HBRP PUBLICATION

Journal of Network Security and Data Mining
Volume 4 Issue 2

Analysis of Stock Market using Data Mining Techniques

***Ashmita Phuyal*[1], *Aditi Pokharel*[2*], *Nilima Dahal*[3], *Sushil Shrestha*[4]**
[1,2,3,4]*Department of Computer Science and Engineering,*
Kathmandu University, Dhulikhel, Nepal.

***Corresponding Author**
***E-Mail Id:-**aditipokh13@gmail.com*

ABSTRACT

Long term investments are one of the major leading investment strategies in the modern financial market. However, calculating intrinsic value of some company and evaluating shares for long term investment is not easy, since analysts have to visualize a large number of financial indicators and evaluate them in the right manner. The prediction of stock markets is considered as one of the major challenging tasks of financial time series. Due to the presence of non-linear data sets and dynamic nature, there is an increasing demand in analysis of the market and prediction of future stock trends. In this paper we present a data mining and machine learning aided approach to evaluate the equity's future price over the long term. However, the main objective of this paper is to find the best algorithm for prediction to predict the values of the stock market.

***Keywords:—** stock market, analysis, data mining, algorithm, machine learning*

INTRODUCTION

Share market plays an important role in the economic growth. It is a common platform for companies to raise funds for a company by allowing customers to buy shares at an agreed price [1].

Trading stocks on the stock market is one of the major investment activities. In the context of Nepal, the stock market index is taken as a barometer of an economy. The index of growth in the stock is regarded as good as it indicates that the investors are confident about the future possibility of the economy.

Investors want to know whether the stock will rise or fall over a certain period of time. In order to predict how some company, in which investors want to invest, would perform in future, they developed a number of analysis methods based on current and past financial data and other information about the company [1]. Many researchers in the past try to predict stock prices using statistical and graphical approaches but data mining techniques can now uncover hidden patterns that were previously impossible to find using traditional approaches. [2].

Problem Statement

The investors investing in the stocks are unfamiliar with the behaviour of the market. Being able to predict the market trend can be valuable as one can gain financial benefits and economic insight. But, it is said to be uncertain to predict the market behaviour due to highly nonlinear nature and complex dimensionality which has led to the importance of analysing the market behaviour of financial institutes over time. For analysis of such large datasets, data mining techniques can be adapted for not only analysis but also for predictions of the market trend. To make such predictions, it is required to select the best possible algorithm to model the data to gain insight whether or not to make smart investments.

Research Problem

The purpose of this research study was to address the following question:

- What would be the best possible algorithm to predict the stock market index?
- What is the correlation among the data sets present in the stock market?
- What could be the parameter used to evaluate the algorithm?

RELATED WORKS

Prediction of the stock market is a challenging task and is uncertain as it varies with time. In recent times, many tasks implemented in the financial sectors have attracted the attention of the researchers mainly the stock price forecasting. There is a certain assumption in the past that available information has predictive relationship to the future stock returns. Several data mining techniques have been used by the researchers to the relationships from the past history.

Ballings and Van den Poel together evaluated multiple classifiers for stock price direction prediction. They compared many learning methods like Random forest, Artificial Neural Network, Logistic regression and found that the random forest has the best performance among all [3].

Weng and Ahmed developed stock market one-day ahead movement prediction using disparate data sources using decision trees, neural networks, and support vector machines. The data in the data source are combined with traditional time series and technical indicators to provide a more up-to-date system for prediction [4].

Chen and Hao developed a feature weighted support vector machine and K-nearest neighbor algorithm in which the input characteristics of the methods were weighted by their importance for stock market indices prediction [5].

Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies, is the title of Chong and Han [6] study. By raw inputs that relate to stock returns of past periods, seven feature sets are produced using three methods to predict future stock returns. Input data is related to the South Korean stock market. The results show that the proposed three-layer neural network method has a better predictive value when used in the self-correlation model.

In the previous studies, the future trend of stock prices was predicted using decision trees, neural networks, support vector machines, random forest, logistic regression, K nearest neighbor algorithm. The focus of those researchers has been on the improvement of the prediction model. With the impact from those researchers and their studies, the focus of this paper is on four data mining algorithms: linear regression, support vector machine, deep learning and random forest as these algorithms have been considered the best approach by the research mentioned above.

METHODOLOGY

This section gives the detailed analysis of each step involved in the process. Each sub section defines stages of the analysis.

Data Source

The primary source of the dataset used is originally from Global IME Bank Ltd. from year 2015 to 2020. This dataset has been extracted from sharesansar site. The data consisted of attributes such as date, high price, open price, low price, close price, change, quantity and turnover. All these data were incorporated to provide a detailed view of stock trends in the near future of the company. The data set consists of 7 columns and 2057 rows containing the following attributes:

Table 1:-*Description of attributes of the data sets*

S.N	Attributes	Attribute Description
1.	Date	Date on which trading occurs.
2.	Open	Price of the first trade on a particular day.
3.	Close	Price at which a stock is traded on the day.
4.	High	Highest price at which a stock is traded during the course.
5.	Low	Lowest price at which a stock is traded during the course.
6.	Turnover	Total value of stocks traded during a specific period of time.
7.	Quantity	Total number of stock bought/sold on a specific date.
8.	Change	Difference between closed price from one day to the next.

Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. The following are the techniques implemented:

Data Cleaning

The datasets require data cleaning as a pre-processing step to acquire more accuracy with less error rate for further use. However, while examining the data of Global IME Bank Ltd. it was found that the data did not contain any missing value but could be further reduced.

Noisy Data

The dataset obtained initially contains a large set of information and could not be interpreted correctly. To acquire more accuracy, the data was smoothened by the help of moving average which smooths the data set from random price fluctuations.

Data Reduction

During data preprocessing, some attributes were found to be irrelevant in the study as it was found that Open, High and Low have perfect correlation with the target attribute closed price. Open price, High and low have correlation 99.32%, 99.66% and 99.62% respectively. These perfect correlation attributes lead to the duplication of targeted attributes and could bias the predictive model. Another at tribute, Change was found to have least correlation of 0.05% with the attribute, making it least significant. Hence, removal of these four columns in the study. Besides these attributes, Date, Quantity and Turnover were found to have high correlation of 73.38%, 71.33%, 72.27% which was selected for further analysis.

a) Attributes Selected

After reduction, the data set contained 4 columns and 2057 rows and in order to predict the future trend of the stock, the profit or loss is generally calculated by considering the closed price data of a stock. Therefore, the closed price is considered as a target attribute which has high correlation with date attribute. Hence, these two attributes were selected in the study.

- Date
- Close

Data Visualization Tools

The tools and programming language supported to analyze different data mining algorithms with relevant accurate results are Python along with a user-friendly requisite library and RapidMiner, a visualization tool used in deploying predictive models.

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis technique used to analyze the data sets of stock with visual methods are: Histogram, Scatter plot, Line Graph and Moving Average.

a) Histogram

Histogram is used to visualize the frequency distributions. It inspects the data for normal distribution, skewness, outliers, etc. From the figure, it can be seen that it has wider distribution indicating more volatility.

b) Scatter Plot

A scatter plot, also called scatter graph, is a graph in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present. The graph was made to detect any outliers [7]. From the figure, it can be seen that the relationship between date and close price of stock is strong and linear.

c) Line Graph

A line graph, also known as a line chart, is a graph used to visualize the information that changes continuously over time. It consists of a horizontal x-axis and a vertical y-axis. From the figure, it shows a decrease in close price with time.

d) Moving Average

Moving Average is a simple technical analysis tool that smooths up price by creating constantly updated average prices. The average is taken over a specific period of time. The figure shows the removal of noisy data using moving average technique.

ALGORITHMS

Data analysis requires different types of algorithms to identify the character of data available. In stock market analysis, there are multiple approaches or techniques that one can adopt to be able to understand the data. Such techniques are divided into two types i.e. classification and regression. For the problem identified in the paper, classification is not advisable as it is applicable for categorical data. The data collected falls under time series data and the output from regression models is often used for forecasting. It can be used to build predictive models. The following regression algorithms are used in this study:

1. Linear Regression
2. Support Vector Machine (SVM)
3. Random Forest (RF)
4. Deep Learning

1) Linear Regression: Linear regression is a linear approach to modeling the relationship between a scalar response (dependent variable) and explanatory variables (independent variables). It attempts to model the relationship between two variables by fitting a linear equation to observed data [8]. In this regression approach, linear predictor functions are used to model the correlations and the unknown parameters of the functions are evaluated by the data. These are linear models.

2) Support Vector Machine (SVM): In time series analysis, the application of SVMs used in regression analysis is called Support Vector Regression (SVR) where each data item is plotted as a point in n-dimensional space and distinct classification is performed by finding the hyper-plane and maximizing the distance between the plane and nearest input data points [9] To train the support vector regression method, sequential minimal optimization algorithm is used. The utilization helps to replace all the missing values and transform the nominal attributes to binary values. Meanwhile, it helps to normalize the attributes by default

values [10].

3) Random Forest (RP): Random Forest is an ensemble algorithm for classification and regression that creates a decision tree from a randomly selected trained model to get more accurate prediction. The trees in random forest are parallel to each other without any interaction between them. It solves the problem of overfitting of the training set and can also be modeled for categorical values. Larger the number of trees, the more accurate the result is [11].

4) Deep Learning: Deep Learning is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation (multilayer perceptron). A multilayer perceptron (MLP) is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron with a nonlinear activation function. MLP utilizes back propagation for training the network [12]. This class of networks consists of multiple layers of computational units, usually interconnected in a feed forward way.

RESULTS

This section contains the result obtained from exploratory data analysis and the implementation of data mining algorithms. From [14,15], it was found that the Root Mean Squared Error (RMSE) and Mean Average Error (MAE) the best prediction algorithm is determined.

Exploratory Data Analysis Result

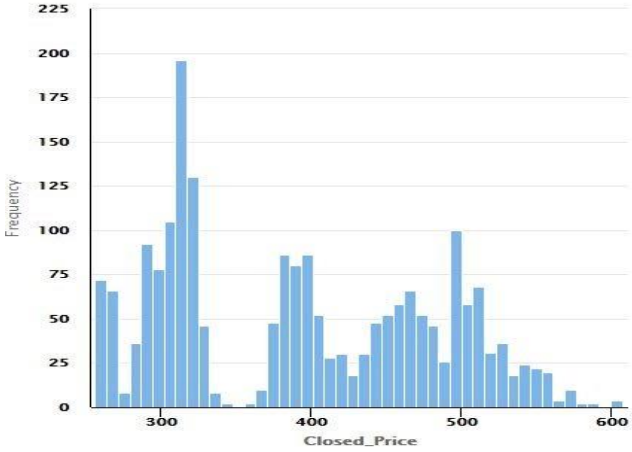

Fig.1:-*Histogram*

A histogram is a graphical method of representing data using bars of different heights. The taller bars show that more data falls in that range [16]. The graph (Figure 1) shows the distribution of close price throughout the data sets. The histogram is Skewed left.

Table 2:-*Close price parameter used for histogram*

Minimum	245
Maximum	600
Mean	393.84
Standard Deviation	88.168

Table 2 shows the values of close price parameters i.e. Minimum value, Maximum value, Mean, Median and Standard Deviation used for histogram.

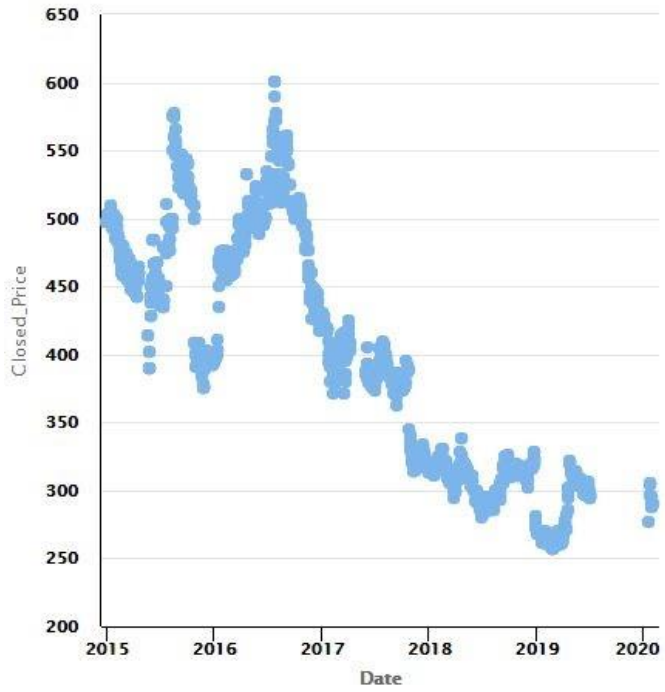

Fig.2:-*Scatter plot of date (x-axis) and Close Price (y-axis)*

The above scatter plot shows that the relationship between date and close price of Stock has a strong and linear relationship.

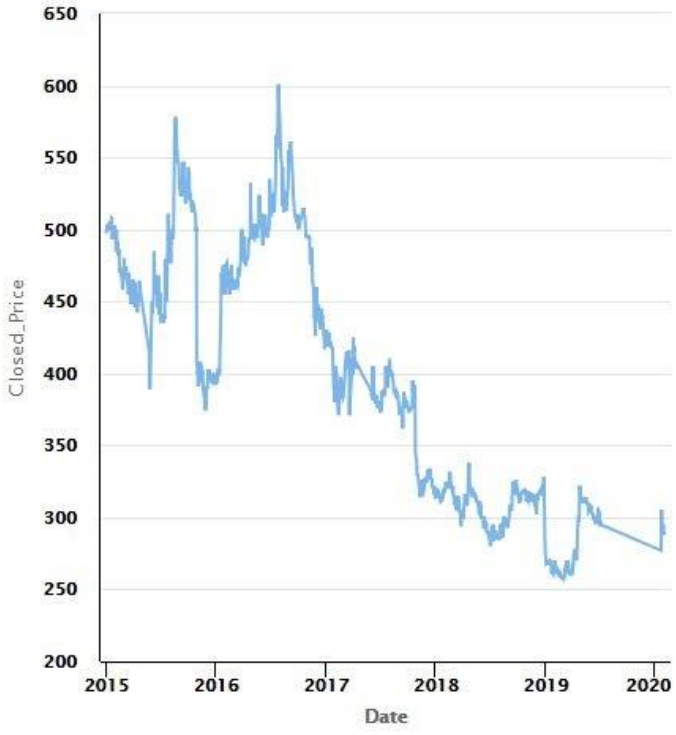

Fig.3:-*Line Graph of date (x-axis) and closed price (y-axis)*

Line Graphs are frequently used to represent changes in the prices with time. The graph (Figure 3) shows a decrease in close price with time.

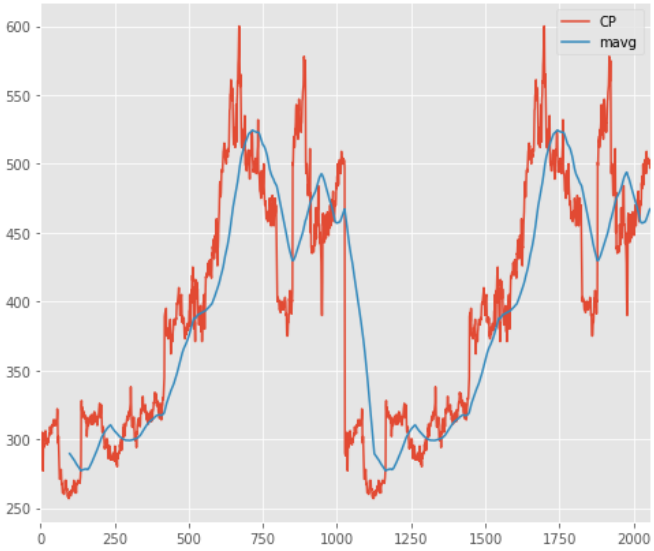

Fig.4:-*Smoothing of noisy data*

The above line graph (Figure 4) contains noises which are removed by the moving average technique and plotted against the noise graph.

Algorithms Result

To evaluate and compare the performance of the algorithms we calculated the following metrics:

1) ***Mean Absolute Error (MAE):*** MAE measures the average magnitude of the errors in a set of predictions, without considering their direction.

2) ***Root Mean Squared Error (RMSE):*** RMSE is a quadratic scoring rule that also measures the average magnitude of the error

3) ***Relative Error:*** Relative error is a measure of the uncertainty of measurement compared to the size of the measurement.

Table 3:-*Comparison of parameters of algorithm used*

Model	**Root Mean Squared Error (RMSE)**	**Mean Absolute Error (MAE)**	**Relative Error**
Linear Regression	43.32	33.823	7.8%
Support Vector Machine	14.823	9.475	2.3%
Random Forest	**8.885**	**5.664**	**1.3%**
Deep Learning	15.432	11.961	3.1%

The above table represents the comparison of parameters with the algorithms used for the prediction of stock. The table shows that the Random Forest has the least value in comparison to other algorithms.

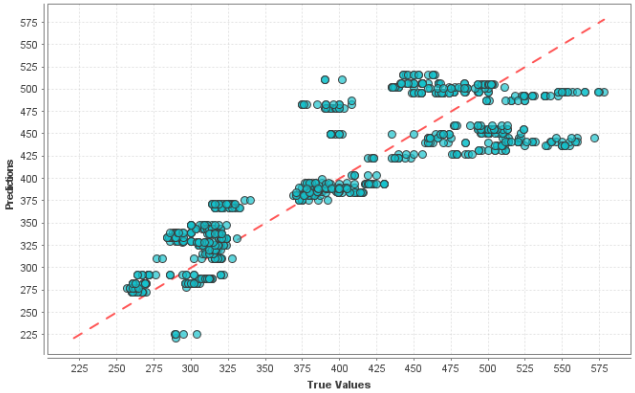

***Fig.5**:-Scatter plot of True value (x-axis) and Prediction value (y-axis) of closed price using Linear regression algorithm*

The above Figure shows the prediction of closed price using a linear regression algorithm plotting the train data against test data. It can be seen in the graph (Figure 5) that most points scattered do not lie on the line of best fit but cluster around them with many visible outliers.

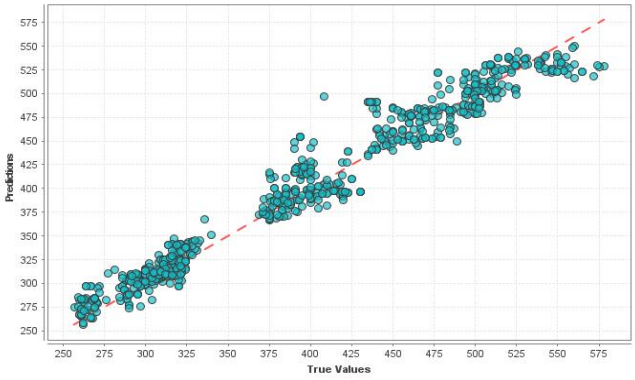

***Fig.6:-**Scatter plot of True value (x-axis) and Prediction value (y-axis) of closed price using deep learning algorithm*

The above Figure shows the prediction of closed price using a deep learning algorithm plotting the train data against test data. In the graph, it can be seen that most points cluster around the line of best fit with some visible outliers.

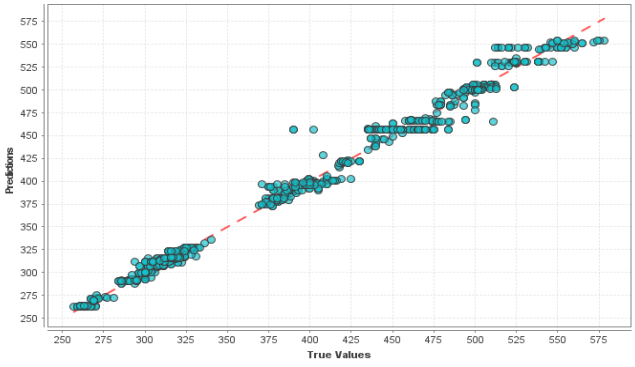

***Fig.7:-**Scatter plot of True value (x-axis) and Prediction value (y-axis) of closed price using random forest algorithm*

The above Figure shows the prediction of closed price using a random forest algorithm plotting the train data against test data. It can be seen in the graph (Figure 7) that most points lie in the line of best fit with some visible outliers.

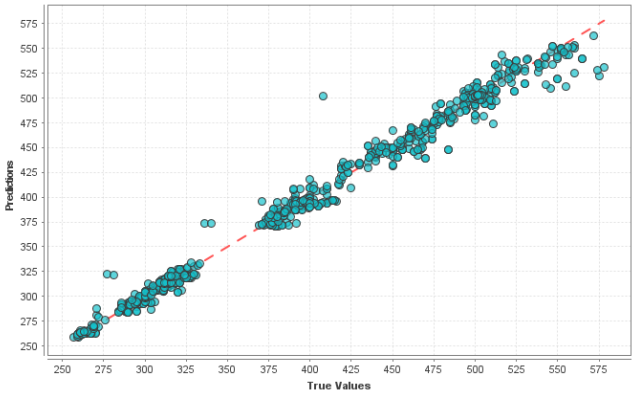

***Fig.8**:-Scatter plot of True value (x-axis) and Prediction value (y-axis) of closed price using support vector machine algorithm*

The above Figure shows the prediction of closed price using a support vector machine algorithm plotting the train data against test data. In the graph most of the points fall on or cluster around the line of best fit with some visible outliers.

DISCUSSION AND CONCLUSION

Discussion

During analysis, four regression tools were used to analyse the stock market of Global IME bank: Linear Regression, Support Vector Machine, Random Forest and Deep Learning. In the data, the close price was taken as a basis to make future predictions for the company. The test dataset was used as the train datasets for prediction to view the similarity of the patterns in the data. Using the data, three factors were calculated which are: Absolute error, Root Mean Squared Error, relative error percentage for the close price.

The two common metrics used to measure accuracy for continuous variables are Absolute Error (MAE) and Root mean squared error (RMSE). RMSE was taken as a parameter to reduce large errors for precision; it was found that using linear regression value of RMSE of 43.32. Similarly, for SVM the RMSE value to be 14.823, random forest to be 8.885 and Deep learning to be 15.432.

In the above Figure 5-8, it can be seen that the graph of actual closed price against predicted price is plotted with the line of best fit. The line of best fit denotes the x= y equation where, if the points fall under the line, it means the predicted value is as close to the actual value [17]. The accuracy of prediction is deemed to be high if most of the points fall on the line.

In Figure 5, it can be seen that most of the points scatter around the graph rather than cluster near the best fits showing there are multiple outliers. It means that the predicted value deviates from the actual value of the closed price making the algorithm of linear regression that has the RMSE value of 43.32 with relative error of 7.8% least suitable algorithm for the case.
In the case of Figure 6., it can be seen that most do cluster around the best fits line but still showing some outliers deviate from the line. Though most points cluster around the best fit, it doesn't completely lie on the line. So, the algorithm of deep learning shows better prediction than linear regression graph with the RMSE value of 15.432 and relative error of 3.1%.

In the case of Figure 7, it can be seen that most points of the predicted price plotted actual price not only cluster around the line of best fit but also completely lie on the line. It means that the predicted value of the closed price was similar to that of the actual price making the algorithm of Random Forest the most suitable algorithm for the case with RMSE value of 8.885 and relative error of 1.3%. The graph shows only minor deviations from the line of best fits with only few outliers visible.

In the case of *Fig. 8*, it can be seen that most points cluster around the line of best fit with minor deviation and only few outliers visible. It means that the predicted value of the closed price was similar to that of the actual price making the algorithm of SVM better for prediction of the closed price but less than that of Random Forest Algorithm. In this case, for the RMSE value of 14.823 and relative error of 2.3% making the algorithm slightly better than deep learning algorithm.

Among the algorithms used Random forest yields the most accurate result as it contains the least value of RMSE whereas, linear regression is found to be least appropriate for the data we have. The graph of random forest shows that the train data and test data are most similar in nature compared to the other graphs of the different algorithms.

CONCLUSION

In the stock market, there are larger historical datasets. Different data mining regression techniques are used for the prediction of future stock value trends. This paper considers four regression techniques for comparative analysis based on various parameters. The prediction by the different algorithms is calculated to identify the best model for the prediction. According to the study, Random Forest was found to be the most effective algorithm for the prediction. The predicted stock price provides aid to the investors, stakeholders and data analysts to predict the trend of the market. Since the stock market tends to be highly sensitive and volatile, determining the trend is difficult.

REFERENCES

1. J. Lu, K. Dang & K. Sakakibara (2020). Parameters for Stock Market Prediction. *Semanticscholar.org*, 2020. [Online]. Available: https://www.semanticscholar.org/paper /Parameters-for-Stock-Market-Prediction-N'guyenLu/be9880553b0e423b64b9a2b2e27f07a71d7661e7.[Accessed: 28-Jan-2020]
2. Prasanna, S., & Ezhilmaran, D. (2013). An analysis on stock market prediction using data mining techniques. *International Journal of Computer Science & Engineering Technology (IJCSET)*, *4*(3), 49-51.
3. Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert systems with Applications*, *42*(20), 7046-7056.
4. Weng, B., Ahmed, M. A., & Megahed, F. M. (2017). Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, *79*, 153-163.
5. Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, *80*, 340-355.
6. Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, *83*, 187-205.

7. ”A Complete Guide to Scatter Plots”, Chartio, 2020. [Online]. Available: https://chartio.com/learn/charts/what-is-a scatter-plot/. [Accessed: 14- Mar 2020].
8. ”Linear Regression”, Stat.yale.edu, 2020. [Online]. Available: http://www.stat.yale.edu/Courses/199 7- 98/101/linreg.htm. [Accessed: 26-Feb 2020].
9. ”Support Vector Machine — Introduction to Machine Learning Algorithms”, Medium, 2020. [Online]. Available: https://towardsdatascience.com/suppor t vector-machine-introduction-to-machine learning-algorithms-934a444fca47. [Accessed: 23- May-2020].
10. U. code), “Understanding Support Vector Machines(SVM) algorithm (along with code)”, Analytics Vidhya, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog /20 17/09/understaing-support-vector-machine example-code/. [Accessed: 7-Mar 2020].
11. D. Elsinghorst, “Machine Learning Basics - Random Forest”, Shirin’s playgRound, 2020. [Online]. Available: https://www.shirin glander.de/2018/10/ml$_b$*asics*$_r$*f*/.[*Access ed* : 23 − *M ay* − 2020].
12. “A Beginner’s Guide to Multilayer Perceptrons (MLP)”, Pathmind, 2020. [Online]. Available: https://pathmind.com/wiki/multilayer perceptron. [Accessed: 7- Mar 2020].
13. “Developing a Conceptual Framework for Research (Sample)”, Scribbr, 2020. [Online]. Available: https://www.scribbr.com/dissertation/ conceptual-framework/. [Accessed: 15- May 2020].
14. “Root-Mean-Squared Error - an overview — ScienceDirect Topics”, Sciencedirect.com, 2020. [Online]. Available: https://www.sciencedirect.com/topics/ engi neering/root-mean-squared-error. [Accessed: 14- May- 2020].
15. N. Watson, “Using Mean Absolute Error to Forecast Accuracy”, Contemporary Analysis, 2020. [Online]. Available: https://canworksmart.com/using-mean absolute-error-forecast-accuracy/. [Accessed: 13- May- 2020].
16. “How to make a histogram — Data displays — Statistics (video) — Khan Academy”, Khan Academy, 2020. [Online]. Available: https://www.khanacademy.org/math/a p-statistics/quantitative-dataap/histograms-stem-leaf/v/histograms-intro. [Accessed: 10- Mar- 2020].
17. “Line of Best Fit (Least Square Method)”, Varsitytutors.com, 2020. [Online]. Available: https://www.varsitytutors.com/h otmath/hotmath$_h$*elp/topics/line* − *of* − *best* − *f it.*[*Accessed* : 14 -May -2020].