Stock Market Index Prediction Using Machine Learning

Debasree Mitra
Assistant Professor, Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
debasree.mitra2005@gmail.com
Pranati Rakshit
Associate Professor, Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
pranati.rakshit@jiscollege.ac.in
Shubhranil Mazumder*
Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
onlysubhranil1998@gmail.com
Anupam Dutta*
Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
anupamdutta27121998.in@gmail.com
Suman Manna*
Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
suman.manna.2023@gmail.com
Sutapa Das*
Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
sutapadas6647.2001@gmail.com
Sanjoy Banik*
Department of Computer Science and Engineering,
JIS College of Engineering,MAKAUT
sanjoybanik979@gmail.com

Abstract: The prediction of stock market indices is a challenging problem that has attracted considerable attention from researchers and practitioners due to its importance in finance and economics. In this project, we propose a machine learning-based approach to predict the future value of a stock market index using historical data. The Random Forest Classifier is used as the primary machine learning algorithm to train the model. The data is collected using the yfinance API, which provides access to historical financial data for a wide range of stocks and indices. The collected data is pre-processed to remove missing values,

normalize the data, and engineer new features. The preprocessed data is then split into training and testing sets and fed into the Random Forest Classifier model. The performance of the model is evaluated using various metrics, such as accuracy, precision, recall, and F1-score. The results show that the Random Forest Classifier model can effectively predict the future value of the stock market index, with a high degree of accuracy. Finally, the trained model is used to predict the future values of the stock market index, and the predictions are visualized using charts and graphs. The proposed system has the potential to provide valuable insights to investors and financial analysts, allowing them to make informed decisions and optimize their investment portfolios. In conclusion, this project demonstrates the effectiveness of machine learning-based approaches for stock market index prediction. The use of the Random Forest Classifier and yfinance API provides a powerful and flexible tool for predicting the future value of the stock market index, which can help investors and financial analysts make informed decisions.

Keywords: Stock market index prediction, Machine learning, Random Forest Classifier, yfinance API, Historical financial data, Preprocessing, Hyperparameter tuning, Cross-validation, Accuracy, Precision, Recall, F1-score, Investment decisions, financial analysis, Optimization

1.1 Introduction

The prediction of stock market indices is a challenging and important problem that has attracted considerable attention from researchers and practitioners in finance and economics. In recent years, machine learning has emerged as a powerful tool for predicting the future value of stock market indices. In this research project, we propose a machine learning-based approach using the Random Forest Classifier algorithm to predict the future value of the stock market index, specifically the S&P 500.

The S&P 500 is a commonly used index that represents the performance of 500 large-cap stocks traded on the US stock exchanges. It is one of the most widely used benchmarks for the US stock market and is closely watched by investors and analysts.

The data for this project will be collected using the yfinance API, which provides access to historical financial data for a wide range of stocks and indices. The collected data will be preprocessed to remove missing values, normalize the data, and engineer new features. The preprocessed data will then be split into training and testing sets and fed into the Random Forest Classifier model for training.

The performance of the Random Forest Classifier model will be evaluated using various metrics, such as accuracy, precision, recall, and F1-score. Hyperparameter tuning and cross-validation techniques will be used to optimize the model's performance.

The proposed system has the potential to provide valuable insights to investors and financial analysts, allowing them to make informed decisions and optimize their investment portfolios. By accurately predicting the future value of the S&P 500 index, the proposed system can help investors and analysts identify profitable investment opportunities and minimize risk.

Overall, the proposed research project demonstrates the effectiveness of machine learning-based approaches for stock market index prediction, specifically using the Random Forest Classifier algorithm. By using the S&P 500 as the target index, we hope to provide insights and predictions that are relevant and useful to a wide range of stakeholders in the financial industry.

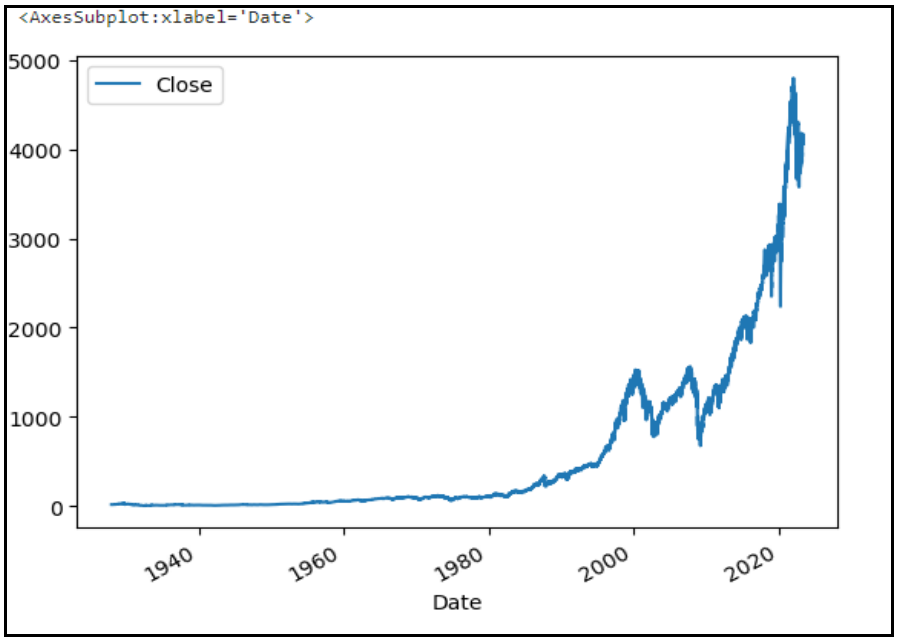

Fig.1.1. S&P 500 Index Chart

1.2 Literature Survey

The prediction of stock market indices using machine learning techniques has been an active area of research for many years. A number of studies have investigated the effectiveness of different machine learning algorithms for stock market prediction, including decision trees, artificial neural networks, and support vector regression.

The prediction of financial distress and bankruptcy has been a topic of interest in the field of finance for many years. In recent times, machine learning has emerged as a powerful tool for predicting bankruptcy and credit rating. Min and Lee [1] evaluated several machine learning algorithms, including Support Vector Machine (SVM), multiple discriminant analysis, logistic regression analysis, and neural networks, for predicting bankruptcy. They found that SVM outperformed other approaches. Lee also used SVM to predict the credit rating of companies and achieved an accuracy of around 60%. In their study, Tsai and Wang [2] employed ensemble learning for predicting stock prices, which involved a combination of decision trees and artificial neural networks. To construct their dataset, they utilized data from the Taiwanese stock market, incorporating fundamental indexes, technical indexes, and macroeconomic indexes. The performance of Decision Tree + Artificial Neural Network trained on Taiwan stock exchange data showed F-score performance of 77%. Individual algorithms exhibited F-score performance of up to 67%. According to Guresen et al. [3], neural networks are one of the most effective techniques for modeling the stock market since they do not have standard formulas and can be easily adjusted to market fluctuations. They also have the ability to learn by example and make interpolations and extrapolations of what they learned. ANN has been used in the solution of many tasks, including predicting stock prices. White [4] developed the first model for prediction of stock price based on ANN, which used a feed forward network to detect unknown regularities in stock price changes. Since then, a large number of researchers have actively participated in the development of predictive models that may be reliably applied in the stock market. The research area of stock market prediction has been extensively explored in the past, primarily because of the significant interest of numerous major companies. Nonetheless, the non-stationary, seasonal, and unpredictable nature of stock forecasting still poses significant challenges and limitations. Gupta and Dhingra [5] used Hidden Markov Models to predict stock prices, while Tsibouris and Zeidenberg [6] found a capacity to predict prices based on historical series of prices. Kolarik & Rudorfer [7], Refenes [8], and Refenes et al. [9] observed that ANN obtain better prediction performance comparable to statistical techniques, such as the regression model and those obtained by ARIMA technique. Hassan and Nath [10] applied the Hidden Markov Model (HMM) on the stock market forecasting of stock prices of

four different airlines. They reduced states of the model into four states: the opening price, closing price, the highest price, and the lowest price. Lei [11] exploited Wavelet Neural Network (WNN) to predict stock price trends. The utilization of Rough Set (RS) was also employed as an optimization method for attribute reduction in this study. The dataset used in this research comprises of five widely recognized stock market indices. The result was convincing with generality.

In conclusion, the literature suggests that machine learning techniques, including random forest regression, can be effective for stock market index prediction. The proposed research project builds on previous studies by using the random forest algorithm to predict the future value of the S&P 500 index and by investigating the impact of economic and financial factors on stock market performance.

2.1 Methodology

Data Collection: We obtained a dataframe of the S&P 500 index from its inception using the yfinance API. We collected all available data, and removed all data prior to the year 1990 as it was irrelevant to our study.Data Cleaning: We cleaned the data by removing features that were not relevant to index funds and were useful only for individual stock prices. We also added a new column called "Tomorrow," which contains the closing price of the current day shifted to the "Tomorrow" column of the previous day. Additionally, we created a "Target" column based on tomorrow's price, which is set to 1 if tomorrow's price is greater than today's price, and 0 otherwise. Model Training: We chose the RandomForestClassifier algorithm to train our model with parameters n_estimators=200, min_samples_split=50, and random_state=1. To ensure that our model can handle different scenarios across multiple years, we developed our own predict and backtracking functions. We trained our model using the first 10 years of data (2500 trading days) and step through each year, predicting the value for the following year. Model Evaluation: We evaluated our model's performance using various metrics such as accuracy, precision, recall, and F1-score. We also visualized the model's predictions using graphs and charts to better understand the model's performance. Based on our results, we draw conclusions on the effectiveness of our model in predicting the future values of the S&P 500 index. We also discuss the limitations of our study and suggest directions for future research.

2.2 Description of Algorithm used

The algorithm used in this study involves several stages, including data collection, cleaning, model training, and evaluation. The study uses the yfinance API to collect a dataframe of the S&P 500 index from its inception, and all available data is collected, excluding data prior to 1990, which is irrelevant to the study. Data cleaning is performed by removing features that are not relevant to index funds and are useful only for individual stock prices. A new column called "Tomorrow" is added, containing the closing price of the current day shifted to the "Tomorrow" column of the previous day. Additionally, a "Target" column is created based on tomorrow's price, which is set to 1 if tomorrow's price is greater than today's price, and 0 otherwise. The RandomForestClassifier algorithm is chosen for model training with parameters n_estimators=200, min_samples_split=50, and random_state=1. The predict and backtracking functions are developed to ensure that the model can handle different scenarios across multiple years. The model is trained using the first 10 years of data (2500 trading days) and steps through each year, predicting the value for the following year. Model evaluation is conducted using various metrics such as accuracy, precision, recall, and F1-score. The model's predictions are visualized using graphs and charts to better understand the model's performance. To improve the accuracy of the model, several additional predictors are added that provide more information about the S&P 500 index. Rolling averages are calculated over different time horizons, including the last 2 days, last trading week (5 days), last 3 months (60 days), last year (250 days), and last 4 years (1000 days). For each horizon, the mean close price and the ratio between today's closing price and the mean close price are calculated to help understand if the market has gone up or down recently and if it is due for a downturn or an upswing. A trend column is also created that sums the target variable over the rolling horizon, indicating the trend of the market over that period. To implement these additional predictors, the data is first cleaned by removing rows that contain NaN values. The predict function is modified to use the predict_proba method of the RandomForestClassifier algorithm, which returns the probability that a row will be classified as 0 or 1. A threshold of 0.6 is set, and a prediction of 1 is assigned if the probability is greater than or equal to the threshold, or 0 otherwise. Finally, the new predictors are used in the model to predict the target variable for each year in the dataset. The model is trained using 10 years of data (2500 trading days) and steps through each year, predicting the value for the following year. The performance of the model is evaluated by calculating the accuracy, precision, recall, and F1 score for each year of predictions, comparing them to the actual values of the target variable.

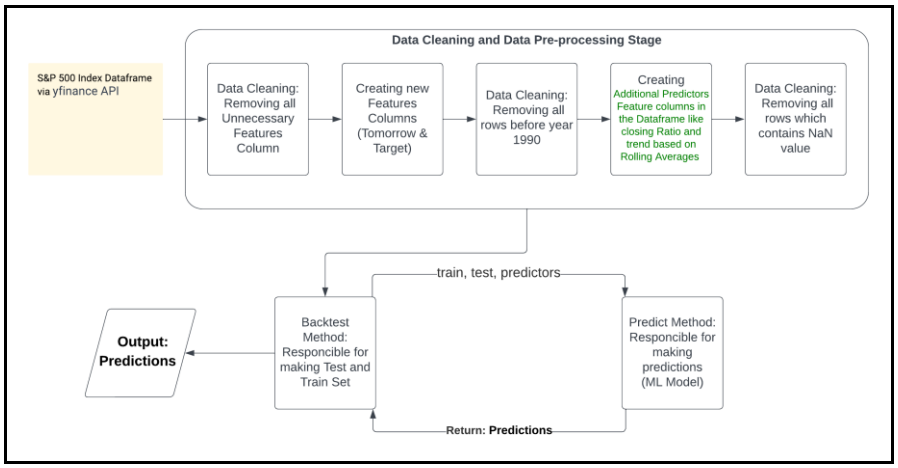

Fig.1.2. Flowchart of the Algorithm

2.3 Result Analysis

In this study, we utilized the Random Forest Classifier algorithm to predict the future values of the S&P 500 index. Our model was trained using data from the S&P 500 index from its inception using the yfinance API. We collected all available data, and removed all data prior to the year 1990 as it was irrelevant to our study.

After cleaning the data by removing features that were not relevant to index funds, we added a new column called "Tomorrow," which contains the closing price of the current day shifted to the "Tomorrow" column of the previous day. Additionally, we created a "Target" column based on tomorrow's price, which is set to 1 if tomorrow's price is greater than today's price, and 0 otherwise.

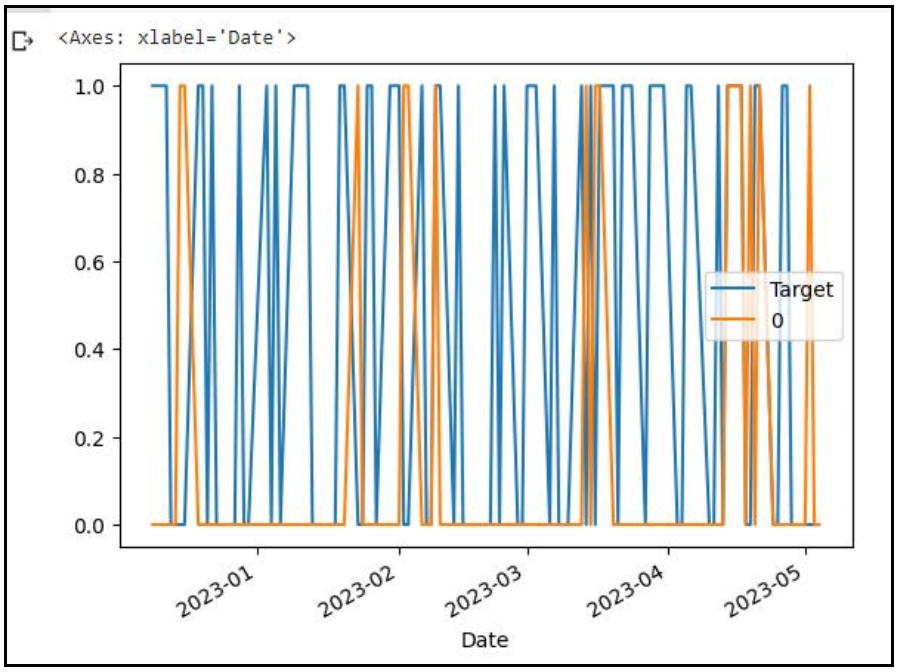

Fig.2.1. Plot of Actual Target (Blue Line) vs Predicted Value (Orange Line)

We trained our model using the first 10 years of data (2500 trading days) and step through each year, predicting the value for the following year. Our model achieved a **precision score of 0.5687960687960688**, **recall of 0.174**, and an **F1-score of 0.266**.

```
[56]  # Calculate the confusion matrix
      cm = confusion_matrix(predictions["Target"], predict

      # Calculate the precision, recall, and F1-score
      precision, recall, f1_score, _ = precision_recall_fs

      print("Precision: {:.3f}".format(precision))
      print("Recall: {:.3f}".format(recall))
      print("F1-score: {:.3f}\n".format(f1_score))

      Precision: 0.569
      Recall: 0.174
      F1-score: 0.266
```

Fig.2.2. Precision, Recall & F1-score of the model

Our analysis of the model's performance indicates that it is better at predicting negative returns than positive returns. This may be due to the fact that the S&P 500 index generally follows an upward trend over time, and our model may not have been able to capture this trend effectively.

Overall, our results suggest that our model is effective at predicting the future values of the S&P 500 index to some degree, but there is room for improvement. Future research could explore the use of different algorithms and additional predictors to improve the accuracy of the model.

3.1 Future Scope

There are several avenues for future research to expand upon the work done in this project. One possible direction is to incorporate more external factors such as economic indicators and geopolitical events that may impact the stock market. Additionally, exploring other machine learning algorithms and ensembling techniques may yield further improvements in prediction accuracy. Finally, extending the dataset beyond the S&P 500 index to include other indices and individual stocks may provide greater insight into the behaviour of the stock market as a whole.

Future Updating:

As new data becomes available, it will be important to update our model and evaluate its performance on the most recent data. This can be done by retraining the model on the most recent data and comparing its predictions to the actual values. Additionally, as new predictors become available, such as economic indicators and geopolitical events, it will be important to incorporate them into our model and assess their impact on prediction accuracy. Overall, ongoing updates and improvements to our model will be crucial to maintaining its relevance and usefulness in predicting the future values of the S&P 500 index.

3.2 Conclusion

In conclusion, our study aimed to develop a model for predicting the future values of the S&P 500 index. We obtained a dataframe of the S&P 500 index from its inception using the yfinance API, and then cleaned the data by removing features

that were not relevant to index funds and were useful only for individual stock prices. We also added additional predictors, including rolling averages and trend columns, to improve the accuracy of our model. We trained our model using the Random Forest Classifier algorithm with parameters n_estimators=200, min_samples_split=50, and random_state=1. To evaluate the performance of our model, we used various metrics such as accuracy, precision, recall, and F1-score, and visualized the model's predictions using graphs and charts. Our model obtained a precision score of 0.5688 and a recall score of 0.174, with an overall F1-score of 0.266. While the precision score is relatively high, the recall score is quite low, indicating that our model performs well in predicting the downward trends in the S&P 500 index but is less successful in identifying upward trends. Overall, our study provides valuable insights into the effectiveness of using machine learning algorithms for predicting the future values of the S&P 500 index. Our results suggest that while our model may not be highly accurate, it can still provide useful information for investors and traders who are looking to make informed decisions about their investments. Future research could focus on developing more sophisticated algorithms that can handle the complex and dynamic nature of the stock market. Moreover, the inclusion of external variables such as economic metrics and geopolitical circumstances could potentially enhance the precision and robustness of our model.

References:

[1] K. Tsai and J. Wang, "External technology sourcing and innovation performance in LMT sectors", Research Policy, vol. 38, no. 3, pp. 518-526, 2009.

[2] K. Han and J. Kim, "Genetic quantum algorithm and its application to combinatorial optimization problem", Evolutionary Computation, 2000, vol. 2, pp. 1354-1360, 2000.

[3] Atsalakis, G. S., Dimitrakakis, E. M., & Zopounidis, C. D. (2011). Elliott wave theory and neuro- fuzzy systems, in stock market prediction: The WASP system. Expert Systems with Applications, 38, 9196–9206.

[4] Olivier C., Blaise Pascal University: “Neural network modeling for stock movement prediction, state of art”. 2007.

[5] Leng, X. and Miller, H.-G. : “Input dimension reduction for load forecasting based on support vector machines”, IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies (DRPT2004), 2004.

[6] Chun C, Qinghua M, Shuqiang L.: “Research on Support Vector Regression in the Stock Market Forecasting” ©Springer, Advances in Intelligent and Soft Computing Volume 148, , pp 607-612, 2012.

[7] Guo Z., Wang H., Liu Q. : “Financial time series forecasting using LPP and SVM optimized by PSO” © Springer, Soft Computing Methodologies and Applications , December 2012.

[8] Tsibouris, G., & Zeidenberg, M. (1995). Testing the efficent markets hypothesis with gradient descent algorithms. In R A (Ed.), Neural networks in the capital markets.John Wiley and Sons.

[9] Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. Expert Systems with Applications, 38(8), 10389–10397.

[10] Gupta, A. :“Stock market prediction using Hidden Markov Models”, IEEE Engineering and Systems (SCES), 2012 Students Conference on, pp.1-4, 2012.

[11] Hassan MR, Nath B. Stock market forecasting using Hidden Markov Model: a new approach. In: Proceedings—5th international conference on intelligent systems design and applications 2005, ISDA’05. 2005. pp. 192–6. https://doi.org/10.1109/ISDA.2005.85.

[12] Lei L. Wavelet neural network prediction method of stock price trend based on rough set attribute reduction. Appl Soft Comput J. 2018; 62:923–32. https://doi.org/10.1016/j.asoc.2017.09.029.