IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Stock price prediction using machine learning

[1]Soham K Patil, [2]Saiprasad A Gavhane, [3]Himanshi Daharwal, [4]Prof. Prachi R Salve

[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1]Computer Engineering Department,
[1]Dr.Dy Patil School Of Engineering, Ambi, Pune, India

Abstract: The Stock Market is a large collection of markets and exchanges where equities of publicly held companies can be traded. Trading in stocks has many benefits, and can be a lucrative income if the market is well understood. Companies are able to gain access to capital by selling off slices of ownership to investors, and the investors have the opportunity to gain income and assets. The Stock Market can also be one of the better predictors in determining the health and direction of an economy. It can help predict if economic and political policies are paying off, whether or not housing prices are going to rise, and influence the size of a nation's workforce . For these reasons, predicting the Stock Market can be very advantageous. The issue is the market is very volatile and influenced by a large number of factors. Many argue that the market cannot be predicted, and in a strong definition that is likely correct. However, the market often does have concrete trends that can be analyzed, and therefore may still be able to be reasonably predicted in short periods of time. In this project, Machine Learning concepts with a technical analysis of individual stocks in an attempt to predict their stock prices.

Index Terms - **Machine Learning, Linear Regression, Stock Market, Time Series Splitting**

I. INTRODUCTION

Accurate financial prediction is of great interest for investors. This paper proposes use of Data analytics to be used in assist with investors for making right financial prediction so that right decision on investment can be taken by Investors. Two platforms are used for operation: Python and R. various techniques like Arima, Holt winters, Neural networks (Feed forward and Multi-layer perceptron), linear regression and time series are implemented to forecast the opening index price performance in R. While in python Multi-layer perceptron and support vector regression are implemented for forecasting Nifty 50 stock price and also sentiment analysis of the stock was done using recent tweets on Twitter. Nifty 50 (A NSEI) stock indices is considered as a data input for methods which are implemented. 9 years of data is used. The accuracy was calculated using 2-3 years of forecast results of R and 2 months of forecast results of Python after comparing with the actual price of the stocks. Mean squared error and other error parameters for every prediction system were calculated and it is found that feed forward network only produces 1.81598342% error when opening price of stock is forecasted using it. Stock price prediction is a difficult task. It is because there is no certain variable that can precisely predict the stock price every day. Based on Efficient Market Hypothesis (EMH), new information is a significant factor that effects changes of stock price [1]. This information, such as news about company can influence people decision whether or not they will buy the company's stock. More people buy the company's stock, the price are getting higher. People tend to buy a company with good reputation. One way to know company's reputation is by seeing relationship between the company and customer [2]. The explosion of social media usage force many companies to create their official account in social media in order to keep in touch with their customer. This make customer can express their opinion about products easily. One of the social media that commonly used by company is Twitter. There are several researches about how the information from social media can affects the stock price. Based on research conducted by Johan Bollen, et.al[3], it concluded that certain mood states of Twitter data can predict the Dow Jones Industrial Average (DJIA) value with 87.6% accuracy. Another research conducted by Anshul Mittal and Arpit Goel[4], shows that with the DJIA value, calmness and happiness mood states of twitter data on previous days can predict the DJIA value on the current day with 75.56% accuracy. This shows that information from Twitter can really be used to predict stock data. Indonesia is the 5th country with highest number of Twitter active user, especially Jakarta where 2.4% of all Twitter post comes from [5]. Since the previous research mentioned was using English, author was curious about the effects of Twitter data to stock price of Indonesian Company. This problem gives us motivation to conduct this research. The contribution of this research lies in the use of existing classification and prediction algorithm to the dataset. The dataset consists of twitter dataset and stock price dataset. Twitter dataset used was in Bahasa and stock price dataset retrieved from several companies in Indonesia.

II. LITERATURE SURVEY

This paper examines the theory and practice of regression techniques for prediction of stock price trend by using a transformed data set in ordinal data format. The original pre transformed data source contains data of heterogeneous data types used for handling of currency values and financial ratios. The data formats in currency values and financial ratios provide a process for computation of stock prices. The transformed data set contains only a standardized ordinal data type which provides a process to measure rankings of stock price trends. The outcomes of both processes are examined and appraised. The primary design is based on regression analysis from WEKA machine learning software. The stock price movement in Bursa Malaysia is used as our research setting. The data sources are corporate annual reports which included balance sheet, income statement and cash flow statement. The variables included in the data set were formed based on stock market trading fundamental analysis approach. Classifiers in WEKA were used as algorithms to produce the outcomes. This study showed that the outcomes of regression techniques can be improved for the prediction of stock price trend by using a dataset in standardized ordinal data format.[1]

Stock price prediction is a difficult task, since it very depending on the demand of the stock, and there is no certain variable that can precisely predict the demand of one stock each day. However, Efficient Market Hypothesis (EMH) said that stock price also depends on new information significantly.

One of many information sources is people's opinion in social media. People's opinion about products from certain companies may determine the company's reputation and thus affecting people's decision to buy the stock of the company. When using opinion as primary data, it is necessary to make a suitable analysis of it. A famous example using opinion as data is sentiment analysis. Sentiment analysis is a process to determine emotion/feeling within people opinion about something, in this case products of some companies. There are some researches about sentiment analysis used to predict the stock prices. Bollen on his research concludes that people opinion on social media such as Twitter can predict DJIA value with 87.6% accuracy. This shows that there is a relation between sentiment analysis and stock prices. Our purpose on this research is to predict the Indonesian stock market using simple sentiment analysis. Naive Bayes and Random Forest algorithm are used to classify tweet to calculate sentiment regarding a company. The results of sentiment analysis are used to predict the company stock price. We use linear regression method to build the prediction model. Our experiment shows that prediction models using previous stock price and hybrid feature as predictor gives the best prediction with 0.9989 and 0.9983 coefficient of determination [4].

III. EXISTING SYSTEM

People’s opinion about products from certain companies may determine the company’s reputation and thus affecting people’s decision to buy the stock of the company. When using opinion as primary data, it is necessary to make a suitable analysis of it. A famous example using opinion as data is sentiment analysis. Sentiment analysis is a process to determine emotion/feeling within people opinion about something, in this case products of some companies. There are some researches about sentiment analysis used to predict the stock prices.

IV. EXISTING SYSTEM DISADVANTAGE

Since it very depending on the demand of the stock, and there is no certain variable that can precisely predict the demand of one stock each day.

V. PROPOSED SYSTEM

Machine Learning concepts with a technical analysis of individual stocks in an attempt to predict their stock prices. The final implementation will be in the form of a web application that allows users to select up to 4 stock symbols, and the application will then output the predicted results and analysis of each stock. Prices based on learning and training from historical data. The continuous nature of the data makes it particularly suitable for regression type algorithms. After evaluating the types of algorithms available I have decided to use an Ordinary Least Squares Linear Regression algorithm, the K-Nearest Neighbor (KNN) Regressor algorithm, and a Random Forest Regressor algorithm to evaluate which returns the best results. Each of these algorithms are available in the Scikit Learn Python library. Another technique that will be utilized for all three models is Time Series Splitting for cross-validation.

Linear Regression: In this model, one technique that will be utilized is to pass the model what is known as a Standard Scalar. The idea of it is to bring each of the features passed to the linear regression within the same numerical scale as well as try representing the features as a normal distribution. The goal of the Linear Regression model is to fit a line to the data that can best generalize any trends. The output of the model will be an equation that can then be used to make future predictions.

VI. PROPOSED SYSTEM ADVANTAGES

This application will report future stock predictions over a 14-day period in intervals of 7 days. Ex: 1 day out, 7 days out, 14 days out from the current date. All data sets were obtained from Yahoo Finances

VII. SYSTEM ARCHITECTURE

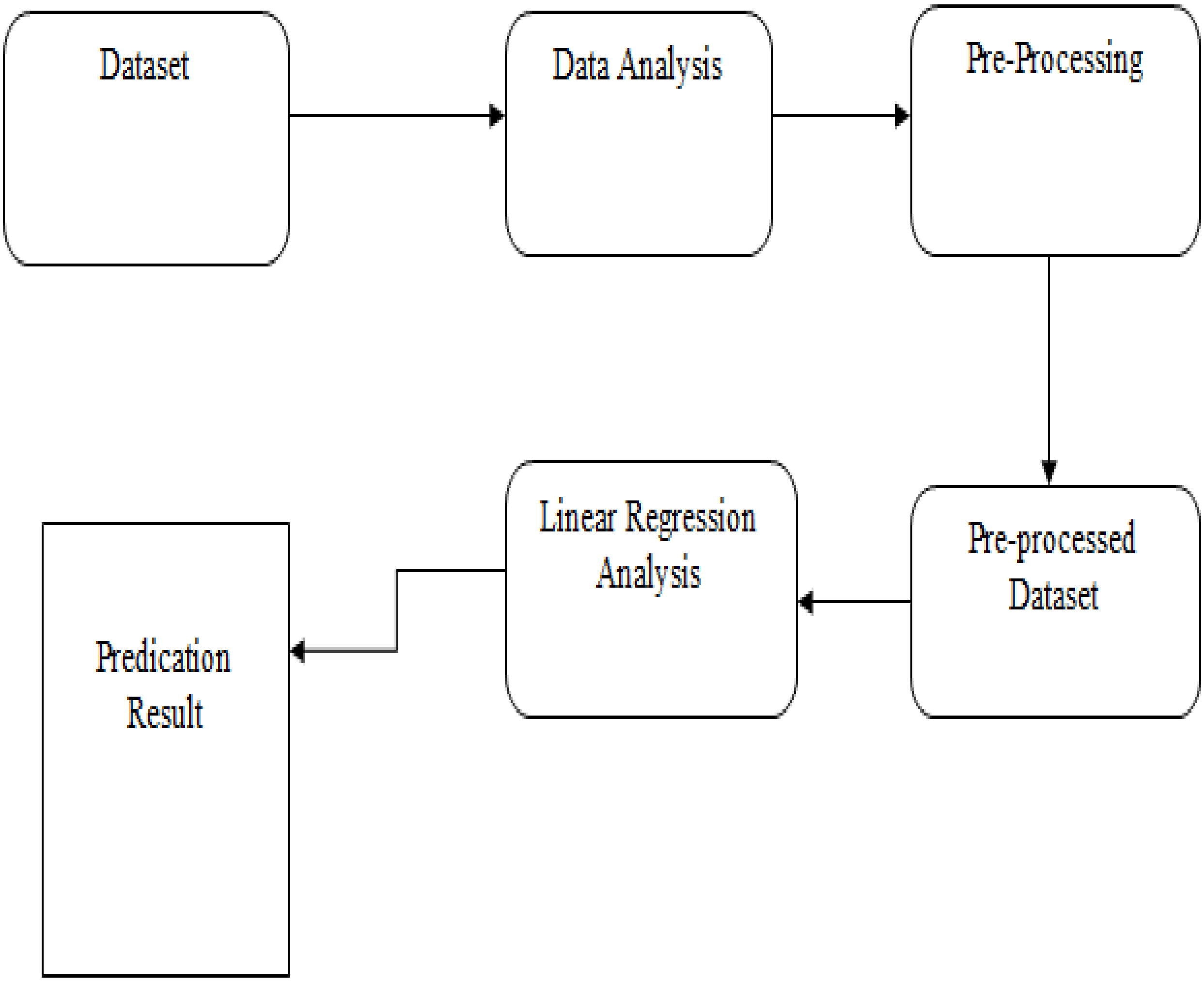

Fig. System architecture diagram of proposed system

VIII. ALGORITHMS AND TECHNIQUES

8.1 Linear Regression

In this model, one technique that will be utilized is to pass the model what is known as a Standard Scalar. The idea of it is to bring each of the features passed to the linear regression within the same numerical scale as well as try representing the features as a normal distribution. The goal of the Linear Regression model is to fit a line to the data that can best generalize any trends. The output of the model will be an equation that can then be used to make future predictions.

8.2 Time Series Splitting

Cross-validation is a useful technique for training and testing a model by utilizing all the available data for testing and training. The goal is to reduce overfitting. However, traditional cross-validation does not preserve data ordering in the process which is important in our problem domain. Time Series Split will preserve the ordering solving this issue.

IX. CONCLUSION

One interesting quality about the problem domain of the project is how dependent a stock prediction is on previous historical dates and values. Given its continuous nature one can reason why predicting the future of a stock over long periods of time purely from technical analysis is a futile task. In my model that surmounts to only a few days. A stocks motion (rise or fall in value) is only evident in comparison to a defined amount of dates prior. This concept is illustrated by the trend noticed in the predicted values. In the final model, the further out the prediction dates are the prediction values start to see a steady decline.

X. Future Enhancements

It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Some of the future enhancements that can be done to this system are:

- As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment.
- Based on the future security issues, security can be improved using emerging technologies like single sign-on.

XI. RESULT

In below Result graph show the results difference between the KNN-Linear Regression Algorithms and Random Forest Linear Algorithm in ms.

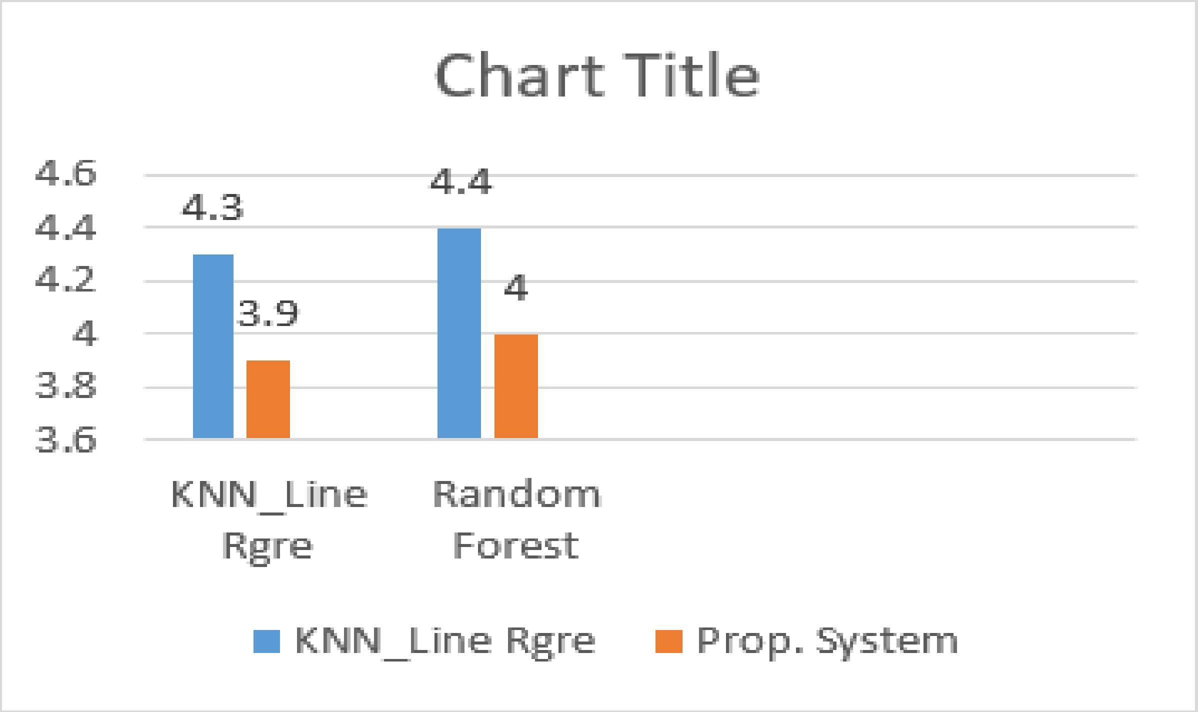

Fig. Result graph

Table: Result table

Type/Time in ms	Ex. System	Prop. System
KNN Linear Regression	4.3	3.9
Random Forest	4.4	4

IJCRT

REFERENCES

[1]Bracey, L. (n.d.). The Importance of Business Reputation. Retrieved Juli 9, 2015, from Business in Focus: http://www.businessinfocusmagazine.com/2012/10/the-importance-ofbusiness-reputation/

[2] Bollen, J., Mao, H., & Zeng, X. J. (2010). Twitter mood predicts the stock market. arXiv .

[3] Mittal, A., & Goel, A. (2009). Stock Prediction Using Twitter Sentiment Analysis. CiteSeerX

[4] Berita 8. (2013, November 21). Ini 5 Negara Pengguna Aktif Twitter Terbanyak di Dunia. Retrieved June 25, 2015, from Berita http://www.berita8.com/berita/2013/11/ini-5-negara-pengguna-aktiftwitter-terbanyak-di-dunia/

[5] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Claypool Publishers. [7] Witten, I. H., Frank, E., &9 Hall, M. A. (2011). Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed). Burlington: Morgan Kaufmann, pp. 191-192

[6] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed). Burlington: Morgan Kaufmann, pp. 90-93

[7] Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering , 1114-1119.

[8] Cutler, A., Cutler, D. R., & Stevens, J. R. (2008). Tree-based Method. In X. Li, & R. Xu, High-Dimensional Data Analysis in Cancer Research (pp. 89-109). New York: Springer Science & Business Media.

[9] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed). Burlington: Morgan Kaufmann, pp. 127-129

[10] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed). Burlington: Morgan Kaufmann, pp. 124-125

IJCRT