Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 04, April-2019

Stock Market Trend Analyzer

Rohan Sethi
Department of Information Technology
Maharaja Agrasen Institute of Technology
Guru Gobind Singh Indraprastha University
New Delhi, India

Sanjam Chhabra
Department of Information Technology
Maharaja Agrasen Institute of Technology
Guru Gobind Singh Indraprastha University
New Delhi, India

***Abstract* - The million-dollar question for stock investors is if the price of a stock will rise or not. The fluctuation of the stock market is violent and there are many complicated financial indicators. This topic has always attracted interest of researchers from different fields. Numerous studies have been conducted in the stock market to predict trends using machine learning algorithms such as Support Vector Machine (SVM) and Reinforcement learning. The prediction of a stock's future price will maximize investors' gains. In this project, we propose a machine learning prediction model that will predict stock prices. We have used Logistic Regression and Random Forest machine learning algorithms that exploit the relation between articles about stocks and stocks prices on a given day to predict the trend (UP/Down) and stock price for future. We combined results from sentiment analysis performed on various articles with historical data to form the input dataset.**

Keywords - Stocks, Reinforcement learning, Random Forest, Market Trend, Logistic Regression

I. INTRODUCTION:

Stock market is a very vast topic and difficult to understand. It is too uncertain to be predictable due to large fluctuations of the market. Stock market prediction divides researchers and academics into two groups, those who believe that market can be predicted through mechanisms and those who believe that the market varies on daily basis and can't be predicted. Investing in a good stock but at bad time can have unprofitable result, while investing in a stock at the right time can bear good profits. Prediction cannot be based on one factor. There are multiple factors that need to be considered while prediction a market, factors like news, social media data, government bonds, historical price and country's economics. Hence, including these factors must increase the accuracy of the model. There are two common methods to predict the stock market prices. One is Fundamental Analysis and second one is technical analysis. Proposed method is built on the principle of technical as well as fundamental principles. The goal of this research work is to build a model which predicts stock trend movement (trend will be up or down) and stock value using historical data and news about the stock.

Proposed model predicts the future trend for the next day, it considers both sentiment and historical data. New York Times archive data has been used to compute the sentiment of the stock. Its outcome with open price, close price of stock with extracted statistical parameters is considered to build the model.

II. METHODOLOGY:

This section deals with different methodologies carried out in this project such as Data interpretation, Data Transformation, Application of Classification Techniques along with other miscellaneous tasks (Most Prominent Features). Fig 1 depicts the bifurcation of the two disparate pathways taken in the process of research

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 04, April-2019

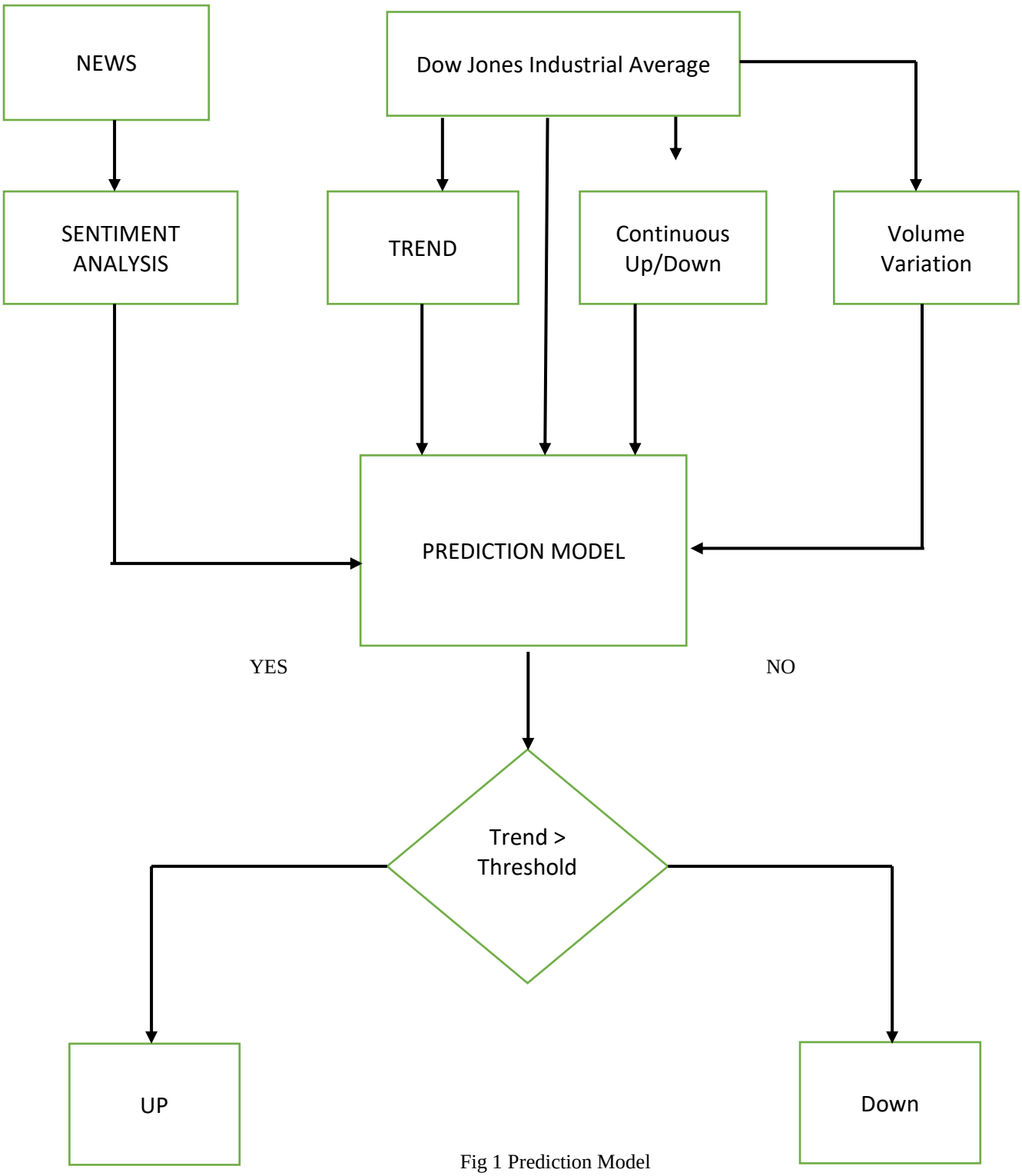

Fig 1 Prediction Model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 04, April-2019

First contribution to the proposed model is that few features has been deduced from the sentimental analysis of news about stocks. Second contribution to the model is historical data available from statistics. One of the statistical parameter considered is relationship between trend of a day and volume of stock traded on that day. Volume traded feature in historical data will reflect both bought and sold stocks on a daily basis. When the trade is down volume traded it might reflect shares bought by traders and when the trend is up volume traded might indicate the sold shares. So, to understand whether the stocks are bought or sold by the trader the above feature along with trend of the day feature can be used. If and only if shares are purchased by the trader big number of volume traded has positive impact. Assumption for the shares purchased is stock transactions are more and trend is down. If volume traded is more and trend is up means, shares are sold to gain money. Prediction model is built by considering historical price dataset and sentiment dataset. Prediction model for the same is as shown in Fig. 1. Fig 1 shows input and output of the prediction model. Patterns like continuous up/down, volume analysis has been derived from DJIA finance data. Sentiment from news dataset is also given as input to the model. Supervised machine learning algorithm is used to train the prediction model. Model is tested on test data, which tells whether threshold is up or down.

III. DATA INTERPRETATION

A. *Data collection:* Data is gathered using following methods.

- News information is collected from New York Times Archive API using python language.
- Historical data is collected from Yahoo Finance

B. *Data Processing*: Collected data is processed using following methods.

- *Lemmatization:* Lemmatization is applied to each row to get the all words to common form which is helpful while assigning polarity to each word. Lemmatization is done with the help of natural language tool kit (NLTK) package which is available in python.

C. *Data Analysis:* Collected data has to be analyzed to get the sentiment on each day.

- *Assigning Polarity*: Polarity has been assigned to each word using SentimentIntensityAnalyzer. Polarity of each item is divided into 3 categories depending upon positive, negative, neutral.

```
In [7]: sentence = 'paris shootout police officer suspected guman dead'
        from nltk.sentiment.vader import SentimentIntensityAnalyzer
        import unicodedata
        sentimentintensity = SentimentIntensityAnalyzer()
        score = sentimentintensity.polarity_scores(sentence)
        score

Out[7]: {'compound': -0.7351, 'neg': 0.554, 'neu': 0.446, 'pos': 0.0}
```

Fig 2 Assigning Sentiment Score

IV. DATASET

A. *Historical prices*

Historical prices are obtained from Yahoo Finance. Each transaction date consists of open price, close price, low price, high price, adjusted close price and volume traded on that day. Adjusted close price and close price depicts the close price of stock on a particular day. Adjusted close price will be adjusted for dividends and splits. Adjusted close price is considered as stock price as in other researches.

B. *Sentiment dataset*

Sentiment dataset has been created by considering news dataset. News has been collected for the same year range for which stock prices were taken. Sentiment of the news is integrated on a daily basis. The classification models are built for stock market data analysis. Performance of the model is evaluated using accuracy metric. Accuracy can be defined as proportion of true results in the test dataset.

V. DATA TRANSFORMATION

Such a plethora of data cannot be used directly to train models. The computational power required is enormous. Times like these call for the need of feature transformation. This is a technique which can bring together data in an optimal format. In this project, we look at a statistical approach to transform our data with techniques - Minimum, Maximum, Mean, Median, Standard Deviation, Skewness, and Kurtosis. Each of these tuples thus provides a representation of prices and sentiment score for one stock.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 04, April-2019

	prices	compound	neg	neu	pos
200	click to scroll output; double click to hide			.749	0.093
2007-01-02	12472	-0.8179	0.114	0.787	0.099
2007-01-03	12474	-0.9993	0.198	0.737	0.065
2007-01-04	12480	-0.9982	0.131	0.806	0.062
2007-01-05	12398	-0.9901	0.124	0.794	0.082
2007-01-06	12406	-0.965	0.134	0.771	0.094
2007-01-07	12414	-0.9975	0.193	0.739	0.069
2007-01-08	12423	-0.9601	0.11	0.793	0.097
2007-01-09	12416	-0.9953	0.103	0.848	0.049
2007-01-10	12442	-0.9534	0.134	0.743	0.123
2007-01-11	12514	-0.998	0.128	0.814	0.057
2007-01-12	12556	-0.9986	0.158	0.784	0.059
2007-01-13	12562	-0.9893	0.146	0.794	0.059
2007-01-14	12569	-0.99	0.178	0.711	0.111

The data generated is classified as training and testing dataset. 60% of total data is considered as training dataset and rest 40% data is considered as testing dataset. The prediction model is built on the basis of the training dataset. Accuracy of the model is tested using the testing dataset.

VI. CLASSIFICATION TECHNIQUE

A. Logistic Regression:

It is a binary classification algorithm used when the response variable is either 1 or 0. It returns the set of probabilities of the target class. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. An Accuracy of 63.04% has been achieved by using a Logistic Regression model. Since the scikit-learn's [3] train_test_split() function uses an attribute called random_state which will affect the accuracy value, the accuracies are bound to change a little each time this function is run before applying the models.

B. Random Forest

Random Forests, in simpler terms, uses an ensemble of trees to make prediction. It uses averaging to improve the predictive accuracy and control overfitting which is why it more generalized and tolerant than decision trees. Random Forest adds additional randomness to the model, while growing the trees. It searches for the best feature among a random subset of features instead of searching for the most important feature while splitting a node. This results in a wide diversity that generally results in a better model. Random Forests model achieved an accuracy of 81.3%.

VII. RESULTS

Stock Market Trend was successfully analyzed and value was predicted for a particular day using the Machine Learning Techniques mentioned previously. The accuracies obtained for Logistic Regression model technique was 63.04% and for Random Forest model technique was 81.3%.

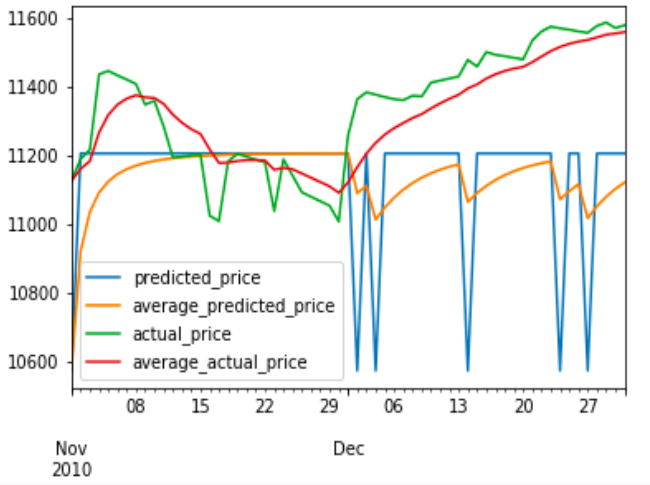

Fig 3 Prediction using Logistic Regression for November - December period.

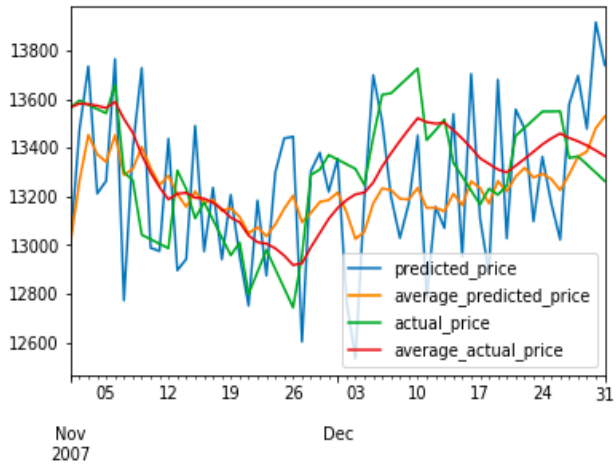

Fig 3 Prediction using Random forest for November - December period.

In this model, Scikit-learn's train_test_split() function has been used for the purpose of splitting the transformed dataset into training and testing data. This function intelligently selects an equal number of data for training the model. The function, however, employs a random number generator due to which accuracies vary a little every time it is run. Hence, a little variance in accuracies is expected when the models is run again for prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 04, April-2019

VIII. CONCLUSION AND FUTURE WORK

Recently, it has been observed that people are putting their money in stocks to get heavy returns and earn some money easily.
At the same time investor also worries to lose all their money. So an efficient predictive model is required. There are many predictive models which tell about the market trend whether it is up or down, but they fail to give accurate results due to the volatile stock market and its fluctuation. An attempt has been made to build a predictive model which can predict the daily trend and stock value by taking into account the historical data as well as sentiment of the news related to the stock. On the dataset considered for testing of this model, Random Forest performed better than Logistic Regression. The sentiment analysis also showed how the political and economic news and influence of the social media affects the future performance and volatility of the markets.

The prediction performance of this model can be improved using model based on probabilistic neural network techniques and reinforcement learning where appropriate policies can be decided and can be used to reward if correct prediction is achieved by model for a value and methods like back propagation can be used in case of neural networks to improve the performance and achieve higher accuracy. In addition, we can include a few other variables that may affect the prediction performance.

IX. REFERENCS

[1] S. Hannon, "5 Rules For Prediciting Stock Market Trends - StockTrader.com", StockTrader.com, 2016. [Online].
[2] Online Stock Trading Guide. Head and shoulders pattern, March 2015.
[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[4] S. A. R. Nai-Fu Chen and Richard Roll, Economic Forces and the Stock Market, The Journal of Business, vol. 59, no. 3, pp. 383–403, (1986). [Online]. Available: http://www.jstor.org/stable/2352710.
[5] P. A. G. Xue Zhang and Hauke Fuehres, Predicting Stock Market Indicators through Twitter I Hope it is not as Bad as I Fear, Procedia – Social and behavioral Sciences, vol. 26, pp. 55–62, (2011).
[6] J. Bollen, H. Mao and X. Zeng, Twitter Mood Predicts the Stock Market, Journal of Computational Science, vol. 2, no. 1, pp. 1–8, (2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187775031100007X
[7] K. Mizumoto, H. Yanagimoto and M. Yoshioka, Sentiment Analysis of Stock Market News with Semi-Supervised Learning, In 2012 IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS), pp. 325–328, May (2012).
[8] Jadav K, Panchal M. Optimizing weights of artificial neural networks using genetic algorithms. Int J Adv Res Comput Sci Electron Eng. 2012;1(10):47–51.
[9] Wang YH. Nonlinear neural network forecasting model for stock index option price: Hybrid GJR–GARCH approach. Expert Syst Appl. 2009;36(1):64–70.
[10] Wing-Keung Wong, Meher Manzur, and Boon-Kiat Chew. How rewarding is technical analysis? evidence from singapore stock market. Applied Financial Economics, 13(7):543–551, 2003.
[11] Burton Gordon Malkiel. A random walk down Wall Street: the time-tested strategy for successful investing. WW Norton & Company, 2003.
[12] Investopedia.com. Apr 2015. URL http://www.investopedia.com/terms/s/slippage.asp.
[13] R. Wilson and R. Sharda, "Bankruptcy prediction using neural networks", Decision Support Systems, vol. 11, no. 5, pp. 545-557, 1994.
[14] N. Lin, J. Yuan, W. Xu, L. Wei and X. Wang, How web News Media Impact Futures Market Price Linkage?, In 2013 Sixth International Conference on Business Intelligence and Financial Engineering (BIFE), pp. 562–566, November (2013)
[15] Mostafa MM. Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait.
[16] Hamed, I.M., Hussein, A.S., Tolba, M.F.: An Intelligent Model for Stock Market Prediction. International Journal Computational Intelligence Systems 5(4), 639–652 (2012)