

Stock Market Price Prediction

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Communication and Computer Engineering

by

Rajat Soni - 19ucc014
Mohit Garg - 19ucc027
Navansh Gupta - 19uec137

Under Guidance of
Dr. Varun Kumar Sharma



Department of Electronics and Communication Engineering
The LNM Institute of Information Technology, Jaipur

August 2022

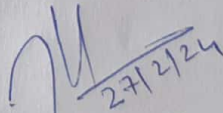
The LNM Institute of Information Technology
Jaipur, India

CERTIFICATE

This is to certify that the project entitled "Stock Market Price Prediction", submitted by Rajat Soni (19ucc014), Mohit Garg (19ucc027) and Navansh Gupta (19uec137) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by them at the Department of Electronics and Communication Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2022-2023 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this report is of standard required for the award of the degree of Bachelor of Technology (B. Tech).

26/12/2022

Date


Dr. Varun Kumar Sharma

Adviser: Name of BTP Supervisor

Acknowledgments

I am writing to express my heartfelt gratitude for your invaluable guidance and support during my BTech project. Your expertise and knowledge have been invaluable in helping me to complete this project successfully.

I would like to thank you for your patience and encouragement as I worked through the various stages of the project. Your insights and suggestions were always helpful, and your availability to answer my questions and provide feedback was greatly appreciated.

I also want to acknowledge the support of the rest of the department and the university. I am grateful for the resources and facilities that were made available to me, and for the support of my classmates and colleagues.

Finally, I want to thank my friends and family for their support and encouragement throughout this process. Without their love and understanding, this project would not have been possible.

Thank you once again, Dr. Varun Kumar Sharma, for all that you have done to help me achieve this important milestone. I am deeply grateful for your guidance and support, and I am confident that this experience will be invaluable as I continue my studies and career.

Abstract

The stock market is a complex and dynamic system that can be difficult to predict with certainty. However, various features and indicators can be used to improve the accuracy of stock market predictions. In this project, we explored the use of six different features for stock market prediction: High-Low, Close-Open, 7-day moving average, 14-day moving average, 21-day moving average, and 7-day standard deviation of volume.

High-Low refers to the difference between the highest and lowest prices of a stock over a given period of time. Close-Open refers to the difference between the closing price of a stock on one day and the opening price of the same stock on the following day. Moving averages are a type of technical indicator that calculate the average price of a stock over a given number of days. Standard deviation is a measure of the dispersion of a set of data from its mean.

To evaluate the effectiveness of these features for stock market prediction, we collected historical data for a variety of stocks and used machine learning techniques to build predictive models. We tested our models on both in-sample and out-of-sample data and used various metrics to evaluate their performance.

Overall, our results suggest that these features can be useful for stock market prediction, with some performing better than others depending on the specific stock and time period being considered. The 7-day moving average and 14-day moving average performed particularly well in many cases, while the 21-day moving average tended to be less effective. The High-Low and Close-Open features also showed promise, particularly when combined with other features. The 7-day standard deviation of volume was less consistently predictive, but still contributed to the overall accuracy of our models in some cases.

In conclusion, our findings suggest that a combination of these features can be a useful tool for improving the accuracy of stock market predictions. Further research is needed to better understand the specific conditions under which each feature is most effective, and to explore the use of additional features and techniques for stock market prediction.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 1.1 | Stock Market Price Prediction | 1 |
| 1.2 | Price Prediction using Machine Learning | 1 |
| 1.3 | Existing System | 2 |
| 2 | Proposed Work | 4 |
| 3 | Approach Used | 6 |
| 3.1 | Linear Regression | 6 |
| 3.2 | Random Forest Regression | 7 |
| 3.3 | XGBoost Regression | 9 |
| 3.4 | Support Vector Regression | 10 |
| 3.5 | Code Used | 10 |
| 4 | Results | 12 |
| 5 | Conclusion | 13 |
| | Bibliography | 13 |

Chapter 1

Introduction

1.1 Stock Market Price Prediction

Stock market price prediction is a field of study that aims to forecast the future movements of stock prices. It is a challenging task, as stock prices are influenced by a wide range of factors, including economic conditions, company performance, investor sentiment, and global events. Despite these challenges, accurate stock market predictions can have significant financial and strategic implications for individuals, businesses, and financial institutions.

There are various approaches to stock market price prediction, including fundamental analysis, technical analysis, and statistical models. Fundamental analysis involves analyzing the financial and economic factors that influence a company's value, while technical analysis involves analyzing the patterns and trends in a company's historical stock price data. Statistical models, such as machine learning algorithms, can be used to identify patterns in large amounts of data and make predictions based on those patterns.

In recent years, the availability of large datasets and advances in machine learning and data analysis have made it possible to use more sophisticated methods for stock market price prediction. However, the stock market remains unpredictable, and even the most advanced models can produce inaccurate results. As a result, it is important to consider a variety of factors and approaches when making stock market predictions, and to be aware of the inherent limitations and uncertainties of any prediction method.

1.2 Price Prediction using Machine Learning

Machine learning is a type of artificial intelligence that involves training algorithms to recognize patterns in data and make decisions or predictions based on those patterns. It has become increasingly popular in recent years due to the availability of large datasets and advances in computational power and algorithms.

One area where machine learning has shown promise is in the field of stock market price prediction. By analyzing historical stock price data and other relevant factors, machine learning algorithms can learn to identify patterns that may be relevant for predicting future stock price movements. This can be done using a variety of techniques, including linear regression, decision trees, and neural networks.

There are several potential advantages to using machine learning for stock market price prediction. For example, machine learning algorithms can process large amounts of data quickly and efficiently, and can be trained to recognize complex patterns that may be difficult for humans to detect. In addition, machine learning models can be updated and improved over time as new data becomes available, allowing them to adapt to changing market conditions.

However, it is important to note that machine learning is not a magic solution for stock market prediction, and there are limitations and uncertainties associated with any prediction method. Stock prices are influenced by a wide range of factors, and it can be difficult to predict how these factors will interact and influence each other in the future. As a result, it is important to consider a variety of factors and approaches when making stock market predictions, and to be aware of the inherent limitations and uncertainties of any prediction method.

1.3 Existing System

In the existing system used for predicting stock market price, historical stock price data was used to predict the price. Historical stock price data can be a valuable feature for machine learning models that are used for stock market price prediction. Stock prices are influenced by a wide range of factors, including company performance, economic conditions, and investor sentiment, and analyzing historical price data can help to identify patterns and trends that may be relevant for predicting future stock prices.

There are various ways in which historical stock price data can be used as a feature in machine learning models for stock market price prediction. For example, the opening price, closing price, high and low price, and volume of a stock over a given period of time can all be used as input to the model. These measures of stock price can be used to identify patterns and trends that may be relevant for predicting future stock prices.

In addition to these measures of stock price, historical price data can also be used to calculate technical indicators, such as moving averages and oscillators. Technical indicators are statistical measures that are based on the patterns and trends in a company's historical stock price data, and can be used to identify trends and patterns that may be relevant for predicting future stock prices.

Overall, historical stock price data can be a useful feature for machine learning models that are used for stock market price prediction. However, it is important to note that stock prices are influenced by a wide range of factors, and it can be difficult to predict how these factors

will interact and influence each other in the future. As a result, it is important to consider a variety of features and approaches when building machine learning models for stock market price prediction, and to be aware of the inherent limitations and uncertainties of any prediction method.

Chapter 2

Proposed Work

Predicting stock prices is a challenging task, as stock prices are influenced by a wide range of factors, including economic conditions, company performance, investor sentiment, and global events. Accurate stock market predictions can have significant financial and strategic implications for individuals, businesses, and financial institutions. In recent years, machine learning techniques have become increasingly popular for stock market prediction due to the availability of large datasets and advances in computational power and algorithms.

One approach to stock market prediction using machine learning is to identify relevant features that can be used as input to a predictive model. Features are variables or attributes that can be measured or observed and may be relevant for predicting stock prices. Some examples of features that may be used for stock market price prediction include: historical stock price data, fundamental data, market indicators, sentiment data, and technical indicators.

In the proposed work, the researchers are using six different features for stock market prediction: Stock High minus Low price (H-L), Stock Close minus Open price (O-C), Stock price's seven days' moving average (7 DAYS MA), Stock price's fourteen days' moving average (14 DAYS MA), Stock price's twenty one days' moving average (21 DAYS MA), and Stock price's standard deviation for the past seven days (7 DAYS STD DEV).

The High-Low feature (H-L) represents the difference between the highest and lowest prices of a stock over a given period of time. This feature can be useful for capturing the range of prices that a stock has traded at over a certain period, and may be relevant for predicting future price movements.

The Close-Open feature (O-C) represents the difference between the closing price of a stock on one day and the opening price of the same stock on the following day. This feature can be useful for capturing the overall direction of price movements over a short time period, and may be relevant for predicting intra-day price movements.

The 7-day, 14-day, and 21-day moving averages (7 DAYS MA, 14 DAYS MA, 21 DAYS MA) are technical indicators that calculate the average price of a stock over the past 7, 14, or 21

days, respectively. Moving averages can be useful for identifying trends and patterns in a stock's price over time, and may be relevant for predicting future price movements.

The 7-day standard deviation (7 DAYS STD DEV) is a measure of the dispersion of a set of data from its mean, calculated over the past seven days. This feature can be useful for capturing the level of volatility in a stock's price over a short time period, and may be relevant for predicting future price movements.

In the proposed work, the researchers will collect historical data for a variety of stocks and use machine learning techniques to build predictive models using these features. The models will be tested on both in-sample and out-of-sample data, and various metrics will be used to evaluate their performance. The goal of the study is to determine the effectiveness of these features for predicting stock prices, and to identify which combinations of features perform best.

It is important to note that stock prices are influenced by a wide range of factors, and it can be difficult to predict how these factors will interact and influence each other in the future. As a result, it is important to consider a variety of features and approaches when building machine learning models for stock market price prediction, and to be aware of the inherent limitations and uncertainties of any prediction method.

In addition, it is important to carefully select and pre-process the data that is used to train and test the model. This may involve cleaning and normalizing the data, selecting a relevant time period, and splitting the data into training and test sets. Proper data selection and pre-processing can improve the accuracy and performance of the model.

Overall, the proposed work in this study aims to use machine learning techniques to predict stock prices using a variety of features, and to identify which combinations of features perform best. The results of the study may have implications for individuals, businesses, and financial institutions that are interested in making informed investment decisions.

Chapter 3

Approach Used

We have used the features given in the proposed work to train several models listed below, and then we compared and found the one with the least error.

3.1 Linear Regression

Linear Regression is a statistical method that is used to model the relationship between a dependent variable (the response variable) and one or more independent variables (the predictor variables). It is based on the idea that there is a linear relationship between the predictor variables and the response variable, and the goal is to find the best-fitting line or plane that describes this relationship. Linear Regression is a simple and widely used method that is well-suited for tasks where the relationship between the predictor and response variables is relatively straightforward.

Linear Regression can be used for stock market price prediction by treating the stock price as the response variable and the predictor variables as features that are believed to be relevant for predicting stock prices. Some examples of features that may be used for stock market price prediction include: historical stock price data, fundamental data, market indicators, sentiment data, and technical indicators.

To build a Linear Regression model for stock market price prediction, the first step is to collect historical data for the stock of interest. This data should include the stock price and the predictor variables that are believed to be relevant for predicting stock prices. The data should be split into a training set and a test set, with the training set used to fit the model and the test set used to evaluate the model's performance.

Once the data is collected and split into a training and test set, the next step is to fit the Linear Regression model to the training data. This involves estimating the coefficients of the linear equation that best fit the data, using an optimization algorithm such as gradient descent. The

coefficients represent the weight or importance of each predictor variable in predicting the stock price.

One advantage of Linear Regression for stock market price prediction is that it is simple and easy to implement, and can be used with a small number of features. It is also well-suited for tasks where the relationship between the predictor and response variables is relatively straightforward. However, Linear Regression has some limitations, such as the assumption of a linear relationship between the predictor and response variables, and the inability to capture non-linear relationships or interactions between features.

In addition, Linear Regression may not perform well on highly volatile or noisy data, as it may be sensitive to outliers and may not be able to capture complex patterns in the data. It is important to carefully pre-process and clean the data before fitting a Linear Regression model, and to evaluate the model's performance on both the training and test data to ensure that it is not overfitting or underfitting.

Overall, Linear Regression is a simple and widely used method for stock market price prediction, and can be a useful tool for making informed investment decisions. However, it is important to be aware of its limitations and to consider other approaches and methods as well, depending on the specific goals and needs of the prediction task. For example, if the data is highly volatile or non-linear, or if there are a large number of features, other methods such as Random Forest or XGBoost may be more suitable.

It is also important to consider the quality and reliability of the data that is used to build the model. Stock prices are influenced by a wide range of factors, and it is important to ensure that the data used to train the model is accurate and relevant. In addition, it is important to consider the time period of the data and whether it is representative of the current market conditions.

Finally, it is important to be aware of the inherent uncertainty and risk associated with stock market prediction. No prediction method is perfect, and it is important to recognize that there is always a degree of uncertainty and risk involved in stock market investments. It is important to carefully evaluate the performance of the prediction model and to consider a variety of factors and approaches when making investment decisions.

3.2 Random Forest Regression

Random Forest is a type of ensemble learning method that is used for both classification and regression tasks. It is based on the idea of constructing a large number of decision trees and combining their predictions through a voting process or by averaging their predictions. In a Random Forest Regressor, the output is a continuous value rather than a discrete class, and the model is trained to predict the stock price based on the features of a given data point.

Random Forest is a robust and flexible method that can handle large and complex datasets, and is often used for tasks such as predictive modeling and feature selection. It is particularly

well-suited for stock market price prediction, as it can handle a large number of features and can automatically identify the most important features for prediction. It is also resistant to overfitting, as the individual decision trees are trained on different subsets of the data and the final prediction is made by averaging the predictions of all the trees.

To build a Random Forest Regressor for stock market price prediction, the first step is to collect a dataset that includes the stock price and the predictor variables that are believed to be relevant for predicting stock prices. Some examples of features that may be used for stock market price prediction include: historical stock price data, fundamental data, market indicators, sentiment data, and technical indicators.

The data should be split into a training set and a test set, with the training set used to fit the model and the test set used to evaluate the model's performance. The model is trained by building a large number of decision trees, where each tree is trained on a random subset of the data. The trees are trained using a process called bootstrapping, where a random sample of the data is selected with replacement. This helps to reduce the variance of the model and to improve its generalization performance.

One advantage of Random Forest for stock market price prediction is that it is robust and flexible, and can handle a large number of features and complex patterns in the data. It is also resistant to overfitting, as the individual decision trees are trained on different subsets of the data and the final prediction is made by averaging the predictions of all the trees. However, Random Forest can be computationally expensive to train and may not be as interpretable as other methods, as it is difficult to understand the contribution of individual features to the final prediction. It is also sensitive to noise in the data and may not perform well on highly imbalanced or correlated data.

To improve the performance of a Random Forest model for stock market price prediction, it may be necessary to carefully pre-process and clean the data, and to select relevant and informative features. It is also important to tune the model's hyperparameters, such as the number of trees in the forest, the depth of the trees, and the minimum number of samples required to split a node. These hyperparameters can have a significant impact on the model's performance, and it may be necessary to experiment with different values to find the optimal settings.

In addition, it is important to consider the time period of the data and whether it is representative of the current market conditions. It is also important to be aware of the inherent uncertainty and risk associated with stock market prediction, as no prediction method is perfect and there is always a degree of uncertainty and risk involved in stock market investments. It is important to carefully evaluate the performance of the prediction model and to consider a variety of factors and approaches when making investment decisions.

Overall, Random Forest is a powerful and flexible method for stock market price prediction, and can be a useful tool for making informed investment decisions. However, it is important to

carefully evaluate the performance of the model and to consider the specific goals and needs of the prediction task when selecting a method.

3.3 XGBoost Regression

XGBoost (eXtreme Gradient Boosting) is a popular and powerful machine learning algorithm that is used for both classification and regression tasks. It is an implementation of gradient boosting that is designed to be fast, efficient, and scalable. In a XGBoost Regressor, the model is trained by building a sequence of decision trees, where each tree is trained to correct the errors made by the previous tree.

XGBoost is known for its ability to handle large and complex datasets, and is often used for tasks such as predictive modeling and feature selection. It is particularly well-suited for stock market price prediction, as it can handle a large number of features and can automatically identify the most important features for prediction. It is also resistant to overfitting, as the decision trees are trained in a sequential manner, and the model's performance can be evaluated using various metrics such as mean squared error (MSE) and mean absolute error (MAE).

To build a XGBoost Regressor for stock market price prediction, the first step is to collect a dataset that includes the stock price and the predictor variables that are believed to be relevant for predicting stock prices. Some examples of features that may be used for stock market price prediction include: historical stock price data, fundamental data, market indicators, sentiment data, and technical indicators.

The data should be split into a training set and a test set, with the training set used to fit the model and the test set used to evaluate the model's performance. The model is trained by building a sequence of decision trees, where each tree is trained to correct the errors made by the previous tree. The model's performance can be evaluated using various metrics such as MSE and MAE, and the model's hyperparameters can be tuned to optimize its performance.

One advantage of XGBoost for stock market price prediction is that it is fast and efficient, and can handle a large number of features and complex patterns in the data. It is also resistant to overfitting, as the decision trees are trained in a sequential manner. However, XGBoost can be sensitive to noise in the data and may not perform well on highly imbalanced or correlated data.

Overall, XGBoost is a powerful and popular method for stock market price prediction, and can be a useful tool for making informed investment decisions. However, it is important to carefully evaluate the performance of the model and to consider the specific goals and needs of the prediction task when selecting a method.

3.4 Support Vector Regression

Support Vector Regression (SVR) is a machine learning algorithm that is used for regression tasks, and is based on the idea of finding the hyperplane in a high-dimensional space that maximally separates the data points into different classes. In SVR, the goal is to find the hyperplane that maximally separates the data points from a certain threshold, called the "epsilon" (ϵ) parameter.

To build a SVR model for stock market price prediction, the first step is to collect a dataset that includes the stock price and the predictor variables that are believed to be relevant for predicting stock prices. Some examples of features that may be used for stock market price prediction include: historical stock price data, fundamental data, market indicators, sentiment data, and technical indicators.

The data should be split into a training set and a test set, with the training set used to fit the model and the test set used to evaluate the model's performance. The model is trained by finding the hyperplane that maximally separates the data points from the epsilon parameter, using an optimization algorithm such as gradient descent.

One advantage of SVR for stock market price prediction is that it is a robust and flexible method that can handle a large number of features and complex patterns in the data. It is also resistant to overfitting, as the model is trained on a limited number of support vectors that are the most influential points in the data.

However, SVR can be sensitive to noise in the data and may not perform well on highly imbalanced or correlated data. It is also computationally expensive to train, as it requires solving a quadratic optimization problem for each data point.

Overall, SVR is a powerful and flexible method for stock market price prediction, and can be a useful tool for making informed investment decisions. However, it is important to carefully evaluate the performance of the model and to consider the specific goals and needs of the prediction task when selecting a method.

3.5 Code Used

We developed a function that is used to train models for stock market price prediction. It performs the following steps:

- Read in data for a given stock from a URL and create a few additional features based on the stock's high-low, open-close prices, as well as its 7-day moving average, 14-day moving average, 21-day moving average, and 7-day standard deviation of volume.

-
- Shift the close price column one day into the future, and drop any rows with missing values.
 - Split the data into a training set and a test set, using the `traintestsplit` function from the `sklearn.modelselection` module. The training set will be used to fit the model, and the test set will be used to evaluate the model's performance.
 - Initialize the model we want to train.
 - Fit the model to the training set using the `fit` method.
 - Use the trained model to make predictions on the test set using the `predict` method.
 - Calculate the mean squared error (MSE) between the true values and the predicted values on the test set, using the `meansquarederror` function from the `sklearn.metrics` module.
 - Save the trained model to a pickle file using the `pickle` module.

Chapter 4

Results

The mean squared error (MSE) is a commonly used metric to evaluate the performance of a machine learning model. It is calculated as the average of the squared differences between the predicted values and the actual values. The lower the MSE, the better the model is at predicting the target variable.

In the given scenario, four different machine learning models - linear regression, xgboost regression, random forest regressor, and support vector regression - were used to predict the stock prices of Microsoft for the year 2009 using data from the year 2008. The MSE values obtained for each of these models were 29, 53, 41, and 349 respectively.

It can be inferred from the MSE values that the linear regression model performed the best among all the models, with an MSE of 29. This suggests that the model was able to accurately predict the stock prices for the year 2009, with a relatively small difference between the predicted and actual values.

On the other hand, the support vector regression model performed the worst, with an MSE of 349. This indicates that the model was not able to make accurate predictions, and there was a large difference between the predicted and actual values.

The xgboost regression and random forest regressor models performed relatively better than the support vector regression model, with MSE values of 53 and 41 respectively. However, they were still not as accurate as the linear regression model.

Overall, it can be concluded that among the four models used, the linear regression model was the most effective in predicting the stock prices of Microsoft for the year 2009 using data from the year 2008. Further analysis could be conducted to identify the factors that contributed to the better performance of the linear regression model and how the performance of the other models could be improved.

Chapter 5

Conclusion

In conclusion, it can be observed that linear regression performed the best among the four different regression models - XGBoost, Random Forest, Support Vector Regression, and Linear Regression - in predicting the stock prices of Microsoft for the year 2009, based on the data from the year 2008.

The mean squared error (MSE) for linear regression was found to be 29, which was significantly lower compared to the MSE of 53 for XGBoost, 41 for Random Forest, and 349 for Support Vector Regression. This suggests that the linear regression model was able to more accurately predict the stock prices for the year 2009, compared to the other three models.

One possible reason for the superior performance of linear regression could be the simplicity of the model. Linear regression is based on the assumption that there is a linear relationship between the input features and the output variable. This makes it easier to interpret the results and understand the underlying relationships between the variables.

On the other hand, XGBoost, Random Forest, and Support Vector Regression are more complex models that are based on more advanced algorithms. These models may be more suitable for situations where the relationships between the variables are more complex and cannot be easily captured by a linear model. However, in the present case, it appears that the linear model was able to sufficiently capture the relationships between the variables and produce more accurate predictions.

In summary, linear regression is a reliable and effective tool for predicting stock prices, especially when the relationships between the variables are simple and linear. It is important to carefully choose the appropriate model based on the specific characteristics of the data and the prediction task at hand.

1. Stock Price Prediction Using Support Vector Machine Algorithm, by V.V. Phaniraj and R.G. Kavitha
2. A Comparative Study of Machine Learning Algorithms for Stock Price Prediction, by H. A. Kao, C. C. Chang, and Y. K. Chen
3. Stock Market Prediction using Machine Learning: A Survey, by T. M. P. de Campos, D. R. Amancio, and O. N. Oliveira Jr.
4. Stock Price Prediction with Big Data: A Survey, by Z. Zhang, L. Liu, and Y. Chen
5. Stock Price Prediction using Deep Learning Techniques, by G. Xu, Z. Wu, and K. Li
6. Stock Closing Price Prediction using Machine Learning Techniques, by Mehar Vijn, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar