Applying Machine Learning Techniques to Investigate the Behaviour of Stocks

Author : Sudarshan Govindaprasad

Thesis Advisor : Professor Robert J Brunner

Table of Contents

Introduction

Background

There is a moving trend in the finance industry towards automated stock trading. A lot of research is being done today, in both an academic and industry context, with the goal of having a better understanding of the factors that affect the stock market. In order to build a successful stock trading business, one has to be able to build models that understand the current market sentiment and be able to accurately predict the future market sentiment. This starts with understanding factors that affect the market sentiment and performing a deeper analysis on how changes in these factors affect the market.

One of the big challenges present in the analysis of stock data is the fact that the datasets are massive in size. Let us take a look at the New York Stock Exchange (NYSE) as example. NYSE trades stocks for over 3000 companies and about 2 billion stocks are traded on a daily basis. This results in a massive amount of data being collected everyday. In order to build models to predict changes in stock prices, we would generally have to analyze historical market data over a long period of time, making problem of stock prediction a Big Data problem. In order to tackle this challenge, more research has been dedicated on finding ways to apply Machine Learning techniques on these datasets to create more accurate and scalable and robust systems that can perform real time analysis.

Senior Thesis Goals

The aim of this Senior Thesis is to utilize Applied Machine Learning techniques to perform in-depth analysis on various companies' stocks. I am going to be focusing closely on studying the correlation between multiple pairs of stocks and using the insights gained from this study to create models to predict the fluctuation in stock prices. This research paper is divided into 3 different sections. Section 1 analyses the behaviour of a single stock by using methods such as rolling averages. Section 2, an extension of the first section, that looks more closely into the the relationship between two stocks by performing correlation analysis. Section 3 focuses on applying techniques such as Logistic Regression and Neural Networks, in combination with the insights learned from the earlier sections, to build models to understand the fluctuation in stock prices and predict future stock prices.

Dataset Description

We have access to minute wise stock data from **2010-10-01** to **2010-10-29** for 2191 different stocks. For a given day, we have the stock price by the minute from **09:30** to **16:00**. This results in our dataset having 8211 data points for each of the 2191 stocks with a total of 17990301 rows of data.

Section 1 : Analysis on a Single Stock

In this section, we are going to be performing analysis on a single stock of our choosing, BP. British Petroleum (BP), is a British multinational oil and gas company. We are going to analyse how the stock price changes over a month, daily and calculate rolling averages to find at trends.

BP Stock Price over a Month

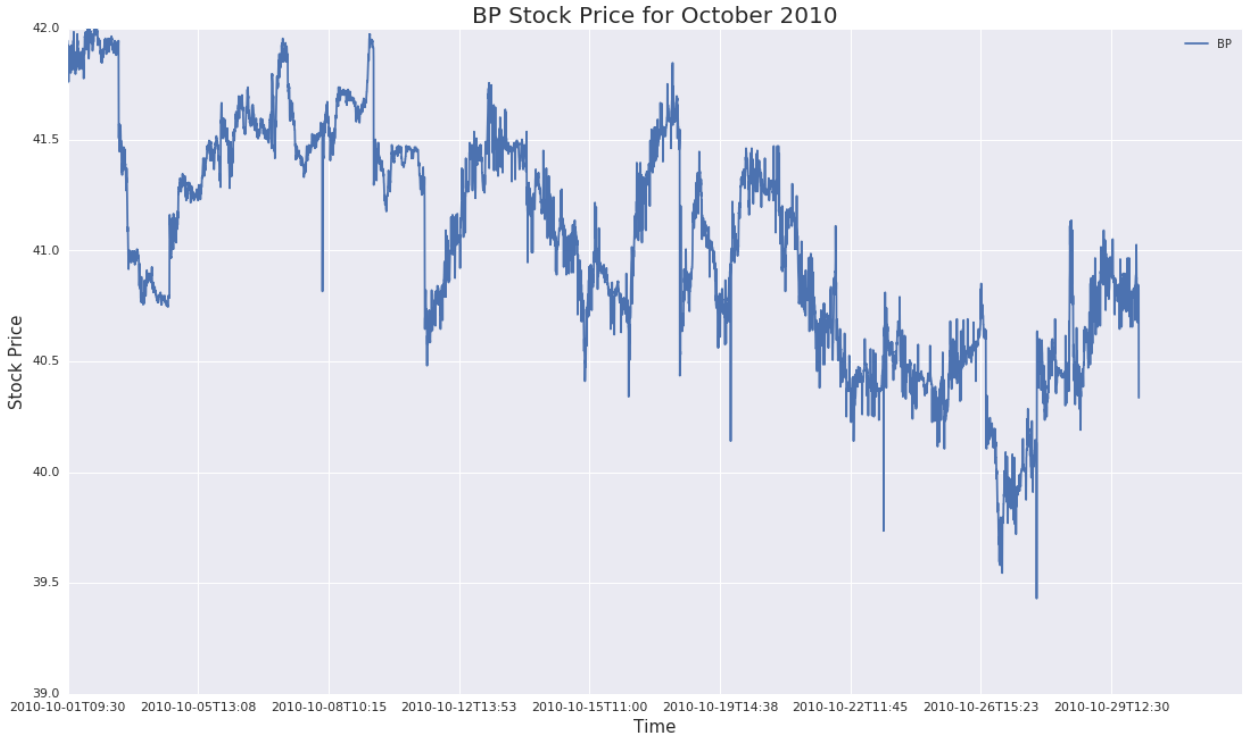

Figure 1 : BP Stock Price for October 2010

From the graph in Figure 1, we can observe that there has been a drop of about $2.00 in the stock price between 1st October 2010 and 29th October 2010. However, it is slightly difficult to take a look at more granular changes in the stock price.

BP Day Wise Stock Price over a Month

This prompts us to dive in deeper and look at a slightly more granular data. Here is a graph that shows the day wise plot of the BP stock price for October 2010.

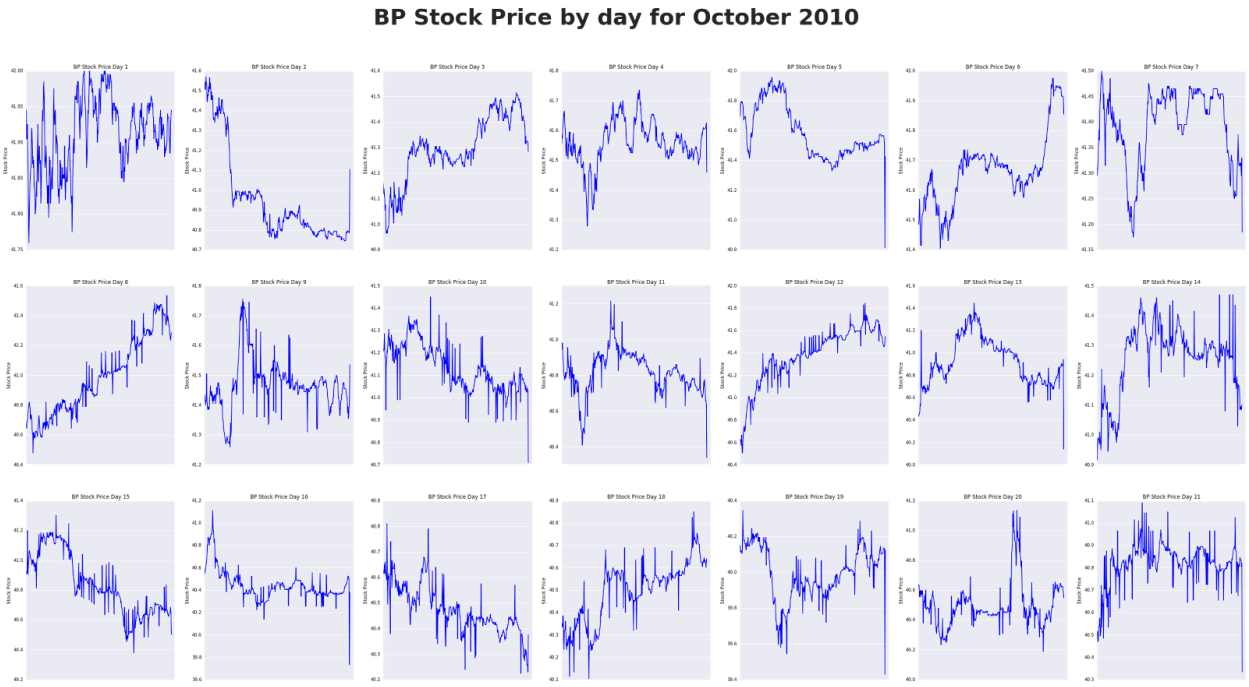

Figure 2 :BP Stock Price by day for October 2010

In the graph plot above, we can see that the stock price fluctuates quite a bit even in one day. The magnitude of change in BP stock price for a given day varies from anywhere between $1.00 and $2.00, which is significant. In the real world, a lot of stock transactions are made within a given day. Most automated stock trading algorithms make trades in the minutes. As such, it is vital to capture these patterns that we see in a given day to create models that are able to learn from these patterns. This would help us build a better model that can more accurately predict an upward or downward movement of the price of a stock.

BP Stock Price with Rolling Averages over a Month

It is often times hard to observe trends in the change in stock prices by just looking at minute wise stock data. In order to better observe trends, we need to performing rolling average for varying time windows. Since we only have a month of stock data, we are going to be calculating the 1 Day Rolling Average, 3 Day Rolling Average and 7 Day Rolling Average.

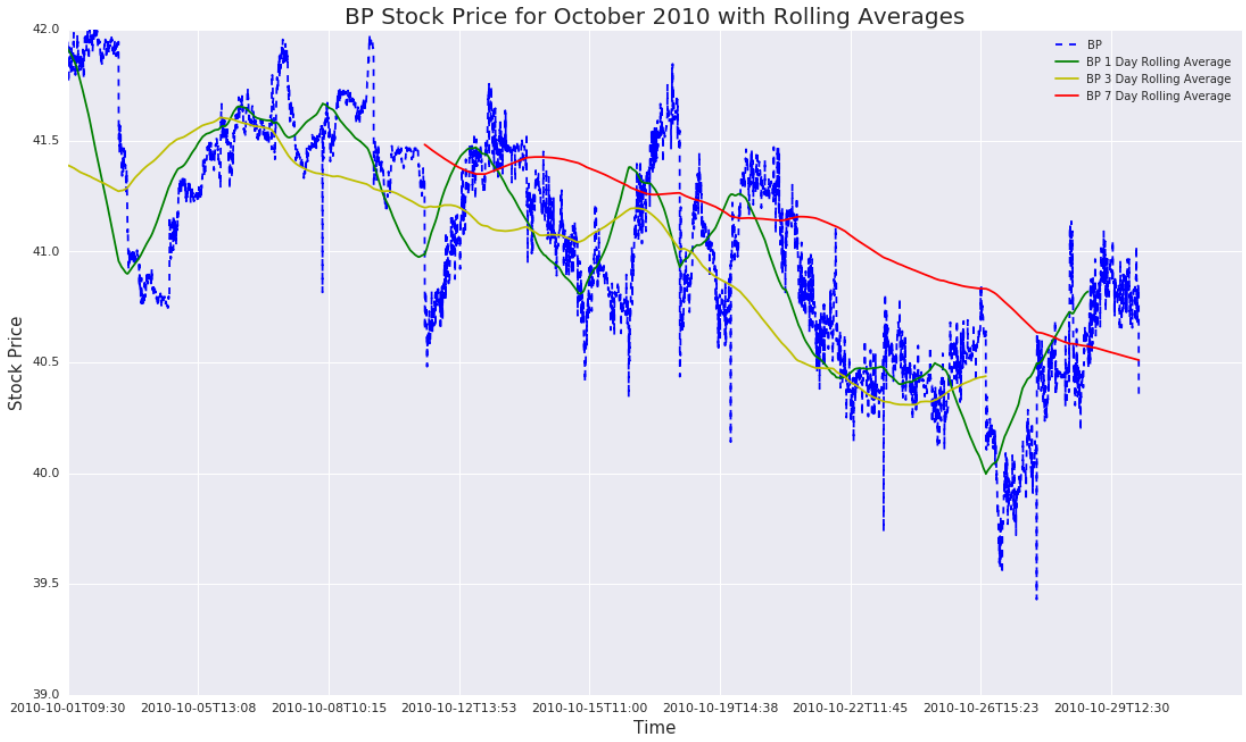

Figure 3 : BP Stock Price for October 2010 with Rolling Averages

From the graph in Figure 3, we can observe that the rolling average graphs provide a smoother graph with discernable trends. For example, the 1 Day Rolling Average allows us to observe how the stock performs over a given day. As we increase the length time window for the rolling average, the graphs become less granular and more smooth. We can observe from the 7 Day Rolling Average that the BP Stock in general has decreased over October 2010. Thus, we learn that it is important to include insights and patterns obtained from varying rolling average in creating our model to predict whether a particular stock's price would go up or down.

Section 2 : Analysis on the Relationship between Two Stocks

In this section, we are going to be performing analysis on a pair of stocks. The aim of this section is to figure out patterns in stock prices that affect correlation between two stocks. The correlation between two stocks is defined to be the measure of the degree to which two stocks price move together. A positive correlation signifies that the a change in direction of the first stock's price is also reflected by the change in the second stock's price. A negative correlation represents the inverse relation. In this section, we measure correlation using Spearman's rank correlation coefficient.

The change in price of a stock does not necessarily result in an immediate change in price of a correlated stock. It might take period of time for the change to propagate. In order to take this under consideration, we are also going to be utilizing the percentage change (pandas.pct_change) in Sections 2 and 3 of this paper. The percentage change function takes a series of data, shifts it by a period and calculates the relative difference between the shifted data and the original data. For example, let us take 60 minutes of stock data for BP. If we were to calculate the percentage change by a period of 5 minutes, we would take the shift forward the original data by 5 minutes and calculate the relative difference between these two series of data.

It is important to understand the pairwise correlation between stocks. In general, in order have a diversified portfolio, stocks for a portfolio are chosen with a metric of minimizing pairwise correlation. However, in our case, using a pair of stocks that have a high correlation would help in predicting the stock price of second stock by using the first stock as a basis to train. We will be exploring more about this in this section.

Comparing PG and XOM Stock Price

Procter & Gamble Company (PG) is one of the world's largest consumer goods companies by market cap. During the 2015 fiscal year, PG has reported a total revenue of PG reported total revenue of **$76.28 billion** and net income of **$7.04 billion**.

Exxon Mobil Corporation (XOM) is the fifth-largest publicly traded company in the world with a market capitalization of **$389.82 billion** as of July 22, 2016. Moreover, Exxon is the world's largest company in the major integrated oil and gas industry. XOM is also in one of the most coveted groups of stocks known as the **S&P 500 dividend aristocrats**.

Although these two stocks are from different sectors, XOM : Basic Materials and PG : Consumer Goods, we performed analysis between PG and XOM to find observable trends.

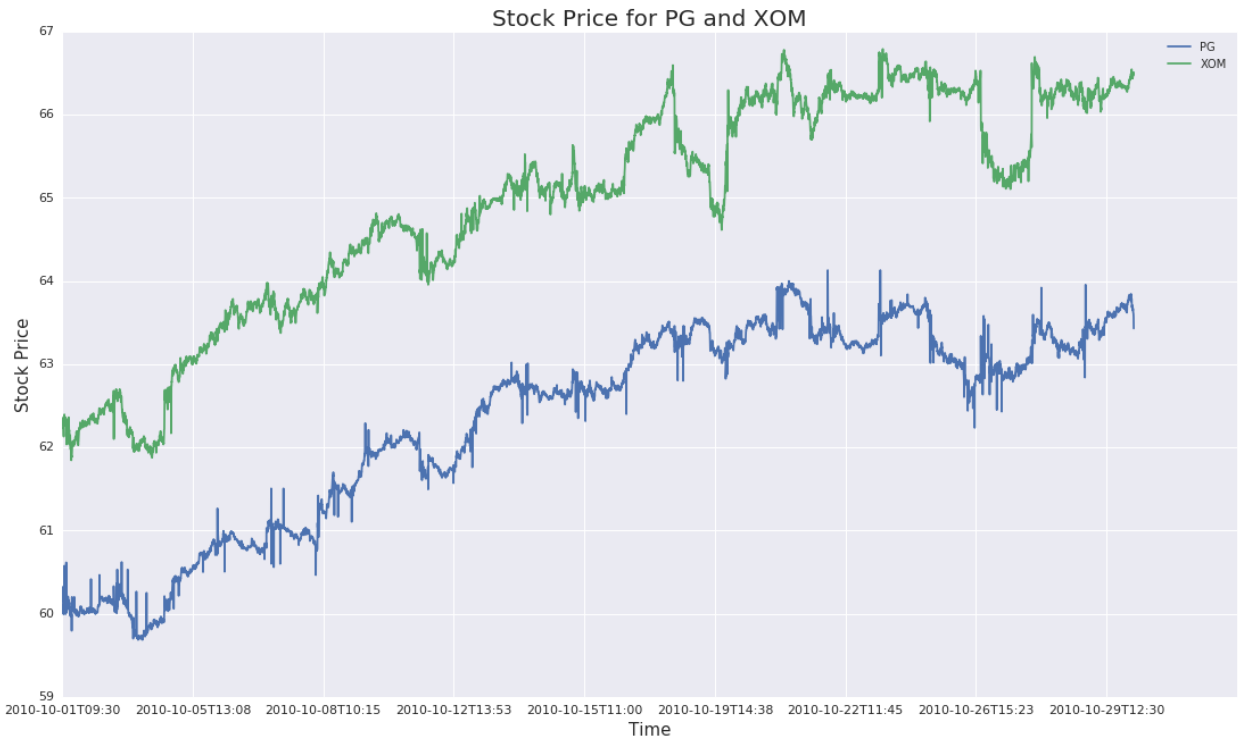

Figure 4 : Stock Price for PG and XOM

From the graph in Figure 4, we can observe a general trend in the change in stock price between PG and XOM. In general, we can see that when PG goes up, XOM also goes up and vice versa. However, we are unable to observe more granular trends since both stocks are of varying magnitude. Instead, we are going the difference in percentage between two stocks to look at how different changes between the two stocks are to see if we can get more discernable trends.

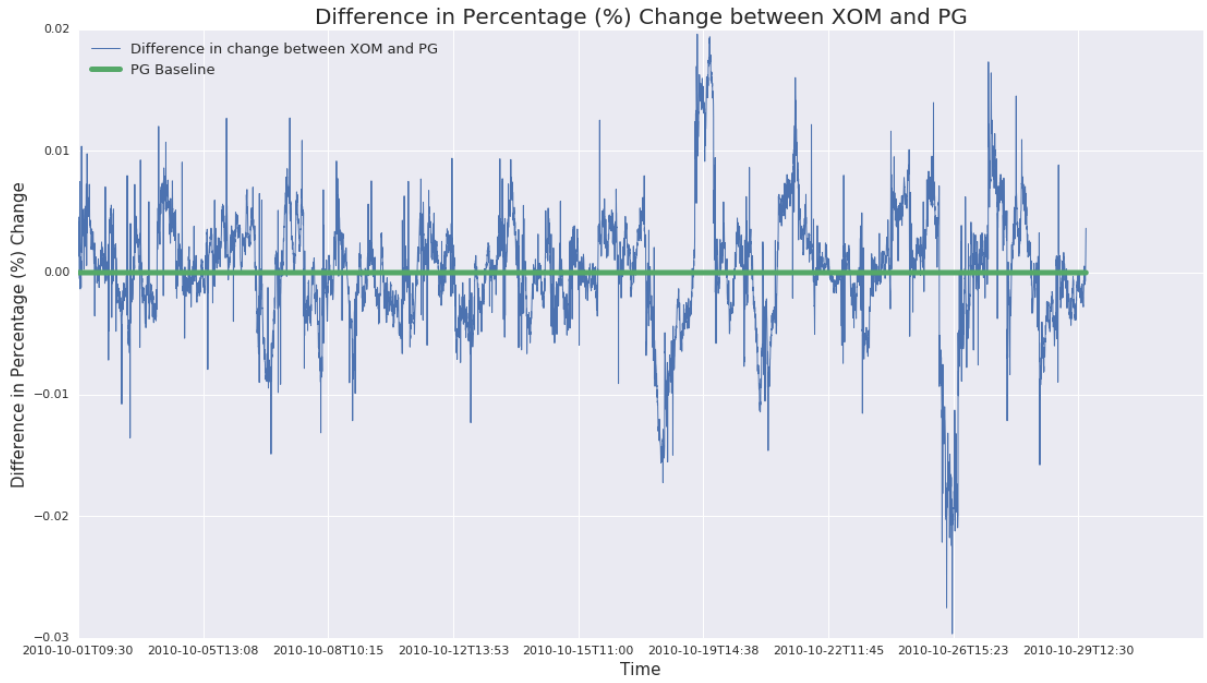

Figure 5 : Difference in Percentage (%) Change between XOM and PG

The graph in Figure 5 shows the Difference in Percentage (%) Change between XOM and PGM. We used the percentage change for PG as a baseline and subtracted that from the percentage change of XOM. From this graph, we can clearly see that the difference in percentage change between PG and XOM is small with a mean percentage change difference of 0.0001383%. The next intuitive thing would be calculate the correlation between the two stocks.

Using the Pearson Coefficient, we calculated the correlation coefficient between PG and XOM stock to be **0.934**. This shows that there is a high correlation between PG and XOM stock price as we observed in the percentage change graph above. This indicates that both PG and XOM are good candidates to use as an example to train our prediction models in future sections.

Correlation between Stocks from Different Sectors

In the above example, we observed a high correlation between the PG and XOM stock although the stocks are from different sectors. It is often the case that fluctuations in a stock's price is the result of changes in stocks both in the same sector as well as in sectors that are complementary to its business. PG is a Consumer Goods stock while XOM is a basic materials stock. It is the case that when the demand for consumer goods goes up, the demand for raw materials involved also goes up and vice versa. This is a good example of how correlation can exist across stocks from different sectors.

In this section, we are going to extend this by performing correlation analysis on a larger set of stocks from different sectors. We have picked 3 top stocks from the following sectors: Financial, Healthcare, Industrial Goods, Basic Materials, Consumer Goods, Services and Technology. We are going to calculate the pairwise correlation using Spearman for every pair in the table below.

Financial

- Bank of American Corporation (BAC)
- JPMorgan Chase & Co (JPM)
- Visa Inc (V)

Health Care

- Johnson & Johnson (JNJ)
- UnitedHealth Group Inc (UNH)
- Novo Nordisk A/S (NVO)

Industrial Goods

- General Electric Co (GE)
- 3M Co (MMM)
- Caterpillar Inc (CAT)

Basic Materials

- Exxon Mobil Corp (XOM)
- Chevron Corp (CVX)
- BP PLC (BP)

Consumer Goods

- Procter & Gamble Co (PG)
- PepsiCo Inc (PEP)
- Philip Morris International Inc (PM)

Services

- Wal-Mart Stores Inc (WMT)
- The Home Depot Inc (HD)
- McDonald's Corp (MCD)

Technology

- AT&T Inc (T)
- Verizon Communications Inc (VZ)
- Taiwan Semiconductor Manufacturing Co Ltd (TSM)

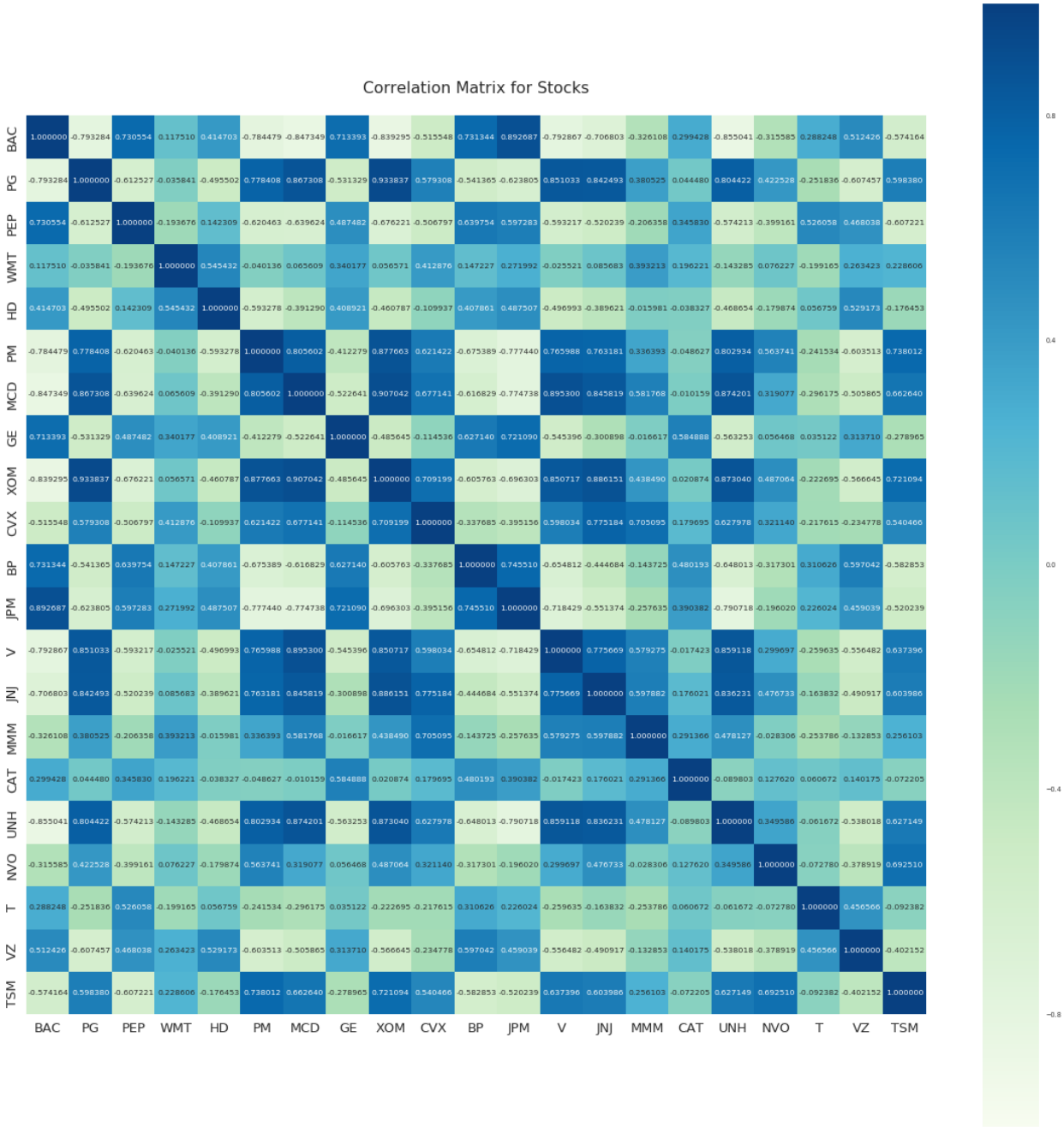

Figure 6 : Correlation Matrix for Stocks from Different Sectors

The correlation heatmap in Figure 6 is a representation of the correlation matrix. The correlation coefficient ranges from -1.0 to 1.0. After ranking the pairs in the correlation matrix, we were to obtain the top 10 pairs of correlated stocks.

Table of Top 10 Pairs of Stocks

Stock 1	Stock 2	Correlation Coefficient
UNH	V	0.859118
MCD	PG	0.867308
UNH	XOM	0.873040
MCD	UNH	0.874201
XOM	PM	0.877663
XOM	JNJ	0.886151
JPM	BAC	0.892687
MCD	V	0.895300
MCD	XOM	0.907042
XOM	PG	0.933837

Table 1 : Table of Top 10 Pairs of Stocks

From Table 1, we can observe that stocks do not necessarily be from the same stock sector to be highly correlated. Often times, it is the case that stocks from different sectors tend to have a higher correlation. In our set of stocks, XOM and PG has the highest correlation coefficient of 0.93. We are going to use this information to build a Logistic Regression model to predict the price of the PG stock.

Section 3 : Approach to Stock Prediction

In this section we will be covering 3 different applied machine learning methods to predict the changes in stock price. First, we will be looking at how we can utilize Logistic Regression to predict a binary outcome of whether the stock price for next time period would go up or down. The next two methods look at how we can apply Neural Networks to predict the actual stock price. The second method looks at how we can utilize Multilayer Perceptron Models to predict stock prices. Thirdly, we are going to look at how we can apply LSTM Recurrent Neural Network on the same problem.

Throughout all three methods, we are going to model the stock prediction model as a recurrence relation. We have the knowledge (data) of how the stock has been performing up to time t. We need to be able to predict what the price of the stock would be at time t+1. Thus, the recurrence relation is $T_n \rightarrow T_{n+1}$.

In order to achieve this, we have created a *create_lookback_dataset* function that takes in a parameter look_back. This function takes in a dataset and appends look_back number of minutes of data into a row. Through this process, you are able to have each row of data being passed into your dataset being a 'mini' timeseries on its own. This helps to build the regression that we are trying to solve since we are using look_back number of minutes of data (T_n) to predict whether the stock price would go up for down for (T_{n+1}).

```
def create_lookback_dataset(dataset, look_back=15):
    dataX = []
    for i in range(len(dataset)):
        if (i + look_back) < (len(dataset)):
            dataX.append(dataset.ix[i:i+look_back].values.flatten())
    return np.array(dataX)
```

Using Logistic Regression to Predict Rise/Fall Stocks

We are going to use the insights we gained from our analysis on stocks to try and predict the increase or decrease of a stock. In our example in Section 2, we found that correlation coefficient between XOM and PG is the highest in our set of stocks at 0.934. We are going to train our Logistic Regression model on XOM stock data. We are then going to use that model to predict the increase or decrease in stock price for PG.

We observed in our analysis between PG and XOM that calculating percentage change helps to provide more granular trends in prices changes. We are going to calculate percentage change in stock price for 1 minute, 15 minutes and 30 minutes. We observed in our day-by-day plot of BP stock that prices fluctuate a lot in a given day. Thus, we decided to pick small time windows for calculating percentage in stock price.

We are also going to utilize the *create_lookback_dataset* defined with a look_back = 15 minutes. We created a ylabel array which tells you whether for a given set of figures, the price of the stock will go up or down the next minute.Once we calculated this data, we trained our Logistic Regression model with XOM and predicted with PG.

We are going to be using sklearn.linear_model's Logistic Regression Model. After training on XOM data, we were able to predict with 73.4% accuracy whether or not the PG stock will go up or down. In order to visualize what this looks like, I have created a line plot below of PG stock price for October 2010.

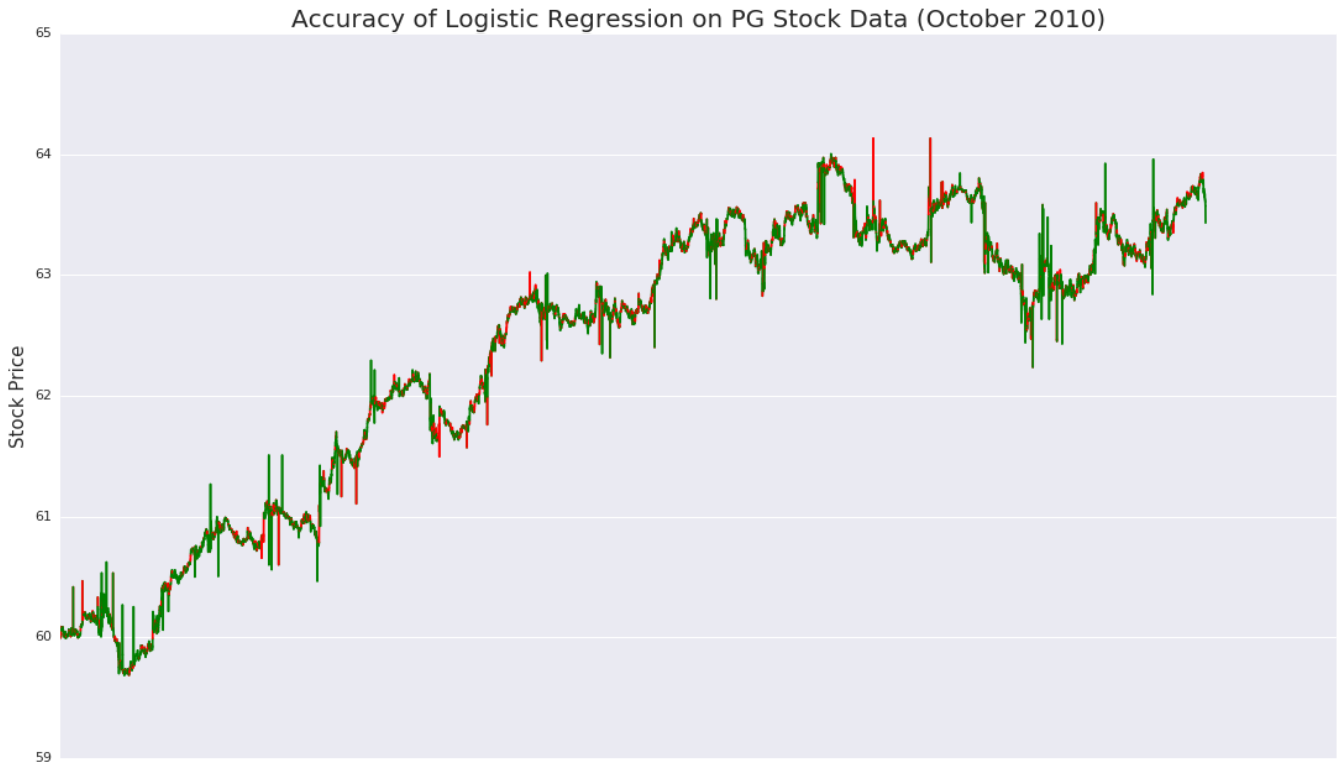

Figure 7 :Accuracy of Logistic Regression on PG Stock Data (October 2010)

This line plot is helpful in visualizing the accuracy of our Logistic Regression as is colors every point that we accurately predicted green and every point we mis-predicted red. We are able to observe that the Logistic Regression model did a good job predicted small changes, most of the largest magnitude changes in stock were not predicted accurately.

Multilayer Perceptron Model to Predict Stock Price

In this section, we are going to use keras, a deep learning library, to perform analysis on stock data. We are building a Simple Multilayer Perceptron Model using keras that predicted the actual price of the stock for the next time segment (i.e. the next minute). We are going to be using British Petroleum (BP) stock to perform this analysis.

The aim of this test is to understand how stock data performs with deep learning techniques in order to gauge the feasibility of applying deep learning techniques on stock data. We are going to split the dataset into to test and train datasets using a 67% test/train split, where we are training on 67% of the data and testing on the rest of the data. We are also creating a time lag by using the *create_lookback_dataset* function which creates a time lag of size look_back (Refer to Section 3 for the *create_lookback_dataset* function) .

Train Prediction

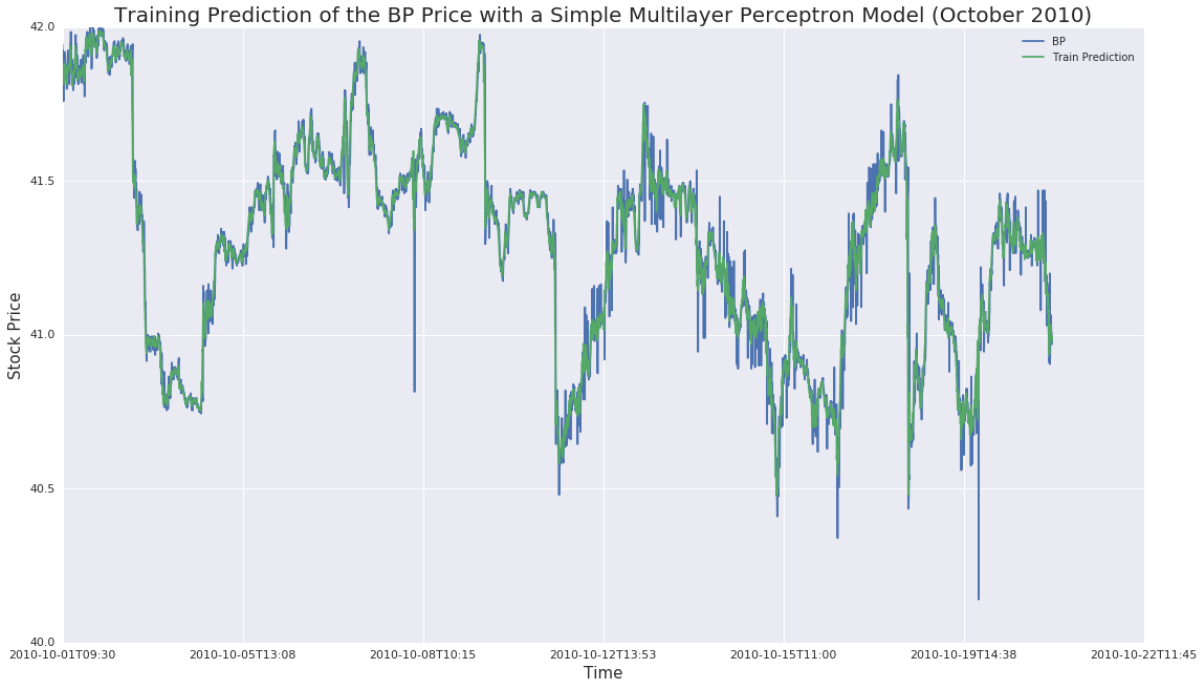

Figure 8 :Training Prediction of the BP Price with a Simple Multilayer Perceptron Model (October 2010)

The graph in Figure 8 shows us the training performance of the model. The green line plot is the plot of the Train Prediction while the blue line plot is the actual price of the BP stock. It might be difficult to decipher how accuracy of the model from the above plot due to the lack of granularity. Instead we will look at the percentage change in price for both the BP stock price (true value) and the prediction in Figure 9.

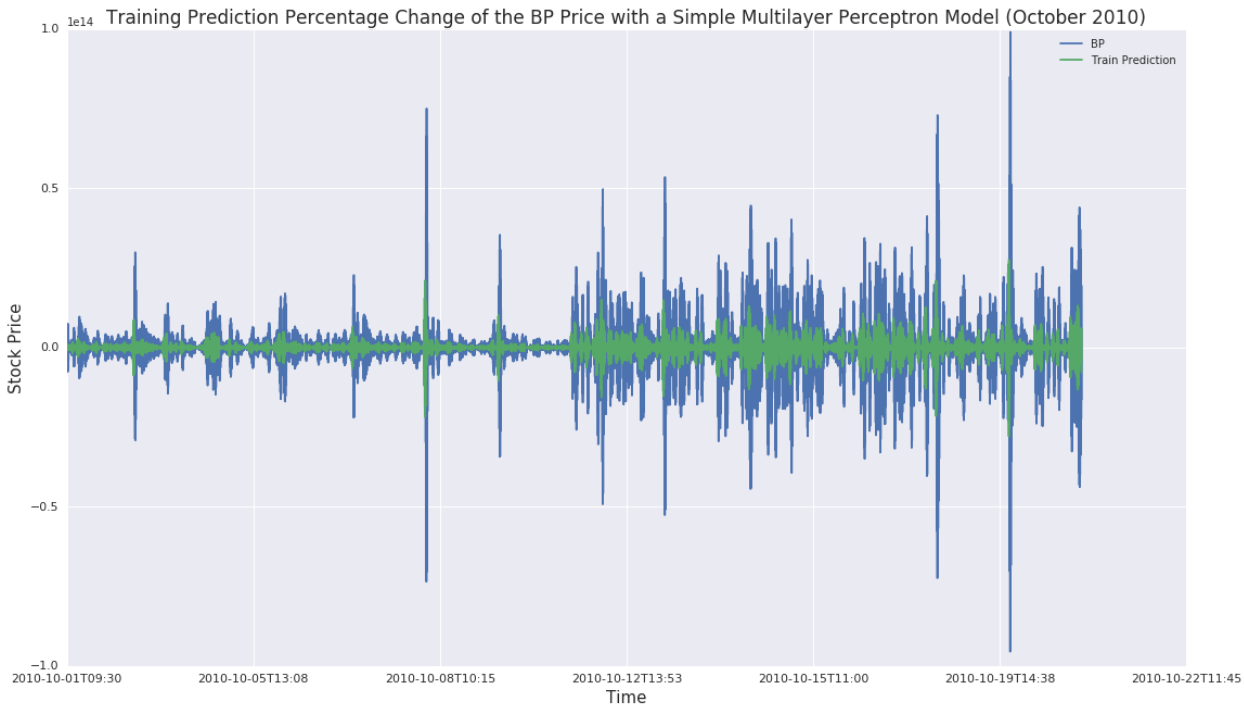

Figure 9 :Training Prediction (Pct. Change) of the BP Price with a Simple Multilayer Perceptron Model (October 2010)

From the graph in Figure 9, we can see that model is able to perform well in predicting the price of the stock. We are also able to observe that the model does fare well in predicting the magnitude of the change due to the fact that the larger amplitude spikes are not accurately predicted. We are going to be using the mean squared error as a metric to measure how well our model has performed. The mean squared error for the train is 0.002159.

.

Testing Prediction

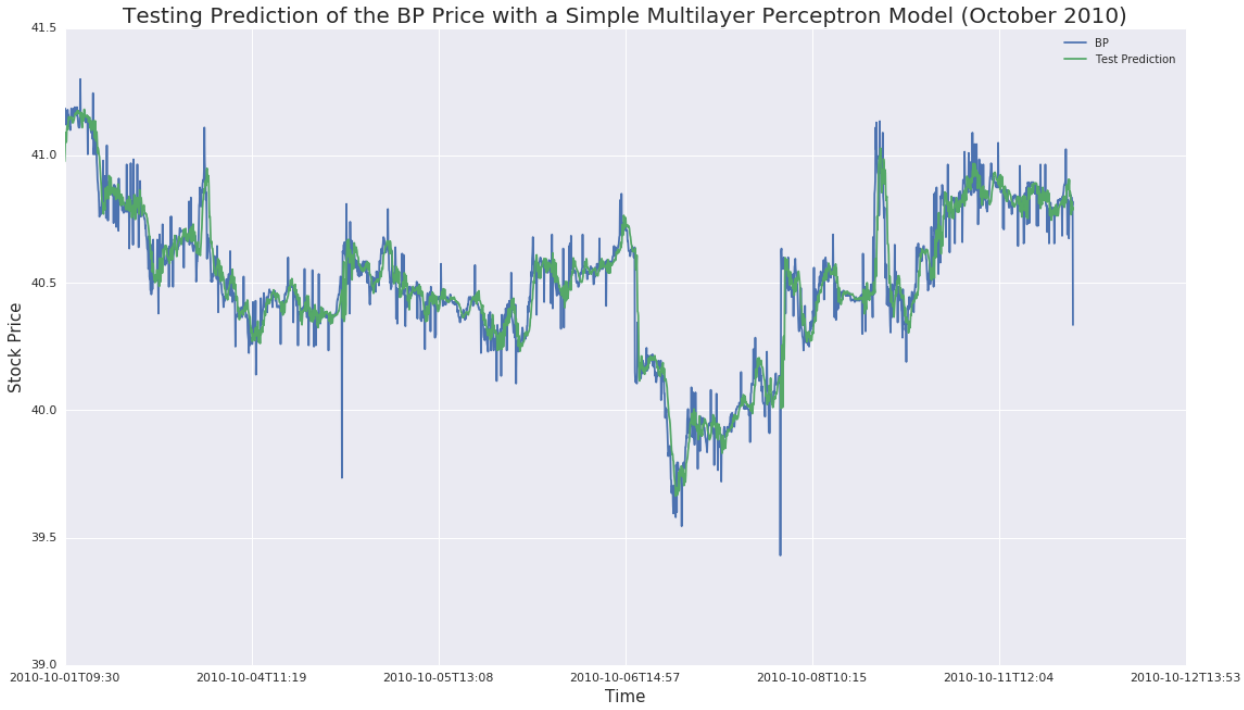

Figure 10 : Testing Prediction of the BP Price with a Simple Multilayer Perceptron Model (October 2010)

The graph in Figure 10 shows us the testing performance of the model. The green line plot is the plot of the Test Prediction while the blue line plot is the actual price of the BP stock. Similar to the train prediction, it might be difficult to decipher how accuracy of the model from the above plot due to the lack of granularity. Instead we will look at the percentage change in price for both the BP stock price (true value) and the prediction in Figure 11.

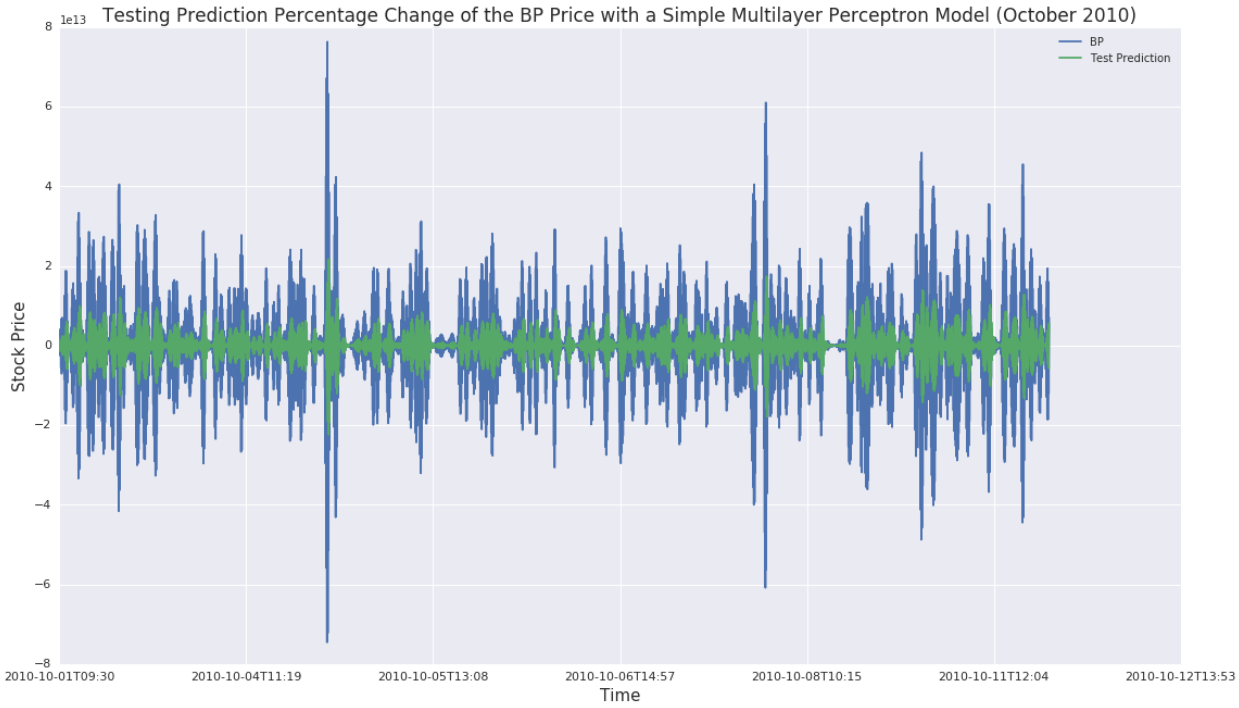

Figure 11 :Testing Prediction (Pct. Change) of the BP Price with a Simple Multilayer Perceptron Model (October 2010)

From Figure 11, we can see that the model's performance on the test data is similar to that of the train data. It is unable to accurately predict the amplitude of change in stock price as seen by the large spikes in blue lines accompanied with smaller spikes in green lines. The mean squared error of testing is 0.007431, which is significantly higher than the training mean squared error.

Optimizing Look Back Parameter

We have structured the problem of stock prediction as a regression problem. The look back function helps us to create this recurrence relation. Thus, is it also important for us to figure out the ideal look_back number that helps to reduce testing. In order to figure out the right look_back value, we need to repeat the train and test we did above for varying values of look_back. Figure 12 is a graph of Look Back against Mean Squared error for Look Back ranging from 1 to 38.

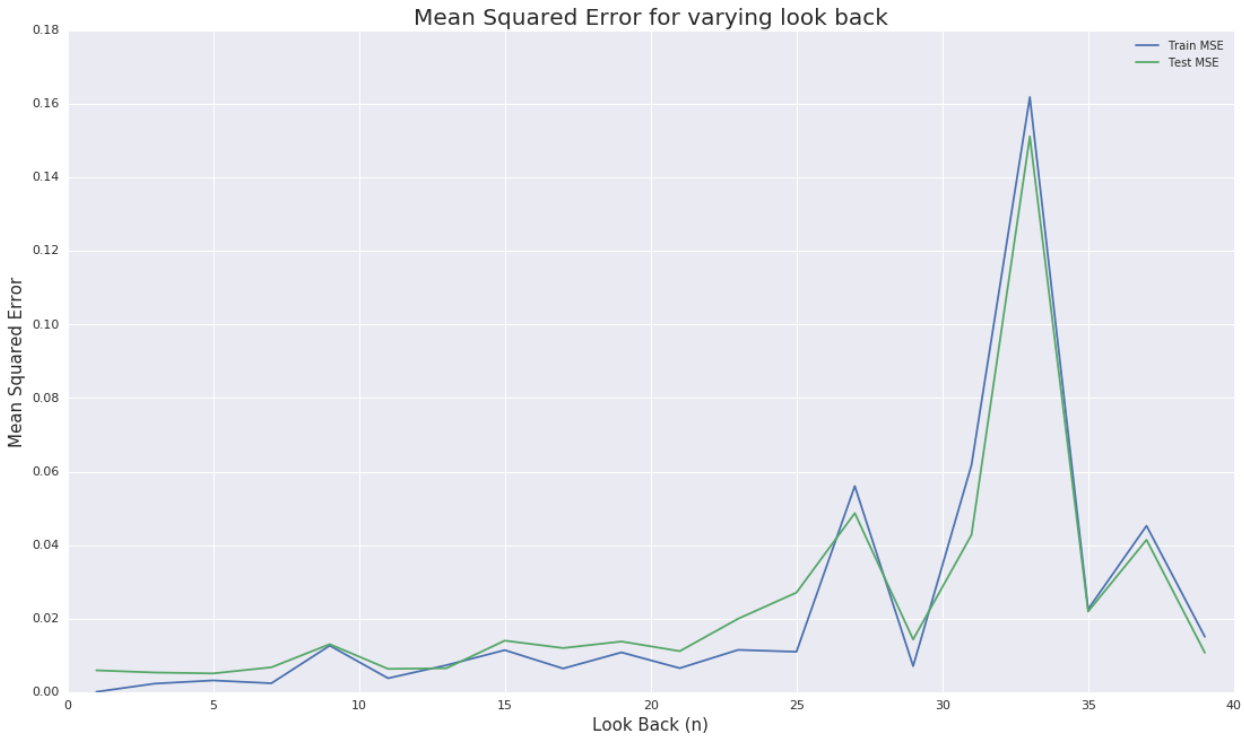

Figure 12 : Mean Squared Error for varying look back

We need to look for the minimum Test Mean Squared Error in Figure 12. Since this happens at look_back = 5, we should use 5 as the look_back value in future neural network models.

LSTM Recurrent Neural Network to Predict Stock Price

A recurrent neural network has connections with loop, adding feedback and memory to the networks over time. This makes it work well with sequence problems, like our stock price prediction problem. In this section, we are going to attempt using the Long-Short-Term Memory Network (LSTM). The LSTM network is a recurrent neural network that is trained using Backpropagation Through time and works to overcome the vanishing gradient problem. Thus, we can use this technique to transform a difficult sequence recurrence problem into a deep learning problem. Below is a basic example of how we can LSTM to predict the price of BP stock. We are going to be using look_back = 5 as explained in the previous section.

Train Prediction

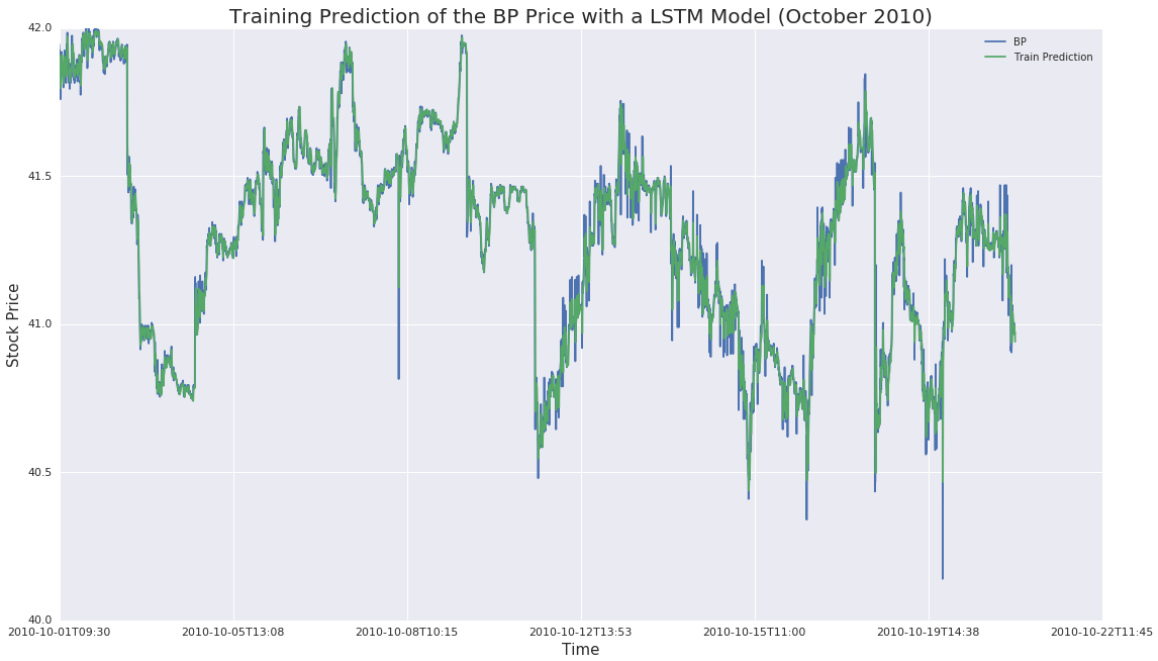

Figure 13 : Training Prediction of the BP Price with a LSTM Model (October 2010)

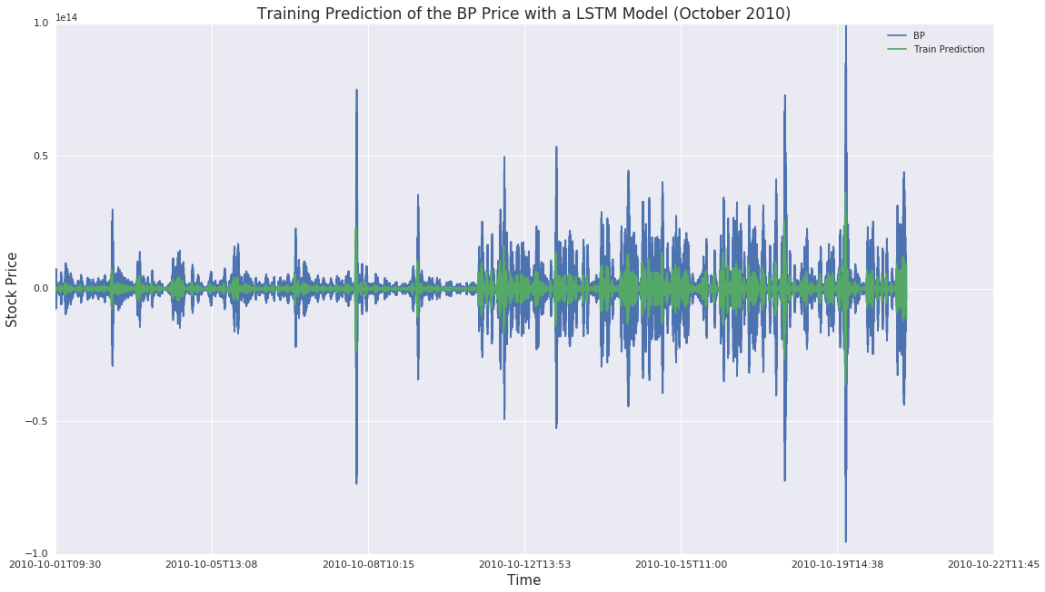

Figure 14 : Training Prediction (Pct. Change) of the BP Price with a LSTM Model (October 2010)

Figure 13 and 14 show the training prediction and Training Prediction (Pct. Change) for the BP stock using a LSTM Model. The Mean Squared Error of Training = 0.00351 is slightly higher than that of the Perceptron train Mean Squared Error.

Test Prediction

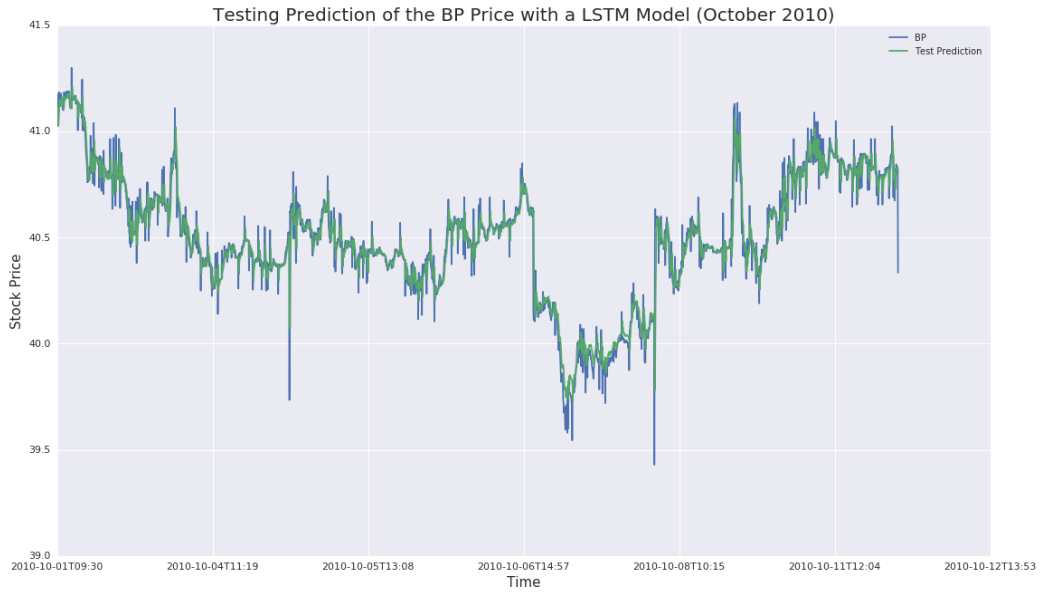

Figure 15: Testing Prediction of the BP Price with a LSTM Model (October 2010)

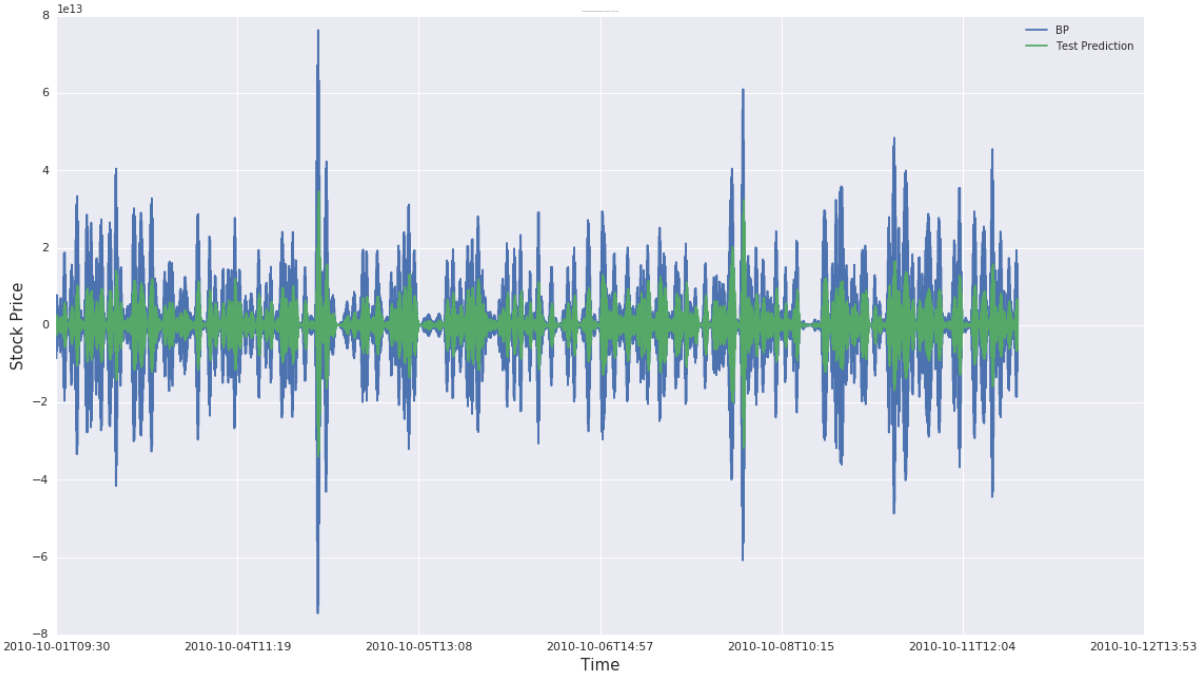

Figure 16 : Testing Prediction (Pct. Change) of the BP Price with a LSTM Model (October 2010)

Unlike the training mean squared error, the testing mean squared error = 0.004908 which is significantly lower than that of the testing mean squared error for the Perceptron Model. This is also visible is how the test prediction in Figure 16 better models the percentage change magnitudes of the BP stock.

Future Work

In this thesis, we performed detailed analysis on how we can utilize Applied Machine Learning techniques to leverage correlation between pairs of stocks to create models to understand fluctuations in stock prices and predict future stock prices. The research we have done so far has given us insights into some of the methods of approaching the problem of stock prediction. The lessons we have learned in this research would act as a starting point for future projects.

In order to create better performing stock prediction models, we need more data. In this thesis, we only analyzed a snapshot of stock price data from **2010-10-01** to **2010-10-29.** As such, we would need to have data for a longer period of time and have access to other features like Volume of Stock traded. This would help us in training our machine learning models better and create more realistic models.

We should also look into alternative data sources that contain preprocessed curated data. There are services, like Quandl, that allow developers to have access to live datasets that have been processed. Along with historical stock prices, Quandl has alternative datasets such as the FinSentS Web News Sentiment dataset and Alpha One Sentiment Dataset that provide sentiment scores based on articles aggregated from news sources. Furthermore, Quandl has curated live data about how company assets are managed and data about Long Term Growth Estimates History. The fact that these datasets are easily accessible allows researchers to focus more on performing analysis on the data rather than spending time to preprocess the data.

We should also extend this research to created a backtesting framework, utilizing libraries such as Zipline and Pyfolio, to automate the testing of the machine learning models that we create. Moreover, we should perform a deeper analysis on how we can perform time series analysis using libraries such as tsfresh and PyFlux.

Sources

1. Jan Ivar Larsen. Predicting Stock Prices Using Technical Analysis and Machine Learning (Link : http://www.diva-portal.org/smash/get/diva2:354463/fulltext01.pdf), 1-32.
2. Jason Brownlee. Develop Your First Neural Network in Python With Keras Step-By-Step (Link : http://machinelearningmastery.com/tutorial-first-neural-network-python-keras/)
3. Jason Brownlee. Crash Course in Recurrent Neural Networks for Deep Learning (Link : http://machinelearningmastery.com/crash-course-recurrent-neural-networks-deep-learning/)