Recent Developments in Electronics and Communication Systems
KVS Ramachandra Murthy et al. (Eds.)

doi:10.3233/ATDE221289

Study of Linear Regression Prediction Model for American Stock Market Prediction

KARAN CHAUHAN and NITIN SHARMA
Department of Electronics and Comm. Chandigarh University, Gharuan
Department of Electronics and Comm. Chandigarh University, Gharuan

Abstract. A difficult task in the financial market is to make accurate stock market predictions. This is because there are plenty of factors that affect the price of a company's stock. Even a single bad tweet can affect the price of a stock. As a result, making an exact prediction is a difficult task. Many scientists are working to find a solution that can withstand a wide range of factors and still provide an accurate result. Supervised learning model called linear regression have produced excellent predictions in a variety of fields over the last few decades. In these studies, a model is designed to predict the stock market prediction by using linear regression. The model is evaluated by using the dataset of three best companies that is (Walmart, Tesla, and Amazon) listed on the American stock exchange called NASDAQ and the result is analyzed in terms of root mean squared error. Then the following results are compared with other machine learning models that is Random Forest and Support Vector Machine. In all cases, linear regression gives the best results.

Keywords. Linear Regression, Machine Learning, Stock Market Prediction.

1. Introduction

Stock market prediction is very demanding in the financial market as it provides direct profit to the investors. With the advancement in technology, there are numerous apps available through which anyone can easily buy and sell stocks, but to earn profit, one should know the future situation of the stock. This is where stock market prediction takes place. Supervised machine learning for predictive analysis has gained popularity in recent decades. It is used for prediction in various fields such as healthcare, communication, military, weather forecast etc. In these studies, we are trying to predict stock market prediction by using a machine learning model called linear regression. First, let us see how machine learning can be used in predictive analysis Figure 1. demonstrates the block diagram of machine learning predictive analysis.

1.1. Process of Machine Learning.

The first step is problem statement in this the problem is define that what type of problem is there. What type of approach can be used to solve it what type of dataset is need, next is data collection for stock market prediction historical dataset of particular company is employed. Indian companies stock markets historical data can be found on the National

Stock Exchange of India (NSE) [1]. Similarly, via yahoo finance, we can retrieve the NASDAQ index data.

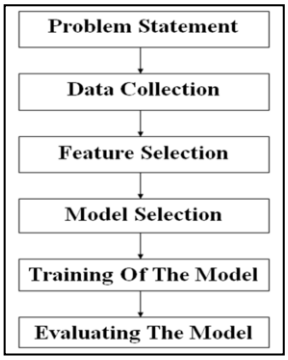

Figure 1. Process Of Machine Learning

After achieving the input dataset, the next step is feature selection It decreases the number of input variables. Feature selection means examining the common features in the given dataset. This step increases the system's efficiency and performance. After selecting the best features next step is Choosing an algorithm is a critical task as there are number of different types of machine learning models, each having their own advantages and disadvantages while dealing with different types of datasets and problems. Next is training of model It is called learning from the model. During the training period, past event data is analyzed and re-treated until the model learns all about the given input data and the difference between the actual and anticipated values of the target variable is as small as equal. At last, the evaluation of the trained model is done by using the unknown dataset.

2. Literature review

Various machine learning techniques have been employed to forecast the stock market in a variety of studies. In paper [2] the author design machine learning technique for stock market prediction that is Linear Regression, Genetic Algorithm, Support Vector Machine (SVM), K-Nearest Neighbor, Neural Network and Random Forest. Another author uses machine learning approach for sentiment analysis for Apple stock market is shown in this research. The historical data used in this is taken from Yahoo Finance. A machine learning model, SVM, is utilized in this stock market predictions. The model achieves 75.22% accuracy in learning and 76.68% when training data is utilized. More accuracy can be achieved if more input data is treated in this system model [3]. In paper [4] a model is created in this research by studying and merging two distinct elements. Forecasts for stock price trends and feature selections Both the entities are evaluated by different random forest machine learning models, and the end results are in combined form. The model is fit for long-term stock market predictions. By using this RF-RF approach, the maximum annual return is 29.51%. The feature selection method can be improved. This model performs better than other existing models, as its calculative result is good in terms of long-term forecasting. [5] A stock market prediction android app by using multiple linear regression model is designed. The researcher analyses Indonesian stock market by using the following app Machine learning algorithms like Random

Forest and Support Vector Machines are used in this study to develop an approach for anticipating stock market fluctuations. Classification and regression have been quite effective using the Random Forest model. In paper [6] The researcher designs and compares random forest, linear regression and journalized linear regression in order to forecast 10 different stocks of firms where the linear regression model predicts the best outcome. In a recent study, a researcher used eight different machine learning methods to predict the stock market. The researcher analyzed the data using the NIFTY historical data and found that a linear regression model performed best in most cases[7]. A linear regression model is used in various different fields also. For example, a researcher used linear regression model to predict weather forecasting in his research. The researcher designed a model in which the feature selection of the model is done twice. The results are impressive then other existing models [8].

3. Methodology

We are trying to predict the three best companies' stock that are (Walmart, Tesla, and Amazon) listed on the American stock exchange by using the linear regression model. The following companies are NASDAQ-listed firms. Recent 1-year data from May 30, 2021, to May 30, 2022 is downloaded from internet which consist a variety of variables, such as stock daily open cost, stock daily close cost, stock daily lowest cost, stock daily highest cost, etc. From this data close prize is utilized as input data. Now to fit this dataset into the designed model the input data set is converted into the form of an array. To accomplish that, the min-max scalar is utilized, having a feature ranging from zero to one. The formula for the min-max scalar is as follows:

$$X^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

here, denoted x is real value of close cost on particular time x max, x min are the maximum and minimum value of close cost and x* represents array value.

The designed mode used seventy percent of the data as training data. To test the trained model, unseen data is required and are termed as "testing data." Last thirty percent of the data from the input dataset is used as the testing data. With that, the moving window approach is used to form a data matrix having a length of 10. That means, to anticipate the 11th value of the set, the prior 10 values are responsible. And to anticipate the twelfth number, the value of the preceding 10 values (2 to 11) is employed.

3.1. Linear regression

Here we used Amazon stock market data. The linear regression model initially uses Seventy percent of the input data for the training of the model. Figure 2 represent the code used. The outcomes throughout the training procedure are indicated in Figure 3 (a). We kept the model as simple as possible. So, the hyper-parameters of the designed model remain constant.

```
[ ]  lr = LinearRegression()
     lr.fit(X_train_reg, y_train)

     lr_predict = lr.predict(X_test_reg)

[2]  math.sqrt(mean_squared_error(y_test, lr_predict))

[ ]  lr_train_predict = lr.predict(X_train_reg)

     math.sqrt(mean_squared_error(y_train, lr_train_predict))
```

Figure 2. Algorithm used for Linear Regression.

The model attains high predicting accuracy during the training process. After accomplishing the training, the testing of the model is done by utilizing the unseen thirty percent of data. The outcome during testing is indicated in the Figure. 3(b). where Tangerine lines represent the model's predicted cost, and the blue line reflects the actual cost. As the data is prepared by using a moving window consisting of a window length of ten results, the first ten values are absent in predicting the plot.

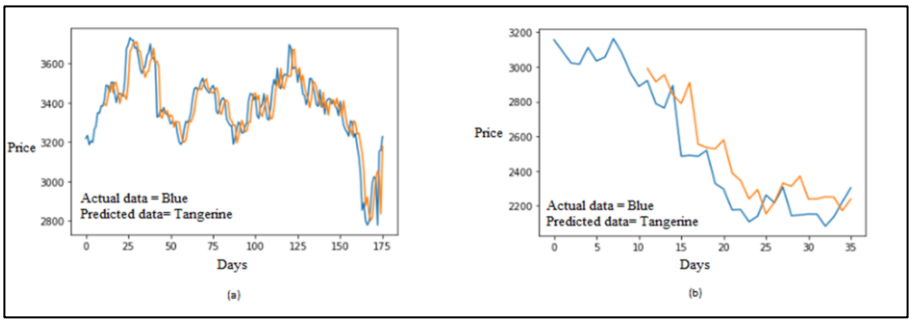

Figure 3. (a) Training Prediction Plot (b) Testing Prediction Plot of Linear Regression Model.

To calculate the performance of a model, different researchers use different formulas like MAE mean absolute error, MAPE mean absolute percentage error MSE means mean squared error, RMSE (root mean squared error) etc. We used RMSE as it is used in most of the regression based predictive analysis. The formula of RMSE is:

$$RMSE = \sqrt{1/N\sum_{i=1}^{n} = (d_i - z_1)^2} \tag{2}$$

Here di represents the actual cost and zi represents the predicted cost n represents number of samples in a set [9].

3.2. Support Vector Mechanism and Random Forest

Random forest and support vector mechanism are also supervised machine learning models which are used in different predictive analyses. To frame a comparative analysis, the same input data matrix is utilized during linear regression are analyze over the random forest and support vector machine model without varying the hyper parameter. The x_train_reg data is directly applied to the model for training. Once the model is trained then the testing of the model is done by using x_test_reg. At last, RMSE is

calculate in order to analyses the performance of the model. The designed Random Forest model consist a negligible variation in hyperparameter where max_depth=3 and n_estimator= 500. The x_train_reg data is applied to the model for training. Once the model is trained then the testing of the model is done by using x_test_reg. The output graph of the training and testing results of both models is shown in figure 4 and 5.

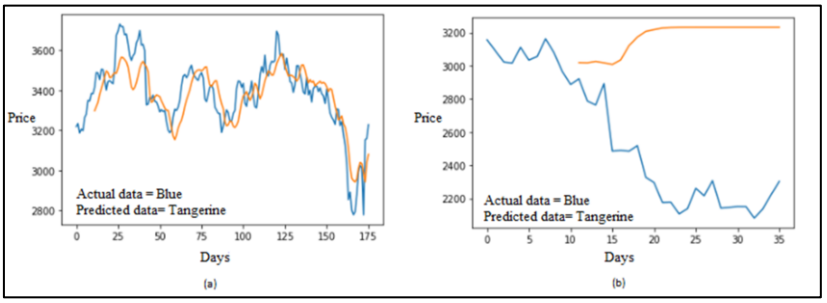

Figure 4. (a)Training Prediction Plot (b) Testing Prediction plot of Support Vector Mechanism

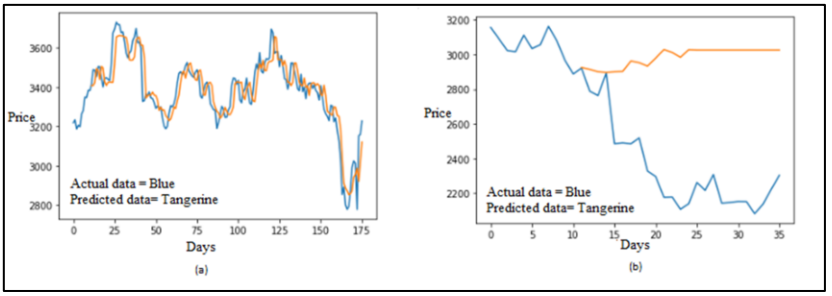

Figure 5. (a) Training Prediction Plot (b) Testing Prediction Plot of Random Forest

Tangerine lines represent the models' predicted cost, and the blue line reflects the actual cost. Table 1 shows output results for all three companies by using different models.

Table 1. Observed RMSE result of support vector machine random forest and linear regression.

S. No	Company Name	SVM (RMSE)	RANDOM FOREST (RMSE)	LINEAR REGRESSION (RMSE)
1	Walmart	0.3354	0.2289	0.0986
2	Amazon	0.5314	0.4176	0.0859
3	Tesla	0.1583	0.0924	0.0709

4. Conclusion

Prediction of the stock market is a difficult task. However, with the advancement of technology, machine learning can provide a reliable path for the prediction of the stock market. In this research, three distinct machine learning models is developed for stock

market prediction. We did not vary any hyper-parameters of the designed models, so it can be considered that all the models equal. The stock data used is taken from the American stock exchange. Based on the data presented in table 1, it is evident that the designed linear regression model can give better prediction accuracy than the support vector mechanism and random forest. The least RMSE achieved is of 0.0709 for Tesla. It is observed that the accuracy of different models may vary as the dataset changes. The drawbacks of the individual model can be resolved by implying hybrid model. A hybrid model by using linear regression and random forest can give better results in stock market prediction.

References

[1] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "NSE Stock Market Prediction Using Deep-Learning Models," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1351–1362, 2018, doi: 10.1016/j.procs.2018.05.050.

[2] T. J. Strader, J. J. Rozycki, T. H. Root, and Y.-H. (John) Huang, "Machine Learning Stock Market Prediction Studies : Review and Research Directions," *J. Int. Technol. Inf. Manag.*, vol. 28, no. 4, pp. 63–83, 2020, [Online]. Available: https://scholarworks.lib.csusb.edu/jitim/vol28/iss4/3.

[3] R. Batra and S. M. Daudpota, "Integrating StockTwits with sentiment analysis for better prediction of stock price movement," *2018 Int. Conf. Comput. Math. Eng. Technol. Inven. Innov. Integr. Socioecon. Dev. iCoMET 2018 - Proc.*, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICOMET.2018.8346382.

[4] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, "Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market," *IEEE Access*, vol. 8, pp. 22672–22685, 2020, doi: 10.1109/ACCESS.2020.2969293.

[5] A. Izzah, Y. A. Sari, R. Widyastuti, and T. A. Cinderatama, "Mobile app for stock prediction using Improved Multiple Linear Regression," *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, vol. 2018-Janua, pp. 150–154, 2018, doi: 10.1109/SIET.2017.8304126.

[6] M. J. Awan, M. S. M. Rahim, H. Nobanee, A. Munawar, A. Yasin, and A. M. Zain, "Social Media and Stock Market Prediction: A Big Data Approach," *Comput. Mater. Contin.*, vol. 67, no. 2, pp. 2569–2583, 2021, doi: 10.32604/cmc.2021.014253.

[7] D. G. Singh, "Machine Learning Models in Stock Market Prediction," *Int. J. Innov. Technol. Explor. Eng.*, vol. 11, no. 3, pp. 18–28, 2022, doi: 10.35940/ijitee.c9733.0111322.

[8] I. Gupta, H. Mittal, D. Rikhari, and A. K. Singh, "MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of A Day," 2022, [Online]. Available: http://arxiv.org/abs/2203.05835.

[9] M. Waqar, H. Dawood, B. Shahnawaz, and M. A. Ghazanfar, "Prediction of Stock Market by Principal Component Analysis," pp. 599–602, 2017, doi: 10.1109/CIS.2017.00139.