iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: I Month of publication: January 2022

DOI: https://doi.org/10.22214/ijraset.2022.39891

www.ijraset.com

Call: ☎ 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

Stock Price Prediction using Machine Learning Algorithms

Shreya Pawaskar
Computer Engineering Department, Cummins College of Engineering, Pune

Abstract: Machine learning has broad applications in the finance industry. Risk Analytics, Consumer Analytics, Fraud Detection, and Stock Market Predictions are some of the domains where machine learning methods can be implemented. Accurate prediction of stock market returns is extremely difficult due to volatility in the market. The main factor in predicting a stock market is a high level of accuracy and precision. With the introduction of artificial intelligence and high computational capacity, efficiency has increased. In the past few decades, the highly theoretical and speculative nature of the stock market has been examined by capturing and using repetitive patterns. Various machine learning algorithms like Multiple Linear Regression, Polynomial Regression, etc. are used here. The financial data contains factors like Date, Volume, Open, High, Low Close, and Adj Close prices. The models are evaluated using standard strategic indicators RMSE and R2 score. Lower values of these two indicators mean higher efficiency of the trained models. Various companies employ different types of analysis tools for forecasting and the primary aim is the accuracy to obtain the maximum profit. The successful prediction of the stock will be an invaluable asset for the stock market institutions and will provide real-life solutions to the problems of the investors.

Keywords: Stock prices, Analysis, Accuracy, Prediction, Machine Learning, Regression, Finance

I. INTRODUCTION

The modern advances in artificial intelligence have led to the creation of new mathematical tools like deep learning and reinforcement learning. Businesses use data science and analytics to obtain solutions for various business issues. Stock trading is one of the most important activities in the finance world. Stocks are an equity investment that denotes a part of ownership in an organization or a company; it entitles you to be a part of that company's earnings and assets. Stock market prediction can be defined as trying to determine the future value of a stock or other financial instrument that is traded on a financial exchange.

The successful prediction of a stock's future price can lead to hefty profits. Financial information is a significant component of all electronic data. An average stock exchange creates nearly trillions of Gigabytes (GB) of trade and order book data in a month. In recent years, the rising popularity of machine learning in various industries has enlightened many traders to apply machine learning techniques to the field, and some of them have produced quite promising results. Machine Learning (ML) provides a unique perspective to us on understanding the stock market and financial data.

Customers can clearly determine whether it is worth investing in a particular stock. The pricing of the complex financial product needs many statistical computations and simulations based on the data. The computations are ready-to-use data resources and information for future applications. Algorithmic Trading not only just makes purchase or sell decisions but also recommends the product efficiently.

In Indian stock markets, a leading part of trading decisions is made using computer programs. For decades, computer algorithms have been built and tweaked to try to predict the stock market and make the appropriate investment at the right time. Machine Learning can additionally implement algorithms to identify unusual patterns of behaviour based on past behaviours. The stock market has an extremely volatile nature. The major goal is to minimize the uncertainty of the returns by accurately predicting the future stock prices and also identifying their fluctuations in advance to reduce risks.

II. ALGORITHMS USED

A. *Multiple Linear Regression*

Multiple linear regression is an extended version of the simple linear regression algorithm. The objective is the prediction of the value of a variable based on the value of two or more other variables. The independent variables are needed to predict the value of the dependent variable. The input independent variables could be of continuous or categorical type. The variable for the prediction is called the dependent variable. The regression coefficient means how the dependent variable changes due to a unit change in the independent variable.

This statistical method is widely implemented to accurately forecast the data, predict the outputs, correctly analyse the time series, and find the causal effect dependencies between the variables. It can be employed to identify the effect of the independent variables on a dependent variable. Practical applications of multiple regression models can be the prediction of CGPA in college from the CET score, Weather Forecasting, etc. Being a highly established statistical technique, Multiple linear Regression is used in stock market analysis.

B. Polynomial Regression

In statistics, a form of regression analysis in which the relationship between the dependent variable y and the independent variable x is modelled like an nth degree polynomial in x is called a Polynomial Regression. It has extra predictors which are obtained by increasing each of the original predictors to a power. The amount of higher-order terms rises with the increasing value of n. Users can fit a non-linear line to a data set using it. This is done via the use of higher-order polynomials such as square, cubic, quadratic, etc. to one or more predictor variables.

It is used for one predictor and one outcome variable in general cases. It is known to give a great defined relation between the independent and dependent variables. Applications can be the study of health outcomes in medicine, isotopes of the sediments, etc. The dataset used here to train is of a non-linear kind in nature. Considering the real world, the growth of a stock market is never linear like a line; polynomial regression can be used for prediction.

C. Decision Tree Regressor

A Decision Tree is a very prominent practical approach for supervised learning. It can be employed to solve both Regression and Classification tasks. Both continuous and categorical output variables can be predicted. This algorithm is very useful for resolving decision-related problems. Applications are evaluating growth opportunities for businesses, use of demographic data to find potential clients, etc. The features of an object are analysed to train a model as a tree and produce meaningful continuous output.

It mainly uses mean squared error to decide whether to split a node into two or more sub-nodes. The accuracy is dependent on the decision of making strategic splits. The top-most decision node in a decision tree is the root node. The parent node that splits into one or more child nodes is called a decision node. Removing the child nodes of a decision node is known as pruning. Bottom nodes that don’t split any further are the leaf nodes. The plotting is typically done in an upside-down manner so that the root node is at the top and the leaf nodes are at the bottom. When the dependent variable is continuous, this can be implemented.

D. Random Forest Regressor

Random forest is a supervised learning algorithm. It is an ensemble learning method for regression and classification problems. An ensemble method uniquely combines the predictions from many ML algorithms together to provide more accurate predictions than any individual model. You can get higher accuracy through cross-validation. Random forests do not overfit so the user can run as many trees as possible.

Users can work precisely on a large data set with higher dimensionality. Some notable applications are Diabetes Prediction, Product Recommendation, Bitcoin Price Detection, Predicting Loan Defaults, etc. The Meta estimator fits many classifying decision trees on numerous sub-samples of the dataset and uses the technique of averaging to shoot up the accuracy. Output predictions are generated by a combination of outcomes from a sequence of the regression decision trees. An output prediction is a mean of the predictions produced by the trees in the forest.

III.COMPARISON OF THE ALGORITHMS USED

Srno	Algorithms	Advantages	Disadvantages
1.	Multiple linear regression	• Find the relative influence of 1 or more predictor variables on the output variable. • Ability to find out the anomalies or outliers.	• It does not work if there is a correlation between the error terms or independent variables. • Incomplete data can cause many errors.
2.	Polynomial Regression	• It works well on nonlinear problems. • It is efficient on a dataset of any size.	• For a good bias/tradeoff value, an accurate value of ‘n’ needs to be found. • It is very sensitive to outliers.

3.	Decision Tree Regressor	• It is easy to interpret, understandable for beginners, and can visualize faster. • It can easily identify the relationship between dependent and independent variables. • There is no need to use feature scalability. • It can be used for linear as well as nonlinear problems.	• It does not give suitable results if the dataset is small. • Overfitting is possible. • The calculations might be hard so the time complexity might increase. • A small change can cause a big change in the final answer.
4.	Random Forest Regressor	• It is accurate, versatile, and powerful to use. • The performance is good on nonlinear problem statements. • It takes care of the null values.	• Overfitting is possible. • It is difficult to interpret. • You should correctly choose the number of trees. • A large number of trees can make it too slow.

Table 1: Algorithm Comparison

All the algorithms have their benefits and pitfalls. To choose an algorithm that perfectly finds the appropriate value of the stock in the future, it needs to be checked whether it aligns with the goals. The amount of pre-processing, the accuracy, how explainable it is and its speed should be checked before choosing it. It's also necessary to check the time complexity of the algorithm.

IV.METHODOLOGY

The entire code is in the format of the Python Notebook. Python libraries like Pandas, Numpy load the dataset and perform the mathematical calculations on the dataset. Sklearn is used to implement the four different machine learning algorithms. Matplotlib and Seaborn are needed to visualize the data in an interactive way. The historical data of the last 5 years was downloaded from the Yahoo finance website. The stock in consideration is Tata Consultancy Services - TCS.

The dataset available has the following attributes:

1) Date
2) Open
3) High
4) Low
5) Close
6) Volume
7) Adj. Close.

a) *Open Price:* First price of the stock at the beginning of a trading day
b) *Closing Price:* The price of the stock at the end of a trading day
c) *Adj Close:* Stock's value after distributing dividends (True value of the stock)
d) *High Price:* The highest price of the stock in the trading day
e) *Low Price:* Lowest price of the stock in the trading day
f) *Volume:* Number of stocks traded for security in all the markets during a given time.

The closing price over the time periods of 5 years can be seen in Figure 1.

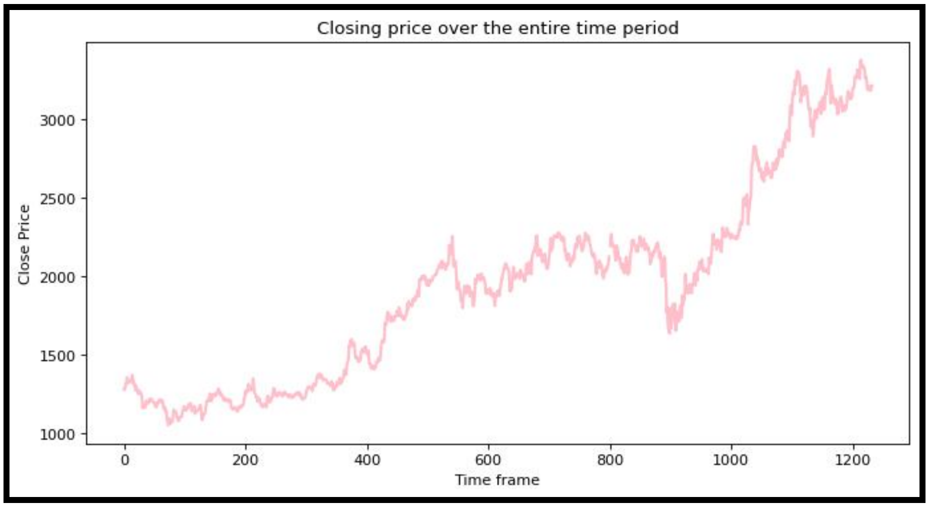

Fig 1: Visualizing closing price over 5 years

The daily change in the stock price over the time period of 5 years can be seen in Figure 2.

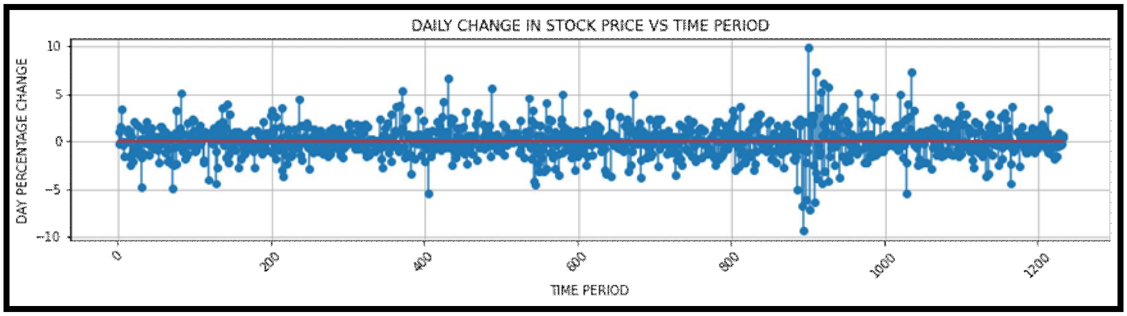

Fig 2: Visualizing Daily Change in Stock Price over 5 years

The histogram of Daily Percentage Change in the stock price can be seen in Figure 3.

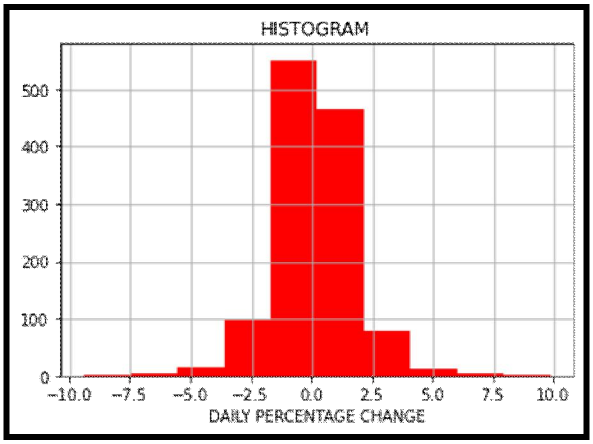

Fig 3: Histogram of Daily Percentage Change

There is no presence of any kind of duplicate rows in the dataset. There are 15 cells with missing values which were filled with values. A new column 'Day_Perc_Change' is created which denotes the percentage change in a day of the closing price.
Another new column 'HL_PCT' is created which denotes the difference of high and low price divided by the Adj Close Price.
Using feature importance, Day_Perc_Change, HL_PCT, Adj Close Price, and Volume columns are chosen as independent predictor variables. The dependent variable is the 'Label' column which is needed to be predicted. 80 per cent of the whole dataset is used to train the model. 20 per cent of the whole dataset is used to test the model.

V. RESULTS

The performance of regression models must be denoted in the form of an error. The performance might change in case of a change of datasets and yield different results. You cannot find the accuracy score. Root mean squared error (RMSE) and the coefficient of determination (R^2 error) can be used in this case.

1) *R^2 error:* It is the proportion of variance in the dependent variable that can be told by the independent variable. It is used to measure the goodness of fit. A larger value means that it is a better model.
2) *RMSE:* It is the standard deviation of the prediction errors i.e residuals. It gives you a relatively high weight to the large errors. A smaller value means it is a better model.

SR.NO	Algorithms	RMSE	R^2 error
1	Multiple linear regression	106.42	0.97
2	Polynomial Regression	102.29	0.97
3	Decision Tree Regressor	0.0	1.0
4	Random Forest Regressor	32.45	0.99

Table 2:Result Accuracy Analysis

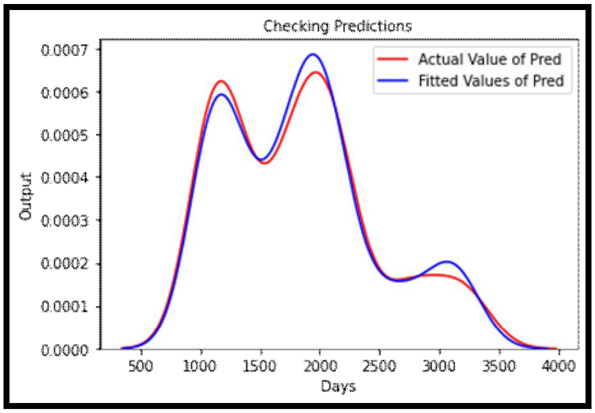

Figure 4: Actual vs Predicted Values DistPlot - Multiple Linear Regression

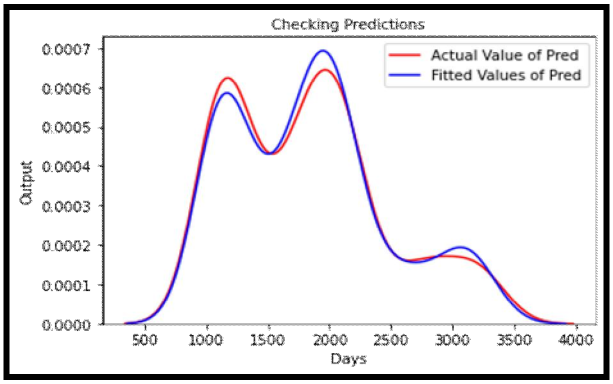

Figure 5: Actual vs Predicted Values DistPlot - Polynomial Regression

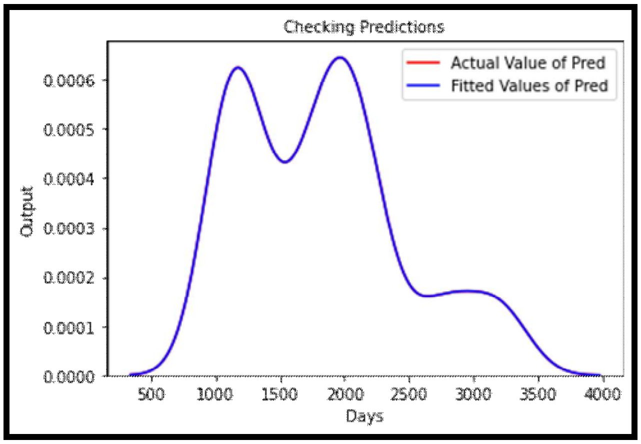

Figure 6: Actual vs Predicted Values DistPlot - Decision Tree Regressor

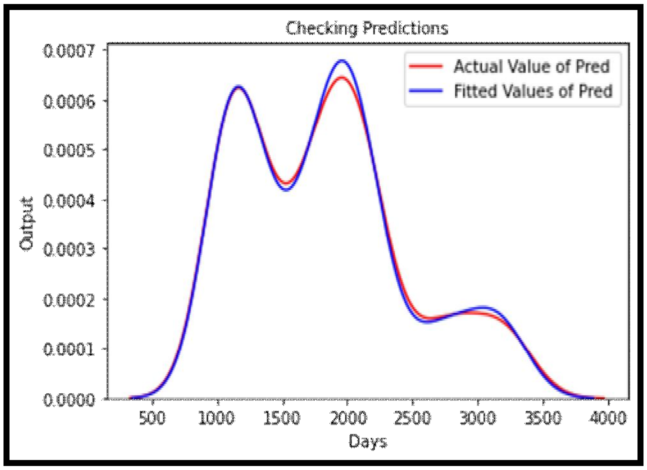

Figure 7: Actual vs Predicted Values DistPlot - Random Forest Regressor

The above four figures are the visualizations of the results found. The distribution plots help to evaluate the distribution of the predicted values of the data. The actual value is known as the value that is obtained by observation or by measuring the available data. The fitted value is known as the value of the variable predicted based on the analysis. The lower value of R Square indicates the weaker the goodness of fit of the model. For a perfect prediction, there are Fitted values = Actual values. In Figure 6: Actual vs Predicted Values DistPlot - Decision Tree Regressor, there is an exact mapping of both the curves. The R^2 error of the Decision Tree Regressor is 1.0 that means the predictions perfectly fit the data.

VI. ADVANTAGES

The prediction will increase the gains of the customer successfully. There is less chance of human error. The data ranges from the years 2016-2021 so the model trained is very reliable. Since 4 different methods are used to predict the same kind of output, the algorithm can be chosen with the highest accuracy and lowest error rate. The user finds it easy to use. It has a faster process completion rate. There is better handling of data and it is better in making decisions. It increases the productivity of the user.

VII. DISADVANTAGES

The accuracy might decrease if there are more levels of stock market movement. The results might be wrong if there is a larger dataset. There might be large variations due to political, economic, and social conditions like Covid-19. Using machine learning algorithms with high complexity can lead to slower training and computations which might upset the customers. Training with a small amount of data can lead to biased results. There is no guarantee that the prediction will be 100% accurate as there is uncertainty in the future.

VIII. FUTURE SCOPE

Even though good scores are achieved using ML algorithms, there can be an improvement. Adding more data helps the algorithm to learn better. Hyperparameter optimization is another method of tuning the hyperparameters to get the best performance on the data set provided. It can be implemented using the Scikit-learn machine learning library. The two famous algorithms which can be used are:

1) *Grid Search:* In this method, a search space as a bounded domain of hyperparameter values is defined then random points are sampled within the bounded domain.
2) *Random Search:* In this method, a search space as a bounded domain of hyperparameter values is defined, and then random points are sampled in the bounded domain.

Deep Learning algorithms can be implemented to predict accurate results. Deep learning is a branch of machine learning where neural networks algorithms are inspired by the human brain. Long-short term memory (LSTM) can be implemented to predict the stock price. LSTM can learn order dependence in sequence prediction problems. Artificial Neural Network (ANN) is also an extremely recognized method for predictive finance. ANNs are multi-layer fully connected neural nets. Convolutional Neural Networks (CNN) are made up of neurons with biases and learnable weights. CNNs, which are designed to map image data to an output variable, can help to improve predictions.

The future prospects include building a Machine learning web app in Python where the user can simply input a stock dataset and get appropriate output with the highest accuracy. The machine app should take in the dataset correctly and choose the algorithm that gives the lowest error rate. The predictions should get printed on the screen. The user interface should be easy and user-friendly for beginners. The app can then be deployed on servers like Heruko to see the model in action.

IX. CONCLUSIONS

Machine learning has applications related to the recommendation of financial products, customer sentiment analysis, etc. In order to predict the prices of stocks, we need the historical data of the stock. This paper analyses the Data Set with 7 attributes and makes a prediction using different regressors to find the future price. It can be seen that the highest accuracy is obtained using the Decision Tree Regressor model with the R^2 error being 1.0 and the RMSE being 0.0. We can summarize by saying that the Decision Tree Regressor gave the best results out of all the models used for the stock price prediction.

REFERENCES

[1] Gareja Pradip, Chitrak Bari, J. Shiva Nandhini, "Stock market prediction using machine learning" International Journal of Advance Research and Development, Volume 3, Issue 10, 2018

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

[3] K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, Sarmistha Padhi, Saurav Sanyal, "Stock Market Prediction Using Machine Learning Algorithms", International Journal of Engineering and Advanced Technology (IJEAT), Volume-8 Issue-4, 2019

[4] Raut Sushrut Deepak, Shinde Isha Uday, Dr. D. Malathi, "MACHINE LEARNING APPROACH IN STOCK MARKET PREDICTION", International Journal of Pure and Applied Mathematics, Volume 115, No. 8, 2017.

[5] Mehtab, S., Sen, J., A robust predictive model for stock price prediction using deep learning and natural language processing. In: Proceedings of the 7th International Conference on Business Analytics and Intelligence

[6] M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques", 2016 3rd International Conference on Computer and Information Sciences (ICCOINS)

[7] Tang, J., Chen, X., Stock market prediction based on historic prices and news titles. In: Proceedings of the International Conference on Machine Learning Technologies (ICMLT)

[8] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017

[9] Sachin Sampat Patil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, “A Survey on Stock Market Prediction Using SVM”, IJCTET 2016.

[10] https://marutitech.com/ai-and-ml-in-finance/

[11] https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

[12] https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a

[13] https://www.javatpoint.com/machine-learning-polynomial-regression

[14] https://www.scribbr.com/statistics/multiple-linear-regression/