Indian Institute of Technology Bombay

Course Code: EE769

Course Name: Introduction to Machine Learning

Stock Market Trading using Machine Learning

Team members:	*Roll Number:*
Harsh Maheshwari	173050015
Kedar Anavardekar	17305R006

May 4, 2018

Contents

1 Introduction

A listing announcement is an attempt to determine the future value of a company's stock or other financial instruments that are traded on a stock exchange. A successful forecast of the value of the future campaign can lead to significant gains. An effective market hypothesis shows that stock prices reflect all currently available information and that any price changes that are not based on disclosure information are therefore unpredictable. Trading on stock exchanges has always been attractive to investors, as it speaks of the development and devaluation of the stock price. When analyzing the balance sheet, profit and development, companies want to predict the future price of the shares in the interests of the buyer and investor.
It is very useful to predict the price of stocks because it shows the financial position of the market so they can know when to buy and sell stocks. However, this machine becomes a serious problem for students using machine learning algorithms to predict the future value of inventories, as the prediction of the results of the reserves depends on a large number of factors. They can also include the voice of investors. Therefore, it is very difficult to choose functions.

2 Data Set

The dataset used for the development of our project is taken from **The Winton Stock Market Challenge**. In main aim of this competition in the challenge is to predict the return of a stock, given the history of the past few days. As input we have data from past 5 days, **D-2, D-1, D, D+1, and D+2** that is provided. The returns from days D-2, D-1 and part of the day D are given and the returns of the rest of the day D and days D+1 and D+2 have to be predicted.

During day D, there is intraday return data i.e. we have been provided the returns at different points in the day. On this day 180 minutes of data is provided, from t=1 to t=180. In the training set, full 180 minutes of data is given, in the test set just the first 120 minutes are provided.

For each 5-day window, 25 features, Feature_ 1 to Feature_ 25 are provided.

2.1 Datafields

- **Feature_ 1 to Feature_ 25**: different features relevant to prediction
- **Ret_ MinusTwo**: this is the return from the close of trading on day D-2 to the close of trading on day D-1 (i.e. 1 day)
- **Ret_ MinusOne**: this is the return from the close of trading on day D-1 to the point at which the intraday returns start on day D (approximately 1/2 day)

- **Ret_ 2 to Ret_ 120**: these are returns over approximately one minute on day D. Ret_ 2 is the return between t=1 and t=2.
- **Ret_ 121 to Ret_ 180**: intraday returns over approximately one minute on day D. These are the target variables you need to predict as id_ 1-60.
- **Ret_ PlusOne**: this is the return from the time Ret_ 180 is measured on day D to the close of trading on day D+1. (approximately 1 day). This is a target variable you need to predict as id_ 61.
- **Ret_ PlusTwo**: this is the return from the close of trading on day D+1 to the close of trading on day D+2 (i.e. 1 day) This is a target variable you need to predict as id_ 62.

3 Algorithms implemented

3.1 Artificial Neural Networks (ANN)

It is a mathematical model established by W.S.Mcculoch and W. Pitts, and it was name as MP model. It was made by simulating biological nervous systems like the brain. It has following functions:

I. Receive inputs

II. Weight assignment to inputs

III. Calculate weighted sum of inputs

IV. Comparing result with threshold

V. Determine output

Many studies and materials prove that the ANN stock market is more favorable according to estimates As another MLT, because in KNN there is the ability to find non-linear associations in academic input For this reason, ANN is ideal for creating non-linear systems like stock market.
In the ANN, we do not accept the working forms of relationships, they have the ability to find relationships. KNN is known as the universal size through data. Using adequate information for modeling using ANN, each association can be done with some specific accuracy. There is also sound tolerance and incomplete data. On the other hand, KNN does not show the value of each attribute and how it weighs independent features The picture shows a very simple way. The author used a very simple artificial nervous system architecture The network made the primary process of data; They used the method of analyzing the relevance properties to remove unwanted properties and then applied normalization x-max. This reduces the error (2).

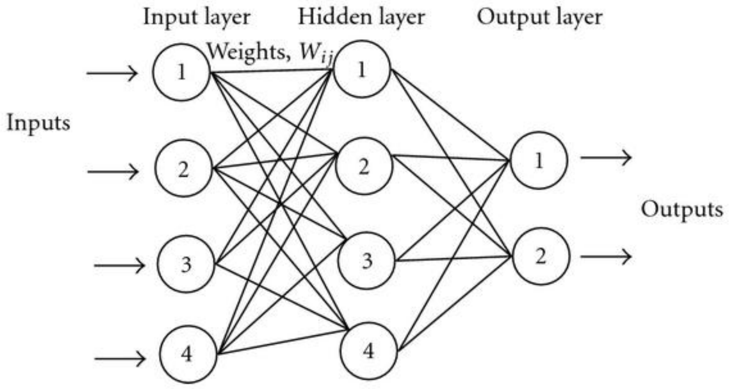

Figure 1: Artificial neural network

3.2 AdaBoost

AdaBoost, is an acronym for "Adaptive Boosting", is a machine learning meta-algorithm developed by Yoav Freund and Robert Schapire who won the Gödel Prize in 2003 for their work. It can be used in conjunction with many other types of learning algorithms to improve the performance of overall procedure. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

Every learning algorithm will tend to suit some problem types better than others, and will typically have many different parameters and configurations to be adjusted before achieving optimal performance on a dataset. AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier. When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder-to-classify examples.(3)

3.3 Future scope:LSTM

We have implemented MLPRegressor and Adaboost. And in future we will implement LSTM.

Long short-term memory (LSTM) block or network is a simple recurrent neural network which can be used as a building component or block (of hidden layers) for an eventually bigger recurrent neural network. The LSTM block is itself a recurrent network because it contains recurrent connections similar to connections in a conventional recurrent neural network.

An LSTM block is composed of four main components: a cell, an input gate, an output gate and a forget gate. The cell is responsible for ”remembering” values over arbitrary time intervals; hence the word ”memory” in LSTM. Each of the three gates can be thought as a ”conventional” artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as regulators of the flow of values that goes through the connections of the LSTM; hence the denotation ”gate”. There are connections between these gates and the cell. Some of the connections are recurrent, some of them are not.

The expression long short-term refers to the fact that LSTM is a model for the short-term memory which can last for a long period of time. There are different types of LSTMs, which differ among them in the components or connections that they have.

An LSTM is well-suited to classify, process and predict time series given time lags of unknown size and duration between important events.

LSTMs were developed to deal with the exploding and vanishing gradient problem when training traditional RNNs. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs, hidden Markov models and other sequence learning methods in numerous applications

4 Results

4.1 Error

The Weighted Mean Absolute Error according to kaggle were.

Adaboost with 50 iterations	Adaboost with 350 iterations	MLPRegression
4828.94362	2044.88105	456780.57527

4.2 Graphs

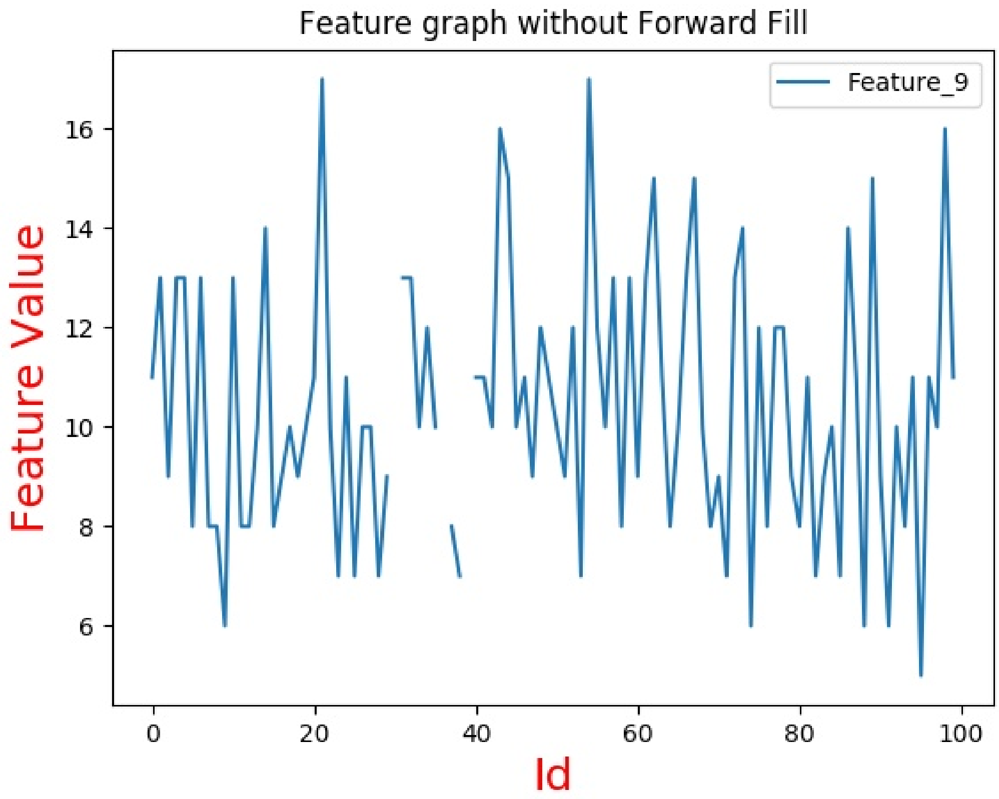

Figure 2: Feature graph without forward fill

This figure 2 represents the feature_9 which has uncleaned data and missing values.

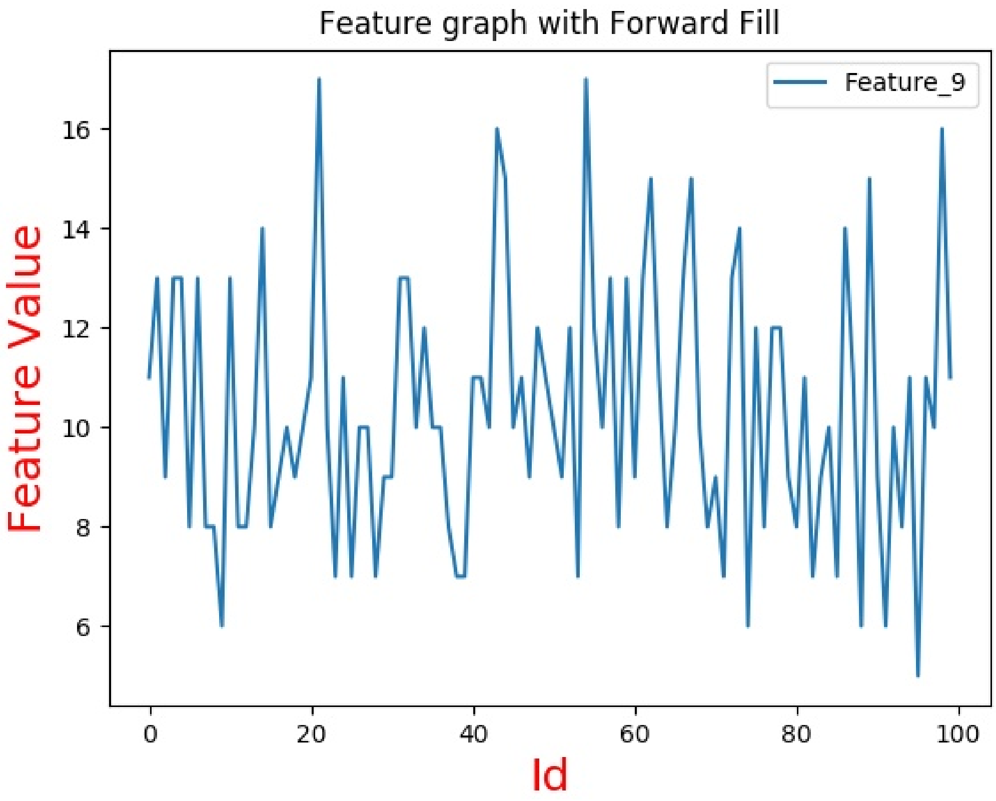

Figure 3: Feature graph with forward fill

This figure 9 represents the feature_9 now with data filled using forward fill technique. In this technique, the missing values are filled by their previous filled values.

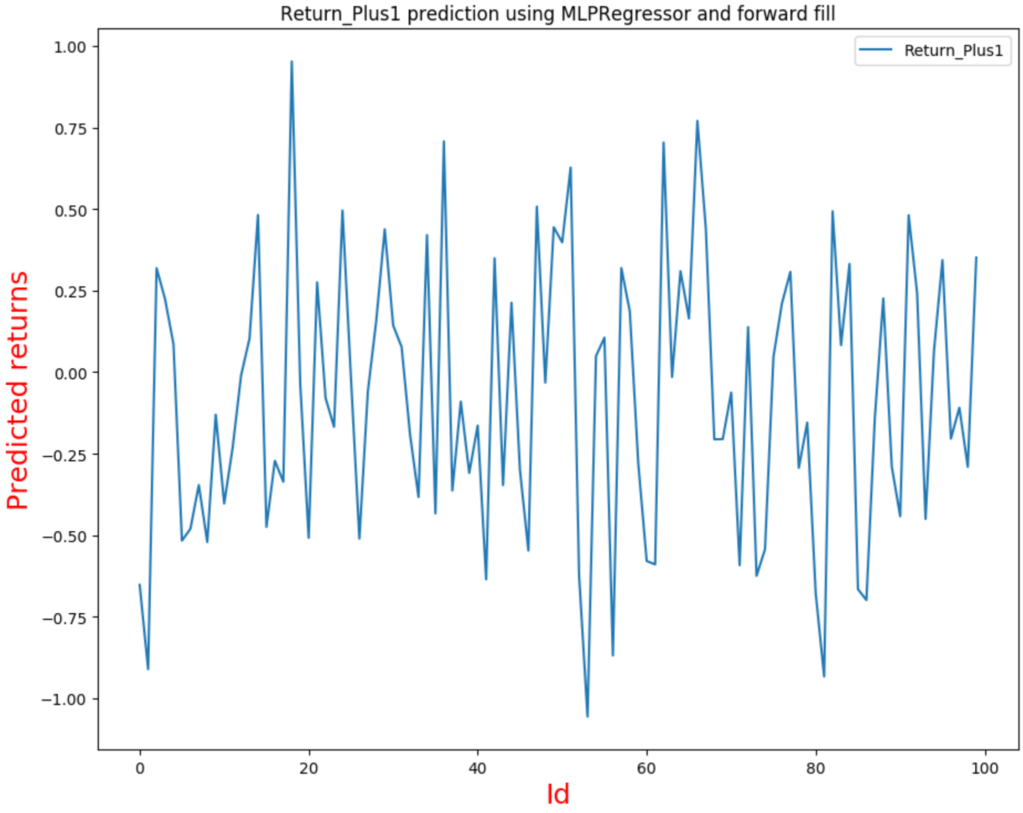

Figure 4: Return_plus1 prediction using MLPRegressor

This figure 4 represents the prediction of the returns calculated by the Multi-layer Perceptron regressor algorithm.

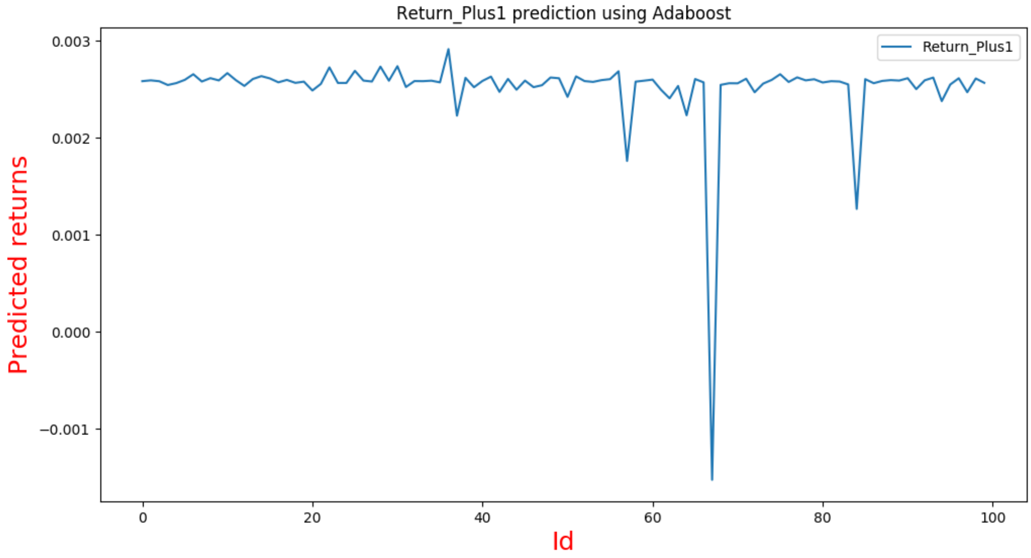

Figure 5: Return_plus1 prediciton using 50 rounds of iterations of Adaboost

This figure 5 represents the prediction of the returns calculated by the Adaboost algorithm with 50 iterations.

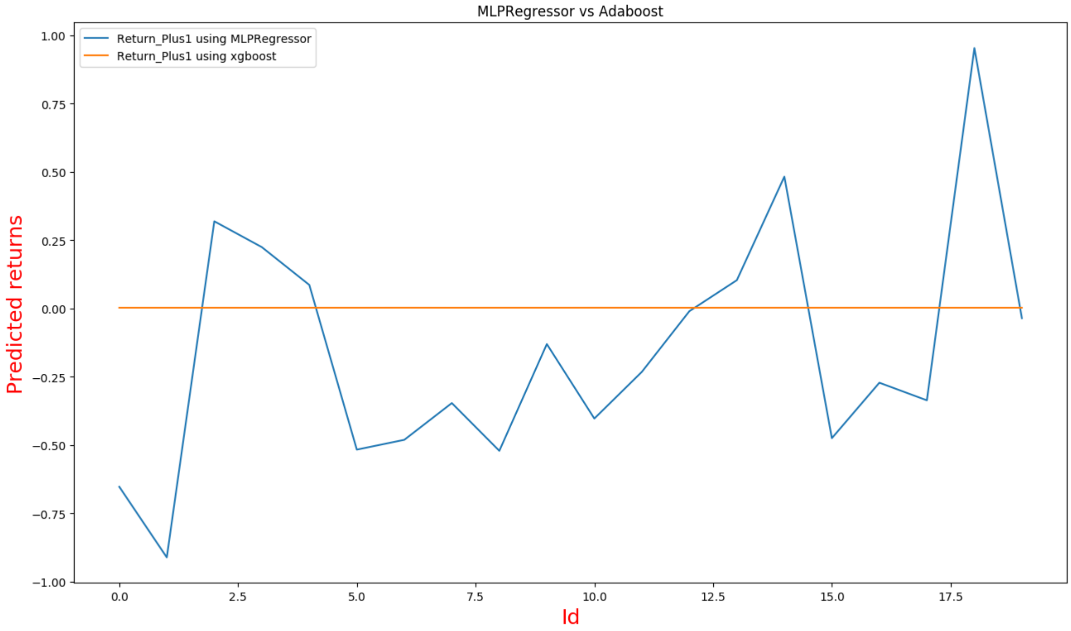

Figure 6: Comparision between Adaboost and MLPRegressor

This figure 6 represents the comparisions between prediction of the returns calculated by the Multi-layer Perceptron regressor algorithm and the Adaboost algorithm.

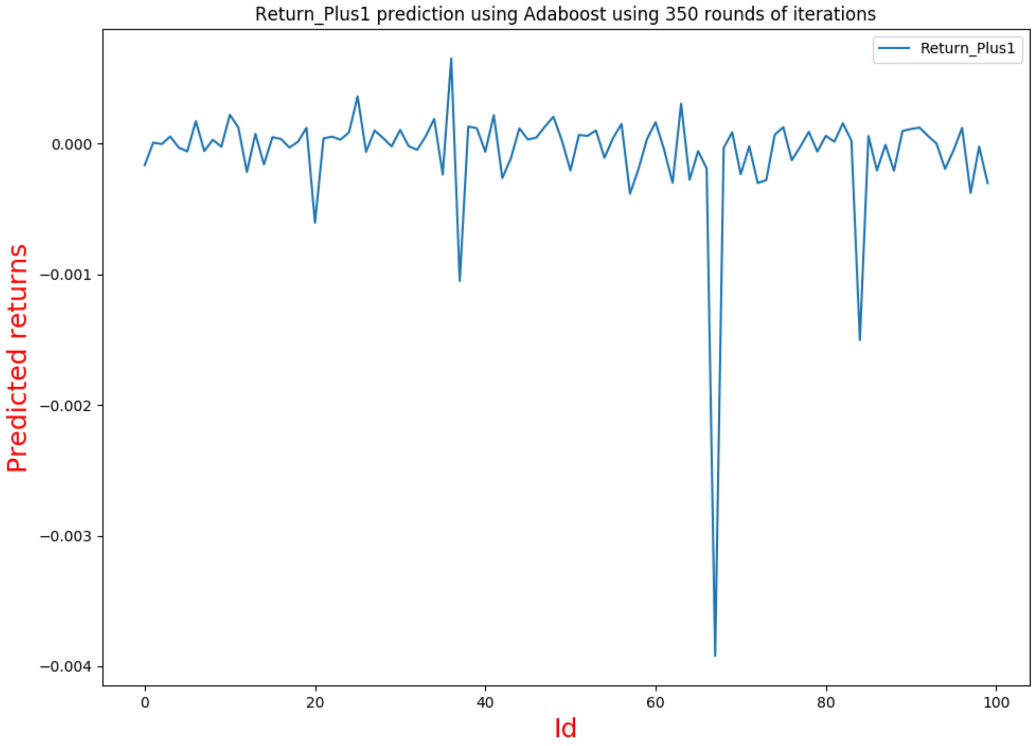

Figure 7: Return_plus1 prediciton using 350 rounds of iterations of Adaboost

This figure 7 represents the prediction of the returns calculated by the Adaboost algorithm with 350 iterations.

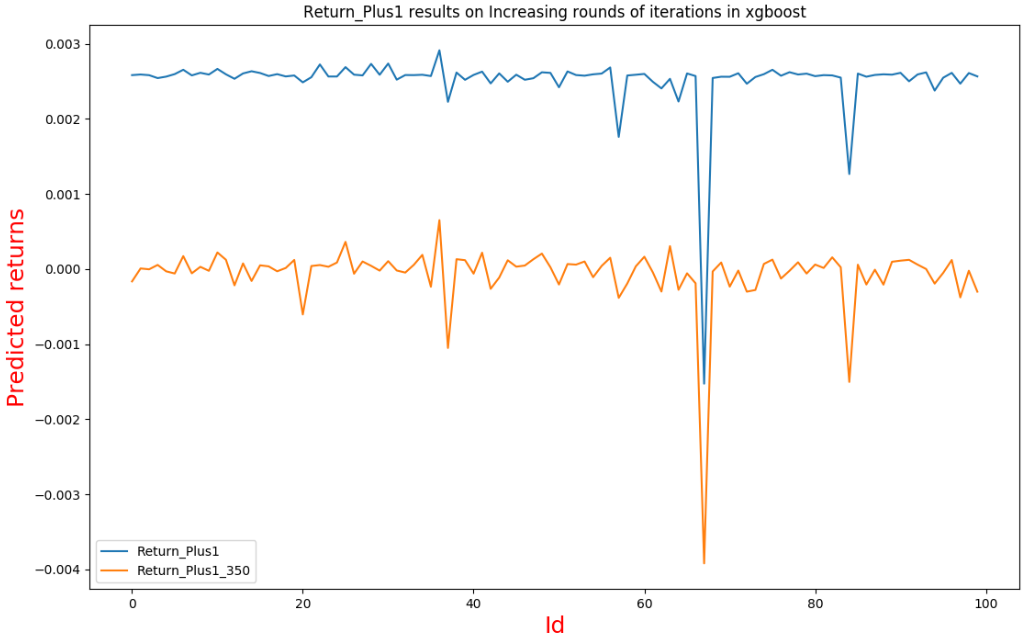

Figure 8: Comparing results of Return_plus1 using Adaboost with 50 iterations and 350 iterations

This figure 8 represents the comparisions between prediction of the returns calculated by the Multi-layer Perceptron regressor algorithm and the Adaboost algorithm.

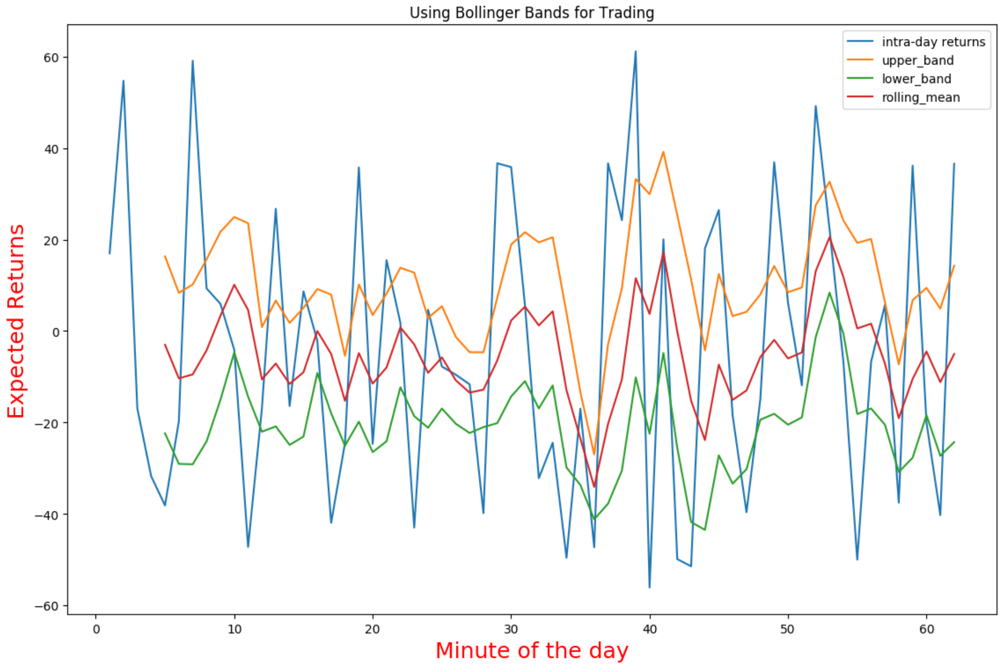

Figure 9: Bollinger band for trading

This figure 9 represents the prediction of the returns calculated by the Bollinger Band algorithm. Bollinger bands are used to calculate the fluctuations in the market. It is also used to predict the points in time when one should buy or sell a stock. It used the concept of rolling mean and standard deviation. Rolling mean is the mean calculated by sliding a window of fixed size over whole data. We have used the window of size 5 minutes. Standard deviation is calculated in the same way.

$$UpperBand = RollingMean + 2 * StandardDeviation$$
$$LowerBand = RollingMean - 2 * StandardDeviation$$

5 Conclusion

In this report, we have used MLPRegressor and Adaboost algorithms to predict the stock returns of the remainder of the D day and the returns at the end of D+1 day and D+2 day. We have used these algorithms because both these algorithms predict the future stock prices. In our project we found Adaboost result more accurate than produced by MLPRegressor. We have also used Bollinger Band algorithm which gives us a band beyond which company should buy or sell stocks.

References

[1] https://en.wikipedia.org/wiki/Stock_market_prediction

[2] Waqas Ahmad NUST, College of Electrical and Mechanical engineering Rawalpindi Pakistan, “Analyzing Different Machine Learning Techniques for Stock Market Prediction”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 12, No. 12, December 2014 .

[3] https://en.wikipedia.org/wiki/AdaBoost

[4] Yusuf Perwej , Asif Perwej , “Prediction of the Bombay Stock Exchange (BSE) Market Returns Using Artificial Neural Network and Genetic Algorithm”, Computer Science & Information System, Jazan University, Jazan, Kingdom of Saudi Arabia (KSA); Department of Management, Singhnia University, Rajasthan, India, February 13th, 2012.

[5] Zahid Iqbal, R. Ilyas, W. Shahzad, Z. Mahmood and J. Anjum, Member, IEEE, “Efficient Machine Learning Techniques for Stock Market Prediction”, Zahid Iqbal et al Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.855- 867.

[6] Waqas Ahmad NUST, College of Electrical and Mechanical engineering Rawalpindi Pakistan, “Analyzing Different Machine Learning Techniques for Stock Market Prediction”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 12, No. 12, December 2014.

[7] https://en.wikipedia.org/wiki/Long_short-term_memory