Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/52/20230758

Predicting the Direction of Stock Price Movement with Machine Learning Algorithms

Minghan Yang[1,a,*]

[1]*Department of Mathematics, University College London, London, The United Kingdom*
a. Minghan.yang.20@ucl.ac.uk
**corresponding author*

Abstract: The application of machine learning algorithms in predicting stock price directional movements has been a widely discussed problem. Due to the chaotic, uncertain, and dynamic characteristics of stock markets, relatively accurate predictions could contribute to financial benefits and risk reduction. This paper aims to assess the prediction performance of five well-known machine learning models, which are Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, and Extreme Gradient Boosting. Data in this study covers the period from 2020-01-02 to 2023-07-03, and assets studied in this paper are Nike, Amazon, Microsoft, Tesco, and Airbnb. After feature selection, data pre-processing, cross-validation, model selection, and evaluation, it is concluded that the Random Forest Classifier tends to perform better in directional predictions, as it demonstrates higher accuracy and precision. This research highlights the application of machine learning algorithms in financial area, especially in the stock price prediction.

Keywords: stock price movement, prediction, machine learning algorithms

1. Introduction

Predicting the stock price movements has always been a widely discussed problem, as predictions could offer insights for decision-making, risk management, and optimising investment returns. However, the financial market represents a multifaceted, dynamic, chaotic, constantly evolving, and nonlinear pattern [1]. Stock market movements are strongly linked to economic environments, policies, and expectations from traders, as examples [2]. Hence it is argued that even a slight improvement in prediction accuracy could lead to considerable benefits and reduce the risk in investment [1,3].

Numerous studies on statistical time series forecasting methods have been discussed to analyze and predict the stock price movements, such as the autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroskedasticity (GARCH) models, however, these models are constructed under the assumption of stationarity and lack the ability to react quickly to the changes in temporary stock price movements, which may result in less satisfactory performance [3]. Some studies also combine traditional statistical models such as ARIMA with Neural Networks, which is said to improve prediction performance [4,5].

The applications of machine learning algorithms in stock price prediction have been studied using a wide selection of methods and models, aiming to improve capacity, ability, and prediction performance. Most of the machine learning algorithms considered in stock price direction forecasting

DOI: 10.54254/2754-1169/52/20230758

are Support Vector Machines, Logistic Regression, Random Forest, Boosting, and Neural Networks [1]. Some popular tree-based ensemble machine learning models have also been studied. In one study, it is suggested that AdaBoost outperforms other classifiers, such as Extra Trees, Base Classifier, Random Forest, XGBoost, and Bagging [6]. Some other research suggested that Support Vector Machines produce decent performance [7]. However, the best-performing models and algorithms for the most recent stock price data have not been explored, and the selected assets in this paper were not compared and combined in previous studies.

Hence, this paper evaluates prediction performances of five popular machine learning models basing on the most recent stock market information, with a focus on stock price directional movements prediction.

In the following parts of this paper, data description, feature selection, and preprocessing will be discussed. After that, the five machine learning algorithms will be introduced in the methodology section. Finally, the methods will be evaluated based on prediction performance, and conclusions will be drawn.

2. Data

2.1. Data Description

This study collected stock price data from Yahoo Finance in five different industries. Most of the five-time series cover the period from 01/01/2020 to 03/07/2023 and include Open-High-Low-Close-Volume (OHLCV) information about the five assets. Descriptions of the five time series are presented in Table 1.

Table 1: Descriptions of five stock price series.

Data set	Industry	Time Frame	Number of Observation
NKE	Manufacturing	2020-01-02 to 2023-07-03	881
AMZN	Specialty Distribution	2020-01-02 to 2023-07-03	881
MSFT	Technology	2020-01-02 to 2023-07-03	881
TSCO.L	Retailer	2020-01-02 to 2023-07-03	881
ANBN	Hotels/Resorts	2020-12-10 to 2023-07-03	643

In this paper, the target variable is the closing price directional movement indicator, calculated as follows:

$$Target = \begin{cases} 1, \ close\ price\ difference > 0, i.e. increase \\ 0, \ close\ price\ difference < 0, i.e. decrease \end{cases} \quad (1)$$

2.2. Feature Extraction and Selection

Feature extraction for machine learning methods hold great importance as the movements of stock prices are affected by various factors, for example historic price movements, periodic patterns, and historic market behaviors [8]. Apart from the basic features of OHLCV stock price features, some widely used technical features include Relative Strength Index (RSI), On Balance Volume (OBV), Moving Average Convergence Divergence (MACD), among others [9]. Based on prior research and common features to measure stock price movement, features including SMA, RSI, MACD, OBV, STDEV, H-L, C-O are constructed by using OHLCV data obtained, as shown in Table 2 [10,11]. The

DOI: 10.54254/2754-1169/52/20230758

features are normalized to assist modelling and reduce the impact of inconsistent scales of the feature values.

Table 2: Features extracted.

H-L	Difference between daily highest price and lowest price
C-O	Difference between daily close price and open price
RSI	A momentum oscillator to evaluate the velocity of stock price movement
SMA	Average of the closing price in the past n days
MACD	A momentum indicator that compares
STDEV	Rolling standard deviation
OBV	On balance volume. A cumulative indicator that estimate the change of price by using the change of volume.

In previous studies, multiple methods for feature selection and dimension reduction were introduced. Correlation is one of the most common and straightforward criteria for feature selection as it measures the relationship between features and the target variable [8,12]. Based on the correlation matrix computed for each of the five stocks, the selected features are shown in Table 3.

Table 3: Features selected for each stock price time series.

Data set	Feature selected based on correlation matrix
NKE	Volume, H-L, C-O, RSI_14, MACDh_12_26_9, STDEV_5, OBV, SMA_50
AMZN	Volume, H-L, C-O, RSI_14, MACDh_12_26_9, STDEV_5, OBV, SMA_100
MSFT	Volume, H-L, C-O, RSI_14, MACDh_12_26_9, STDEV_5, OBV
TSCO.L	Adj Close, Volume, H-L, C-O, RSI_14, MACDh_12_26_9, STDEV_5, OBV, SMA_100
ANBN	Volume, H-L, C-O, RSI_14, MACDh_12_26_9, STDEV_5, SMA_50

The training sets for each time series in this paper are set to cover the period from 2020-01-02 to 2022-12-31, and the test sets include data from 2023-01-01 to 2023-07-03. A five-fold time series cross-validation is conducted to each of the five training sets to select the best models. After that, the models will be evaluated based on accuracy, precision, and recall, which are obtained by comparing the predictions with the test sets.

3. Methodology

In this paper, five machine learning algorithms are considered: Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Extreme Gradient Boosting.

3.1. Logistic Regression Classifier (LGC)

LRC is a classification algorithm utilized in the field of machine learning and statistics, particularly advantageous for binary classification problems. Rooted in the concept of probability, it applies a logistic function to a linear combination of features to estimate the probability of an event. In essence, it models the relationship between a set of independent variables and a binary dependent variable. Its simplicity, interpretability, and computation efficiency make it a widely used approach in many predictive modeling scenarios (See Figure 1).

DOI: 10.54254/2754-1169/52/20230758

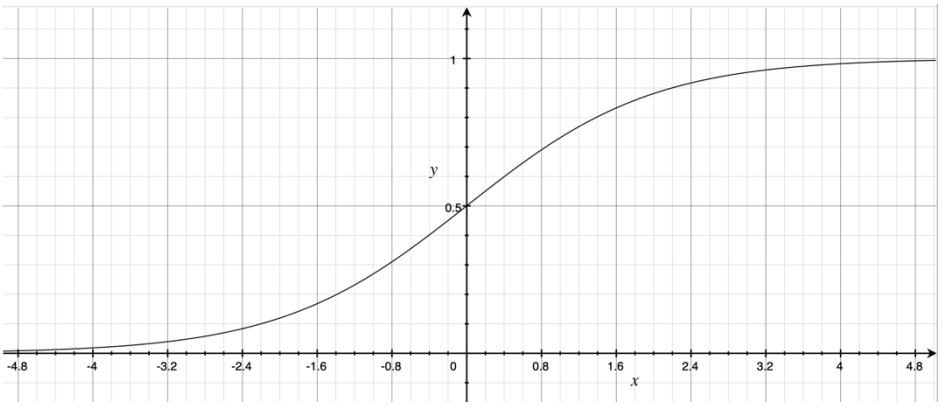

Figure 1: Logistic Regression (Picture credit: Original).

3.2. Support Vector Classifier (SVC)

SVC is another powerful supervised learning model used for classification tasks. It constructs an N-dimensional hyperplane in a transformed problem space to create the optimal decision boundary between different classes, with the goal of maximize the distance between the nearest data points (support vectors) of each class to the decision boundary. Its inherent abilities to manage data of higher dimensions, generating non-linear boundaries using kernel functions, and preventing overfitting make it an invaluable tool in complex classification problems, ranging from image recognition to biological and medical applications. The interpretability and versatility of Support Vector Classifier make it a well-regarded tool in machine learning (See Figure 2).

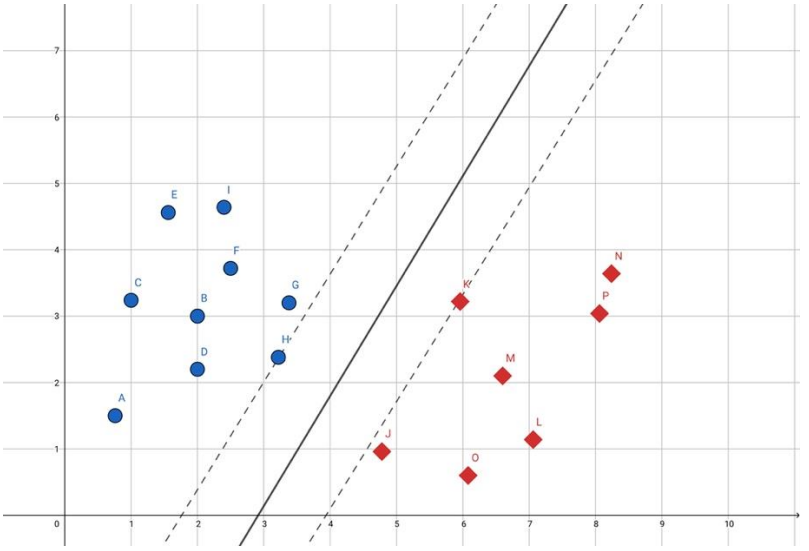

Figure 2: Support Vector Machines (Picture credit: Original).

3.3. Decision Tree Classifier (DTC)

Decision Tree Classifier operates by dividing the feature space into a set of rules, which can be visually represented in tree-like structure of decisions. Every node in the tree stands for a decision based on the value of an input feature. The final decision or prediction is determined by the majority vote or average of the leaf nodes reached by the input data. With the capability to handle both categorical and numerical data, decision trees excel in exploratory knowledge discovery. The

DOI: 10.54254/2754-1169/52/20230758

Decision Tree Classifier, with its interpretability and straightforward implementation, remains a steadfast tool in predictive analytics and machine learning (See Figure 3).

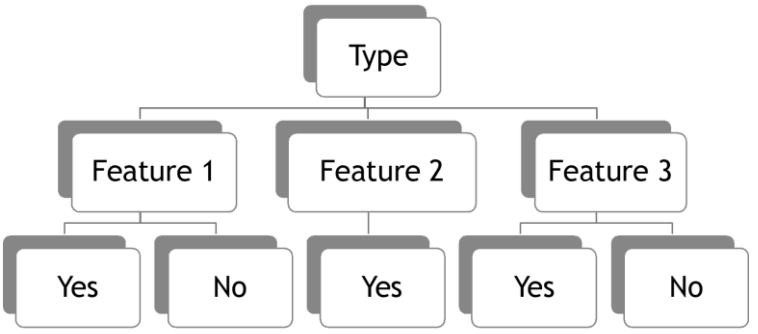

Figure 3: Decision Tree (Picture credit: Original).

3.4. Random Forest Classifier (RFC)

RFC harnesses the power of several decision trees, which are based on randomly chosen subsets of the data and features, and aggregates their outputs, typically through a majority voting system. This method helps mitigate the risk of overfitting and boosts model generalization by introducing randomness into the model creation process and combining the predictions of individually weak, but collectively strong, decision trees. Consequently, the Random Forest Classifier demonstrates exceptional performance across diverse domains, by effectively handling large datasets with high dimensionality. With its inherent feature importance evaluation capability and its robustness to outliers and non-linear data, it constitutes a pivotal component of machine learning (See Figure 4).

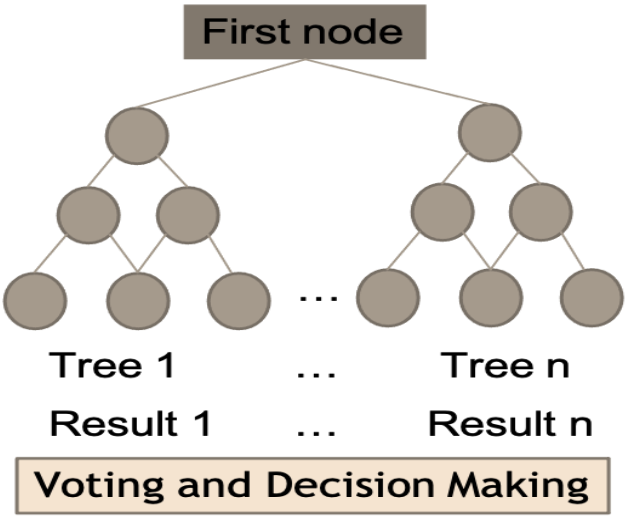

Figure 4: Random Forest (Picture credit: Original).

3.5. Extreme Gradient Boosting Classifier (XGBC)

XGBC is based on the gradient boosting framework. It capitalizes on the principle of building a collection of predictive models sequentially, which are typically decision trees. Each subsequent model in the sequence learns from the errors of its predecessor, aiming to minimize the overall prediction error. XGBoost includes regularization parameters to prevent overfitting, making it a standout in terms of both accuracy and efficiency (See Figure 5).

DOI: 10.54254/2754-1169/52/20230758

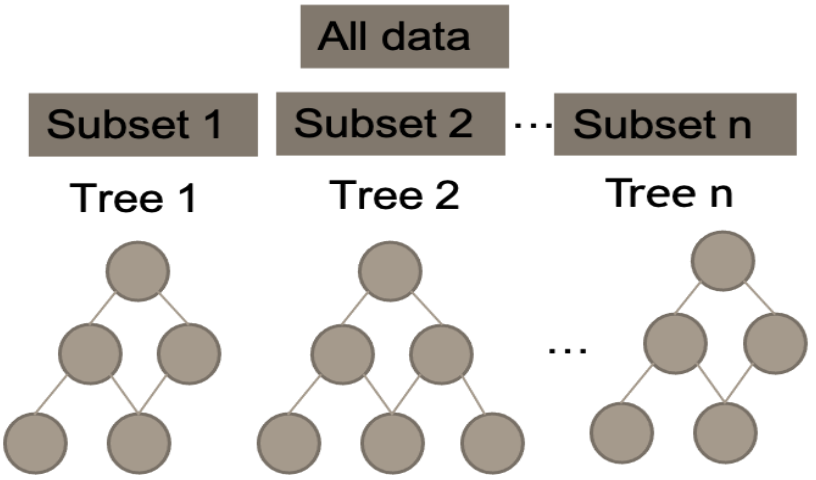

Figure 5: Extreme Gradient Boosting (Picture credit: Original).

4. Results

The model performance is evaluated by three factors: accuracy, precision, and recall. For Random Forest and XGBoost, each of the accuracy, precision and recall values were calculated by taking averages of resulting values from running models 10 times.

4.1. Accuracy

Accuracy measures the ability to correctly predict the decreases and increases, and is calculated by the following formula:

$$accuracy = \frac{correct\ predictions}{all\ predictions} \tag{2}$$

The accuracy scores of the five algorithms for each of the five stocks are presented in Table 4 below.

Table 4: Accuracy values.

Accuracy	Nike	Amazon	Microsoft	Tesco	Airbnb
LRC	0.512	0.592	0.528	0.603	**0.544**
SVC	0.504	0.584	**0.600**	0.571	0.520
DTC	0.552	0.544	0.528	0.595	0.520
RFC	**0.554**	**0.616**	0.592	**0.611**	0.524
XGBC	0.546	0.552	0.544	0.588	0.496

In terms of accuracy, Random Forest Classifier appears to produce a generally better performance than other algorithms, as it achieves higher accuracy values in predicting 3 out of 5 stocks.

4.2. Precision

Precision is evaluated by the following formula:

$$precision = \frac{tp}{tp+fp} \tag{3}$$

DOI: 10.54254/2754-1169/52/20230758

where tp is the true positive, and fp is the false positive. The precision values measure the proportion of true increases among all the predicted increases. Hence higher precision could represent a lower likelihood of false investment choices based on the incorrect predictions of price increases, thereby reducing the risk of losses due to erroneous investments.

From Table 5, it is evident that the Random Forest Classifier generally demonstrates decent performance, achieving the highest precision values in predicting the stock price movements of Amazon, Tesco, and Airbnb. In contrast, the Decision Tree Classifier seems to yield the poorest prediction performance. The highest precision value, 0.71, is observed when forecasting the direction of Tesco's stock prices using the Random Forest Classifier.

Table 5: Precision values.

Precision	Nike	Amazon	Microsoft	Tesco	Airbnb
LRC	0.49	0.59	0.52	0.63	**0.59**
SVC	0.48	0.57	**0.60**	0.64	0.56
DTC	0.50	0.57	0.53	0.60	0.58
RFC	0.50	**0.65**	0.57	**0.71**	**0.59**
XGBC	**0.53**	0.58	0.54	0.67	0.55

4.3. Recall

Recall values are calculated by using the following formula:

$$recall = \frac{tp}{tp+fn} \tag{4}$$

where fn is the false negative. It calculates the proportion of correctly forecasted increases out of the total number of actual increases. Higher recall values indicate a higher ability of capturing the increase in stock prices, and a lower likelihood of missing investment opportunities.

From table 6, it is noticeable that Support Vector Classifier seems to generate an overall better performance, by achieving highest recall values when predicting Nike, Amazon and Airbnb stock price movements. The highest precision value takes place when predicting the Amazon stock price movement direction by using Support Vector Classifier.

Table 6: Recall values.

Recall	Nike	Amazon	Microsoft	Tesco	Airbnb
LRC	0.87	0.87	**0.83**	0.63	0.61
SVC	**0.96**	**0.97**	0.67	0.48	**0.71**
DTC	0.93	0.72	0.73	**0.72**	0.64
RFC	0.69	0.67	0.80	0.47	0.59
XGBC	0.60	0.65	0.73	0.48	0.59

It is highlighted that a high recall value and a high precision value cannot be obtained at the same time in any case. In this paper, precision performance is prioritized over recall performance, as investors often seek to reduce risk rather than missing potential profit opportunities. In light of this, the Random Forest Classifier is likely to outperform other algorithms, which agrees with the findings from previous studies [1,13].

DOI: 10.54254/2754-1169/52/20230758

5. Discussion

In order to produce a more inclusive and general evaluation of the models' performances, it is recommended to consider a broader range of assets in the study. In a study evaluating multiple classifiers for stock price movements, stock price data of 5767 companies were considered in the study, covering 19 industries [1]. Including a larger and more diverse set of assets would allow for a more generalized evaluation and comparison of model performance, even though the assets considered in the study do not encompass the entire selection of all stocks in the financial market. Additionally, it could be advantageous to incorporate other financial indicators such as liquidity, solvency, profitability indicators, and general features such as public debt, GDP, unemployment, and trade balance, into the models [1].

One limitation of this study is that it considered only five most common machine learning algorithms. Hence, in future work, more advanced algorithms such as Neural Networks, AdaBoosting, and other deep learning models could be included in comparison and evaluation.

6. Conclusion

This project evaluates the performances of five machine learning algorithms in predicting stock price directional movements. Five assets, NKE, AMZN, MSFT, TSCO.L, and ABNB are included in this study. Each time series, obtained from yahoo finance, contains OHLCV data within period 2020/01/02 – 2023/07/03. Technical features such as RSI and MACD are constructed and selected based on correlation matrix. After normalization, each time series is split into training and test sets. Then five-fold cross validation is applied to the training set to evaluate the models. Following that, the models are evaluated using accuracy, precision, and recall on the test set. The results indicate that the Random Forest Classifier provides superior prediction performance on the test set compared to the other models.

References

[1] *Ballings, M., Van den Poel, D., Hespeels, N., and Gryp, R. (2015) Evaluating multiple classifiers for stock price direction prediction. Expert systems with Applications, 42(20), 7046-7056.*

[2] *Kumar, M., and Thenmozhi, M. (2006) Forecasting stock index movement: A comparison of support vector machines and random forest. In Indian institute of capital markets 9th capital markets conference paper.*

[3] *Timmermann, A., and Granger, C. W. (2004) Efficient market hypothesis and forecasting. International Journal of forecasting, 20(1), 15-27..*

[4] *Khandelwal, I., Adhikari, R., and Verma, G. (2015) Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. Procedia Computer Science, 48, 173-179.*

[5] *Zhang, G. P. (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175*

[6] *Ampomah, E. K., Qin, Z., and Nyame, G. (2020) Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information, 11(6), 332.*

[7] *Lee, M. C. (2009) Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 36(8), 10896-10904.*

[8] *Htun, H. H., Biehl, M., & Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. Financial Innovation, 9(1), 26.*

[9] *Alsubaie, Y., El Hindi, K., and Alsalman, H. (2019) Cost-sensitive prediction of stock price direction: Selection of technical indicators. IEEE Access, 7, 146876-146892.*

[10] *Khalid, S., Khalil, T., and Nasreen, S. (2014) A survey of feature selection and feature extraction techniques in machine learning. In 2014 science and information conference, 372-378.*

[11] *Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., and Mosavi, A. (2020) Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access, 8, 150199-150212.*

[12] *Nabi, R. M., Saeed, S. A. M., Harron, H. B., and Fujita, H. (2019) Ultimate prediction of stock market price movement. Journal of Computer Science, 15, 1795.*

DOI: 10.54254/2754-1169/52/20230758

[13] Patel, J., Shah, S., Thakkar, P., and Kotecha, K. (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.