

Prediction for Stock Marketing Using Machine Learning

Shubham Jain

Student, Department of Information Technology
Maharaja Agrasen Institute of Technology, Delhi, India
shjain6670@gmail.com

Mark Kain

Student, Department of Information Technology
Maharaja Agrasen Institute of Technology, Delhi, India
markkain24@gmail.com

Abstract – Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. This paper will showcase how to perform stock prediction using Machine Learning algorithms: Linear Regression, Random Forest and Multilayer Perceptron.

Keywords : *Sentiment Analysis , Machine Learning, Support Vector Machine, Linear Regression, Random Forest, Multilayer Perceptron*

I. INTRODUCTION

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. In this paper, we are going to discuss the method to predict stock market values using machine learning algorithms: Linear Regression, Random forest and Multilayer Perceptron.

In statistics, **linear regression** is a linear approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X . The case of one explanatory variable is called *simple linear regression*. For more than one explanatory variable, the process is called *multiple linear regression*.

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

A **multilayer perceptron** (MLP) is a class of feedforward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

The paper focuses on the task of predicting future values of stock market index. Two indices namely, Dow Jones Industrial

Average and NYTimes are selected for experimental evaluation. Experiments are based on 10 years of historical data of these two indices.

The rest of the paper is organised as: Section 2 provides the overview of Literature Survey of Stock Market Prediction, Section 3 describes methodology used in paper. Section 4 shows result analysis. Finally, Section 5 delivers conclusions made through predictions.

II. LITERATURE SURVEY

Kannan, Sekar, Sathik and P. Arumugam in [1] used data mining technology to discover the hidden patterns from the historic data that have probable predictive capability in their investment decisions. The prediction of stock market is challenging task of financial time series predictions.

Jing Tao Yao and chew Lim tan in [2] used artificial neural networks for classification, prediction and recognition. Neural network training is an art. Trading based on neural network outputs, or trading strategy is also an art. Authors discuss a seven-step neural network prediction model building approach in this article. Pre and post data processing/analysis skills, data sampling, training criteria and model recommendation will also be covered in this article.

Tiffany Hui-Kuang and Kun-Huang Huarng in [3] used neural network because of their capabilities in handling nonlinear relationship and also implement a new fuzzy time series model to improve forecasting. The fuzzy relationship is used to forecast the Taiwan stock index. Jigar Patel [7] focuses on the task of predicting future values of stock market index. Two indices namely CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex from Indian stock markets are selected for experimental evaluation. Experiments are based on 10 years of historical data of these

two indices. The paper proposes two stage fusion approach involving Support Vector Regression (SVR) in the first stage.

Ching-Hseue cheng, Tai-Liang chen, Liang-Ying Wei in [4] this paper proposed a hybrid forecasting model using multi-technical indicators to predict stock price trends. They used RST algorithm to extract linguistic rules and utilize genetic algorithm to refine the extracted rules to get better forecasting accuracy and stock return.

Fazel Zarandi M.H, Rezaee B, Turksen I.B and Neshat E in [5] used a type-2 fuzzy rule based expert system is developed for stock price analysis. The purposed type-2 fuzzy model applies the technical and fundamental indexes as the input variables. The model used for stock price prediction of an automotive manufactory in Asia. The output membership values were projected onto the input spaces to generate the next membership values of input variables and tuned by genetic algorithm.

Ingoo Han [6] proposes genetic algorithms (GAs) approach to feature discretization and the determination of connection weights for artificial neural networks (ANNs) to predict the stock price index. In this study, GA is employed not only to improve the learning algorithm, but also to reduce the complexity in feature space.

Time series prediction techniques have been used in many real-world applications such as financial market prediction, electric utility load forecasting, weather and environmental state prediction, and reliability forecasting. Ravi Shankar [8] provides a survey of time series prediction applications using a novel machine learning approach: support vector machines (SVM).

T. Jan [14] surveys machine learning techniques for stock market prediction. He present recent developments in stock market prediction models, and discuss their advantages and disadvantages. In addition, we investigate various global events and their issues on predicting stock markets.

Y. Kara [9] attempted to develop two efficient models and compared their performances in predicting the direction of movement in the daily Istanbul Stock Exchange (ISE) National 100 Index. The models are based on two classification techniques, artificial neural networks (ANN) and support vector machines (SVM).

Bo Qian [10] investigated the predictability of the Dow Jones Industrial Average index to show that not all periods are equally random. He used the Hurst exponent to select a period with great predictability. Some inductive machine-learning

classifiers—artificial neural network, decision tree, and k -nearest neighbor were then trained with these generated patterns. Through appropriate collaboration of these models, he achieved prediction accuracy up to 65 percent.

E. Guresan [11] evaluates the effectiveness of neural network models which are known to be dynamic and effective in stock-market predictions. The models analysed are multi-layer perceptron (MLP), dynamic artificial neural network (DAN2) and the hybrid neural networks which use generalized autoregressive conditional heteroscedasticity (GARCH) to extract new input variables. The comparison for each model is done in two view points: Mean Square Error (MSE) and Mean Absolute Deviate (MAD) using real exchange daily rate values of NASDAQ Stock Exchange index.

B. Nath [12] deals with the application of hybridized soft computing techniques for automated stock market forecasting and trend analysis. We make use of a neural network for one day ahead stock forecasting and a neuro-fuzzy system for analyzing the trend of the predicted stock values.

Cheng-Yi Tesai [13] hybridizes SVR with the self-organizing feature map (SOFM) technique and a filter-based feature selection to reduce the cost of training time and to improve prediction accuracies. The hybrid system conducts the following processes: filter-based feature selection to choose important input attributes; SOFM algorithm to cluster the training samples; and SVR to predict the stock market price index. The proposed model was demonstrated using a real future dataset – Taiwan index futures (FITX) to predict the next day's price index.

III. METHODOLOGY

We first imported package: numpy, pandas and Natural Language Toolkit in Jupyter Notebook. And, read the saved pickled data file in a table. Then, we selected tuples: price and article, and copied them into another table. We added new columns: 'compound'(compound rating of article), 'neg'(negative rating of article), 'pos'(positive rating of article) and 'neu'(neutral rating of article).

We then applied 'SentimentIntensityAnalyzer' from Natural Language Toolkit on our new table and calculated sentiment score for each article. Then, we created two data frames for testing data and training data namely, `y_test` and `y_train` respectively.

Packages 'treeinterpreter' and 'sklearn' are imported in Jupyter Notebook. Then, through `predict()` function of `RandomForestRegressor` class in package 'treeinterpreter'

package, three data frames are created: 'prediction', 'bias' and 'contribution'.

Another package called 'matplotlibs' is imported. A random forest is plotted through pyplot without smoothening.

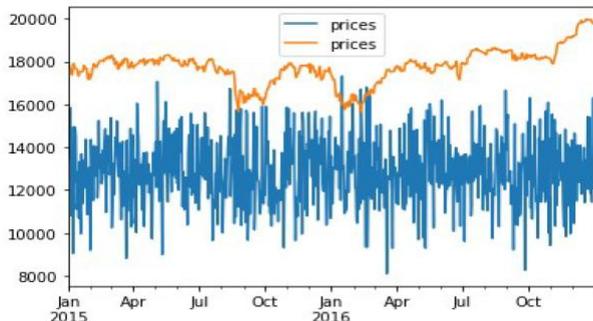


Figure 1. Plotting Random Forest Without Smoothening

Then, we modify the prices by a constant value so that the predicted prices are close to those actual prices of articles. We add a constant 6117 to all the predicted prices. Then, we apply 'Exponential Weighted Mean Average' from pandas to smooth the stock prices. Then, predictions after smoothening are plotted.

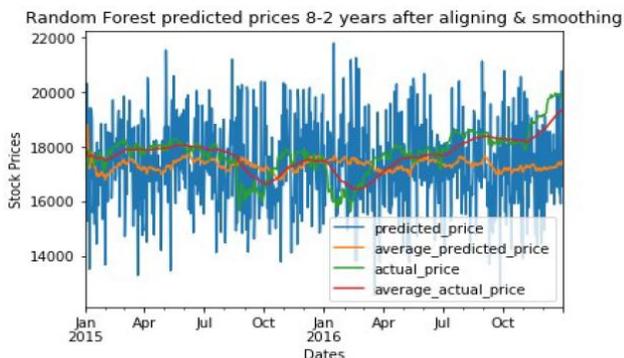


Figure 2. Plotting Random Forest after smoothening and aligning

Similarly, we plot for Linear Regression and Multilayer Perceptron algorithms. Also, To use Linear Regression and Multilayer Perceptron(MLP) algorithm, we import LogisticRegression class and MLPClassifier from 'sklearn'.

IV. ANALYSIS

Using Machine Learning algorithms, we were able to predict stock values based 10 years pickled data, as well as express them by plotting a graph through pyplot of matplotlibs.

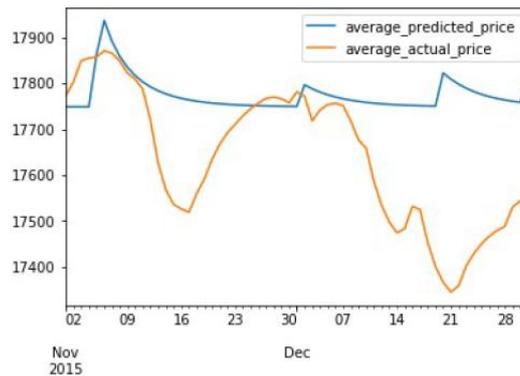


Figure 3. Plotting Logistic Regression on Training Data

In Figure 3, The curve for average predicted price meets with the curve for average actual price are quite close and meet at four points. But as we move rightwards, in the month of December, the difference between the average predicted price and average actual price is quite large.

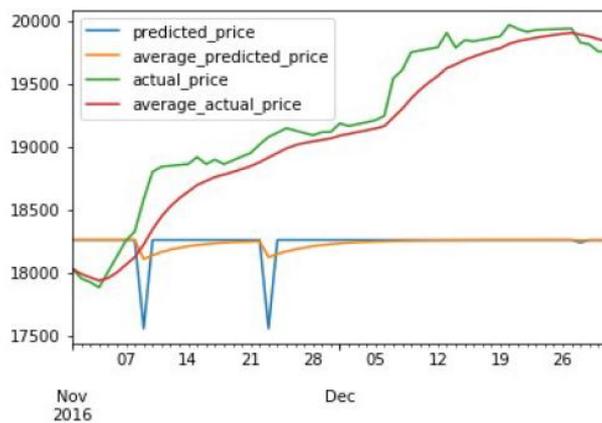


Figure 4. Plotting Logistic Regression on Test Data

In Figure 4, using Logistic Regression, we see that the curves for the average predicted price and predicted price are quite close. Similarly, curves for actual price and average actual price are quite close. But the difference between the actual price and predicted price after 9 November 2016 becomes larger as we move rightwards.



Figure 5. Plotting Random Forest on Training Data

In Figure 5, we can see that difference between average predicted price and average actual price is small compared to Logistic Regression(Figure 3) till date 10 November 2015. After that, Difference between the curves become larger.

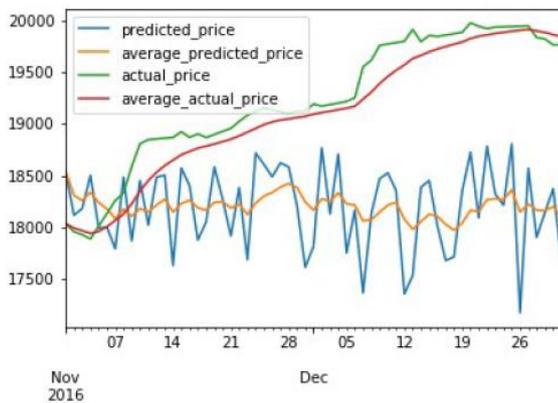


Figure 6. Plotting Random Forest on Test Data

In Figure 6, Difference between actual price and predicted price, as well as, average actual price and average predicted price becomes larger after 11 November 2016. But before that the difference is quite small as compared to that in Logistic Regression

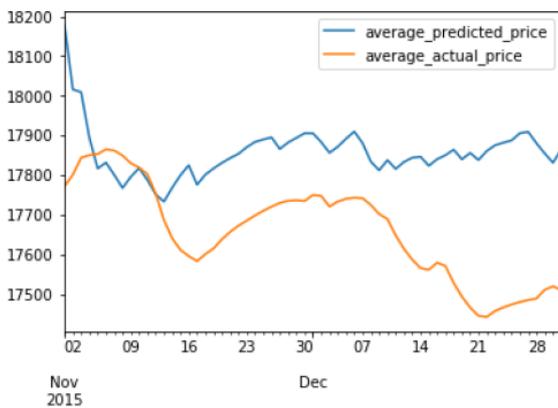


Figure 7. Plotting MLP on Training Data

In Figure 7, after 14 November 2015, difference between average predicted price and average actual price becomes larger. But till 11 December 2015, the difference is quite small as compared to Figure 3 and Figure 5. After that it becomes larger similarly to both the previous figures.

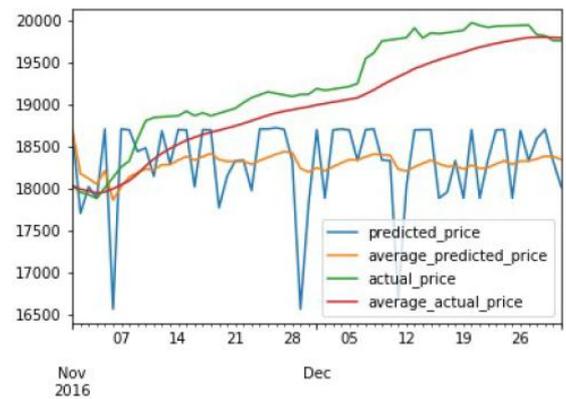


Figure 8. Plotting MLP on Predicted Values

In Figure 8, after 5 December 2016, the difference between the actual price and predicted price becomes quite larger. Before 10 November 2016 the difference is quite small. Between 10 November 2016 and 5 December, it is quite small as compared to Figure 4 and Figure 6.

V. CONCLUSION

Stock market prediction is important factor in finance. It is considered to be dynamic in nature. The paper presented how to predict stock values based on the data of NY Times of 10 years using Machine Learning algorithms: Logistic Regression, Random Forest and Multilayer Perceptron(MLP). We also concluded that MLP is better than the other two algorithms because, within a certain range, the difference between actual price and predicted price is quite small as compared to those in Logistic Regression and Random Forest. Also, Random Forest is better than Logistic Regression, but inferior to MLP, in predicting stock values.

ACKNOWLEDGEMENT

First and foremost, I wish to express my profound gratitude to Ms. Neha Singh (Assistant Professor) for giving me the opportunity to carry out the project . My heartfelt thanks for her invaluable guidance, immense help, support & useful suggestions throughout the course of my project work.

Last but not least I thank the Almighty for enlightening me with his blessings .

REFERENCES

- [1] K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik and P. Arumugam, "Financial stock market forecast using data mining Techniques", Proceedings of the international multiconference of engineers and computer scientists, 2010.
- [2] JingTao YAO, Chew Lim TAN, "Guidelines for Financial Prediction with Artificial neural networks".
- [3] Tiffany Hui-Kuang yu, Kun-Huang Huarng, "A Neural network-based fuzzy time series model to improve forecasting", Elsevier, 2010, pp: 3366-3372..

-
- [4] Ching-Hsue Cheng, Tai-Liang Chen, Liang-Ying Wei, "A hybrid model based on rough set theory and genetic algorithms for stock price forecasting", Pages. 1610-1629, 2010.
- [5] M. H. Fazel Zarandi, B. Rezaee, I. B. Turksen and E.Neshat, "A type-2 fuzzy rule-based experts system model for stock price analysis", Expert systems with Applications, Pages. 139-154, 2009.
- [6] Kyoung-jae Kim, Ingoo Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", Expert Systems with Applications, Vol. 19, Issue 2, Pages. 125-132, 2000.
- [7] J. Patel, S. Shah, P. Thakkar, K. Kotecha, "Predicting stock market index using fusion of machine learning techniques", Expert Systems with Applications, Vol. 42, Issue 4, Pages. 2162-2172, March 2015.
- [8] R. Shankar, N.I. Sapankevych, "Time Series Prediction Using Support Vector Machines: A Survey", IEEE Computational Intelligence Magazine, Vol. 4, Issue 2, Pages 2162-2172, March 2009.
- [9] Y. Kara, M. A. Boyacioglu, Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange", Expert Systems with Applications, Vol. 38, Issue 5, Pages 5311-5319, May 2011.
- [10] Bo Qian, K. Rasheed, "Stock market prediction with multiple classifiers", Applied Intelligence, Vol. 26, Issue 1, Pages 25-33, February 2007.
- [11] E. Guresan, G. Kayakutlu, T. U. Diam,, "Using artificial neural network models in stock market index prediction", Expert Systems with Applications, Vol. 38, Issue 8, Pages 10389-10397, August 2011.
- [12] A. Abraham, B. Nath, P. K. Mahanti,, "Hybrid Intelligent Systems for Stock Market Analysis", Computational Science - ICCS 2001, Pages. 337-347.
- [13] Cheng-Yi Tesai, Cheng-Lung Huang, "A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting", Expert Systems with Applications, Vol. 36, Issue 2, Pages 1529-1539, March 2009.
- [14] P. D. Yoo, M. H. Kim, T. Jan, "Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation", IEEE Computational Intelligence for Modelling, November 2005.