Stock Market Prediction with Sentiment Analysis

Pornima Bansode

College of Engineering
Northeastern University
Boston, USA
bansode.p@husky.neu.edu

Vividh Talesara

College of Engineering
Northeastern University
Boston, USA
talesara.v@husky.neu.edu

***Abstract* - Efficient Market Hypothesis is the popular theory about stock prediction. With its failure much research has been carried in area of prediction of stocks. This project is about taking non-quantifiable data such as financial news articles about a company and predicting its future stock trend with news sentiment classification. Assuming that news articles have impact on stock market, this is an attempt to study relationship between news and stock trend. To show this, we created three different classification models which depict polarity of news articles being positive or negative. Observations show that Logistic Regression and Multilayer Perceptron perform well in all types of testing. Random Forest, Naïve Bayes gives good result but not compared to the other two. Experiments are conducted to evaluate various aspects of the proposed model and encouraging results are obtained in all the experiments. The accuracy of the prediction model is more than 80% and in comparison, with news random labelling with 50% of accuracy; the model has increased the accuracy by 30%**

Keywords- Random Forest, Logistic Regression, EWMA, Naïve Bayes, Polarity check, MLP, Multilayer Perceptron, Stock Price Prediction

I. INTRODUCTION

To get the stock price data we start fetching Stock Price Data for Dow Jones Industrial Average (DJIA) starting the year 2000 till date, New York Times data, and Yahoo finance. "According to the Efficient Market Hypothesis (EMH), stocks always trade at their fair value on stock exchanges, making it impossible for investors to either purchase undervalued stocks or sell stocks for inflated prices. As such, it should be impossible to outperform the overall market through expert stock selection or market timing, and the only way an investor can possibly obtain higher returns is by purchasing riskier investments. Basically, what that means is that the stock prices are hard to predict based on some expertise through previous trends and past stock prices. Stock price fluctuation represents the current market trends and business growth among other factors that could be considered to sell or buy stocks. To analyze the current trends, new company's product information, business growth etc., we could take a look at the daily news which represents factual information about the companies which could be ultimately used to predict the stock prices. Hence, we will be using news articles to predict the change in stock indices rather than predicting the prices by historical stock prices. Our next step is to clean the data and remove the irrelevant data to gain more accurate results. We performed different machine learning algorithms to find the nearest predicted value of the stock prices. To perform sentiment analysis on a particular day we added few columns like neutral, positive and negative. We have used Multilayer Perceptron to get the precise predicted values for the data. Below are the models performed:

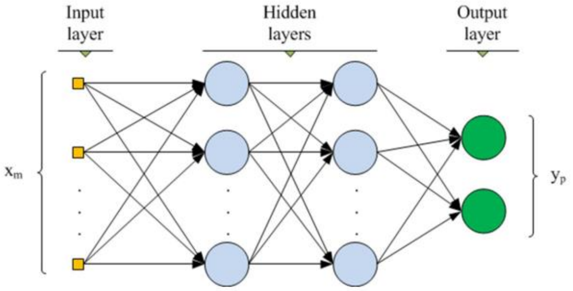
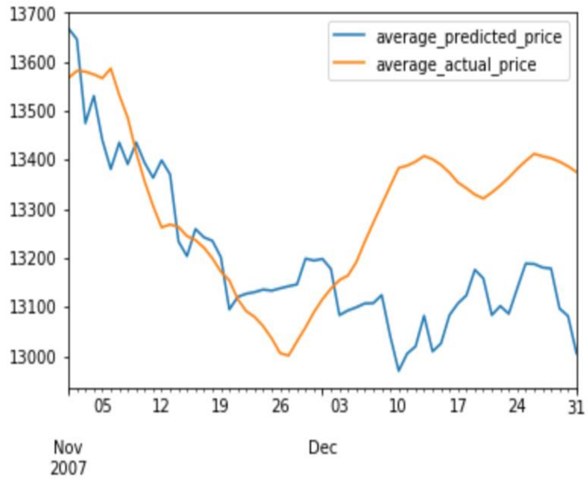


Fig.1. MLP Architecture

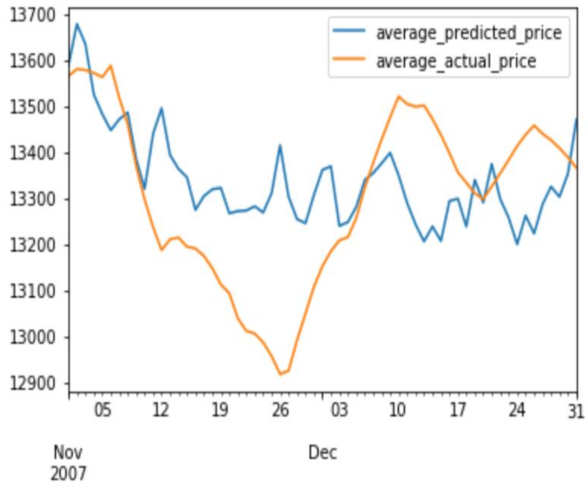
1. Multilayer Perceptron-

MLP is a feedforward neural network with one or more layers between input and output layer. Feedforward means that data flows in one direction from input to output layer. MLP has three layers; an input layer, one more hidden layers and output layer. The input data are fed to the neurons in the input layer and after processing within the individual neurons of the input layer the output values are forwarded to neurons in the hidden layer and finally to the neurons in the output layer. MLPs are widely used for pattern classification, recognition, prediction and approximation. Connections among the neurons are associated by weights and changing the weights in a specific manner results to learning of the associated network. The procedure by which the weight changes take place in the network is called learning or training algorithm. The backpropagation algorithm is the most commonly used learning technique. The technique consist of a forward pass and a backward pass. In the forward pass, an input vector is applied to the nodes of the network and result of which becomes a set of outputs for the network at the output layer. During this phase the weights are all fixed. In the backward pass, the error term is calculated by finding the difference between actual response of the network and desired response specified to the network and is propagated backward through the network. Below figure is our prediction result using MLP.



2. Random Forest Regressor-

In a classification problem, which the target variable is categorical; all variables in a dataset are assigned to the root node. The data is then divided into two child nodes, based on a splitting criterion that splits data characterized by a question. A splitting criterion at each node depends on the single variable value selected from the dataset. Depending on the answer to the question, whether yes or no, data is split into left or right nodes. The splitting of parent nodes continues until the resulting child nodes are pure or until the numbers of cases inside the node reach a predefined number. Thus the tree is constructed by examining all possible splits at each node until maximum depth is reached or no gain in purity is observed with further splitting. Nodes that are pure or homogeneous, which could not be split further, are called terminal or leaf nodes, and they are assigned to a class. In this part, the results of applying technical, fundamental and technical-fundamental indicators to RF are considered. Decision trees can be used for various machine learning applications. But trees that are grown really deep to learn highly irregular patterns tend to overfit the training sets. A slight noise in the data may cause the tree to grow in a completely different manner. This is because of the fact that decision trees have very low bias and high variance. Random Forest overcomes this problem by training multiple decision trees on different subspace of the feature space at the cost of slightly increased bias. This means none of the trees in the forest sees the entire training data. The data is recursively split into partitions. At a particular node, the split is done by asking a question on an attribute. The choice for the splitting criterion is based on some impurity measures such as Shannon Entropy or Gini impurity.



3. Logistic Regression-

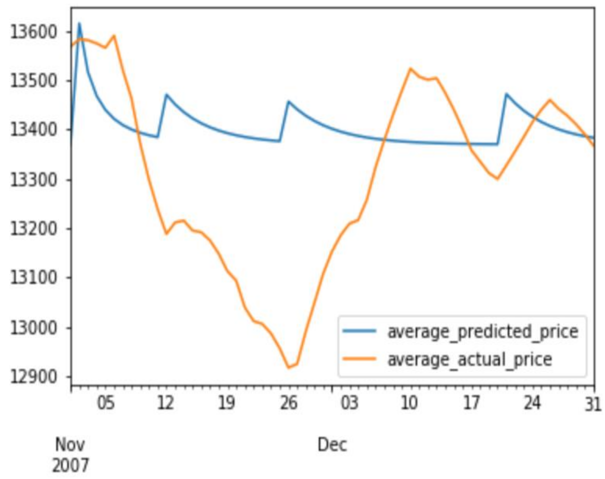
In order to carry out logistic regression analysis, first a method is needed for classifying returns as a "1" or "0" for a given day. In this study we use a method that is simple and objective, if the value of a return in day j is above the return in day j-1, it is noted as a "1"; otherwise, it is classified as a "0". The return was calculated using the following formula:

$$return = \frac{p_j - p_{j-1}}{p_{j-1}} \times 100$$

Where:

pj is the closing price for day j

$pj{-}1$ is the closing price for day j-1



II. METHODS USED

The following methods are used in the project for stock price prediction:

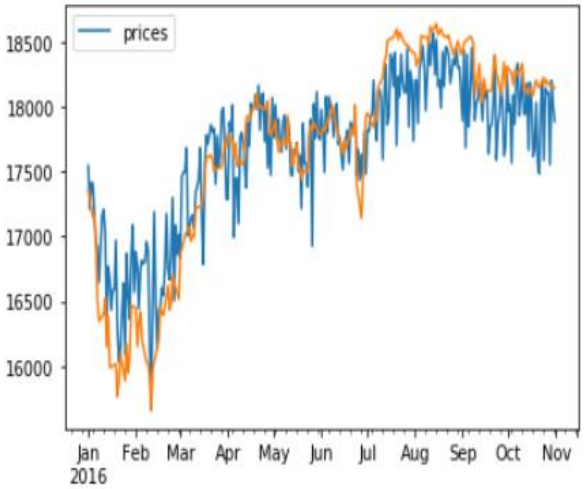
1) NewsAPI – To fetch Data from New York Times
2) Read JSON– To read the JSON format data we got from New York Times API.
3) Interpolate – This method helps in filling the missing values in the data and replace it with the estimated values.
4) EWMA – Exponentially Weighted Moving Average is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time. A moving average takes a noisy time series and replaces each

value with the average value of a neighborhood about the given value. This neighborhood may consist of purely historical data, or it may be centered about the given value.

5) MLP Classifier: This Classifier has been implemented using Keras which runs Tensorflow in the backend. Keras is a deep learning library in python which helps build a CNN for efficient classification of various brand logos. The functions that have been used to accurately detect brand logo images are as follows:
   - Activation function – ReLU, TanH
   - Loss Function – Categorical Cross Entropy
   - Optimization Function – ADAM, LBFG

III. RESULTS

By using the current convolutional model, we have achieved good accuracy. We also tried changing the network architecture and initialization by using different activation and loss functions. The accuracy of the neural network which has “ReLU” as the activation function has an accuracy of 62%. The object localization technique greatly helps us in improving the prediction of the Multilayer Perceptron. it provides the network with only the localized portion where the brand logo is located there by ignoring the other areas insignificant to the network. Our tests determined that using the MLP classifier showed better results than logistic regression and random forest trained models. Although the graphs generated do not shows satisfactory results, further research in the below mentioned areas could lead to better accuracy.



IV. DISCUSSION

We performed different models of machine learning algorithm on our stock price data, predicted and analyzed the data. A different source of stock price data provided variate results which helped in getting precise data. Google Finance API was the initial source to get the data for every company. Since it has been deprecated, we tried other sources like NY times to fetch the data. Our major aim was to work on Kafka and NiFi to introduce the Big Data concept in our project. Even though the Google Finance API is deprecated we can use the HTTP call to Google Finance with stock symbols as query arguments and get the info back. Setting up Kafka cluster on Hortonworks was a challenging task. Different processors helped in retrieving the data from google finance in JSON format. We are still trying to use different models like an ensemble to merge multiple classifiers and create a new model.

V. CONCLUSION AND FUTURE SCOPE

Based on the MLP graphs we could clearly see, MLP gave the highest accuracy score when compared with logistic regression and Random Forest Regressor. The predicted values are close enough to the actual values. We believe that this model would help large companies and individuals to invest in stocks without any hesitation.

REFERENCES

Kafka. (n.d.). Retrieved from

https://community.mapr.com/videos/1198-streaming-stock-market-data-with-apache-spark-and-kafka

multilayer Perception. (n.d.). Retrieved from http://deeplearning.net/tutorial/mlp.html

Nifi. (n.d.). Retrieved from

https://hortonworks.com/tutorial/realtime-event-processing-in-hadoop-with-nifi-kafka-and-storm/

Random Forest Regressor. (n.d.). Retrieved from

http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Sandbox. (n.d.). Retrieved from

https://hortonworks.com/tutorial/getting-started-with-hdf-sandbox/

Dhruv. (n.d.). Dashboard. Retrieved from https://github.com/DhruvKumar/stocks-dashboard-lab

Code Project Articles Retrieved from

https://www.codeproject.com/Articles/1201444/Stock-Predictions-through-News-Sentiment-Analysis