Stock Price Prediction using Machine Learning

Pushkar Patidar
Computer Engineering Department
MS Ramaiah University of
Applied Science.
Bengaluru ,Karnataka
pushkarpatidar400@gmail.com

Karan Patel
Computer Engineering Department
MS Ramaiah University of
Applied Science.
Bengaluru ,Karnataka
kittukpcool@gmail.com

Tanvi Saxena
Computer Engineering Department
MS Ramaiah University of
Applied Science.
Bengaluru ,Karnataka
tanvi.saxena1505@gmail.com

Pranta Paul
Computer Engineering Department
MS Ramaiah University of
Applied Science.
Bengaluru ,Karnataka
prantapaul2002@gmail.com

Ayush Aryan
Computer Engineering Department
MS Ramaiah University of
Applied Science.
Bengaluru ,Karnataka
ayusharyan791@gmail.com

Skanda S Kumar
Computer Engineering Department
MS Ramaiah University of
Applied Science.
Bengaluru ,Karnataka
skanda_2011@hotmail.com

***Abstract* — Stock market prediction is an act of trying to determine the future value of a stock other financial instrument traded on a financial exchange. Prediction of stock market is a long-time attractive topic to researchers from different fields. -In Stock Market Prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. These techniques have proven to be highly effective, yielding maximum accuracy with minimal monetary investment and also saving huge amounts of time. Prediction of stock prices is one of the most researched topics and gathers interest from academia and the industry alike. With the emergence of Artificial Intelligence, various algorithms have been employed in order to predict the equity market movement. Moreover, the behaviour of stock prices is uncertain and hard to predict. For these reasons, stock price prediction is an important process and a challenging one. This leads to the research of finding the most effective prediction model that generates the most accurate prediction with the lowest error percentage.**

In our research, we are going to use Machine Learning Algorithm specially focus on Linear Regression (LR). We obtained data from Yahoo Finance for S&P500 (GSPC) stock after implementation LR. We successfully stock market trend for next month.

Keywords - Machine learning, Random Forest Regressor.

I. INTRODUCTION

Machine learning, a well-established algorithm in a wide range of applications, has been extensively studied for its potentials in prediction of financial markets. ML has been widely used in the financial sector to provide a new mechanism that can help investors make better decisions in both investment and management to achieve better performance of their securities investment. If market prices going up with available stock then stakeholders get profit with their purchased stocks. In other case, if market going down with available stock prices then stakeholders have to face losses. The technical analysis it is an evolution of stocks by the means of studying the statistics generated by market activity, such as past prices and volumes. The vital part of machine learning is the dataset used. The dataset should be as concrete as possible because a little change in the data can perpetuate massive changes in the outcome. In this project, supervised machine learning is employed on a dataset obtained from Yahoo Finance.

In the recent years, increasing prominence of machine learning in various industries have enlightened many traders to apply machine learning techniques to the field, and some of them have produced quite promising results. This dataset comprises of following five variables: open, close, low, high and volume. Open, close, low and high are different bid prices for the stock at separate times with nearly direct names. The volume is the number of shares that passed from one owner to another during the time period. The model is then tested on the test data.

This paper will develop a financial data predictor program in which there will be a dataset storing all historical stock prices and data will be treated as training sets for the program. The main purpose of the prediction is to reduce uncertainty associated to investment decision making.

II. RELATED WORK

Correct Prediction of stock market trends is of great importance for the investors as it helps in determining whether the investment would pay off or not. Many methods have been deployed for the same. Artificial Neural Network based method is the first technique to be used for the stock market trend prediction. Random Forest is another machine learning model used for predicting trend direction of stocks which we are going to use in this paper.In this research paper, the comparative study of the supervised machine learning algorithms.

A. Traditional approaches

This statistical method is commonly used to model the relationship between a dependent variable (i.e., stock price) and one or more independent variables (i.e., economic indicators, company financials, etc.). Ordinary least squares (OLS) regression is a popular technique used for stock price prediction, which assumes a linear relationship between the dependent and independent variables.

B. *Econometric models*

These models use statistical techniques to analyze the relationship between economic variables and stock prices. They attempt to capture the underlying economic factors that influence stock prices, such as interest rates, inflation, and GDP growth.

C. *Expert judgment*

This approach involves using the knowledge and expertise of financial analysts and investors to make predictions about stock prices. Expert judgment can be useful in situations where data is limited or uncertain, but it is subject to biases and errors.

III. METHODOLGY

In this model we are going to detect the stock price of S&P500 for future based on the previous 30 years data. Stock market prediction seems a complex problem because there are many factors that have yet to be addressed and it doesn't seem statistical at first. But by proper use of machine learning techniques, one can relate previous data to the current data and train the machine to learn from it and make appropriate assumptions.

In this model I used historical data in machine learning to recognize trends and understand the current market. Machine learning automates the trading process by using statistical models to draw insights and make predictions. Machine learning can collect and test a large amount of data, both structured and unstructured. It can apply suitable algorithms, transform, search for patterns, and make decisions based on the new data.

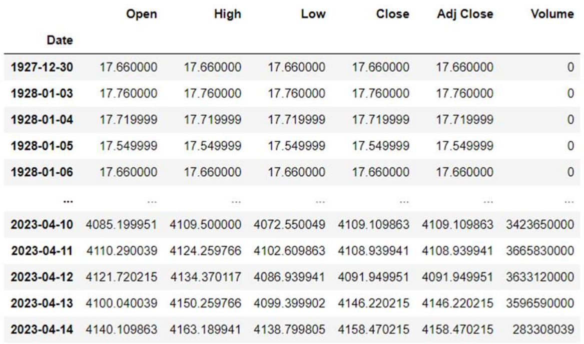
• Opening Price is the first price of any listed stock at the beginning of an exchange on a trading day.
. • High and Low Prices are the highest and lowest price of stock on that day. Generally, these data are used by traders to measure volatility of stock.
• Closing Price is a price of the stock at the close of the trading day.
• Volume is the number of stocks or contracts traded for a security in all the markets during a given time period.
• Adjusted Closed Prices is considered as the true price of that stock, and shows the stock's value after distributing dividends.

*A. **Data Exploration***

Data set in this model is collect from the module yfinance for S&P500 stock. This project shows how various efforts have been taken to apply machine learning in stock forecasting for the S&P500 index. I implemented it in python with open-source libraries. I got the historical data from yahoo finance and applied pre-processing methods to make the data relevant. Also, using the randomized grid search cross-validation, a hyper tuning process to validate the model for building, fitting, and training for prediction. After prediction, error analysis is crucial for identifying how the model performs and how accurate the predicted values.

B. Data Pre-processing

This step is the most important part of this project. Data pre-processing are steps taken to make the data ready for the machine learning model. Pre-processing includes transforming the raw data into a format that the model can take from and operate on. This project aims to have a dataset that the model accepts, and the algorithm understands. In a dataset, values can be missing, and information can be redundant and irrelevant, or noisy. Data cleaning is a form of pre-processing that includes removing missing or inconsistent values and changing the index. As well as feature selection, hyperparameter tuning, and data standardization.



	Open	High	Low	Close	Adj Close	Volume
Date						
1927-12-30	17.660000	17.660000	17.660000	17.660000	17.660000	0
1928-01-03	17.760000	17.760000	17.760000	17.760000	17.760000	0
1928-01-04	17.719999	17.719999	17.719999	17.719999	17.719999	0
1928-01-05	17.549999	17.549999	17.549999	17.549999	17.549999	0
1928-01-06	17.660000	17.660000	17.660000	17.660000	17.660000	0
...	...	...	...	...	...	...
2023-04-10	4085.199951	4109.500000	4072.550049	4109.109863	4109.109863	3423650000
2023-04-11	4110.290039	4124.259766	4102.609863	4108.939941	4108.939941	3665830000
2023-04-12	4121.720215	4134.370117	4086.939941	4091.949951	4091.949951	3633120000
2023-04-13	4100.040039	4150.259766	4099.399902	4146.220215	4146.220215	3596590000
2023-04-14	4140.109863	4163.189941	4138.799805	4158.470215	4158.470215	283308039

Stock data till now

*C. **Feature Engineering***

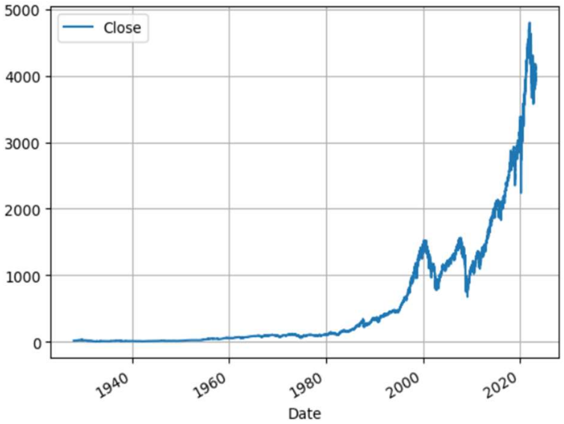
The x and y characteristics are chosen at this point to produce the model's data set. The training and testing data sets each have x and yfeatures specified.

The dataset's columns are called features. One of the fundamental ideas in machine learning applications, feature selection greatly affects the performance of the model. It won't be required to use every column in future selections. These chosen features have an effect and contribute to the outcome of the prediction. The test set performs worse overall because of unnecessary features. Discovering the most crucial aspects and the significance of features is one approach of choosing futures. Feature selector and feature significance modules are available in Sklearn and can be used. Each feature in the data is assigned a score using the feature significance module.

*D. **Model Selection***

- The random forest regression model is used for prediction. This will predict the low and high values of the next trading days, which includes the future prices for the next five days, one month, and one year of the S&P500. The outcome of buying, selling, or holding a stock will be based on the predicted values. The objective of this project is data collection, data processing, and building the trading algorithm for prediction.Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

- The data is got via a python script.Python script is used to gain the data by using yfinance. It will fetch the S&P500 stock data from starting to current date. The downloaded S&P500 stock data is loaded into a data frame and converted into a CSV file (comma separate value). So, that I can store it locally and quickly load it into the algorithm. I stored the data set insp500_data.csv.



E. Model Training and Testing of data set

A subset of the dataset used to create and construct prediction models is called the "training set." Building a training dataset script produces a training set by generating the features of the training set using the input options and the raw stock price data. The model is trained using the data. The model runs on the train set and gains knowledge from the data.

A testing set is a subset of the dataset used to gauge how well a model will perform in the future. It is a useful benchmark for assessing the model. The trained model is tested using the testing set in comparison to the predicted dataset. This section of the set has not been viewed by the model. It serves as an evaluation tool.

Sklearn has a feature called default scaling that is used to standardize a data set. It is known that standardization improves the numerical stability of the model and increases the learning rate.Best practice is to fit the scaler to the training data and then transform the test data. This would prevent data loss during the model testing process. This is useful when you want to compare data for different drives.

F. Model Optimization

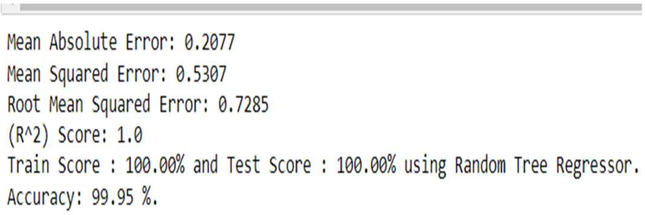
- Hyperparameters are model parameters. It is important to adjust the settings to optimize performance. I set it after training and testing the dataset before adjusting and predicting it. Hyperparameters solve the main problem of machine learning, which is overfitting. In this project, I used random search cross-validation.
- Hyperparameters in a random forest model are used to increase the predictive power of the model or to speed it up.For a random forest regression model, the best parameters to consider are n estimators,max depth,min samples split ,min samples leaf, bootstrap,random state. 'random_state':42,'nestimators':20,'min_samples_split' :2,'min_samples_leaf': , 'max_depth': 14, 'bootstrap': False.
- The data set is now ready for the model. The first step is to choose a value for the random state and build the tree based on the number of random states. Random Forest prevents overfitting by creating random subsets of features and creating smaller trees with those subsets. In order to create a random forest, the data must be trained. Here, too, the parameters of the hyperparameter tuning are used.

G. Performance Evaluation

In stock price prediction, the output variable (stock price) is continuous and there are no classes or categories to predict.

For this tasks,it is more common to use evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2) to assess the performance of the regression model. These metrics provide information about how well the model is able to predict the actual stock prices, in terms of the differences between the predicted and actual prices.

- **Mean absolute error (MEA)** measures the average size of errors in a set of predictions, regardless of their direction. It is the mean absolute difference between the prediction and the actual observation, with all individual differences being given equal weight. First, measure the difference between the actual value and the expected vaue.
- **The R^2** indicates how well the model fits the given dataset. Indicates how narrow the regression line is the predicted and actual values plotted. The maximum value is 1.0. The higher the values, the better the model fits. The regression line fits the data well, and the model works well when r^2 values are between 0.6 and 1.0. Values above 65% are considered good.
- **Mean squared error (MSE)** takes the sum of the absolute value of error. The mean squared error determines the model performance too. In this case, larger errors are well noted, more than that of the MAE. The lower the MSE value, the higher the prediction accuracy.
- **Evaluation Results:**



```
Mean Absolute Error: 0.2077
Mean Squared Error: 0.5307
Root Mean Squared Error: 0.7285
(R^2) Score: 1.0
Train Score : 100.00% and Test Score : 100.00% using Random Tree Regressor.
Accuracy: 99.95 %.
```

Our model accuracy for testing of trained data is 99.9 %

.

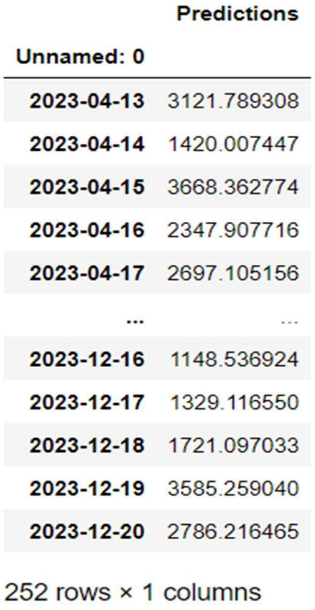
IV. Result And Analysis

I have created data frames with predicted values for the next year, next month and five days. There are 252 trading days in a year, 21 trading days in a month ,8 trading days in 10 days. I extracted the required future days from the forecast of 341 trading days. The dates and prices for these future days are converted to CSV.

To determine supply and demand and maintain prices, investors try to make a profit by selling at the highest price, buying at the lowest price, and holding the price if nothing happens. For example; An investor buys a stock at a minimum of $5 and the price is expected to reach $20 in 21 days. The investor holds his shares until the 21st day to sell them at a profit. In this context, the high price is the ask price and the low is the purchase price.

A. *Prediction of Price for One year :*

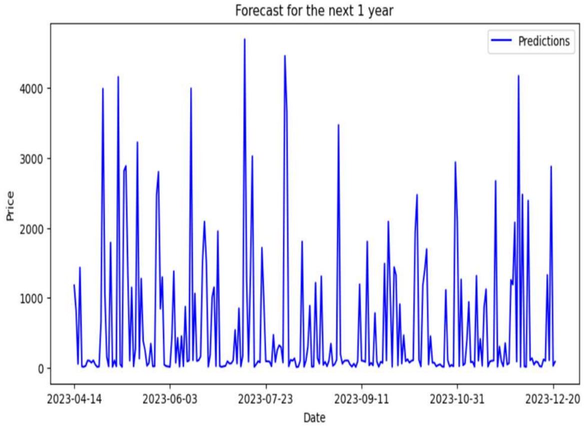
Predicting stock prices is a complex and challenging task, and no single approach or theory can guarantee accurate predictions. However, by combining multiple approaches and using advanced analytical techniques, we improve the accuracy of predictions and identify profitable investment opportunities.



	Predictions
Unnamed: 0	
2023-04-13	3121.789308
2023-04-14	1420.007447
2023-04-15	3668.362774
2023-04-16	2347.907716
2023-04-17	2697.105156
...	...
2023-12-16	1148.536924
2023-12-17	1329.116550
2023-12-18	1721.097033
2023-12-19	3585.259040
2023-12-20	2786.216465

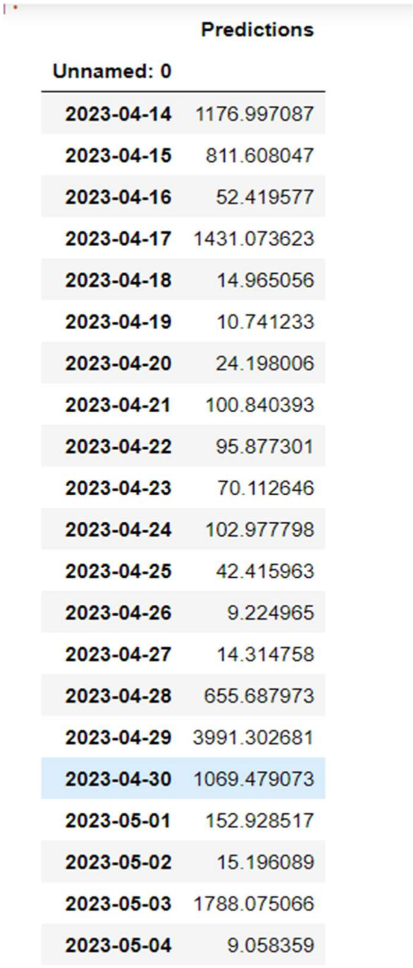
252 rows × 1 columns

Graph :



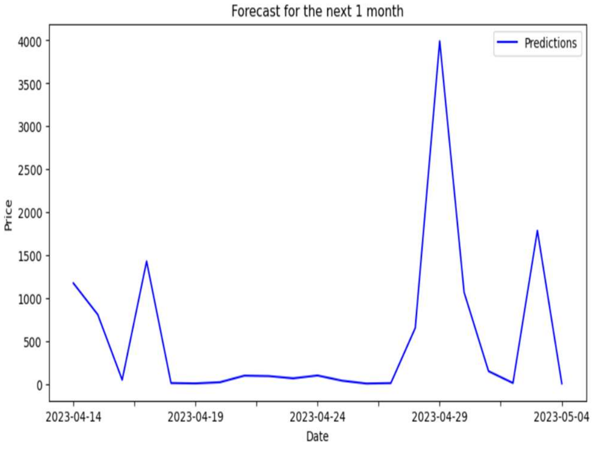
B. *Prediction of Price for One month :*

Prediction of stock price from current date "14-04-2023"to "04-05-2023".based on the historical data.



	Predictions
Unnamed: 0	
2023-04-14	1176.997087
2023-04-15	811.608047
2023-04-16	52.419577
2023-04-17	1431.073623
2023-04-18	14.965056
2023-04-19	10.741233
2023-04-20	24.198006
2023-04-21	100.840393
2023-04-22	95.877301
2023-04-23	70.112646
2023-04-24	102.977798
2023-04-25	42.415963
2023-04-26	9.224965
2023-04-27	14.314758
2023-04-28	655.687973
2023-04-29	3991.302681
2023-04-30	1069.479073
2023-05-01	152.928517
2023-05-02	15.196089
2023-05-03	1788.075066
2023-05-04	9.058359

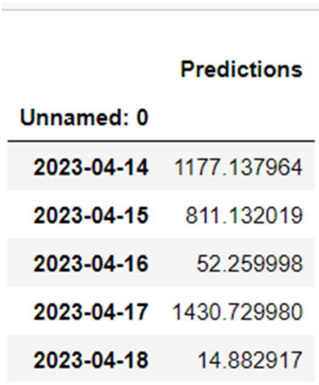
Graph :



C. *Prediction of price for5 days :*

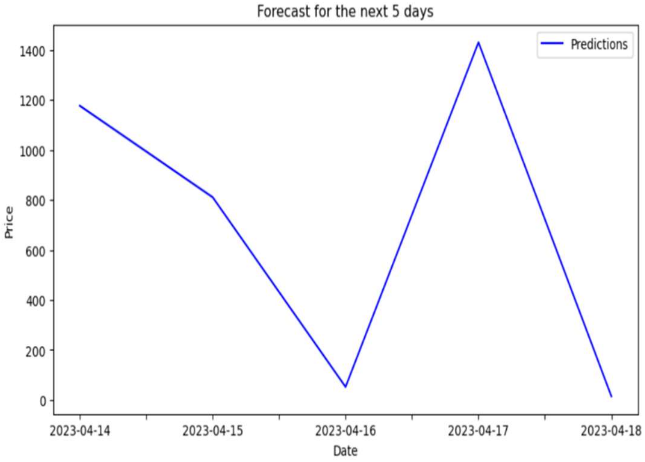
Prediction of stock price from current date “14-04-2023”to “04-05-2023”.based on the historical data.

Data:



	Predictions
Unnamed: 0	
2023-04-14	1177.137964
2023-04-15	811.132019
2023-04-16	52.259998
2023-04-17	1430.729980
2023-04-18	14.882917

Graph:



V. Conclusion

Stock market forecast is the actual demand for a profitable business. Predictions are always helpful to reduce the risk factor in any trading environment. The risk factor can be analyzed based on historical data and previous activitiesTrends. This study was based on multiple results and we used machine learning (ML) algorithm as linear regression(LR) on trade priority. A linear regression was applied to the various data sets obtained Exchange(Yahoo Finance).Yahoo Finance has always been considered the best place on the market to get inventory data on any product.In the project, we proposed using data collected from Yahoo's global financial markets with machine learning algorithms to predict movements of the S&P500 stock index. Our algorithm works on a large amount of data collected from various global financial markets. It also doesn't cause overfitting issues. Several machine learning based models are offered to predict the daily trend of market actions.The numerical results indicate a high efficiency. Convenient trading patterns based on our well-trained predictor. The model generates a higher profit than selected benchmarks and also we calculated a predicted stock price for one year with more accuracy .

In the future, researchers could focus on combining sentiment analysis and stock insights and a numeric value associated with the historical value of the security in the security price prediction. the following Both pieces of information can also be used to construct effective action recommendation systems. deep learning-based approaches can be used to achieve better and more efficient feature extraction techniques

REFERENCES

[1] Alpaydin, E. (2014). Introduction to machine learning. MIT press.

[2] N.K. Chowdhury and C.K.-S. Leung. Improved travel time prediction algorithms for intelligent transportation systems. In Proc. KES 2011, Part II, pp. 355–365

[3] W. Huang et al., “Forecasting stock market movement direction with support vector machine,” Computers & Operations Research, 32, pp. 2513–2522005.

[4] Yuqing Dai, Yuning Zhang, Machine Learning in Stock Price Trend Forecasting.

[5] Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W. and Wills, G., 2020. A survey on machine learning for stock price prediction: algorithms and techniques.

[6] Bohn, Tanner A. "Improving Long Term Stock Market Prediction with Text Analysis."

[7] K. Raza, "Prediction of Stock Market performance by using machine learning techniques," 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), Karachi.

[8] S&P 500 information technology. S&P Dow Jones Indices. http://ca.spindices.com/indices/equity/sp-500-informationtechnology-sector.