Stock Price Trend Forecasting using Supervised Learning Methods.

Mayank Bhayal Bharat Kumar

***Abstract*— The aim of the project is to examine a number of different forecasting techniques to predict future stock returns based on past returns and numerical news indicators to construct a portfolio of multiple stocks in order to diversify the risk. We do this by applying supervised learning methods for stock price forecasting by interpreting the seemingly chaotic market data.**

I. INTRODUCTION

The fluctuation of stock market is violent and there are many complicated financial indicators. However, the advancement in technology, provides an opportunity to gain steady fortune from stock market and also can help experts to find out the most informative indicators to make better prediction. The prediction of the market value is of paramount importance to help in maximizing the profit of stock option purchase while keeping the risk low.

The next section of the paper will be methodology where we will explain about each process in detail. After that we will have pictorial representations of the analysis that we have made and we will also reason about the results achieved. Finally, we will define the scope of the project. We will talk about how to extend the paper to achieve more better results.

II. METHODOLOGY

This section will give you the detailed analysis of each process involved in the project. Each sub section is mapped to one of the stages in the project.

A. Data Pre-Processing

The pre-processing stage involves

- **Data discretization**: Part of data reduction but with particular importance, especially for numerical data
- **Data transformation**: Normalization.
- **Data Cleaning**: Fill in missing values.
- **Data Integration**: Integration of data files.

After the data-set is transformed into clean data-set, the data-set is divided into training and testing sets so as to evaluate. Here, the training values are taken as the more recent values. Testing data is kept as 5-10 percent of the total dataset.

*This work was supported by International Institute of Information Technology

[1]Sharvil Katariya is a student in Computer Science at IIIT Hyderabad, India.

[2]Nikhil Chavanke is a student in Computer Science at IIIT Hyderabad, India.

B. Feature Selection and Feature Generation

We created new features from the base features which provided better insights of the data like 50 day moving average, previous day difference, etc.

To prune out less useful features, in Feature Selection, we select features according to the k highest scores, with the help of an linear model for testing the effect of a single regressor, sequentially for many regressors. We used the **SelectKBest** Algorithm, with f_regression as the scorer for evaluation.

Furthermore, we added **Twitters Daily Sentiment Score**, as an feature for each company based upon the users tweets about that particular company and also the tweets on that companys page.

III. ANALYSIS

For analyzing the efficiency of the system we are used the Root Mean Square Error(RMSE) and r$\hat{2}$ score value.

A. Root Mean Squared Error (RMSE)

The square root of the mean/average of the square of all of the error.

The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions.

Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Fig. 1. RMSE Value calculation

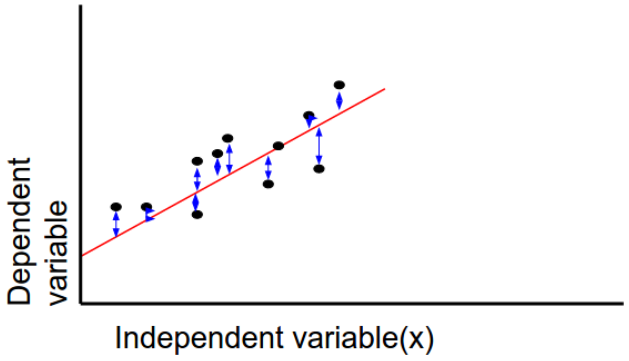

Fig. 2. RMSE Value calculation

B. R-Squared Value($r\hat{2}$ value)

The value of R2 can range between 0 and 1, and the higher its value the more accurate the regression model is as the more variability is explained by the linear regression model.

R2 value indicates the proportionate amount of variation in the response variable explained by the independent variables.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

TABLE I
CLASSIFIER EVALUATION

Algorithm	RMSE Value	R-squared Value
Random Regressor	1.4325434e-07	0.956669
Bagging Regressor	1.329966e-07	0.959771
Adaboost Regressor	2.9882972e-07	0.909611
KNeighbours Regressor	0.00039015	-117.01176
Gradient Boosting Regressor	1.274547e-07	0.961448

IV. GRAPHS

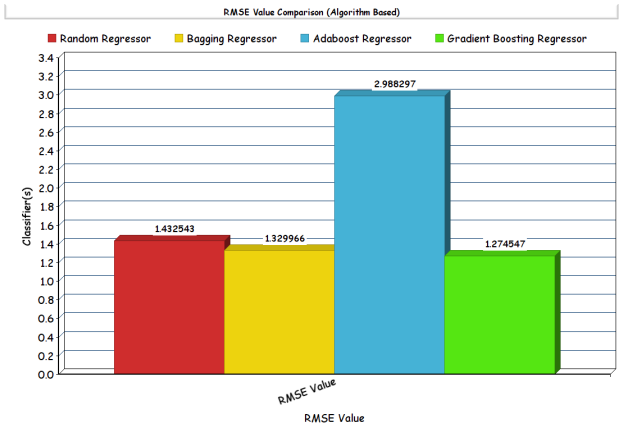

Fig. 3. Comparison Graphs RMSE Value - Different Models

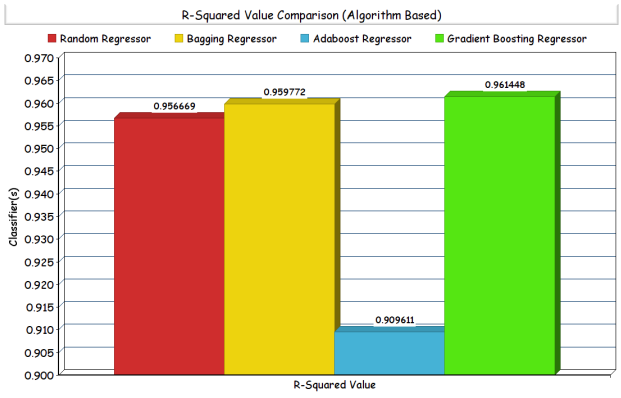

Fig. 4. Comparison Graphs R-squared Value - Different Models

V. RESULTS

Based on the results obtained, it is found that Gradient Boosting Regressor consistently performs the best. This is followed by Bagging Regressor, Random Forest Regressor, Adaboost Regressor and by K Neighbour Regressor.

Bagging Regressor is found to perform good as Bagging (Bootstrap sampling) relies on the fact that combination of many independent base learners will significantly decrease the error. Therefore we want to produce as many independent base learners as possible. Each base learner is generated by sampling the original data set with replacement. From the results, it is safe to say that additional hidden layer(s) improve upon the score of the models.

Random Forest is an extension of bagging where the major difference is the incorporation of randomized feature selection.

ACKNOWLEDGMENT

We would like thank Soham Saha for mentoring our project and introducing us to the new state-of-art technologies and helping us at every stage of this project. We would also like to thank Dr. Bapi Raju, our course instructor for Statistical Methods in AI, and clearing basic concepts required as part of the Project.

REFERENCES

[1] https://en.wikipedia.org/wiki/F-test
[2] http://goo.gl/4OI84b
[3] http://scikit-learn.org/stable/
[4] http://deeplearning.net/software/theano/
[5] http://colah.github.io/posts/2015-08-Understanding-LSTMs/
[6] http://people.duke.edu/ rnau/411arim.htm - []() - []() - []()