# Stock Price Prediction

**N P Samarth, Gowtham V Bhat, Hema N**

***Abstract*: Stock trading is a very crucial activity in the world of Finance and is a supporting structure for many companies. Predicting the future value of a stock is the main goal of stock price prediction project. In this paper, we have used machine learning algorithms to predict future stock prices of a company. Stock prediction by the stock brokers is mainly done using the time series or the technical and fundamental analysis but as these techniques are very unreliable and limited, we propose making use of intelligent techniques such as machine learning. Python is a programming language which can be used to implement machine learning algorithms with its numerous inbuilt libraries. We propose an approach that uses machine learning algorithms and will be trained on the historical stock data that is available and gain intelligence, later it uses the knowledge acquired for predicting the stock prices accurately. Random Forest Regression is one of the machine learning technique that is used for stock price prediction for small and large capitalizations also in different markets employing both up-to-minute and daily frequencies.***

***Keywords*: Machine Learning, Random Forest Regression, Stock Market, Predictions.***

## I. INTRODUCTION

The traders who possess a lot of money buy equities and also stock derivatives when they are available at very cheap cost. Later these stocks are sold back once they attain high prices. Various organizations have been discussing the issue of growing trend in stock market predictions. Before investing in a stock, the investors would analyze the stocks in two ways, one of those techniques is technical analysis which is an evolution of stocks on studying the previous volumes, prices and also market activity that is used to generate the statistics. The other one would be fundamental analysis, in which the investors would look upon the intrinsic value of a stock, economy, performance by the industry and also the political climate. From past few years', traders have shown interest in making use of machine learning for stock price predictions due to its increasing prominence and display of promising results and reliability. Using this, they have been producing very good output with high accuracy. In this paper we discuss about developing predictor programs for the financial market. The dataset consists of previous stock prices, which is used in order to train the program. Reduction the uncertainties and risk factor with respect to decision making is the main goal of stock price predictions.

The very best way of predicting tomorrow's stock price would be by knowing today's stock price. Obtaining an accurate forecast model is very challenging due to the volatility of the stock market. Investor's sentiments are affected by the high fluctuation of stock market indices. The dynamic nature of stock prices is prone to quick changes due to the nature of financial domain and a mixture of other known factors like closing price of previous day, volumes etc., also a few unknown factors such as rumors, elections and many more. Several attempts have been made in the field of predicting the stock prices using machine learning. The target of every research project would vary in two different ways. (a) the change in price can be long term, i.e., months, years. Short term which is a single day to a couple of days. Then it can be near term which is less than a minute to seconds. (b) The predictors used may range from the economic trend and global news, to a pure time series analysis of stock price of the company, to the characteristics of the company for all the stocks.

The goal of predicting the stock price may depend upon the raising or down fall of the market also called as the volatility of the shares. The investment in the stock market is always subjected to market risks.

In our project we have used random forest regression in order to predict the future stock price of the company. The random forest regression which is a collection of decision trees produces different possible prices of the stock. On building several such trees we end up implementing random forest regression which is one among the most effective machine learning regression algorithms.

## II. PROBLEM DEFINITION

Essentially, stock market forecasting is described as attempting to evaluate stock value and providing people with a comprehensive sense of understanding and projection of the economy. The task of predicting the stock market performance is quite a difficult problem. The actions of the stock market are usually determined by thousands of shareholders' opinions. The effects of any events on the investors can predicted based on the analysis of the stock market. The events can be political events such as a speech by a political leader, fake news coverage, etc. It can also be an international event such as fast currency and product movements. All these incidents have an effect on corporate earnings and this once again affects the investor sentiment. To forecast such hyperparameters accurately and consistently is almost unachievable to majority of the investors. It is usually presented with the help quarterly financial ratio data.

**Revised Manuscript Received on December 15, 2019.**

**Mr. N P Samarth**, Department of Information Science and Engineering, RNS Institute of Technology, Bengaluru, Karnataka, India.

**Mr. Gowtham V Bhat**, Department of Information Science and Engineering, RNS Institute of Technology, Bengaluru, Karnataka, India.

**Mrs. Hema N**, Assistant Professor, Department of ISE, Information Science and Engineering, RNS Institute of Technology, Bengaluru, Karnataka, India.

The outcome may be incorrect if the forecast is based on an individual dataset. Therefore, we are considering different data set combinations for the purpose of predicting the variations and demand and the patterns of stocks. We use this dataset to train a model based on a machine learning algorithm. The problem of predicting the stock prices cannot be solved unless and until the proposed algorithm is accurate and reliable. Considering all of these factors it is clear that the prediction of stock prices is not an easy task.

## III. LITERATURE SURVEY

We have collected information of the current methods practiced in the field of stock market prediction during the process of literature survey.

In [1] the author conducts a survey for the purpose of stock market prediction. In the present time, the prediction of stock market has gained a lot of importance. Technical analysis is one of the techniques used which is basic in nature and without the guarantee of results that are accurate. Hence, there is a great need to develop techniques that are more reliable. A large portion of the investments are dependent on these predictions which takes into account all the influencing variables. The method used in this case was a regression. Because financial market marks at any given time produce huge amounts of data, a large volume of information has to be processed before an estimate can be made.

Each regression listed technique has some advantages and disadvantages when compared with other techniques. Linear regression is such a notable technique that is listed. This technique generally operates by fitting the least square method. They can operate by decreasing a disabled variation of the least square loss function. In addition, the solution obtained from least square can be suited to models that are non-linear.

In [2] the author discusses the increasing trend in the application of artificial intelligence and machine learning fields in stock market predictions. More and more data scientists are spending their time each day in finding ways to arrive at approaches that can further boost the market forecast model's accuracy. There are many number of ways this problem can be approached, so it is difficult to find a straight forward and universal method which suits all genres of stock data. Even if the same data set is used, the output varies for each technique. The quoted paper tries to predict the stock price using the previous quarter's financial data and by application of Random Forest algorithm.

This is not the only factor influencing the stock prices. Factors such as corporate public sentiment, investors' opinions, different news agencies and other events such as elections, wars, etc., influence the stock market. Even the changes in the company management, changes in the services offered by the company greatly influences the company's stock prices. The accuracy of the prediction for stock prices can be increased by accurate dataset and a model which takes these factors into consideration.

According to [3], the stock market prediction is a very difficult and challenging problem. This paper proposes a new approach of using the "modern web" as a tool to solve this problem. Because of the interconnected nature of the web, extraction of data influencing the stock market is made easier. It also helps in introducing relationships between the investment patterns and the influencing factors. Investment trends from different companies show correlation, and the trick to forecasting the stock market effectively is by leveraging the similar properties displayed by different data sets. By making use of new tools such as sentiment analyzers instead of depending on the orthodox methods of analysis of historic data, we can derive a significant connection between the emotions of the people and the various factors which influence them. Another significant segment of the prediction process was the effective usage of web as a source for extracting events which influenced the stock market patterns.

In [4] author discusses the volatility in the methodology of forecasting the stock market is filled and the several variables which may affect it. Therefore, in business and finance, the stock market plays an important role. The scientific and basic research is conducted through the method of emotional study. Thanks to its expanded use, social media information has a high impact and can be useful in forecasting the stock market price. Using machine learning algorithms on historical stock price data, technical analysis is done. Usually, the method involves collecting data from different social media platforms, extracting individual sentiments from news outlets. Information such as stock prices from the past year are also considered. This model takes into account the interdependency of the numerous data points and the output is provided based on all these factors.

In [5] author discusses many financial companies and retailers have developed private applications to move ahead in the competition of stock market prediction for the purpose of increasing profits, but in recent times nobody has reached consistently higher than normal rates of productivity. Nevertheless, the challenge of predicting stock prices is so involving because the improvement in just some factors can substantially benefit these organizations and bring large figure of profit.

Lot of research is being done in financial institutions on the problem of prediction of stock prices using time series. Specific computational models of the time series are based on machine learning. The SVM was built for non-linear classification and time series analysis to solve regression problems. The error of generalization in the model is reduced by making use of minimization theory by making use of approximation methods [6]. Therefore, the ICA methodology extracts from the database many important features. The prediction of the time series is based on SVM.

Different types of analysis tools including Time series analysis is being used by many stock brokers for the purpose of stock market predictions. But these methods are not completely reliable, so there was a need to introduce a more reliable alternative. So, a new methodology was developed using AI and machine learning [7]. This method was coupled with a supervised classifier in order to find accurate results.

The results obtained were verified using binary classification and the classifier available in Support Vector Machines and using a different set of attributes.

Most of the Machine Learning method has derived from empirical approaches that have omitted Artificial Intelligent, even when there were appropriate protocols for individual issues. The SVM parameters were easily handled by using Cuckoo scan, Swarm Intelligence method used for optimization. In contrast to ANN, the hybrid CS-SVM strategy proposed displayed promising results with high accuracy. Similarly, in predicting the stock values, the CS-SVM screen performed better. Predicting inventory costs used parse documents to measure the prediction, provide the data to the consumer but the exchanges of stocks were completely automatized.

In [8] the author proposes that an organizations' development can be predicted by closely monitoring the patterns in communication. It introduces a methodology which helps in effectively predict a company's performance. This methodology tries to find interconnected factors in the various emails that are sent between important employees and the stock value of the company. The author also proposes to conduct a test of the open source data mining algorithm applied to the Enron Corp dataset, which is open to interested researchers. The company under consideration has itself invested in energy, other commodities and services. The dataset mentioned above is open for everybody.

## IV. PROPOSED SYSTEM

The paper proposes the use of Random Forest Algorithm which is a machine learning regression algorithm in the field of stock market forecasting. We have used the Random Forest algorithm to predict the stock market value of the Walt Disney Company. The model displayed high accuracy. We were able to train the machine on different attributes of the historical data of the company, in this proposed system and predict future stock market prices of the company. We've taken stock information from the year 2013 to 2017 to train the model. To solve the problem, we primarily used two machine learning libraries. The dataset was first operated on by the numpy library, which is used to clean and modify the dataset and make it a ready-to-analyze data. The other library used was SciKit. It provides a large number of inbuilt machine learning modules and different methods to evaluate and process the dataset. The data set was downloaded from an online database which is open to the public. The data set was divided into 8:2 ratio. The first 80 percent was utilized to train the regression model and the rest 20 percent was used to check the accuracy of the trained model. The basic approach of the supervised learning system is to learn from the training set and then replicate the patterns and associations in the data for the test data. For data processing, we used the python "pandas" library which merged different datasets into a data frame. The updated data frame allowed us to prepare the data for extraction of the functionality. We checked the accuracy of the model by comparing the predicted values against the original values.

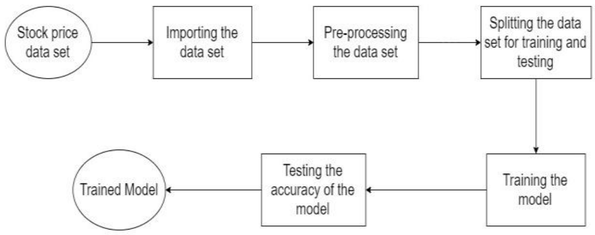

Stock price data set
Importing the data set
Pre-processing the data set
Splitting the data set for training and testing
Training the model
Testing the accuracy of the model
Trained Model

**Figure IV.1 System Architecture**

## V. METHODOLOGY

The Random Forest Regression algorithm is used for the prediction of the stock prices in this project.

### I. Decision Trees

Decision Trees are classifiers. They are a directed tree structure consisting of a group of nodes. The topmost node is called the "root". There are no incoming edges for this node. The nodes other than the root has one incoming node and these nodes are divided into "internal nodes" and "leaves". As the name suggests, the decision trees output a decision which is the result of a sequence of tests performed at each level of the tree.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where, E(S) – entropy of the sample,
$P_i$ – probability of occurrence of event "i".

### II. Random Forest

Random Forest is also known as random decision forests. It is a popular algorithm which can be applied to problems of classification and regression alike. It is a variant of ensemble method. The ensemble methods make use of more than one learning model and by integrating these models it is able to come at a better result than all the individual models.

Random Forest is a supervised learning algorithm. This algorithm creates a large number of decision trees. All the trees are uncorrelated with each other. It makes use of a statistical technique called bagging, formally known as Bootstrap Aggregation.

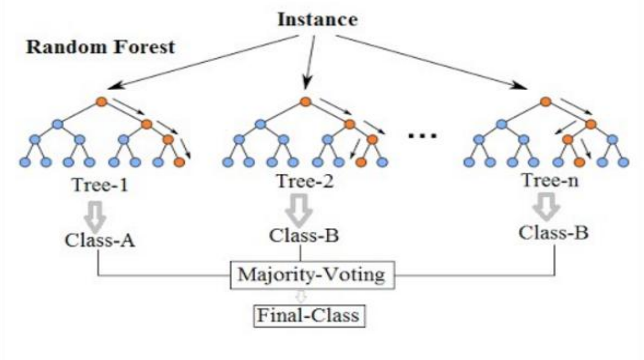


**Figure V.1 Random Forest**

As shown in the Figure 5.1 The algorithm operates by constructing multiple decision trees during the time of training. Each decision tree is trained on different sample of the original dataset. Hence the decision trees become independent of each other. By combining the results from all these decision trees, the final output will be more reliable than any individual decision tree.

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j\in all\ features, k\in all\ trees} normfi_{jk}}$$

Where,
RFfi$_i$ = from all trees in the Random Forest model importance of feature i is calculated.
normfi$_{ij}$ = in tree j the importance of normalized feature for i

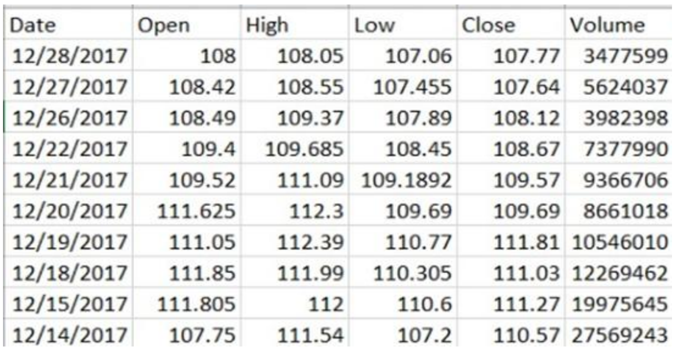
## III. Importing the required libraries

We are making use of several built-in libraries available in Python language. SciKit-learn library includes algorithms of various types such as regression, classification and also clustering algorithms. It provides algorithms such as Random Forest, Support Vector Machines, Naïve Bayes, Gradient Boosting etc.,

We are further making use of NumPy library to manage large matrices and arrays of multiple dimensions. It also includes a large collection of useful functions to operate on the arrays. We are further making use of Pandas, which are used for operations on time series and numerical tables. Matplotlib is used for plotting the data by the help of general-purpose GUI toolkits. We import all libraries mentioned above make use of the required classes.

## IV. Importing and preprocessing the data

Datasets are an integral part of field of machine learning and play a important role in training a efficient model. As shown in Figure 5.2 our dataset consists of 1090 stock records of the Walt Disney Company. Each record consists of 6 values respective to high, low, open, close, date and volume.



| Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|
| 12/28/2017 | 108 | 108.05 | 107.06 | 107.77 | 3477599 |
| 12/27/2017 | 108.42 | 108.55 | 107.455 | 107.64 | 5624037 |
| 12/26/2017 | 108.49 | 109.37 | 107.89 | 108.12 | 3982398 |
| 12/22/2017 | 109.4 | 109.685 | 108.45 | 108.67 | 7377990 |
| 12/21/2017 | 109.52 | 111.09 | 109.1892 | 109.57 | 9366706 |
| 12/20/2017 | 111.625 | 112.3 | 109.69 | 109.69 | 8661018 |
| 12/19/2017 | 111.05 | 112.39 | 110.77 | 111.81 | 10546010 |
| 12/18/2017 | 111.85 | 111.99 | 110.305 | 111.03 | 12269462 |
| 12/15/2017 | 111.805 | 112 | 110.6 | 111.27 | 19975645 |
| 12/14/2017 | 107.75 | 111.54 | 107.2 | 110.57 | 27569243 |

**Figure V.2 Snapshot of the dataset**

Each tuple represents the data of a particular day. In any tuple "Open" attribute refers to the opening price of the stock on that particular day. "Close" attribute refers to closing price. "High" refers to the highest price achieved by the stock on that day, "Low" refers to the lowest price of the stock. Volume represents the number of shares transacted on that particular day.

We import the data set and check the shape of the dataset. By doing so we get an idea about the size and dimensions of the data. We further proceed to clean the data by removing any anomalies in the data and also removing any outlier data points. In case of missing data, we make use of inbuilt functions to fill an assumed data in the respective vacant place.

The dataset is divided in 8:2 ratio, where the larger part is used to train the Random Forest Model. The rest 20% of the dataset is used to check the accuracy of the model. So, this part of the dataset plays no role to in training of the model.

## V. Training the model

Supervised learning is a method where both the inputs and outputs are provided for training the model. After the input is processed, we obtain some output which is compared with the desired output. The errors generated propagates back through the system. This causes the system to adjust the weights and hence control training process of the model.

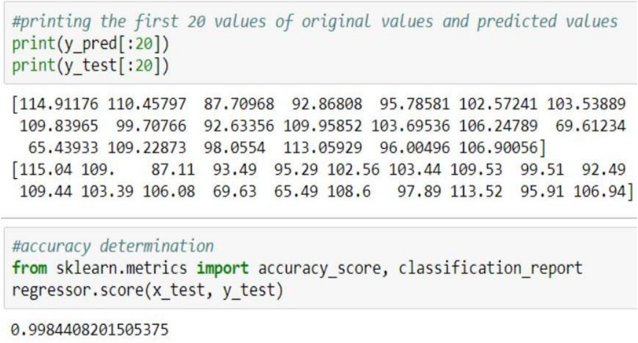
The first 80% of the data is the training set. We are creating a "RandomForestRegressor" object that makes use of 1000 decision trees. So, the final output of the model would the mean of all these decision trees. In general, with the increase in the number of decision trees, the output will become more reliable and accurate. But this may not be the case for all datasets.

We train the model by passing the both input variables and target variable of the training dataset. The model after training is used to predict future values.

After the training completes, we pass the rest 20 percent of input variables to the model and ask it to predict the respective target variable for them. Whatever target variable it predicts is the output generated by the trained model.

Then to check the working efficiency of the model we need to compare the predicted values obtained against the actual real-world values. We check the accuracy of our model by making use the inbuilt function 'score()' of a Random Forest Regressor model. We have considered the threshold value of 90% accuracy.

## VI. RESULT SNAPSHOT



```
#printing the first 20 values of original values and predicted values
print(y_pred[:20])
print(y_test[:20])
```

```
[114.91176 110.45797  87.70968  92.86808  95.78581 102.57241 103.53889
 109.83965  99.70766  92.63356 109.95852 103.69536 106.24789  69.61234
  65.43933 109.22873  98.0554  113.05929  96.00496 106.90056]
[115.04 109.     87.11  93.49  95.29 102.56 103.44 109.53  99.51  92.49
 109.44 103.39 106.08  69.63  65.49 108.6   97.89 113.52  95.91 106.94]
```

```
#accuracy determination
from sklearn.metrics import accuracy_score, classification_report
regressor.score(x_test, y_test)
```

```
0.9984408201505375
```

**Figure VI.1 Snapshot of the predicted values and accuracy**

We can see that the Random Forest Algorithms is best suitable for the dataset that we have considered. The algorithm is successful in predicting the stock prices very close to the original values of the test data.

As shown in the Figure 6.1 the model has attained an accuracy of 99.84%.

## VII. CONCLUSION

In the proposed model we have processed the stock data of the Walt Disney Company. This model is successful in predicting the futures stock prices with a good accuracy rate depending on the various attributes given by the user in the collected data set, thereby eliminating the human error as the decision process is successfully automated.

It shows Random Forest modeling is fairly suitable for the given dataset. For datasets with a small number of variables it is a viable method. This project is a good introduction to training and test datasets which are very important components not just to data science but to statistical learning overall.

## REFERENCES

1. Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017.
2. Loke.K.S. “Impact of Financial Ratios and Technical Analysis On Stock Price Prediction Using Random Forests”, IEEE, 2017.
3. Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, “Stock Market Prediction via Multi-Source Multiple Instance Learning.” IEEE 2018.
4. Vivek Kanade, Bhausaheb Devikar, Sayali Phadatare, Pranali Munde, Shubhangi Sonone. “Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2017.
5. Sachin Sampat Patil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, “A Survey on Stock Market Prediction Using SVM”, IJCTET 2016 Carlos A. Coello, Gary B. Lamont, David A. van Veldhuizen: “Evolutionary Algorithms for Solving Multi-Objective Problems”,Springer, 2007.
6. Hakob GRIGORYAN, “A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)”, DSJ 2016.
7. Raut Sushrut Deepak, Shinde Isha Uday, Dr. D. Malathi, “Machine Learning Approach In Stock Market Prediction”, IJPAM 2017.
8. Pei-Yuan Zhou , Keith C.C. Chan, Member, IEEE, and Carol Xiaojuan Ou, “Corporate Communication Network and Stock Price Movements: Insights From Data Mining”, IEEE 2018.
9. Chun C, Qinghua M, Shuqiang L.: “Research on Support Vector Regression in the Stock Market Forecasting” ©Springer, Advances in Intelligent and Soft Computing Volume 148, , pp 607-612, 2012.
10. Guo Z., Wang H., Liu Q. : “Financial time series forecasting using LPP and SVM optimized by PSO” © Springer, Soft Computing Methodologies and Applications , December 2012.
11. Xie, G.: “The Optimization of Share Price Prediction Model Based on Support Vector Machine”, International Conference on Control, Automation and Systems Engineering (CASE), pp.1-4., 30-31 July 2011.
12. Huang C.; Huang L.; Han T.: “Financial time series forecasting based on wavelet kernel support vector machine” IEEE, Eighth International Conference on Natural Computation (ICNC), 2012.

## AUTHORS PROFILE

**Mr. N P Samarth**, is an engineering graduate and obtained his degree from RNS Institute of Technology in Information Science and Engineering. He is actively involved in projects and research activities

**Mr. Gowtham V Bhat**, is an engineering graduate and obtained his degree from RNS Institute of Technology in Information Science and Engineering. He is actively involved in research projects.

**Mrs. Hema N**, currently working as an Assistant Professor, Dept. of ISE, RNSIT. She is having Teaching experience of 10 years and pursuing Ph.D. in the area of Medical Image Processing.