, algorithmic techniques can be introduced to approximate nonlinear computational effects. This approach effectively reduces noise interference and achieves high predictive accuracy, enabling reliable forecasts of ETF price fluctuations.

However, as the number of ETFs continues to expand and the volume of data increases, there is a growing need for more efficient algorithms to predict ETF prices. Today, machine learning algorithms have gained widespread popularity among scholars in quantitative analysis, owing to their powerful data processing capabilities, ability to identify meaningful signals within high-dimensional data, and their capacity for maximizing prediction accuracy.

Given the volatility of financial markets and the impact of market sentiment on ETFs' prices, a growing number of scholars are adopting machine learning techniques for stock price prediction. In the context of ETFs' price forecasting, Jim found that Random Forest Model and Support Vector Machine (SVM) demonstrate excellent performance in long-term ETFs' price prediction [2]. However, for short-term price forecasting, feature engineering should be applied to adjust the datasets to achieve better predictive results. Furthermore, Perry observed that the Random Tree model achieved prediction accuracy exceeding 80% in forecasting the price of a clean energy ETF, significantly outperforming Logit regression in this application [3].

Currently, while machine learning delivers strong performance in predicting ETFs within specific categories—particularly demonstrating excellent capabilities in medium to long-term forecasting—significant research gaps remain concerning its application across the entire ETF market and especially in short-term price prediction.

This study will establish feature variables from the datasets from Kaggle through feature engineering and employ machine learning algorithms, including the random forest algorithm and the XGBoost model, to predict the price of ETFs in the U.S. market. By leveraging machine learning algorithms, this research aims to assist active traders in the secondary market in better assessing the future value changes of ETFs.

2 Methodology

This study aims to predict the future price of ETFs. Before constructing the prediction model, feature variables will be engineered using adjusted closing prices instead of absolute closing prices to improve forecasting accuracy.

2.1 Datasets

This dataset is sourced from US ETF Prices on Kaggle [4]. This study extracts nine ETFs from this dataset to form the raw ETF datasets for the research, as the Table 1, using these nine ETFs to represent the broader market for price prediction purposes.

Table 1. ETFs datasets

Field	Ticker symbol
Semiconductor	SMH
Nasdaq-100 Index	QQQ
Gold	GLD
Biotechnology	IBB
Cloud Computing	SKYY
S&P 500	VOO
Network Security	CIBR
Russel 2000	IWM
Dow Jones	DIA

2.2 Random Forest Model

Random Forest is a Bagging (Bootstrap Aggregating) ensemble algorithm. It operates by constructing multiple decision trees and aggregating their results to make predictions, making it particularly suitable for computational quantitative finance. The method effectively handles

high-dimensional features, demonstrates robustness to outliers and missing values, and performs well on noisy datasets. Notably, it exhibits strong resistance to overfitting.

In this paper, model hyperparameters are optimized through pruning techniques, with a focus on tuning two key parameters: the maximum depth of each tree and the size of the feature subset. This helps prevent excessive branching in the trees, reduces the influence of noise in the training data, and ultimately produces simpler trees with stronger generalization ability. As a result, it enhances both the interpretability and computational efficiency of the Random Forest model.

2.3 XGBoost Model

XGBoost (Extreme Gradient Boosting) is currently a highly popular Boosting algorithm. It trains a series of weak learners (decision trees), with each subsequent tree correcting the residuals of the previous one, gradually reducing prediction errors. The final model is obtained through a weighted summation of these learners, resulting in a highly efficient predictive system.

XGBoost approximates the loss function using a second-order Taylor expansion and employs a histogram-based algorithm to optimize split point finding, thereby improving computational efficiency. When applied to complex financial data, it maintains generalization capability through the adjustment of regularization parameters. In this study, time-series cross-validation combined with grid search was used to tune the model's regularization parameters to achieve optimal performance.

2.4 Feature Engineering

This study selected indicators such as moving averages, volatility, momentum, RSI, and Bollinger Bands to construct the feature set. Within the feature engineering process, moving averages, RSI, and Bollinger Bands will serve as the primary focus of this paper, and their performance will be validated against expectations through feature importance evaluation.

2.4.1 Moving Average (MA)

The raw ETFs' price data, when applied in machine learning models, exhibits characteristics of high noise and non-stationarity. Transforming the closing prices into moving averages can effectively mitigate the risk of overfitting and enhance the model's generalization capability. This study will construct moving averages with lookback periods of 5, 10, 20, 50, 100, and 200 observations, respectively.

2.4.2 Bollinger Bands (BBs)

The Bollinger Bands consist of three lines: an upper band, a middle band, and a lower band. They are constructed using two parameters: the moving average period and the standard deviation multiplier. The Bollinger Bands indicator enhances machine learning models by providing the relative position of prices based on statistical distribution. This enriches the model's feature set and improves its capability for dynamic analysis.

2.4.3 Relative Strength Index (RSI)

The RSI (Relative Strength Index) is a momentum oscillator that measures the magnitude and velocity of recent price changes to evaluate whether an asset is overbought or oversold. Its core concept involves comparing the average magnitude of gains and losses over a specified period.

3 Results & Discussion

By comparing the performance of these two models, the study aims to identify the more suitable approach for forecasting ETFs' prices. As shown in Fig. 1, the study found that the XGBoost Model performs exceptionally well in the short term. However, its performance declines significantly as the forecast period extends. In contrast, the Random Forest model demonstrates relatively stable performance over short to medium terms. Moreover, under the influence of positive drift observed in some ETFs' data over time, their performance even improves to some extent. Samraj indicates that XGBoost demonstrates remarkable performance in short-term stock price prediction under conditions of low volatility and minimal noise interference [5]. However, it exhibits significant prediction errors in medium to long-term forecasts. Overall, the XGBoost model achieves an R-squared value of approximately 0.89, slightly outperforming the Random Forest model, which yields an R-squared value of around 0.72. In this study, the performance of the two models is similar to the result of Samraj's study. Overall, both models show strong predictive capability for ETFs' prices, with R^2 values consistently above 50% for 5 days to 150 days. This indicates that either model can be effectively used to construct a reliable forecasting framework.

ISSN 2959-6130

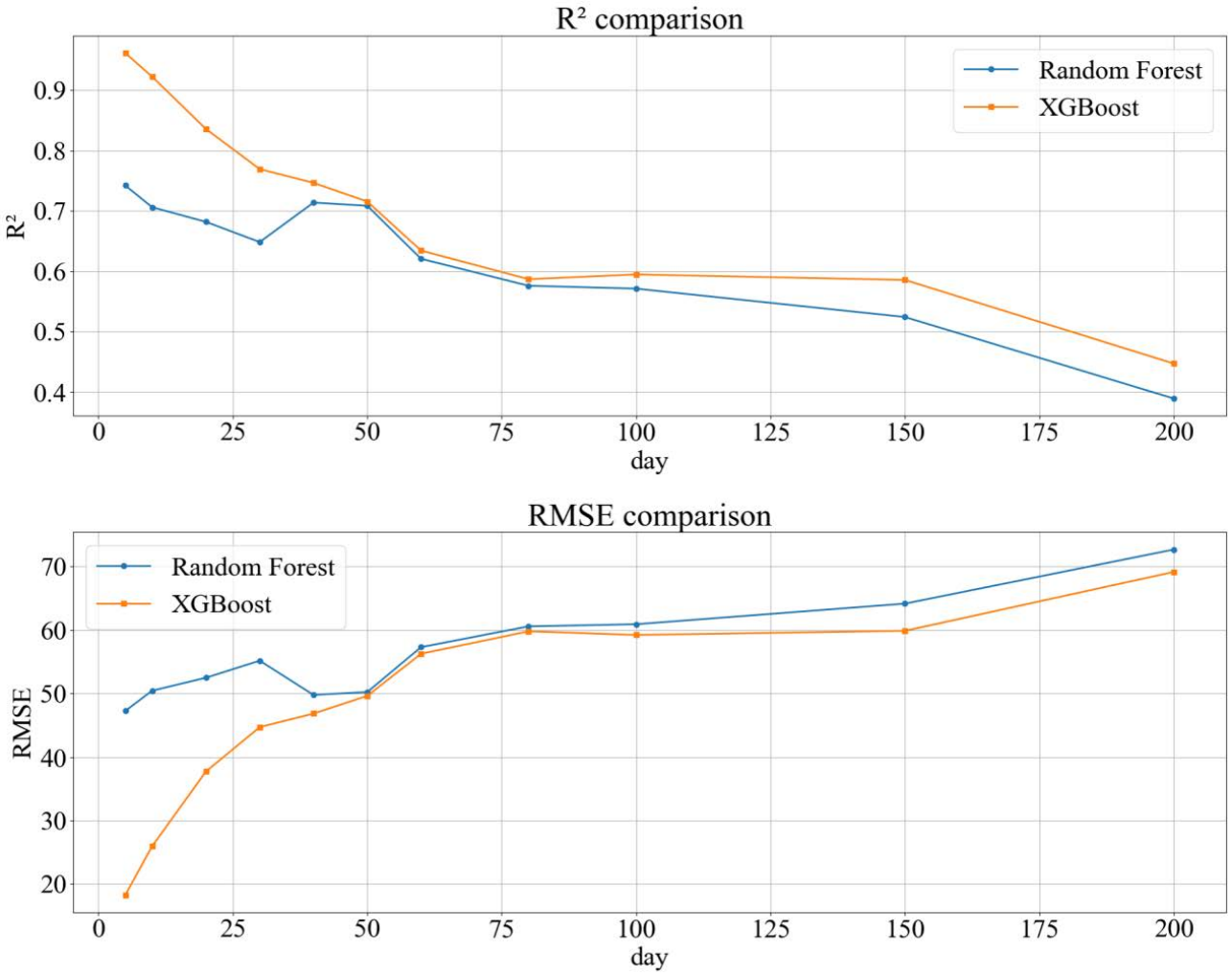

Fig. 1 Rsquare and RMSE of two models (Photo/Picture credit: Original).

To evaluate the accuracy of the testing results, the widely-traded ETF GLD—which has a large sample size in the original database—was selected for separate application of the two price prediction models to validate their forecasting effectiveness. Subsequently, ABEQ, a smaller and higher-risk ETF established for managing high-liquidity hedge funds in recent years and not included in the original datasets, was further introduced. Its predictive outcomes can better illustrate the accuracy of the two forecasting models under more challenging conditions.

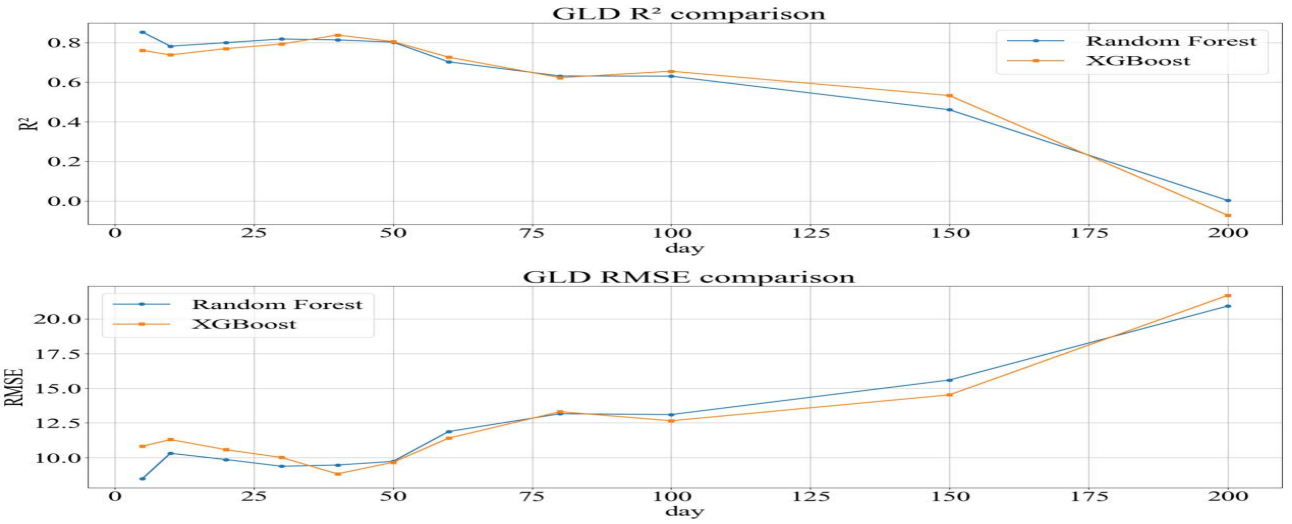

Fig. 2 R-squared of the two models with GLD (Photo/Picture credit: Original).

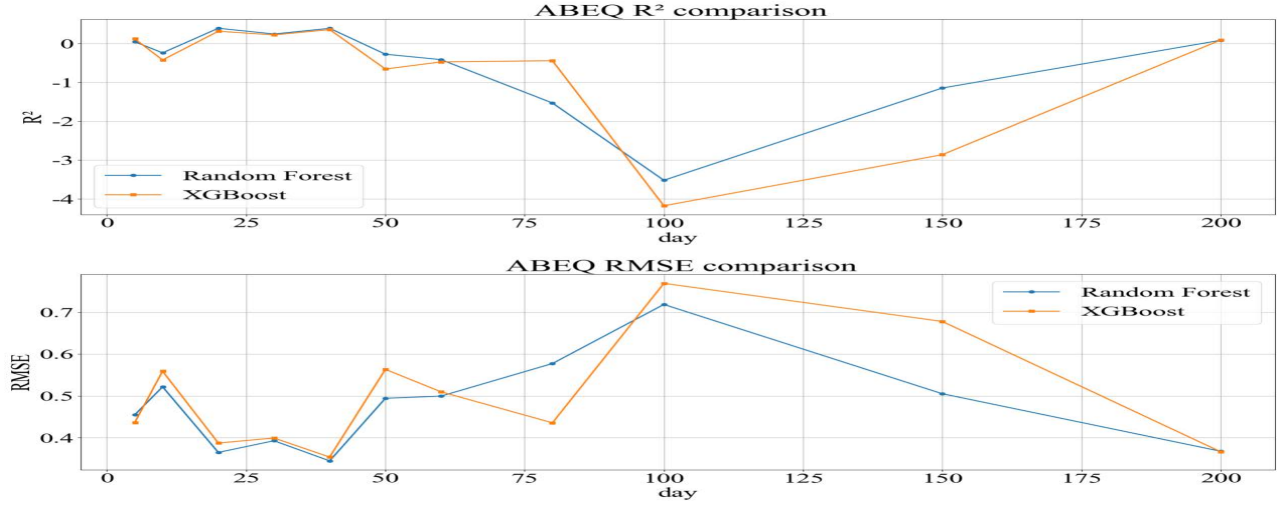

Fig. 3 R-squared of the two models with ABEQ (Photo/Picture credit: Original).

As shown in Fig. 2, the XGBoost model's performance is weaker than the Random Forest Model in short-term predictions of GLD prices. However, as the sample size increased, its predictive performance gradually surpassed that of the Random Forest Model. This suggests that the XGBoost Model exhibits greater potential in large-sample forecasting scenarios. Furthermore, although GLD—being a gold-related ETF—experiences substantial price volatility, both models achieved a goodness-of-fit exceeding 70%, indicating reliable performance even under highly fluctuating market conditions.

As illustrated in Fig. 3, in the specific period, both models demonstrated noticeably weaker performance on the smaller datasets compared to the previous results, particularly in medium-term forecasts, where their predictive accuracy was less satisfactory. This may be attributed to the inherent risk and volatility of ABEQ. These findings suggest that the performance of machine learning algorithms in predicting prices of ETFs is highly influenced by market sentiment requires further optimization—such as enhanced feature engineering or hyper-parameter tuning—in subsequent research.

And then, through feature importance analysis, we aim to identify the key indicators that contribute to the predictive performance of the models. In subsequent research, feature engineering can be applied to increase the weight of these critical variables, thereby further enhancing model performance. Moreover, obtaining feature importance is relatively straightforward for both the Random Forest and XGBoost models.

ISSN 2959-6130

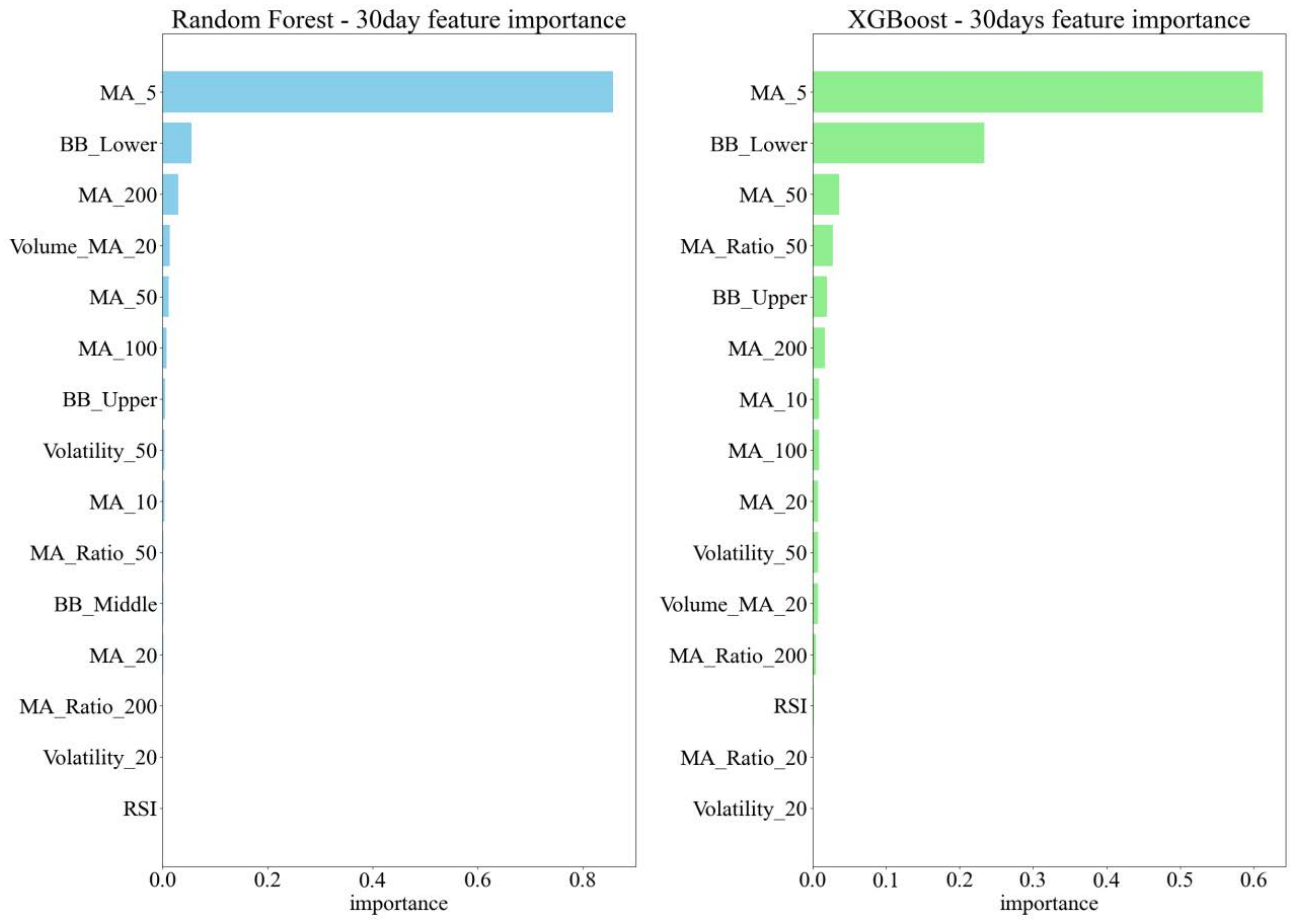

Fig. 4 Feature importance of the two models (Photo/Picture credit: Original).

As shown in Fig. 4, we selected the feature importance values at the 30-day mark for in-depth analysis. The results indicate that the 5-day moving average (MA5) is the most significant feature for both predictive models. Furthermore, an interesting observation emerged: although BB_lower ranked as the second most important feature in both models, it accounted for a substantially larger proportion in the XGBoost model. This suggests that Bollinger Bands-related indicators could be further optimized in subsequent feature engineering to considerably improve prediction accuracy. Additionally, since volatility indicators and the RSI exhibited relatively low importance in both models, these parameters could potentially be removed in future model optimization efforts.

Finally, Zimo noted that the Bayesian optimization algorithm can enhance the performance of models such as LightGBM, XGBoost, and Random Forest, with a particular improvement of up to 6% in accuracy observed for the XGBoost model [6]. To further explore the application of machine learning algorithms in ETF price prediction, this paper employs a Bayesian optimization algorithm to optimize the parameters of the Random Forest model.

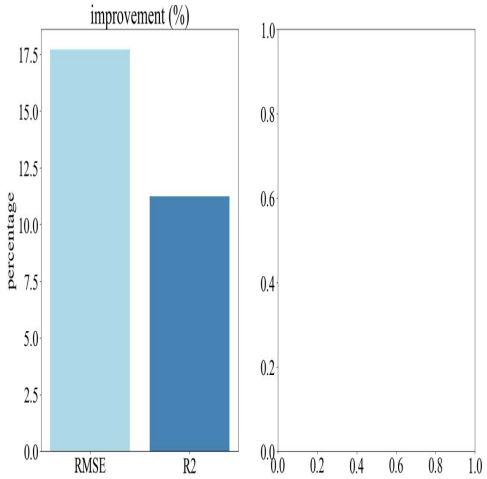

Fig. 5 Improvement after Bayesian (Photo/Picture credit: Original).

As shown in Fig. 5, the Bayesian-optimized Random Forest model demonstrates an overall performance improvement of over 10% compared to its non-optimized baseline. This result indicates that the Bayesian optimization algorithm exerts varying degrees of enhancement across different predictive models, and its effectiveness in

optimizing performance depends on both the model architecture and the characteristic variables involved.

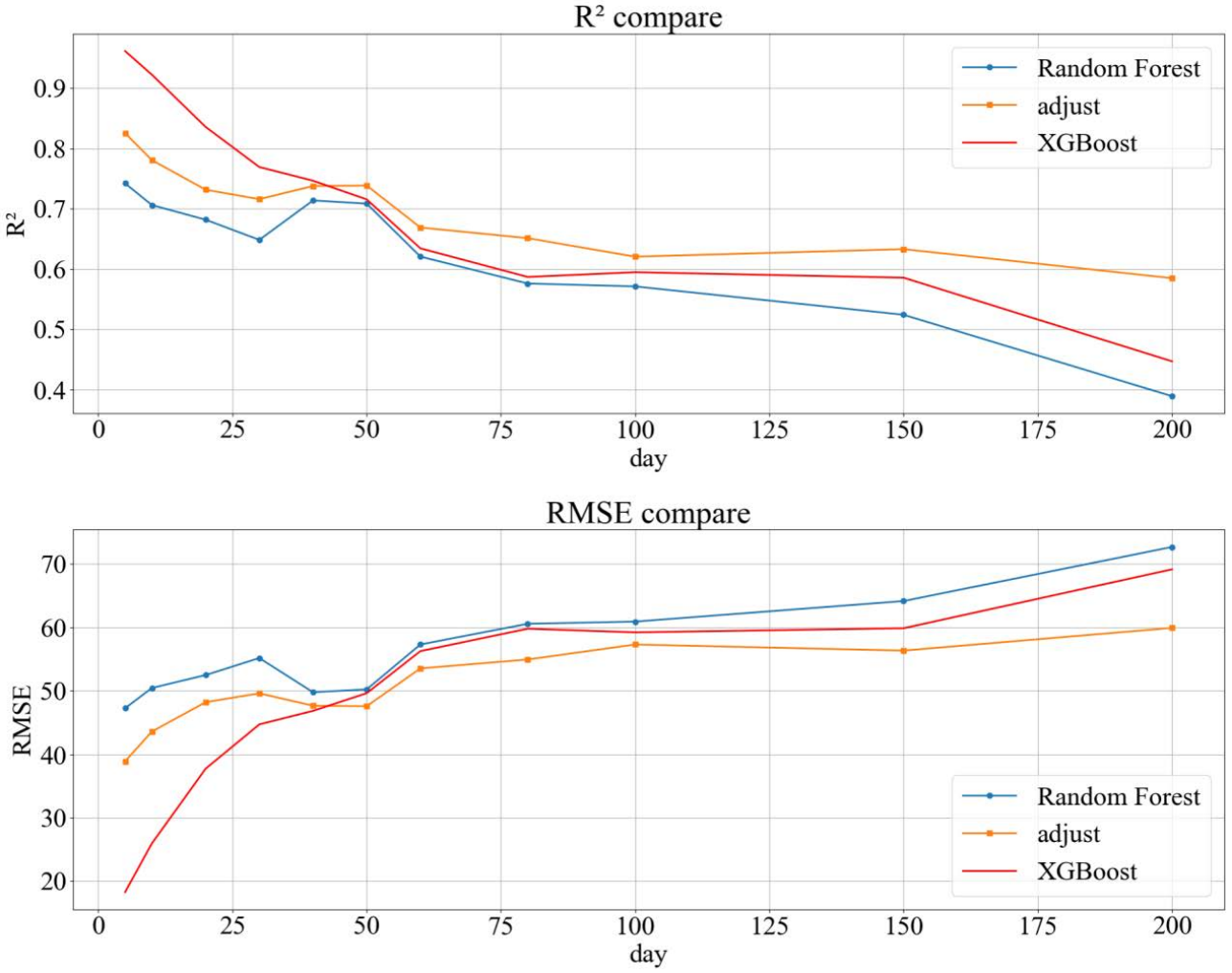

Fig. 6 R-squared and RMSE of the two original models and the adjusted random forest model (Photo/Picture credit: Original).

As illustrated in Fig. 6, the optimized Random Forest model after Bayesian optimization still underperforms the XGBoost model in short-term predictions, yet it demonstrates superior forecasting results in medium to long-term scenarios. This indicates that the Bayesian optimization algorithm significantly enhances the performance of machine learning models. Another critical observation is that the prediction trends of the Bayesian-optimized model remain highly consistent with those before optimization. Therefore, although the Random Forest model exhibits excellent performance in medium to long-term price forecasting, it still shows noticeable limitations in short-term predictions.

4 Conclusion

By evaluating two high-performing machine learning algorithms—Random Forest and XGBoost, this research has predicted the prices of US. ETF Through feature engineering and careful datasets construction, meanwhile, achieved strong predictive performance and obtained valuable outcomes.

First, it is evident that both algorithms demonstrate excellent performance in predicting prices over a predefined window of 5 to 100 days. However, both exhibit limitations in predicting long-term price movements. This is attributable not only to the limited sample size inherent for long-term forecasting but also to the limited presence of positive drift in short-term data, factors that currently constrain the effectiveness of machine learning in very long-term forecasting.

Furthermore, by conducting comparative analysis using individual datasets from the database, we observed that the performance of both models varies across ETFs from different markets or with different investment objectives. XGBoost significantly outperforms Random Forest under

ISSN 2959-6130

moderate sample sizes, whereas it is slightly weaker than in short-term predictions—especially for highly volatile ETFs such as ABEQ. The performance of both models under very large sample sizes warrants further investigation. Nonetheless, it is undeniable that both offer considerable predictive capability for ETF prices. Additionally, the results suggest that more complex and volatile ETFs may require more diverse datasets to optimize prediction accuracy.

The importance of feature engineering in future machine learning optimization cannot be overstated. Feature importance analysis revealed that only a small subset of features significantly influences model performance. Ideally, feature importance should be more evenly distributed rather than dominated by a single variable accounting for over 70% of the weight. Moving forward, incorporating non-technical features—particularly quantitative measures of market sentiment and social media influence—could substantially enhance the predictive power of machine learning models in financial market applications.

In summary, machine learning algorithms already play a pivotal role in today's financial markets. To better support institutional and individual investors in making informed decisions, there remains substantial room for optimization. We will continue to explore and advance the application of machine learning in the field of quantitative finance.

References

[1] Banhi G, Gautam B. Gold price forecasting using ARIMA model. Journal of Advanced Management Science, 2016, 4(2): 117–121.

[2] Jim K S. Forecasting ETFs with machine learning algorithms. Johns Hopkins Carey Business School Version 1.3. 2017. http://etfprediction.pythonanywhere.com/.

[3] Perry S. A random forests approach to predicting clean energy stock prices. Journal of Risk and Financial Management, 2021, 14(2): 48.

[4] Stefano L. US funds dataset from Yahoo Finance. 2021-01-01 [2025-09-20]. https://www.kaggle.com/datasets/stefanoleone992/mutual-funds-and-etfs/data.

[5] Samraj G, Sanchal N, Nirmala P. Stock market time series forecasting using comparative machine learning algorithms. Procedia Computer Science, 2025, 252: 893–904.

[6] Zimo L, Weijia X, Aihua L. Research on multi-factor stock selection model based on LightGBM and Bayesian optimization. Procedia Computer Science, 2022, 214: 1234–1240.