Stock Price Trend Forecasting using Supervised Learning Methods

KUNAL RAJ (22BCS13039)

Problem Statement

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

Motivation

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 Solution

1. Preprocessing and Cleaning
2. Feature Extraction
3. Twitter Sentiment Analysis and Score
4. Data Normalization
5. Analysis of various supervised learning methods
6. Conclusions

Structure of Data

We have taken these Historical Daily Stock Market Data for 500 companies of S&P 500 Market since the company came into existence.

The base features in this data that we collected were :

Open, High, Low, Close, Volume, Adjusted Close for each company and S&P index.

Feature Selection

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

This is done in 3 steps:

1. Start with a constant model, M0
2. Try all models M1 consisting of just one feature and pick the best according to the F statistic.
3. Try all models M2 consisting of M1 plus one other feature and pick the best ...

A F-test (Wikipedia) is a way of comparing the significance of the improvement of a model, w respect to the addition of new variables.

Twitter Sentiment Score as new feature

Social media plays important role in predicting the stock market return values.

So, we then appended our data with one more feature -

Twitter’s Daily Sentiment Score for each company based upon the user’s tweets about that particular company and also the tweets on that company’s page.

Once we were ready with complete set of features, we normalized our data for better results.

Models Used

We have used the following regression models for forecasting and comparison :

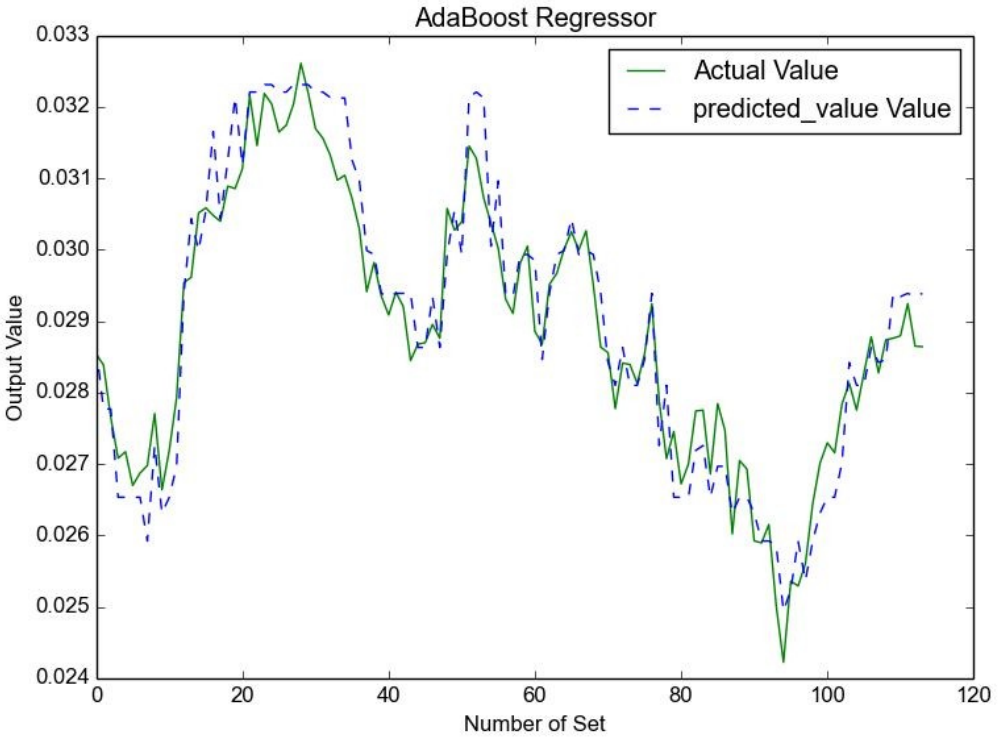
1. Neural Networks
2. Bagging Regressor
3. Random Forest Regressor
4. Gradient Boosting Regressor
5. Adaboost Regressor
6. K Neighbour Regressor

Results : AdaBoost Regressor

American Airlines Group (AAL)

R-squared : 2.988 e-07

RMSE : 0.909

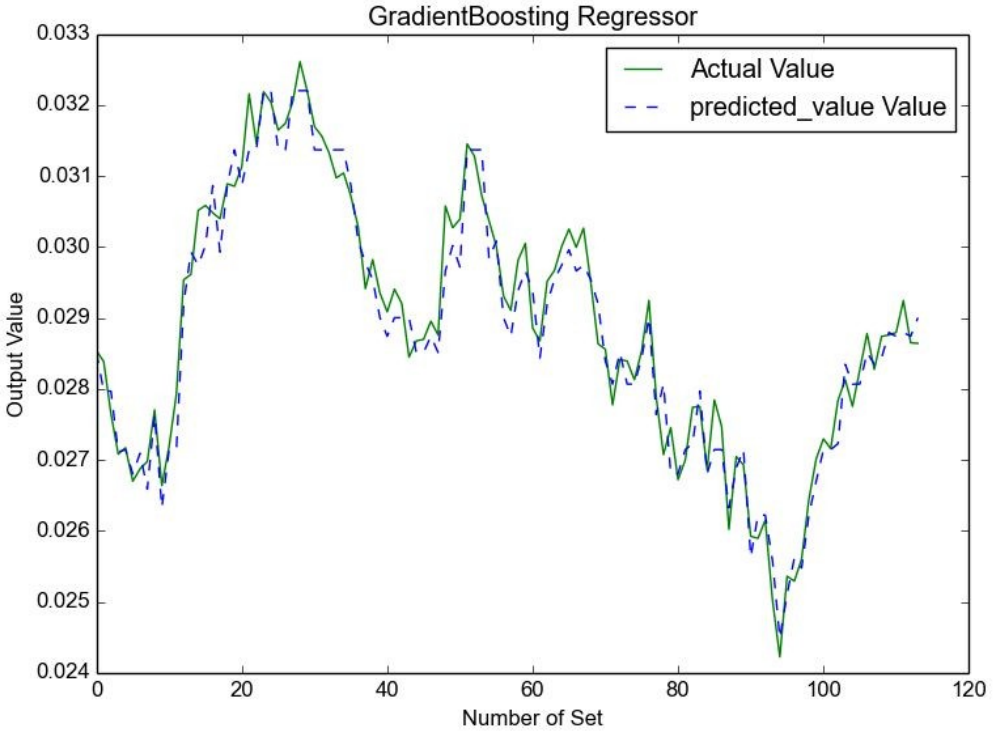


Results : Gradient Boosting Regressor

American Airlines Group (AAL)

R-squared : 1.275 e-07

RMSE : 0.961

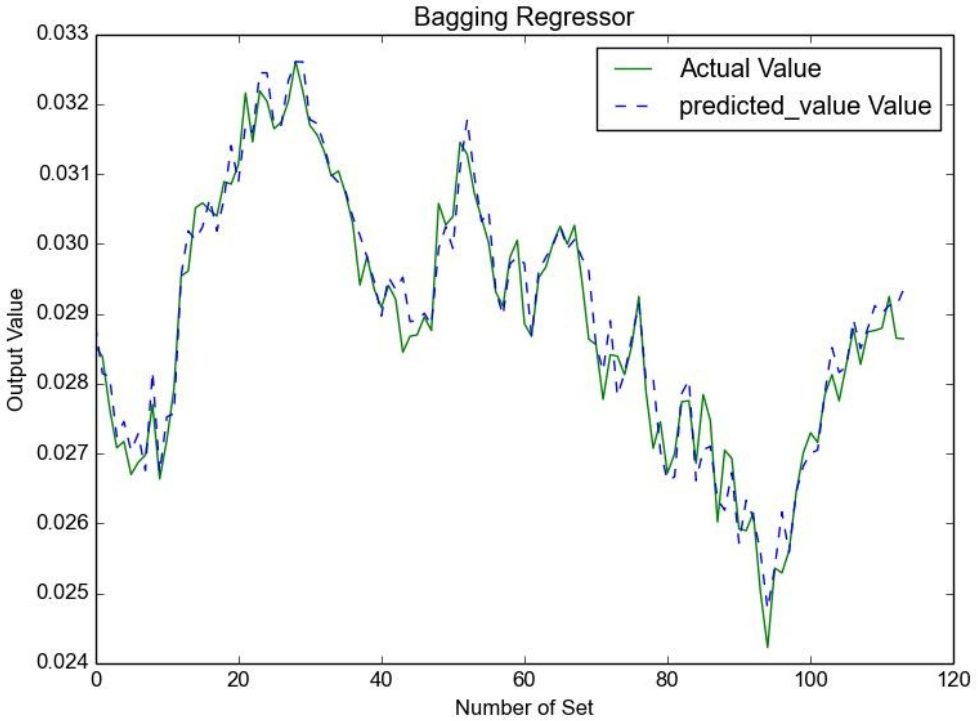


Results : Bagging Regressor

American Airlines Group (AAL)

R-squared : 1.329 e-07

RMSE : 0.959

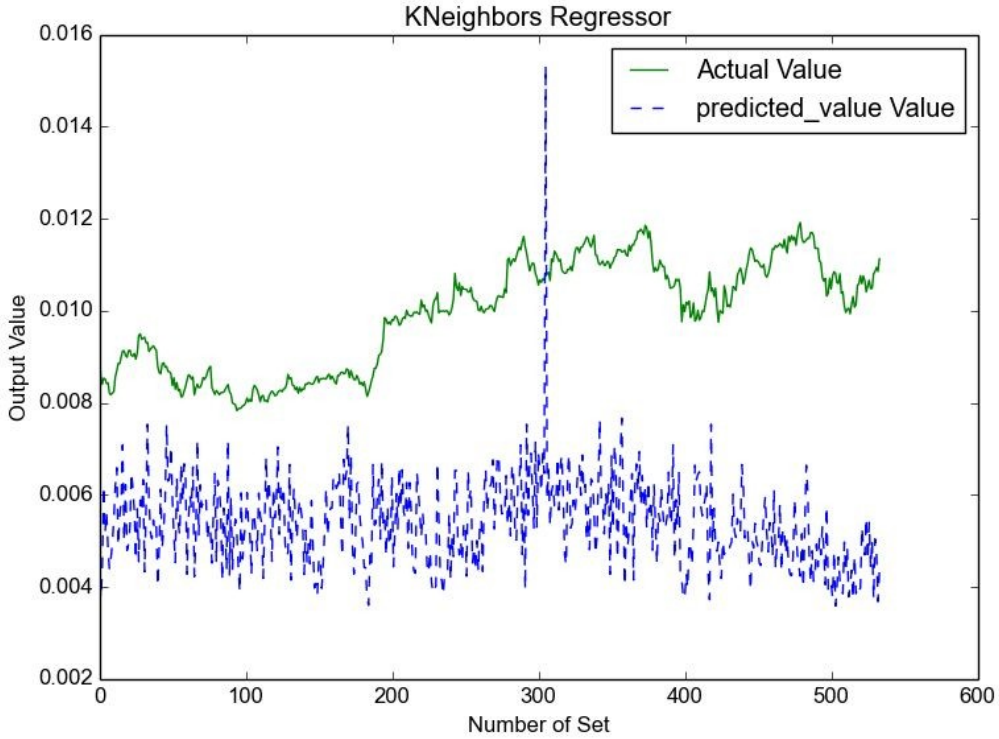


Results : K Neighbours Regressor

American Airlines Group (AAL)

R-squared : 0.00039

RMSE : Too High

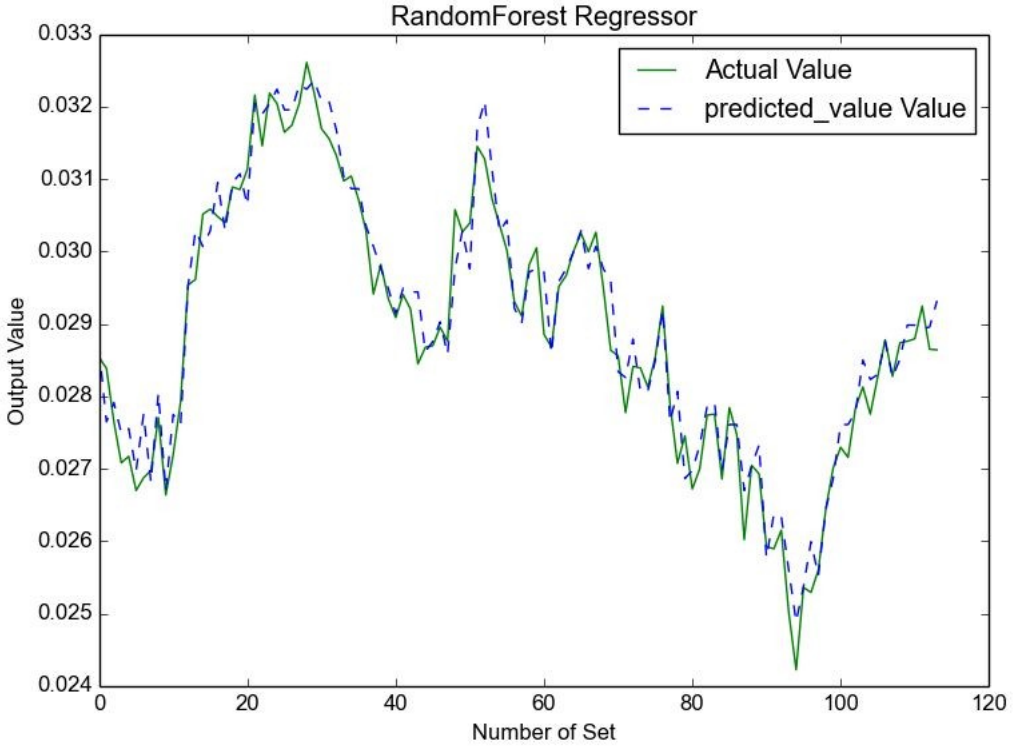


Results : Random Forest Regressor

American Airlines Group (AAL)

R-squared : 1.432 e-07

RMSE : 0.956



Results

- Gradient Descent Regressor: Gradient Descent + Boosting
- Bagging Regressor is found to perform well 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.

Bagging vs Boosting

Bagging:

1. **Parallel** ensemble: Each model is built independently
2. Aim to **decrease variance**, not bias
3. Suitable for high variance low bias models (complex models)
4. An example of a tree based method is **random forest**, which develop fully grown trees (note tha
   RF modifies the grown procedure to reduce the correlation between trees)

Boosting:

1. **Sequential** ensemble: try to add new models that do well where previous models lack
2. Aim to **decrease bias**, and not variance
3. Suitable for low variance high bias models
4. An example of a tree based method is **gradient boosting**

Milestones

18 March 2016

Collection and Preprocessing of data, Feature Extraction and Data Normalization

30 March 2016

Include Social Media Sentiment Analysis as a feature to predict stock trend more accurately.

12 April 2016

Implementation of various regression models and classification based upon the result.

15 April 2016

Comparison and Analysis of Different Models


Thank You