Comparing the Performance of Four Regression Models in Predicting Stock Returns

Zimu Tang

Department of International Business and Trade, Capital University of Economics and Business, Beijing, 100070, China

32021030226@cueb.edu.cn

Abstract. In recent years, stock market investment has seen rapid growth, yet many investors may lack sufficient relevant knowledge. This article aims to help investors achieve higher returns by comparing the predictive results of several models. Using four regression models in ML algorithms, namely LightGBM, decision tree, XGBoost, and CatBoost, to predict the returns of 1500 Japanese stocks. By analyzing the RMSE and MAE, the errors are evaluated to assess the accuracy of the models. LightGBM and XGBoost are gradient boosting-based models offering high training speed and accuracy, suitable for large datasets. Decision trees are easy to interpret but prone to overfitting. CatBoost handles categorical variables seamlessly. Comparing RMSE and MAE, all models perform similarly, with XGBoost showing superior performance. This research contributes to stock market prediction by analyzing model strengths and weaknesses, offering insights for future research.

Keywords: Stock Returns Prediction; LGBM; Decision Tree; XGBoost; CatBoost

1 INTRODUCTION

Against the backdrop of rapid economic market development, the per capita income and the wariness of public investment and financial management increasing swiftly. Over the past few years, the stock as a tool of investment and financial management captured people’s attention.

Stock, as an indicator of economic development, plays an important role in the financial market. Indeed, while stocks are considered a potentially high-return investment, it is widely acknowledged to come with high risks at the same time. Therefore, making informed judgments about the stock market trends and understanding the changes in stock prices is crucial for investors, given the inherent risks involved. Therefore, it is crucial to use professional expertise to make reasonable predictions about the prices and trends of stocks [1]. The reasonable predictions significantly reduce investors’ investment risks. By incorporating predicted stock prices into their investment strategies, investors can maximize their investment returns.

A. Bhunia et al. (eds.), *Proceedings of the 2024 2nd International Conference on Finance, Trade and Business Management (FTBM 2024)*, Advances in Economics, Business and Management Research 304,
https://doi.org/10.2991/978-94-6463-546-1_4

In recent years, the enduring topic of stock return forecasting has become increasingly popular. With the development of data science, Machine-Learning (ML) regression models are being used to predict stock returns. ML models can automatically handle large amounts of data without the need for human intervention [2]. ML models exhibit strong robustness when faced with various types and scales of data, capable of handling complex data patterns and noise. Through continuous learning and optimization, ML models can gradually improve the accuracy and effectiveness of predictions, demonstrating a certain level of intelligence. Based on these properties, ML models are a rational approach for predicting stock prices.

There is some relative research that is helpful to choose suitable models in different areas to predict reference, and also others in stock price prediction research are helpful to have an overview of the prediction and discussion of the results. The LGBM model is also used in the construction field, especially in machinery operations. In the construction field, by adjusting the model, it can predict and support the control of machinery [3]. The decision tree model is widely applied in the medical field, particularly in COVID-19, to assist in the early prediction, diagnosis, and subsequent treatment of the disease [4]. XGBoost also has numerous applications in the field of health monitoring, such as utilizing specific methods for health detection and studying the interactions of various factors on health [5]. Another study utilizes the CatBoost model to predict financial distress, which plays a crucial role for banks and investors in making credit decisions [6]. The successful applications of these ML models in various fields in previous studies demonstrate their potential. This paper will investigate the applications of these four models in the stock market domain.

This article, which is based on historical data for a variety of Japanese stocks and options from January 2017 to December 2021, analyzes four different ML regression models and evaluates their predictive accuracy to identify the most accurate prediction model. The purpose is to help the i. The four regression models are LightGBM models, decision tree models, XGBoost models, and CatBoost models. The article first analyzes the errors of four models using Root Mean Squared Error (RMSE), and then proceeds with Mean Absolute Errors (MAE) analysis. This study aims to assist investors in selecting appropriate stock prediction models to enhance the accuracy of their investment decisions and achieve higher stock returns.

2 METHODOLOGY

2.1 LightGBM Model

The LightGBM model is an ML model based on the gradient boosting algorithm, which is widely used in time series data prediction. The LightGBM model utilizes a tree-based learning approach to predict future values by constructing multiple decision trees. It has efficient training speed and accurate prediction capability, particularly suitable for handling large-scale datasets and high-dimensional features. The LightGBM model offers adjustable parameters that can be fine-tuned to maximize model performance [7].

The training process of the LGBM model can be represented as follows:

$$\widehat{Y_l} = \sum_{K-1}^{K} wk \cdot I(xi\epsilon Rk) \tag{1}$$

Including:

- $\widehat{Y_\iota}$ indicates the predicted value of sample I;
- $K$is the number of trees;
- wk is the weight of the leaf node;
- I(xi∈Rk) is the indicating function that returns 1 if sample i belongs to leaf node Rk, otherwise.

The LGBM model optimizes the weights of the leaf nodes by minimizing the loss function, thereby obtaining the best model parameters. By training many decision trees and combining them, the LGBM model can effectively capture the complex relationships between data, achieving accurate time-series predictions.

2.2 Decision Tree Model

The decision tree model is a commonly used predictive model that constructs a tree-like structure based on a series of decision rules for classifying or predicting input data. The construction process begins with a root node and involves recursively partitioning the data, with each node representing a feature and each branch representing a feature value, until reaching leaf nodes, which correspond to prediction outcomes. During prediction, input data is traversed through the tree according to the decision rules starting from the root node until reaching a leaf node, where the prediction result is obtained. Decision tree models are simple and interpretable but can be prone to overfitting [8].

The decision tree model has the advantage of being easy to understand and interpretable, allowing for an intuitive representation of relationships between features. However, when dealing with time series data, the decision tree model may suffer from overfitting and poor generalization, especially when the data volume is low or the feature dimension is high. Therefore, when using the decision tree model for time series prediction, it is important to adjust the model parameters to improve its generalization ability and adopt appropriate feature engineering methods to enhance the model's performance

2.3 XGBoost Model

The XGBoost model is a high-performance ML model in predictive analytics. XGBoost adopts a tree-based learning strategy, where it builds a sequence of decision trees one after another, with each tree refining the mistakes of its predecessor. This process leads to the creation of a strong predictive model capable of capturing intricate patterns and relationships within the data.

XGBoost is a powerful machine-learning algorithm designed for structured or tabular data. It enhances both speed and performance through the implementation of gradient-boosted decision trees. Renowned for its scalability, parallelization, efficiency, and speed, XGBoost is extensively utilized for feature selection [9].

Furthermore, XGBoost offers flexibility in parameter tuning, allowing practitioners to optimize model performance based on specific requirements and constraints. Its scalable and distributed implementation makes it suitable for large-scale datasets, enabling efficient training and prediction even in resource-constrained environments.

2.4 Catboost Model

The CatBoost model is widely recognized for its effectiveness in time series data prediction and other tasks. Similar to XGBoost and LightGBM, CatBoost utilizes a gradient boosting framework but with a unique feature that handles categorical variables seamlessly without the need for extensive pre-processing [10].

One of the distinguishing features of CatBoost is its ability to automatically handle categorical features, which often pose challenges in traditional ML models. By employing an efficient algorithm for encoding categorical variables, CatBoost can effectively utilize categorical information during the training process, leading to improved predictive performance without the risk of data leakage.

Additionally, CatBoost incorporates advanced regularization techniques and robust loss functions to prevent overfitting and enhance model generalization. This guarantees that the model can efficiently capture intricate relationships and patterns within the data, all the while upholding superior performance when presented with new, unseen data.

Furthermore, CatBoost offers scalability and efficiency, making it suitable for handling large-scale datasets with high-dimensional features. It also provides flexibility in parameter tuning, allowing users to optimize model performance based on specific requirements and constraints.

2.5 Data

The data set is cited from the Kaggle which was published by Japan Exchange Group in 2022, this data set contains the historical data for a variety of Japanese stocks and options from 2017-01-04 to 2021-12-03 and shows the dimensions of 2332531×12, which means that the row of data is2332531and column is 12. This combination of data set and notebook is the latest study published in Kaggle, which cited the newest data on stock prices. In data exploration, this paper cited the notebook studied by the Competition Notebook. After loading the dataset and uploading it into the Python notebook and some simple data processing, it selected the training data after 2020-12-3 and handled the missing values.

The results show that the train data sit become 462000×15 and there are 0 for the Missing values in Target, which means no missing values exist in the data set.

Then, in the data preprocessing, checking for missing values is important.

2.6 Exploratory Data Analysis

After handling missing values, this study visualizes the returns of all stocks.

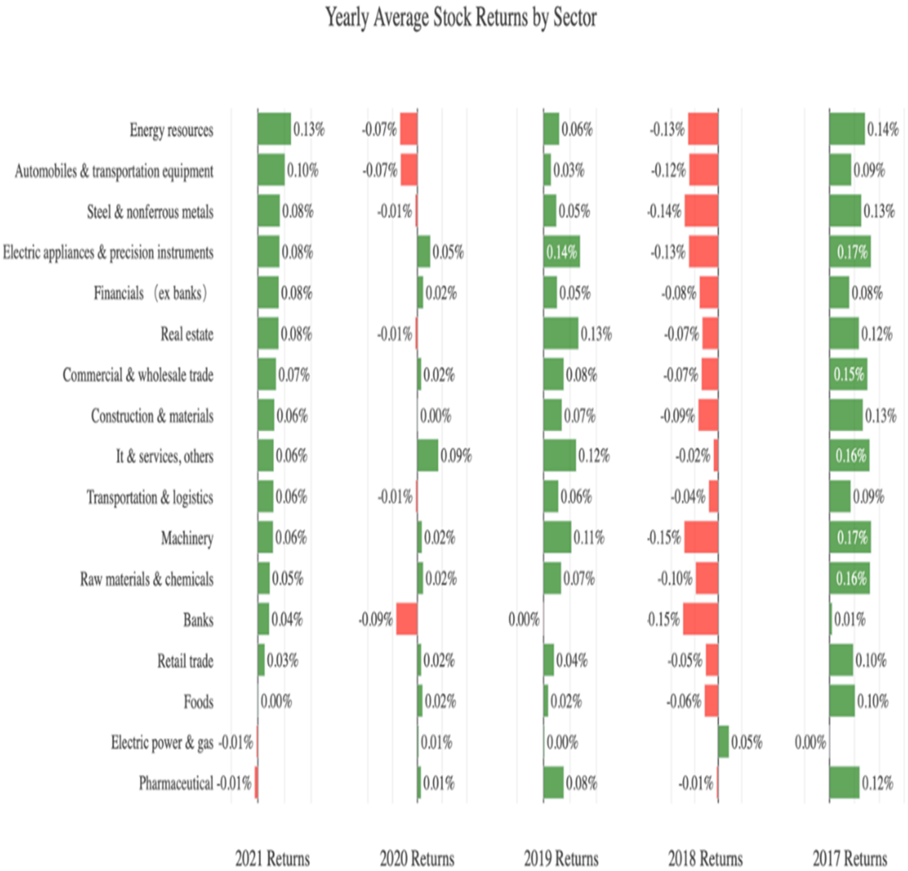

Fig. 1. Yearly stock returns by sector (Photo credit: Original).

Fig. 1 illustrates the average stock returns of various sectors each year. Each subplot represents a specific year and showcases the distribution of sector-wise returns through horizontal bar charts. Green bars denote positive returns, while red bars indicate negative returns. Hovering over each bar reveals the average return for that sector in the respective year. Figure 1 provides insights into the performance trends of different sectors across different years.

Fig. 2 illustrates the distribution of target data across different sectors, depicted using box plots. Each box plot represents a sector, where the height of the box plot indicates the range of target data for stocks in that sector, and the position of the median and quartiles depicts the central tendency and spread of the data. The color of the box plots varies with each sector, facilitating a visual comparison between different sectors.

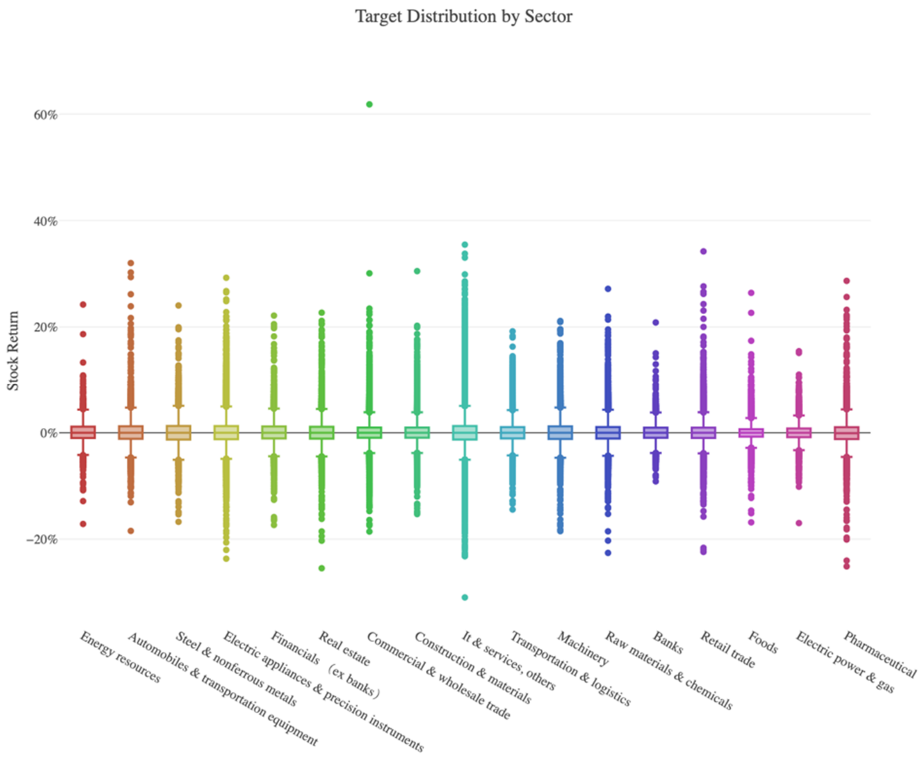

Fig. 2. Target distribution by sector (Photo credit: Original).

3 PREDICTION RESULTS

The RMSE is derived by computing the square root of the mean of squared variances between actual and predicted values. On the other hand, the MAE calculates the average of absolute differences between actual and predicted values. This assessment method is suitable for scenarios where the impact of outliers on evaluation needs to be minimized.

Table 1. RMSE values for the four models.

Models	RMSE										AVE
LGBM	0.0200	0.0230	0.0242	0.0253	0.0204	0.0309	0.0274	0.0236	0.0215	0.0222	0.0239
DT	0.0197	0.0228	0.0240	0.0252	0.0203	0.0304	0.0272	0.0235	0.0215	0.0221	0.0237
XGBoost	0.0196	0.0228	0.0240	0.0251	0.0202	0.0304	0.0272	0.0235	0.0214	0.0221	0.0236
CatBoost	0.0197	0.0228	0.0240	0.0251	0.0203	0.0206	0.0272	0.0235	0.0214	0.0222	0.0237

Table 2. MAE values for the four models.

Models	MAE										AVE
LGBM	0.0128	0.0153	0.0164	0.0169	0.0140	0.0202	0.0190	0.0161	0.0149	0.0151	0.0145
DT	0.0126	0.0151	0.0162	0.0168	0.0139	0.0200	0.0189	0.0160	0.0149	0.0151	0.0160
XGBoost	0.0125	0.0151	0.0162	0.0168	0.0139	0.0199	0.0189	0.0160	0.0149	0.0150	0.0124
CatBoost	0.0126	0.0152	0.0162	0.0168	0.0139	0.0201	0.0189	0.0160	0.0149	0.0151	0.0159

Table 1 displays the results of the four models after 10-fold cross-validation using RMSE, and Table 2 presents the outcomes of the four models following 10-fold cross-validation, utilizing MAE metrics correspondingly, along with their average values.

According to the table, the four regression methods have RMSE of approximately 0.0239, 0.0237, 0.0236, and 0.0237 respectively, which are quite similar. This evaluates the accuracy and adequacy of the models by quantifying the average squared difference between the model's predicted values and the actual observed values. Taking XGBoost as an example, the mean squared error is 0.0236. This means that, on average, its predictions deviate from the actual values by 0.0236 units. The variances among the other models are not substantial, yet all are greater compared to the XGBoost model.

The MAE of the four models vary widely, but according to the data, XGBoost remains close to the true values. MAE measures the average magnitude of the errors between predicted and actual values, using the same units as the data. It is insensitive to outliers as it only considers the absolute differences.

4 CONCLUSION

All four models have effectively predicted the trend of stock returns, Overall, this passage introduces four commonly used ML models for time series data prediction: LightGBM, decision trees, XGBoost, and CatBoost. LightGBM and XGBoost are both gradient-boosting-based models that predict future values by constructing multiple decision trees. These models offer higher training speed and accurate prediction capabilities, especially suitable for handling large-scale datasets and high-dimensional features. Decision tree models have the advantage of being easy to understand and interpret, but they may suffer from overfitting and poor generalization when dealing with time series data. CatBoost is a powerful gradient boosting-based ML algorithm similar to XGBoost and LightGBM, but it has a unique feature for handling categorical variables seamlessly without extensive preprocessing. Comparing the predicted results with the true values using both RMSE and MAE methods, all four models show relatively small differences, indicating good performance. Among them, XGBoost demonstrates superior performance with smaller errors compared to the other models.

While the four models used in this study may not provide more accurate predictions for more complex data, this research still makes significant contributions to the field of stock market prediction. Through comparing and analyzing different ML models, this paper delves into the strengths and weaknesses of these models in handling stock data, providing valuable insights and guidance for future research and serving as a reference for constructing better predictive models.

REFERENCES

1. Avramov, D., Chordia, T.: Predicting stock returns. Journal of Financial Economics, 82(2), 387-415 (2006).
2. Jo, T.: Machine learning foundations. Machine Learning Foundations. Springer Nature Switzerland AG (2023).
3. Jabeur, S. B., Gharib, C., Mefteh-Wali, S., et al.: CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658 (2021).
4. Chen, H., Li, X., Feng, Z., et al.: Shield attitude prediction based on Bayesian-LGBM ML. Information Sciences, 632, 105-129 (2023).
5. Dong, W., Huang, Y., Lehane, B., et al.: XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Automation in Construction, 114, 103155 (2020).
6. Huyut, M. T., Üstündağ, H.: Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees ML model: a retrospective observational study. Medical gas research, 12(2), 60 (2022).
7. Abou Omar, K. B.: XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. Preprint Semester Project (2018).
8. Veen, A., Kaur, R., Kaur, K.: Apply ML Techniques to Predict Stock Market Dynamics and Trends (2024).
9. Vuong, P. H., Dat, T. T., Mai, T. K., et al.: Stock-price forecasting based on XGBoost and LSTM. Computer Systems Science & Engineering, 40(1) (2021).
10. Huang, G., Wu, L., Ma, X., et al.: Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. Journal of Hydrology, 574, 1029-1041 (2021).