Stock Market Prediction Using Machine Learning Algorithms

K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, SarmisthaPadhi, Saurav Sanyal

***Abstract:** The main objective of this paper is to find the best model to predict the value of the stock market. During the process Of considering various techniques and variables that must be taken into account, we found out that techniques like random forest, support vector machine were not exploited fully. In, this paper we are going to present and review a more feasible method to predict the stock movement with higher accuracy. The first thing we have taken into account is the dataset of the stock market prices from previous year. The dataset was pre-processed and tuned up for real analysis. Hence, our paper will also focus on data preprocessing of the raw dataset. Secondly, after pre-processing the data, we will review the use of random forest, support vector machine on the dataset and the outcomes it generates. In addition, the proposed paper examines the use of the prediction system in real-world settings and issues associated with the accuracy of the overall values given. The paper also presents a machine-learning model to predict the longevity of stock in a competitive market. The successful prediction of the stock will be a great asset for the stock market institutions and will provide real-life solutions to the problems that stock investors face.*

***Keywords:** Machine Learning, Data Pre-processing, Data Mining, Dataset, Stock, Stock Market.*

I. INTRODUCTION

The stock market is basically an aggregation of various buyers and sellers of stock. A stock (also known as shares more commonly) in general represents ownership claims on business by a particular individual or a group of people. The attempt [3] to determine the future value of the stock market is known as a stock market prediction. The prediction is expected to be robust, accurate and efficient. The system must work according to the real-life scenarios and should be well suited to real-world settings. The system is also expected to take into account all the variables that might affect the stock's value and performance. There are various methods and ways of implementing the prediction system like Fundamental Analysis, TechnicalAnalysis, Machine Learning, Market Mimicry, and Time series aspect structuring. With the advancement of the digital era, the prediction has moved up into the technological realm. The most prominent and [3] promising technique involves the use of Artificial Neural Networks, Recurrent Neural Networks, that is basically the implementation of machine learning. Machine learning involves artificial intelligence which empowers the system to learn and improve from past experiences without being programmed time and again. Traditional methods of prediction in machine learning use algorithms like Backward Propagation, also known as Backpropagation errors. Lately, many researchers are using more of ensemble learning techniques. It would use low price and time [3] lags to predict future highs while another network would use lagged highs to predict future highs. These predictions were used to form stock prices. [1]

Stock market price prediction for short time windows appears to be a random process. The stock price movement over a long period of time usually develops a linear curve. People tend to buy those stocks whose prices are expected to rise in the near future. The uncertainty in the stock market refrain people from investing in stocks. Thus, there is a need to accurately predict the stock market which can be used in a real-life scenario. The methods used to predict the stock market includes a time series forecasting along with technical analysis, machine learning modeling and predicting the variable stock market. The datasets of the stock market prediction model include details like the closing price opening price, the data and various other variables that are needed to predict the object variable which is the price in a given day. The previous model used traditional methods of prediction like multivariate analysis with a prediction time series model. Stock market prediction outperforms when it is treated as a regression problem but performs well when treated as a classification. The aim is to design a model that gains from the market information utilizing machine learning strategies and gauge the future patterns in stock value development. The Support Vector Machine (SVM) can be used for both classification and regression. It has been observed that SVMs are more used in classification based problem like ours. The SVM technique, we plot every single data component as a point in n-dimensional space (where n is the number of features of the dataset available) with the value of feature being the value of a particular coordinate and, hence classification is performed by finding the hyperplane that differentiates the two classes explicitly.

Revised Manuscript Received on April 25, 2019.

Mrs. K. Hiba Sadia, Asst. Prof., Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Aditya Sharma, B.Tech Student**,** Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Adarrsh Paul , B.Tech Student**,** Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram, Chennai, India.

SarmisthaPadhi, B.Tech Student**,** Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Saurav Sanyal, B.Tech Student**,** Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Predictive methods like Random forest technique are used for the same.The random forest algorithm follows an ensemble learning strategy for classification and regression.The random forest takes the average of the various subsamples of the dataset, this increases the predictive accuracy and reduces the over-fitting of the dataset.

II. PROBLEM DEFINITION

Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the stock prices. It is generally presented using the quarterly financial ratio using the dataset. Thus, relying on a single dataset may not be sufficient for the prediction and can give a result which is inaccurate. Hence, we are contemplating towards the study of machine learning with various datasets integration to predict the market and the stock trends.

The problem with estimating the stock price will remain a problem if a better stock market prediction algorithm is not proposed. Predicting how the stock market will perform is quite difficult. The movement in the stock market is usually determined by the sentiments of thousands of investors. Stock market prediction, calls for an ability to predict the effect of recent events on the investors. These events can be political events like a statement by a political leader, a piece of news on scam etc. It can also be an international event like sharp movements in currencies and commodity etc. All these events affect the corporate earnings, which in turn affects the sentiment of investors. It is beyond the scope of almost all investors to correctly and consistently predict these hyperparameters. All these factors make stock price prediction very difficult. Once the right data is collected, it then can be used to train a machine and to generate a predictive result.

III. LITERATURE SURVEY

During a literature survey, we collected some of the information about Stock market prediction mechanisms currently being used.

1.Survey of Stock Market Prediction Using Machine Learning Approach

The stock market prediction has become an increasingly important issue in the present time. One of the methods employed is technical analysis, but such methods do not always yield accurate results. So it is important to develop methods for a more accurate prediction. Generally, investments are made using predictions that are obtained from the stock price after considering all the factors that might affect it. The technique that was employed in this instance was a regression. Since financial stock marks generate enormous amounts of data at any given time a great volume of data needs to undergo analysis before a prediction can be made. Each of the techniques listed under regression hasits own advantages and limitations over its other counterparts. One of the noteworthy techniques that were mentioned was linear regression. The way linear regression models work is that they are often fitted using the least squares approach, but they may alternatively be also be fitted in other ways, such as by diminishing the "lack of fit" in some other norm, or by diminishing a handicapped version of the least squares loss function. Conversely, the least squares approach can be utilized to fit nonlinear models. [1]

2.Impact of Financial Ratios and Technical Analysis on Stock Price Prediction Using Random Forests

The use of machine learning and artificial intelligence techniques to predict the prices of the stock is an increasing trend. More and more researchers invest their time every day in coming up with ways to arrive at techniques that can further improve the accuracy of the stock prediction model. Due to the vast number of options available, there can be n number of ways on how to predict the price of the stock, but all methods don’t work the same way. The output varies for each technique even if the same data set is being applied. In the cited paper the stock price prediction has been carried out by using the random forest algorithm is being used to predict the price of the stock using financial ratios form the previous quarter. This is just one wayof looking at the problem by approaching it using a predictive model, using the random forest to predict the future price of the stock from historical data. However, there are always other factors that influence the price of the stock, such as sentiments of the investor, public opinion about the company, news from various outlets, and even events that cause the entire stock market to fluctuate. By using the financial ratio along with a model that can effectively analyze sentiments the accuracy of the stock price prediction model can be increased. [2]

3.Stock Market Prediction via Multi-Source Multiple Instance Learning

Accurately predicting the stock market is a challenging task, but the modern web has proved to be a very useful tool in making this task easier. Due to the interconnected format of data, it is easy to extract certain sentiments thus making it easier to establish relationships between various variable and roughly scope out a pattern of investment. Investment pattern from various firms show sign of similarity, and the key to successfully predicting the stock market is to exploit these same consistencies between the data sets. The way stock market information can be predicted successfully is by using more than just technical historical data, and using other methods like the use of sentiment analyzer to derive an important connection between people’s emotions and how they are influenced by investment in specific stocks. One more important segment of the prediction process was the extraction of important events from web news to see how it affected stock prices. [3]

4. Stock Market Prediction: Using Historical Data Analysis

The stock market prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the stock market plays an important role in business and finance. The technical and fundamental analysis is done by sentimental analysis process. Social media data has a high impact due to its increased usage, and it can [6] be helpful in predicting the trend of the stock market. Technical analysis is done [6] using by applying machine learning algorithms on historical data of stock prices. The method usually involves gathering various social media data, news to extract sentiments expressed by individuals. Other data like previous year stock prices are also considered. The relationship between various data points is considered, and a prediction is made on these data points. The model was able to make predictions about future stock values.

5. A Survey on Stock Market Prediction Using SVM

The recent studies provide a well-grounded proof that most of the predictive regression models are inefficient in out of sample predictability test. The reason for this inefficiency was parameter instability and model uncertainty. The studies also concluded the traditional strategies that promise to solve this problem. Support vector machine commonly known as SVM provides with the kernel, decision function, and sparsity of the solution. It is used to learn polynomial radial basis function and the multi-layer perceptron classifier. It is a training algorithm for classification and regression, which works on a larger dataset. There are many algorithms in the market but SVM provides with better efficiency and accuracy. The correlation analysis between SVM and stock market indicates strong interconnection between the stock prices and the market index.

6. Predicting Stock Price Direction Using Support Vector Machines

Financial organizations and merchants have made different exclusive models to attempt and beat the market for themselves or their customers, yet once in a while has anybody accomplished reliably higher-than-normal degrees of profitability. Nevertheless, the challenge of stock forecasting is so engaging in light of the fact that the improvement of only a couple of rate focuses can build benefit by a large number of dollars for these organizations. [6]

7. A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)

The time series prediction problem was researched in the work centers in the various financial institution. The prediction model, which is based on SVM and independent analysis, combined called SVM-ICA, is proposed for stock market prediction. Various time series analysis models are based on machine learning. The SVM is designed to solve regression problems in non-linear classification and time series analysis. The generalization error is minimized using an approximate function, which is based on risk diminishing principle. Thus, the ICA technique extracts various important features from the dataset. The time series prediction is based on SVM. The result of the SVM model was compared with the results of the ICA technique without using a preprocessing step.

8. Machine Learning Approach In Stock Market Prediction

The vast majority of the stockbrokers while making the prediction utilized the specialized, fundamental or the time series analysis. Overall, these techniques couldn't be trusted completely, so there emerged the need to give a strong strategy to financial exchange prediction. To find the best accurate result, the methodology chose to be implemented as machine learning and AI along with supervised classifier. Results were tried on the binary classification utilizing SVM classifier with an alternate set of a feature list. The greater part of the Machine Learning approach for taking care of business [2] issues had their benefit over factual techniques that did exclude AI, despite the fact that there was an ideal procedure for specific issues. Swarm Intelligence [2] optimization method named Cuckoo search was most easy to accommodate the parameters of SVM. The proposed hybrid CS-SVM strategy exhibited the performance to create increasingly exact outcomes in contrast with ANN. Likewise, the CS-SVM display [2] performed better in the forecasting of the stock value prediction. Prediction stock cost utilized parse records to compute the predicted, send it to the user, and autonomously perform tasks like buying and selling shares utilizing automation concept. Naïve Bayes Algorithm was utilized. [8]

9. Corporate Communication Network and Stock Price Movements: Insights from Data Mining

This paper tries to indicate that communication patterns can have a very significant effect on an organization’s performance. This paper proposed a technique to reveal the performance of a company. The technique deployed in the paper is used to find the relationships between the frequencies of email exchange of the key employees and the performance of the company reflected in stock values. In order to detect association and non-association relationships, this paper proposed to use a data mining algorithm on a publicly available dataset of Enron Corp. The Enron Corporation was an energy, commodities, and services company based in Houston, Texas whose stock dataset is available for public use. [9]

IV. DISADVANTAGES OF THE EXISTING SYSTEM

- The existing system fails when there are rare outcomes or predictors, as the algorithm is based on bootstrap sampling.
- The previous results indicate that the stock price is unpredictable when the traditional classifier is used.
- The existence system reported highly predictive values, by selecting an appropriate time period for their experiment to obtain highly predictive scores.
- The existing system does not perform well when there is a change in the operating environment.
- It doesn’t focus on external events in the environment, like news events or social media.
- It exploits only one data source, thus highly biased.

- The existing system needs some form of input interpretation, thus need of scaling.
- It doesn't exploit data pre-processing techniques to remove inconsistency and incompleteness of the data.

V. PROPOSED SYSTEM

In this proposed system, we focus on predicting the stock values using machine learning algorithms like Random Forest and Support Vector Machines. We proposed the system "Stock market price prediction" we have predicted the stock market price using the random forest algorithm. In this proposed system, we were able to train the machine from the various data points from the past to make a future prediction. We took data from the previous year stocks to train the model. We majorly used two machine-learning libraries to solve the problem. The first one was numpy, which was used to clean and manipulate the data, and getting it into a form ready for analysis. The other was scikit, which was used for real analysis and prediction. The data set we used was from the previous years stock markets collected from the public database available online, 80 % of data was used to train the machine and the rest 20 % to test the data. The basic approach of the supervised learning model is to learn the patterns and relationships in the data from the training set and then reproduce them for the test data. We used the python pandas library for data processing which combined different datasets into a data frame. The tuned up dataframe allowed us to prepare the data for feature extraction. The dataframe features were date and the closing price for a particular day. We used all these features to train the machine on random forest model and predicted the object variable, which is the price for a given day. We also quantified the accuracy by using the predictions for the test set and the actual values. The proposed system touches different areas of research including data pre-processing, random forest, and so on.

VI. METHODOLOGIES

1. Classification

Classification is an instance of supervised learning where a set is analyzed and categorized based on a common attribute. From the values or the data are given, classification draws some conclusion from the observed value. If more than one input is given then classification will try to predict one or more outcomes for the same. A few classifiers that are used here for the stock market prediction includes the random forest classifier, SVM classifier.

Random Forest Classifier

Random forest classifier is a type of ensemble classifier and also a supervised algorithm. It basically creates a set of decision trees, that yields some result. The basic approach of random class classifier is to take the decisionaggregate of random subset decision tress and yield a final class or result based on the votes of the random subset of decision trees.

Parameters

The parameters included in the random forest classifier are n_estimators which is total number of decision trees, and other hyper parameters like oob-score to determine the generalization accuracy of the random forest, max_features which includes the number of features for best-split. min_weight_fraction_leaf is the minimum weighted fraction of the sum total of weights of all the input samples required to be at a leaf node. Samples have equal weight when sample weight is not provided.

SVM classifier

SVM classifier is a type of discriminative classifier. The SVM uses supervised learning i.e. a labeled training data. The output are hyperplanes which categorizes the new dataset. They are supervised learning models that uses associated learning algorithm for classification and as well as regression.

Parameters

The tuning parameters of SVM classifier are kernel parameter, gamma parameter and regularization parameter.

- Kernels can be categorized as linear and polynomial kernels calculates the prediction line. In linear kernels prediction for a new input is calculated by the dot product between the input and the support vector.
- C parameter is known as the regularization parameter; it determines whether the accuracy of model is increases or decreases. The default value of c=10.Lower regularization value leads to misclassification.
- Gamma parameter measures the influence of a single training on the model. Low values signifies far from the plausible margin and high values signifies closeness from the plausible margin.

2. Random Forest Algorithm

Random forest algorithm is being used for the stock market prediction. Since it has been termed as one of the easiest to use and flexible machine learning algorithm, it gives good accuracy in the prediction. This is usually used in the classification tasks. Because of the high volatility in the stock market, the task of predicting is quite challenging. In stock market prediction we are using random forest classifier which has the same hyperparameters as of a decision tree.The decision tool has a model similar to that of a tree. It takes the decision based on possible consequences, which includes variables like event outcome, resource cost, and utility. The random forest algorithm represents an algorithm where it randomly selects different observations and features to build several decisiontree and then takes the aggregate of the several decision trees outcomes. The data is split into partitions based on the questions on a label or an attribute. The data set we used was from the previous year's stock markets collected from the public database available online, 80 % of data was used to train the machine and the rest 20 % to test the data. The basic approach of the supervised learning model is to learn the patterns and relationships in the data from the training set and then reproduce them for the test data.

3. Support Vector Machine Algorithm

The main task of the support machine algorithm is to identify an N-dimensional space that distinguishably categorizes the data points. Here, N stands for a number of features. Between two classes of data points, there can be multiple possible hyperplanes that can be chosen. The objective of this algorithm is to find a plane that has maximum margin. Maximizing margin refers to the distance between data points of both classes. The benefit associated with maximizing the margin is that it provides is that it provides some reinforcement so that future data points can be more easily classified. Decision boundaries that help classify data points are called hyperplanes. Based on the position of the data points relative to the hyperplane they are attributed to different classes. The dimension of the hyperplane relies on the number of attributes, if the number of attributes is two then the hyperplane is just a line, if the number of attributes is three then the hyperplane is two dimensional.

VII. SYSTEM ARCHITECTURE

Kaggle is an online community for data analysis and predictive modeling. It also contains dataset of different fields, which is contributed by data miners. Various data scientist competes to create the best models for predicting and depicting the information. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges. The dataset used in the proposed project has been downloaded from Kaggle. However, this data set is present in what we call raw format. The data set is a collection of stock market information about a few companies.

The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So the first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction starts from an initial state of measured data and builds derived values or features. These features are intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction is a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for ease of management, while still precisely and totally depicting the first informational collection.

The feature extraction process is followed by a classification process wherein the data that was obtained after feature extraction is split into two different and distinct segments. Classification is the issue of recognizing to which set of categories a new observation belongs. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model. The splitting is done in a way that training data maintain a higher proportion than the test data.

The random forest algorithm utilizes a collection of random decision trees to analyze the data. In layman terms, from the total number of decision trees in the forest, a cluster of the decisiontrees look for specific attributes in the data. This is known as data splitting. In this case, since the end goal of our proposed system is to predict the price of the stock by analyzing its historical data.

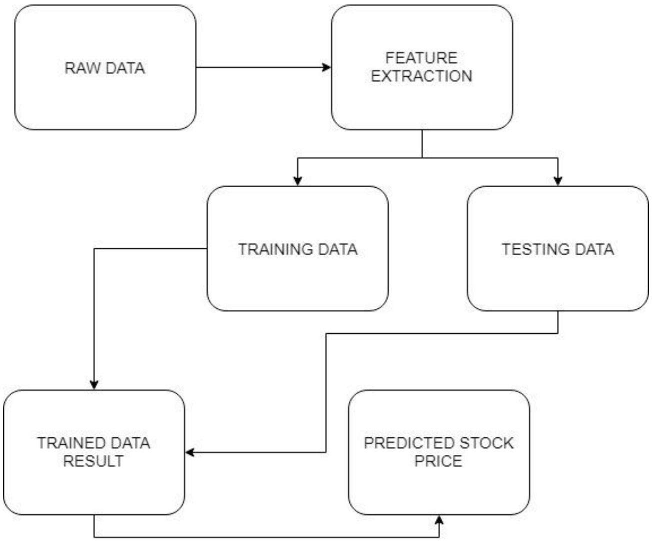


Fig 1 System Architecture

VIII. MODULE IDENTIFICATION

The various modules of the project would be divided into the segments as described.

I. Data Collection

Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects. Data collection also complements to enhance the dataset by adding more data that are external. Our data mainly consists of the previous year stock prices. Initially, we will be analyzing the Kaggle dataset and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

II. Pre Processing

Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of variables so that they can be compared on common environs.

III. Training the Machine

Training the machine is similar to feeding the data to the algorithm to touch up the test data. Thetraining sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyperparameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyperparameter values**.**

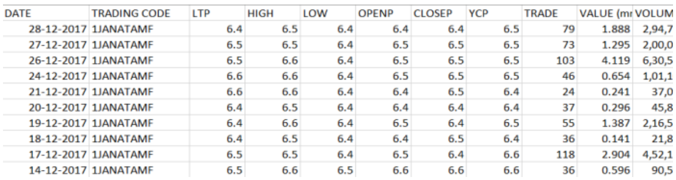
Finally, we will calculate a cross-validated score, for individual sets of hyperparameters. Then, we select the best hyperparameters. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data.

IV. Data Scoring

The process of applying a predictive model to a set of data is referred to as scoring the data. The technique used to process the dataset is the Random Forest Algorithm. Random forest involves an ensemble method, which is usually used, for classification and as well as regression. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

IX. EXPERIMENTAL RESULTS

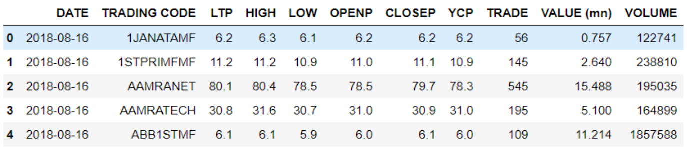
The xlxs file contains the raw data based on which we are going to publish our findings. There are eleven columns or eleven attributes that describe the rise and fall in stock prices. Some of these attributes are (1) HIGH, which describes the highest value the stock had in previous year. (2) LOW, is quite the contrary to HIGH and resembles the lowest value the stock had in previous year (3) OPENP is the value of the stock at the very beginning of the trading day, and (4) CLOSEP stands for the price at which the stock is valued before the trading day closes. There are other attributes such as YCP, LTP, TRADE, VOLUME and VALUE, but the above mentioned four play a very crucial role in our findings.



DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mı	VOLUM
28-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.4	6.5	79	1.888	2,94,7
27-12-2017	1JANATAMF	6.5	6.5	6.4	6.5	6.5	6.5	73	1.295	2,00,0
26-12-2017	1JANATAMF	6.5	6.6	6.4	6.5	6.5	6.5	103	4.119	6,30,5
24-12-2017	1JANATAMF	6.6	6.6	6.4	6.5	6.5	6.5	46	0.654	1,01,1
21-12-2017	1JANATAMF	6.6	6.6	6.4	6.4	6.5	6.4	24	0.241	37,0
20-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.4	6.4	37	0.296	45,8
19-12-2017	1JANATAMF	6.4	6.6	6.4	6.5	6.4	6.5	55	1.387	2,16,5
18-12-2017	1JANATAMF	6.4	6.5	6.4	6.4	6.5	6.4	36	0.141	21,8
17-12-2017	1JANATAMF	6.5	6.5	6.4	6.5	6.4	6.6	118	2.904	4,52,1
14-12-2017	1JANATAMF	6.5	6.6	6.5	6.6	6.6	6.6	36	0.596	90,5

Fig 2 Raw Data

This is a pictorial representation of the data present in our xlxs file. This particular file contains 121608 such records. There are more than ten different trading codes available in the dataset and some of the records do not have relevant information that can help us train the machine, so the logical step is to process the raw data. Thus we obtain a more refined dataset which can now be used to train the machine.



	DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mn)	VOLUME
0	2018-08-16	1JANATAMF	6.2	6.3	6.1	6.2	6.2	6.2	56	0.757	122741
1	2018-08-16	1STPRIMFMF	11.2	11.2	10.9	11.0	11.1	10.9	145	2.640	238810
2	2018-08-16	AAMRANET	80.1	80.4	78.5	78.5	79.7	78.3	545	15.488	195035
3	2018-08-16	AAMRATECH	30.8	31.6	30.7	31.0	30.9	31.0	195	5.100	164899
4	2018-08-16	ABB1STMF	6.1	6.1	5.9	6.0	6.1	6.0	109	11.214	1857588

Fig 3 head()

This is the result of using the head(). Since we are using the pandas library to analyse the data, it returns the first five rows. Here five is the default value of the number of rows it returns unless stated otherwise. The trading code in the processed data set is not relevant so we use the strip() to remove it and replace all of the trading codes with a value 'GP'

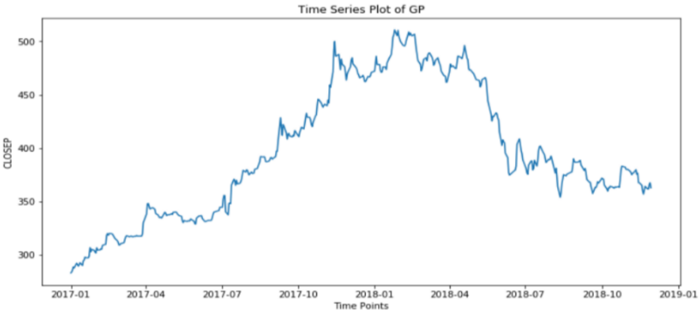


Fig 4 Time series plot of GP

This is a time series plot generated from using the "matplotlib.pyplot" library. The plot is of the attributes "CLOSEP" vs "DATE". This is to show the trend of closing price of stock as time varies over a span of two years. The figure provided below is the candle stick plot, which was generated using the library "mpl_finance". The candle stick plot was generated suing the attributes 'DATE', 'OPENP', 'HIGH', 'LOW','CLOSEP'.

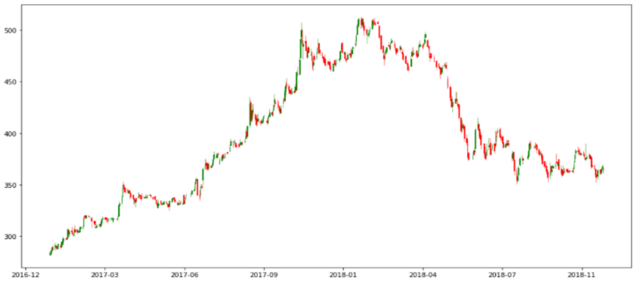


Fig 5 Candlestick plot

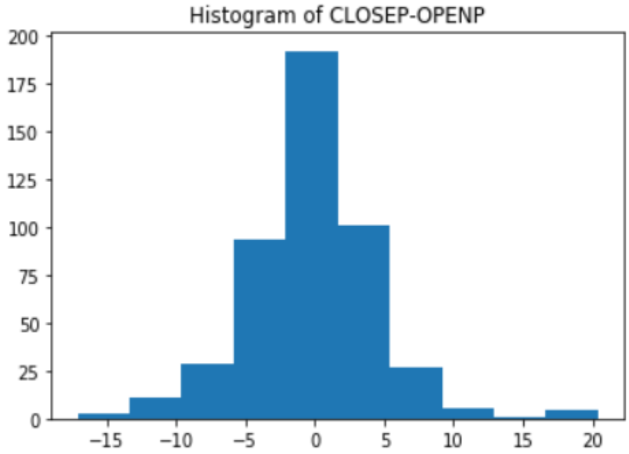


Fig 6 Histogram of CLOSEP-OPENP

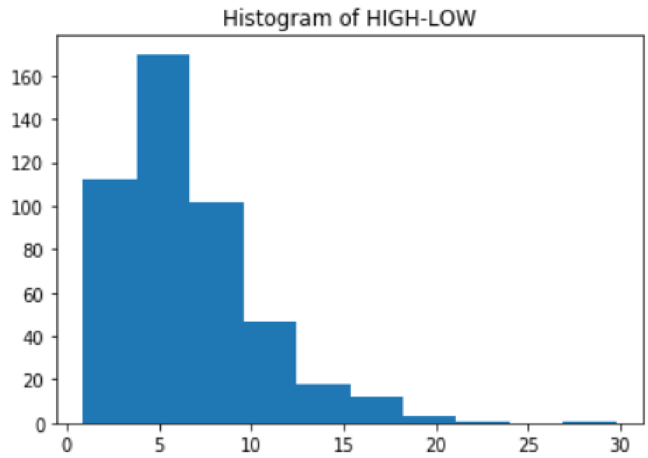
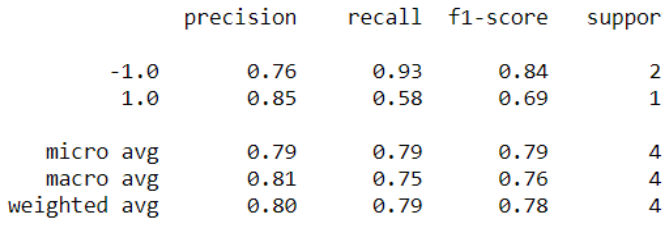


Fig 7 Histogram of HIGH-LOW

The above two figures are histograms plotted between ‘CLOSEP’ and ‘OPENP’ and the attributes ‘HIGH’ and ‘LOW’. This is done because we believe today’s closing price and opening price along with the high and lowest price of the stock during last year will affect the price of the stock at a later date. Based on such reasoning we devised a logic “if today’s CLOSEP is greater than yesterday's CLOSEP then we assign the value 1 to DEX or else we assign the value -1 to DEX. Based on such the whole data set is processed and upon using the head() we get a glimpse of the data obtained thus far.

The next step entailed the setting of feature and target variable, along with the setting of train size. Using the sklearn libraries we import SVC classifier and fit it with the training data. After training the model with the data and running the test data through the trained model the confusion matrix obtained is shown below.



```
              precision    recall  f1-score   suppor

        -1.0       0.76      0.93      0.84         2
         1.0       0.85      0.58      0.69         1

   micro avg       0.79      0.79      0.79         4
   macro avg       0.81      0.75      0.76         4
weighted avg       0.80      0.79      0.78         4
```

Fig 9 Confusion Matrix

Along with this, we use the same dataset to train another model. This model utilises the Random Forest Classifier belonging to the ensemble technique. The decision trees have the default values so that leaves the “n_estimator” value to be 10 since this is version 0.20. However, the value of “n_estimator” will change to 100 in the version 0.22. After fitting the model with the data and running it against predicted data we find that this has an accuracy score of 0.808.

To sum it up, the accuracy of the SVC Model in Test Set is 0.787 whereas the accuracy score of the random forest classifier is calculated to 0.808.

IX. CONCLUSION

By measuring the accuracy of the different algorithms, we found that the most suitable algorithm for predicting the market price of a stock based on various data points from the historical data is the random forest algorithm. The algorithm will be a great asset for brokers and investors for investing money in the stock market since it is trained on a huge collection of historical data and has been chosen after being tested on a sample data.The project demonstrates the machine learning model to predict the stock value with more accuracy as compared to previously implemented machine learning models.

X. FUTURE ENHANCEMENT

Future scope of this project will involve adding more parameters and factors like the financial ratios, multiple instances, etc. The more the parameters are taken into account more will be the accuracy. The algorithms can also be applied for analyzing the contents of public comments and thus determine patterns/relationships between the customer and the corporate employee. The use of traditional algorithms and data mining techniques can also help predict the corporation’s performance structure as a whole.

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. VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. “Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2017.
5. SachinSampatPatil, Prof. Kailash Patidar, Asst. Prof. Megha Jain, “A Survey on Stock Market Prediction Using SVM”, IJCTET 2016.
6. https://www.cs.princeton.edu/sites/default/files/uploads/Saahil_magde.pdf
7. Hakob GRIGORYAN, “A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)”, DSJ 2016.
8. RautSushrut Deepak, ShindeIshaUday, Dr. D. Malathi, “Machine Learning Approach In Stock Market
9. Prediction”, IJPAM 2017.
10. Pei-Yuan Zhou , Keith C.C. Chan, *Member, IEEE*, and Carol XiaojuanOu, “Corporate Communication Network and Stock Price Movements: Insights From Data Mining”, IEEE 2018.