Project Proposal: Stock Market Prediction Using Classical Machine Learning (ML) Algorithms

March 2023

Motivation

As an area of great interest to investors and traders, as well as to researchers and analysts seeking to better understand its behavior and predict future trends, the stock market has long captivated attention. While deep learning models such as Convolutions Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have shown promising results, their use in stock market prediction can be limited by their complexity and lack of interpretability [11].

Therefore, the motivation behind this research project is to explore the use of classical machine learning algorithms for stock market prediction, with the goal of providing more accessible and interpretable models that can help in-vestors and traders make informed decisions. By focusing on classical machine learning algorithms, which have been shown to perform well in certain contexts [5], we aim to provide a simpler and more practical solution for those who do not have access to enormous amounts of data, or the computational resources required for deep learning.

Overall, the motivation for this research project stems from the desire to provide more transparent and understandable models for stock market predic-tion, and to explore the potential of classical machine learning algorithms in this domain. By doing so, I hope to contribute to the development of effective and practical tools for investors and traders in the stock market.

Research Question

What is the accuracy of different classical machine learning algorithms for predicting stock prices, and which classical algorithm(s) perform the best?

Initial Literature Review

Traditional machine learning techniques have limitations when it comes to processing data in its raw form. They require significant knowledge and expertise in feature selection and engineering. However, deep learning is an advanced machine learning approach that enables computers to automatically extract, analyze, and comprehend useful information from raw data. As a result, deep learning often yields superior results compared to traditional machine learning techniques [3].

A few studies have explored the use of classical machine learning algorithms for stock market prediction, with varying degrees of success.

In [12], Mehak et al utilized support vector machines (SVM) to forecast stock prices in the Karachi Stock Exchange (KSE). Their study revealed that SVM outperformed Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP), and Radial Basis Function (RBF) algorithms on the training data. However, MLP algorithm demonstrated superior performance on the test data. Nitin Kumar [3], compared the accuracy of various machine learning algorithms and concluded that SVM outperformed Logistic Regression, Neural Networks, and K-NN in terms of test accuracy, while performing equally as well as the other three algorithms on cross-validation.

One of the strengths of Support Vector Machines (SVM) is that it can effectively handle high-dimensional data by finding a hyperplane that separates the data into different classes with a maximum margin. This can help to avoid overfitting by reducing the impact of noise in the data and providing a simple decision boundary [7].

The primary challenge associated with Support Vector Machines (SVM) is the substantial amount of historical data required for accurate prediction. This large data set demands significant data processing time, which can be a major drawback. Additionally, the accuracy of SVM heavily relies on the precision and quality of the input data [5].

In a study on regression analysis conducted by Nusrat and colleagues [10], the researchers focused on predicting the stock price of a particular company, several days ahead. The findings of the study indicate that their model was able to achieve an impressive level of accuracy, with a reported success rate of 85 percent.

Other classical machine learning algorithms have been used for stock market prediction with mixed results. Some perform well in certain contexts while others do not significantly improve over traditional statistical methods. Despite limitations, these algorithms remain a promising area of research. This study will explore their potential for stock market prediction and evaluate performance using various methods.

- Data Sources

To conduct this research project, historical stock market data will be collected from Yahoo Finance. Yahoo Finance was selected due to its vast data coverage, which includes stock market data dating back to the 1950s. The data will be obtained through an interactive REST API that allows for the retrieval of data in JSON format. The dataset will include daily stock prices for IBM, Apple, and Intel Corp, and will incorporate a range of relevant financial indicators such as opening price, daily high and low prices, closing price, adjusted closing price, trading volume, ticker symbol, dividends, and stock splits. Specifically, the fol-lowing data will be collected: date of stock price, opening price, highest price of the day, lowest price of the day, closing price, adjusted closing price, volume of shares traded, stock symbol, dividends paid per share, and details of stock splits.

To capture news items critical for stock analysis and prediction, we will ob-tain data from the News API at https://newsapi.org/. This API provides access to various news articles from major news sources worldwide. We will extract relevant news articles and merge them with the historical stock price data using the Date column as the join key.

This will provide us with a comprehensive data set to train and test our models. In addition to news articles, market indicators such as interest rates, inflation rates, and exchange rates can have a significant impact on the stock market. We will capture this information from the Fed-eral Reserve Economic Data (FRED) API. FRED is a comprehensive database of economic data, maintained by the Federal Reserve Bank of St. Louis. We will extract the relevant data and join it with our main dataset using the Date column as the join key. This will provide us with a more complete picture of market conditions to improve the accuracy of our models.

To further improve the accuracy of our models, we will scrape through Ya-hoo Finance for information on company performance. This information may include earnings reports, revenue, and profit margins. By adding this infor-mation to our data set, we can better understand the financial health of the companies we are analyzing and make more accurate predictions.

In summary, we will collect news articles from the News API, market indica-tors from the FRED API, and information on company performance from Yahoo Finance. We will merge these datasets together based on the Date column and use them to build, train, test, and evaluate our classical machine learning mod-els for stock market prediction. This approach will help us to create a more comprehensive and accurate model for stock market analysis and prediction.

• Identification of Machine Learning Methods

5.1 Introduction

Predicting stock prices is a complex task that requires the use of sophisticated machine learning algorithms to accurately forecast future stock prices. In this research, I plan to use five classical machine learning algorithms to predict stock prices and determine which algorithm provides the best fit for different market conditions.

5.2 Method and Rationale

Multiple Linear Regression This algorithm is used to model the linear re-lationship between multiple independent variables (market trends, company performance, and economic indicators) and a single dependent variable (stock price). This model was chosen because it can be a good fit for stock predictions due to its ability to capture the linear relationship between multiple predictor variables and a single dependent variable, which in this case is stock prices.[6]

The multiple linear regression predicts the future value of variable (Y)with re-spect to other variables (Xi) using Eq 1.

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (1)$$

Where $\beta_0, \beta_1, \beta_2, \ldots \beta_n$ are co-efficient that can be calculated using Eq 2.

$$\beta = (X^T X)^{-1} X^T Y \; 2 \quad (2)$$

Decision Trees work by dividing the data into smaller subsets based on a series of questions, making it useful for analyzing complex data sets. This method was chosen because it can handle large data sets with many features and provide insights into which variables are most important in predicting stock prices. The Decision Tree algorithm is a highly effective and interpretable machine learn-ing tool that has gained widespread adoption due to its simplicity and ease of implementation in various domains, including image processing, data mining, machine learning, and pattern recognition [8].

Random Forest is an extension of Decision Trees, which creates multiple decision trees and combines their results to make a prediction. Was chosen be-cause it can reduce

overfitting and improve accuracy by combining the results of multiple decision trees. Random Forest is an ensemble learning algorithm that builds a forest consisting of many decision trees for classification during training. In this algorithm, each node of a decision tree represents a feature, and the root node of the tree consists of the best feature that splits the train-ing samples. Decision nodes, also known as internal nodes, represent tests on features [1]. Figure 1 shows a sample random forest decision tree structure:

Figure 1: A sample random forest decision tree structure.

Support Vector Machines (SVM), a powerful algorithm, finds the best line or plane that separates the data into different classes. This method was chosen because it is effective in high-dimensional spaces and can handle complex data sets with many features. It acts as a linear separator between two data points to classify them into different classes in a multidimensional environment [9].

k-Nearest Neighbors (k-NN) - finds the k closest data points to a new data point and uses their average as the prediction. This method was chosen be-cause it is a simple and effective algorithm for classification and regression tasks and can provide insights into which data points are most similar. Stock predic-tion can be treated as a similarity-based classification problem where historical and test data are converted into vectors with N dimensions, and a similarity metric such as Euclidean distance is used to select the k closest records in the training set. The majority vote of the selected records is then used to assign a class label to the query record [2].

5.3 Summary

By comparing the performance of these 5 algorithms in three different data sets of the same domain, we can determine which one is the most accurate and re-liable for predicting stock prices under different market conditions. This will enable investors to make informed decisions and improve their portfolio returns.

- # Identification of Evaluation Methods

To evaluate the performance of our algorithms, I will use three evaluation met-rics, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).

I chose these evaluation metrics for the following reasons:

Mean Squared Error (MSE) is a commonly used evaluation metric for regression problems. It measures the average of the squared differences between the predicted values and the actual values. By using MSE, we can quantify the average magnitude of the errors made by each algorithm. A lower MSE indicates a better-performing algorithm. This metric is particularly useful for evaluating models like Linear Regression, SVM, and k-NN, which are all regression-based. MSE is a suitable objective measure of model performance for data that follows a normal distribution [4].

Root Mean Squared Error (RMSE) is like MSE but takes the square root of the average squared difference between the predicted values and the ac-tual values. RMSE has the advantage of being in the same unit as the target variable, making it easier to interpret the results. Like MSE, a lower RMSE indicates a better-performing algorithm. We will use this metric to evaluate the performance of our models in a more interpretable metric.

R-squared (R2) measures the proportion of variance in the target variable explained by the model. It is a widely used metric for evaluating the goodness of fit of a regression model. R2 ranges from 0 to 1, where a higher value in-dicates a better-performing algorithm. We will use this metric to evaluate the performance of our models in terms of how well they capture the variability of the stock prices.

Using these evaluation metrics allows us to compare the performance of each algorithm and determine which one performs the best. By using these metrics, we can evaluate the algorithms' performance in terms of accuracy, precision, and recall, which are all crucial factors in stock prediction. This approach allows us to provide a comprehensive evaluation of our models, which is important in determining which algorithm to use in practice.

References

1. Rebecca Abraham, Mahmoud El Samad, Amer M. Bakhach, Hani El-Chaarani, Ahmad Sardouk, Sam El Nemar, and Dalia Jaber. Forecasting a stock trend using genetic algorithm and random forest. Journal of Risk and Financial Management, 15(5), 2022.

2. Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi, and Mohammed K Ali Shatnawi. Stock price prediction using k-nearest neighbor (knn) algorithm. International Journal of Business, Humanities and Technology, 3(3):32–44, 2013.

3. Nitin Kumar Chauhan and Krishna Singh. A review on conventional ma-chine learning vs deep learning. In 2018 International Conference on Com-puting, Power and Communication Technologies (GUCON), pages 347–352, 2018.

4. Timothy O Hodson, Thomas M Over, and Sydney S Foks. Mean squared error, deconstructed. Journal of Advances in Modeling Earth Systems, 13(12):e2021MS002681, 2021.

5. Mohit Iyer and Ritika Mehra. A survey on stock market prediction. In 2018 Fifth International Conference on Parallel, Distributed and Grid Comput-ing (PDGC), pages 663–668 , 2018.

6. Abidatul Izzah, Yuita Arum Sari, Ratna Widyastuti, and Toga Aldila Cin-deratama. Mobile app for stock prediction using improved multiple linear regression. In 2017 International Conference on Sustainable Information Engineering and Technology (SIET), pages 150–154, 2017.

7. Akshit Kurani, Pavan Doshi, Aarya Vakharia, and Manan Shah. A com-prehensive comparative study of artificial neural network (ann) and sup-port vector machines (svm) on stock forecasting. Annals of Data Science, 10(1):183–208, 2023.

8. Cheesun Lee, Peck Cheang, and Massoud Moslehpour. Predictive analytics in business analytics: Decision tree. Advances in Decision Sciences, 26:1– 30, 03 2022.

9. Isaac Kofi Nti, Adebayo Felix Adekoya, and Benjamin Asubam Weyori. Efficient stock-market prediction using ensemble support vector machine. Open Computer Science, 10(1):153–163, 2020.

10. Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich, and Hee-Cheol Kim. Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. Electronics, 10(21), 2021.

11. Pushpendra Singh Sisodia, Anish Gupta, Yogesh Kumar, and Gaurav Ku-mar Ameta. Stock market analysis and prediction for nifty50 using lstm deep learning approach. In 2022 2nd International Conference on Innova-tive Practices in Technology and Management (ICIPTM), volume 2, pages 156–161, 2022.

12. Mehak Usmani, Syed Hasan Adil, Kamran Raza, and Syed Saad Azhar Ali. Stock market prediction using machine learning techniques. In 2016 3rd International Conference on Computer and Information Sciences (IC-COINS), pages 322–327, 2016.