CITY
UNIVERSITY OF LONDON
EST 1894

MSc in Data Science
INM430: Principles of Data Science Coursework

Using Random Forest Regression to build a portfolio of stocks based on Modern Portfolio Theory

Submitted by:

Aishwarya Chandra Shekhar

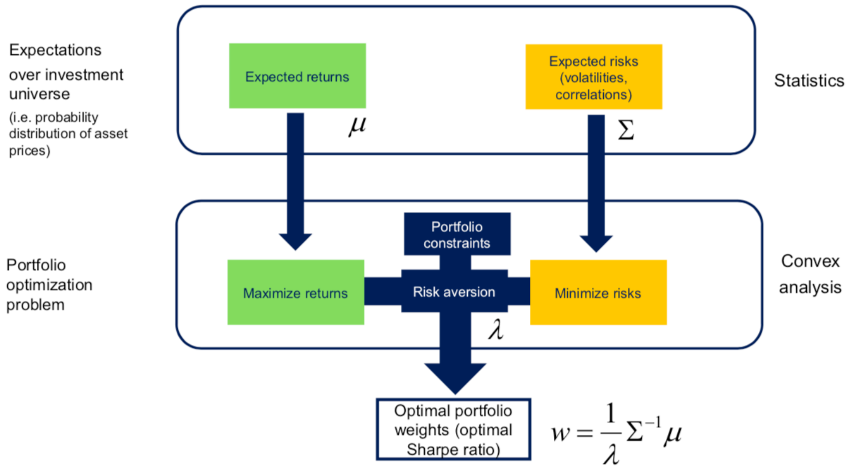
1 Analysis Domain, Questions, Plan

1.1 Introduction

Modern Portfolio Theory (MPT) was pioneered by Harry Markowitz in 1952. MPT describes how risk-averse investors can construct portfolios to optimize expected return based on a given level of market risk, accentuating the fact that risk is an inherent part of higher reward. [1]

MPT makes the assumption that investors are risk-averse. Another factor that comes in to play in MPT is “diversification”. By investing in more than one stock, an investor can reduce risk.

MPT can be explained by a flow diagram as below [2]:



Expectations over investment universe
(i.e. probability distribution of asset prices)
Expected returns
μ
Expected risks (volatilities, correlations)
Σ
Statistics
Portfolio optimization problem
Portfolio constraints
Maximize returns
Risk aversion
Minimize risks
λ
Convex analysis
Optimal portfolio weights (optimal Sharpe ratio)
$w = \frac{1}{\lambda}\Sigma^{-1}\mu$

Figure 1: Modern Portfolio Theory (1952)

1.2 Questions and Plan

The aim of this research is to form a portfolio of 5 stocks (out of 500 S&P stocks) and compare the return of this portfolio against the S&P index return.

In the process of achieving this aim, the following questions will be addressed:

A. Which sectors and individual stocks provide greatest return?
   A.1. Which are the top 5 sectors providing greatest return?
   A.2. Which are the top 5 stocks within each sector that provide the greatest return?
   A.3. Which are the top 5 stocks that provide the greatest return overall

B. What is the correlation between the stocks and the sectors?
   B.1. What is the correlation between stocks of each sector?
   B.2. Which stocks are correlated across the sectors?
   B.3. Which sectors are correlated with each other?
   B.4. What is the correlation between the 5 selected stocks?

C. Compare cumulative return of selected portfolio against that of S&P index.

The plan to achieve the above stated aim and answer the aforesaid questions is as follows:

1. Classify the stocks into their respective sectors and calculate stock-wise average daily returns.

2. Based on these returns the 5 best performing sectors and top 5 stocks in the S&P index (505 stocks) will be identified.
3. Within the 5 identified sectors, top 5 stocks will be identified and correlation between these stocks will be calculated.
4. The identified top 5 stocks from the S&P index will form the desired portfolio.
5. The monthly return of this portfolio will be compared against that of S&P index using Random Forest Regression model.

1.3 Data

The analysis data set pertains to financial domain and consists of daily prices of stocks listed on Standard & Poor's (S&P) index. Another data set contains the daily closing price of S&P index. The S&P index is an American stock market index based on market capitalizations of 505 large companies. The analysis data have been taken from online public data platform, Kaggle [3].

2. Data, Findings and Reflections

2.1 Data

The period of principal data set named 'all_stocks_5yr.csv' ranges from February 2013 to February 2018. Since the returns of individual stocks are compared with returns of S&P index, the period range of the two should be same. The period range of S&P index data ('SPIndex.csv') is January 2000 to December 2017. Thus, the common period ranging from February 2013 to December 2017 is considered for analysis.

2.2 Findings and Reflections

2.2.1 Average daily return

I. In order to identify the top 5 performing sectors, average of mean daily return of all stocks was consid ered. Based on these returns, the following 5 sectors have been identified as the sectors providing greate st average daily return:

Rank	Sector	Average daily return (%)
1.	Consumer Discretionary	0.066
2.	Information Technology	0.060
3.	Industrials	0.054
4.	Energy	0.057
5.	Financials	0.048

Table 1: Top 5 best performing sectors

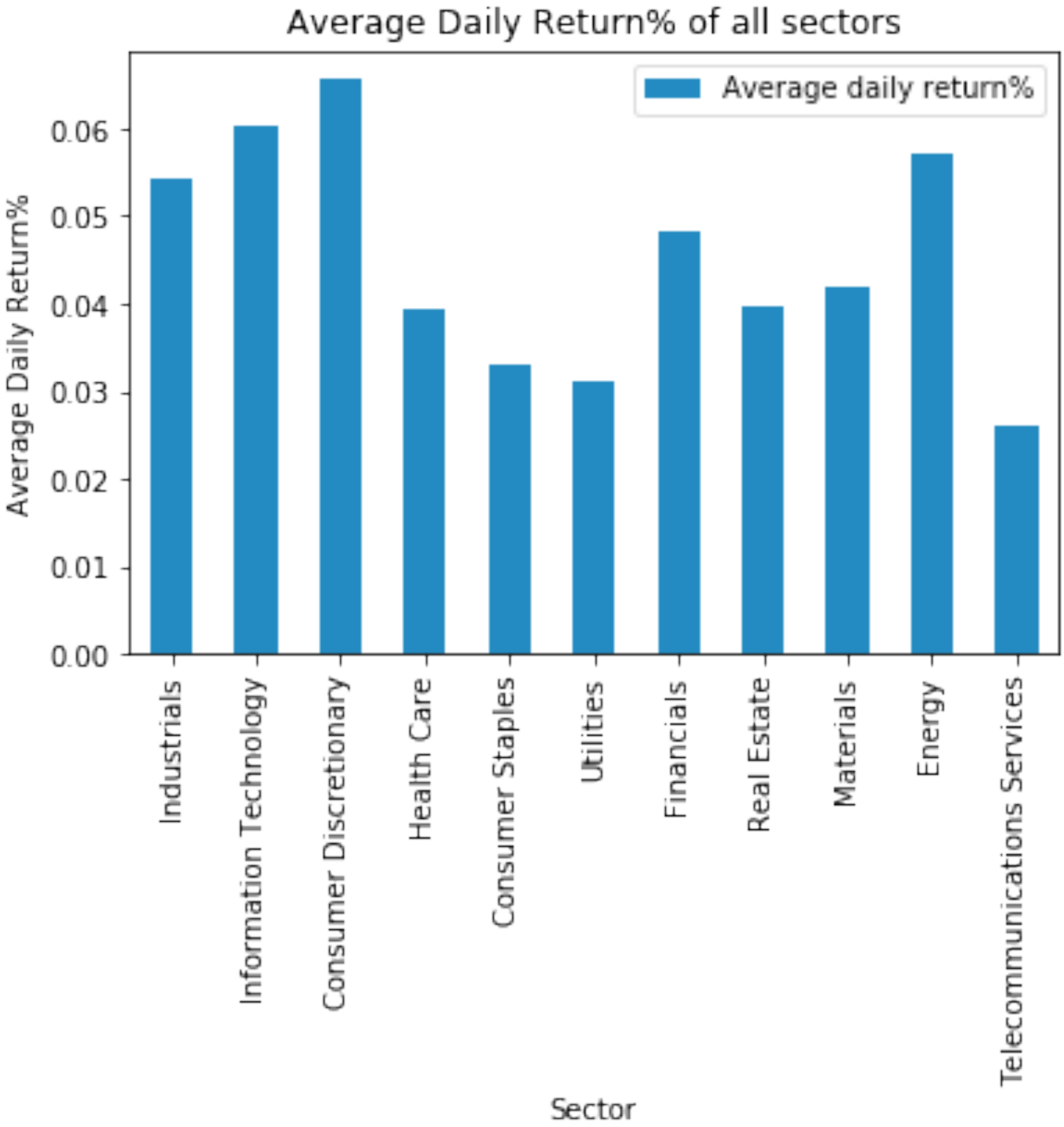


Figure 2: Average daily return of sectors

II. The top 5 performing stocks within the top 5 sectors is pictorially depicted as follows:

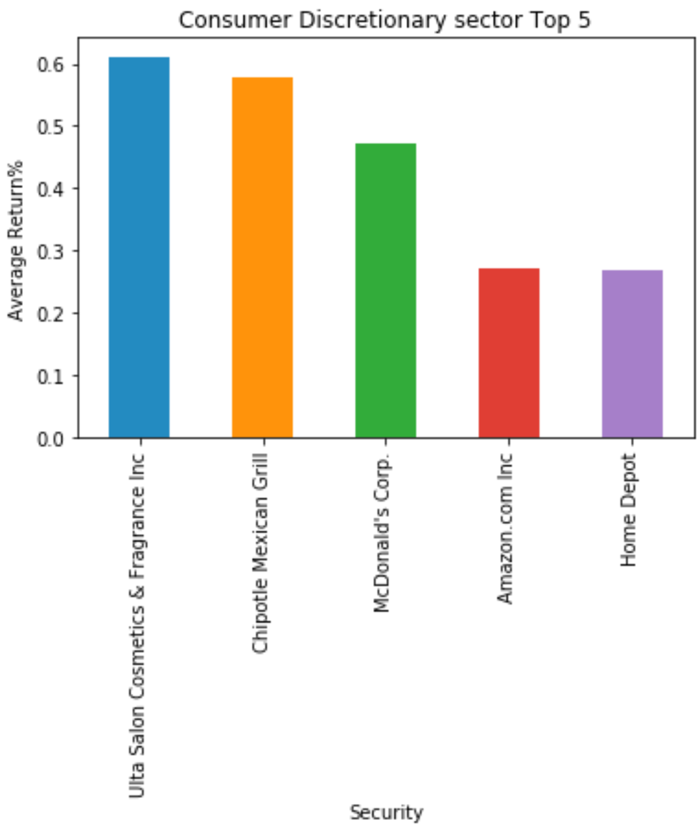


Figure 3: Top 5 stocks of Consumer Discretionary sector

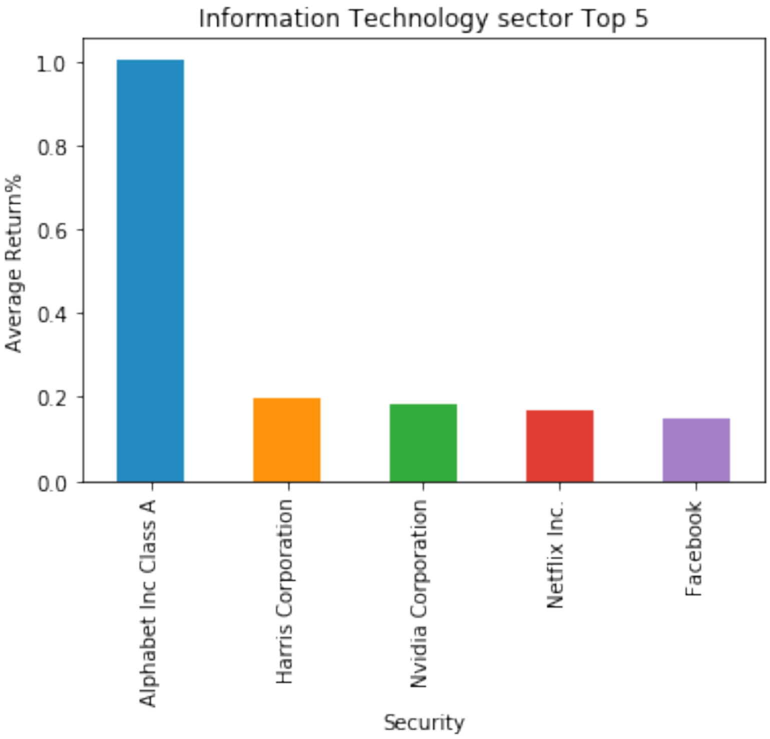


Figure 4: Top 5 stocks of Information Technology sector

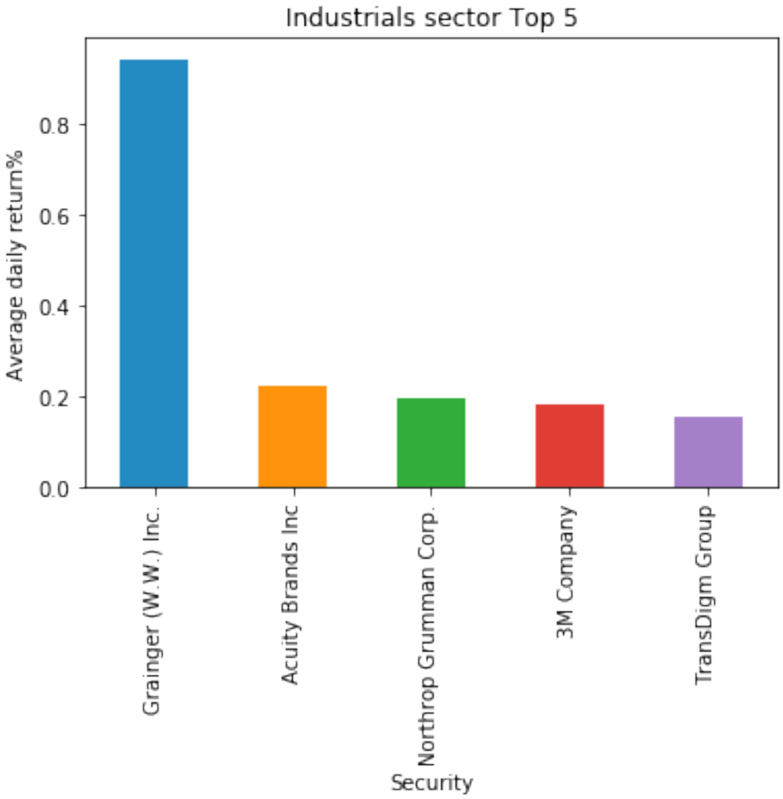


Figure 5: Top 5 stocks of Industrials sector

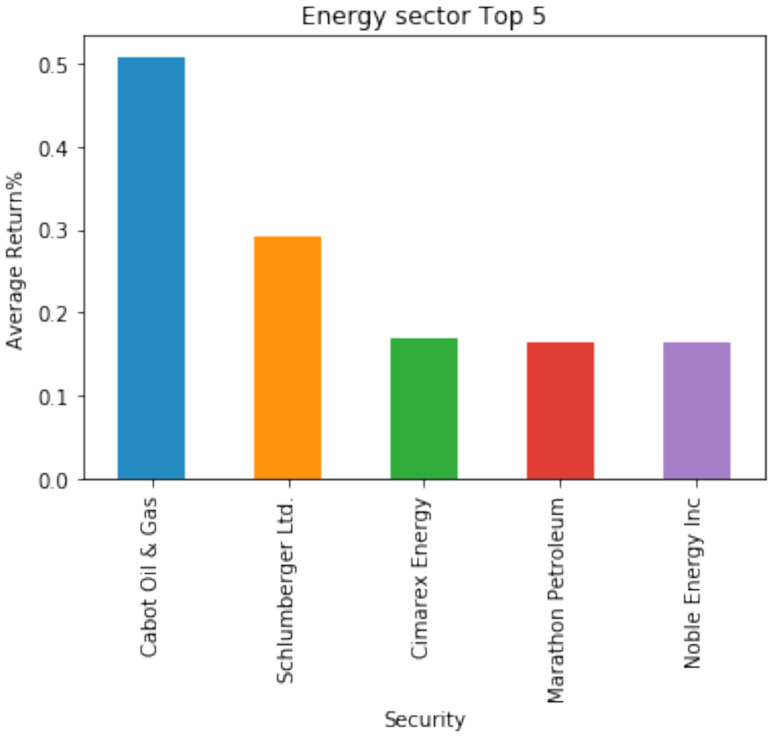


Figure 6: Top 5 stocks of Energy sector

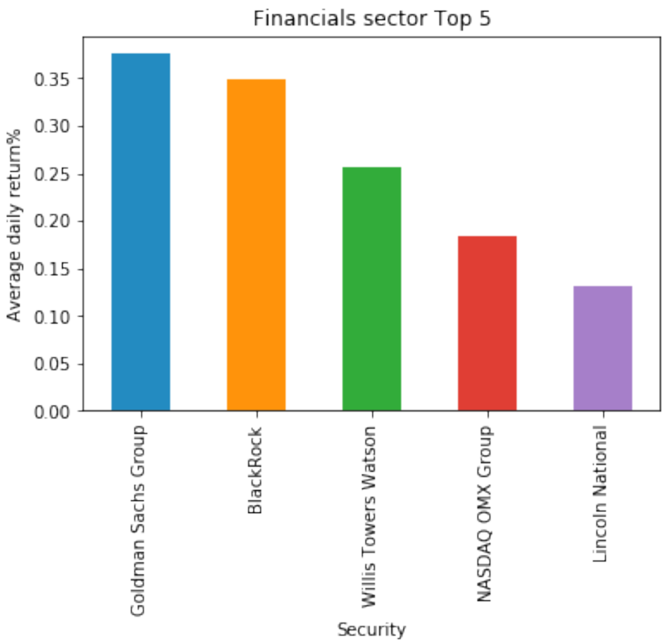


Figure 7: Top 5 stocks of Financials sector

III. The top 5 stocks overall that provide the greatest average daily return are:

Rank	Stock name	Average daily return (%)	Sector
1.	Alphabet Inc Class A	1.004	Information Technology
2.	Grainger (W.W.) Inc.	0.94	Industrials
3.	Ulta Salon Cosmetics & Fragrance Inc.	0.61	Consumer Discretionary
4.	Chipotle Mexican Grill	0.58	Consumer Discretionary
5.	Cabot Oil & Gas	0.51	Energy

Table 2: Top 5 best performing stocks

It can be seen that 2 of the top 5 performing stocks are from Consumer Discretionary sector.

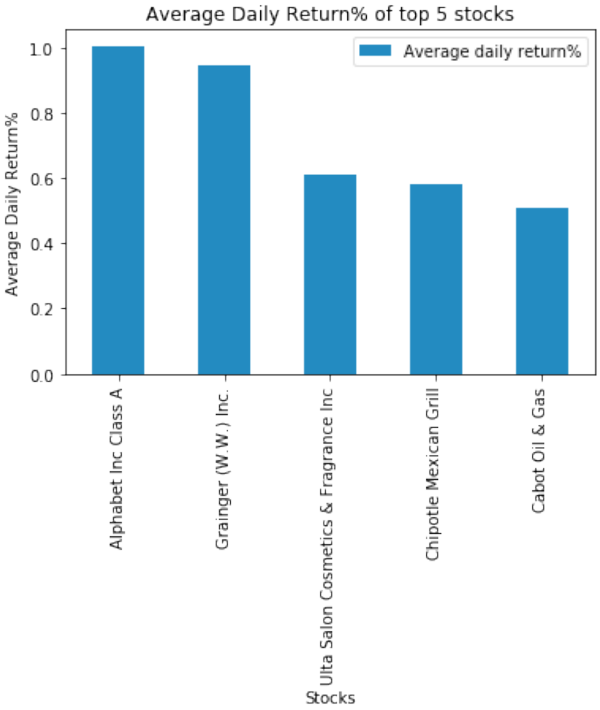


Figure 8: Average daily return of top 5 stocks

2.2.2 Correlation Matrix

I. The correlation between stocks of top 5 performing sectors is depicted using heatmaps as follows:

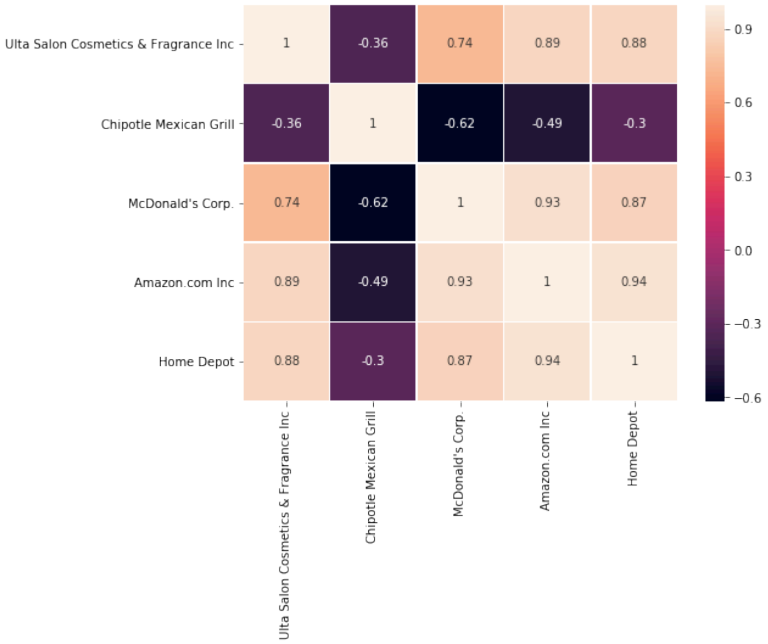


Figure 9: Correlation between stocks of Consumer Discretionary sector

The stocks of Consumer Discretionary sector are highly positively correlated with each other. Only Chipotle Mexican Grill is found to be negatively correlated with all other stocks of Consumer Discretionary sector. Strong negative correlation between Chipotle and McDonald’s could be attributed to the fact that they both are competitors in the fast-food chain industry.

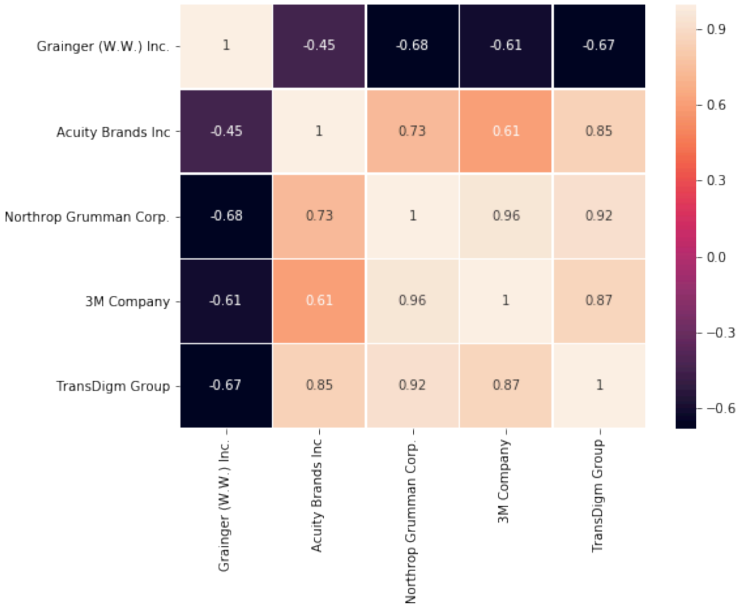


Figure 10: Correlation between stocks of Industrials sector

It can be seen that the top performing stock “Grainger (W.W.) Inc.”, which is a residential property business is negatively correlated with other stocks. Northrop Grumman and TransDigm Group, both manufacture defence components, are seen to have a strong positive correlation with each other.

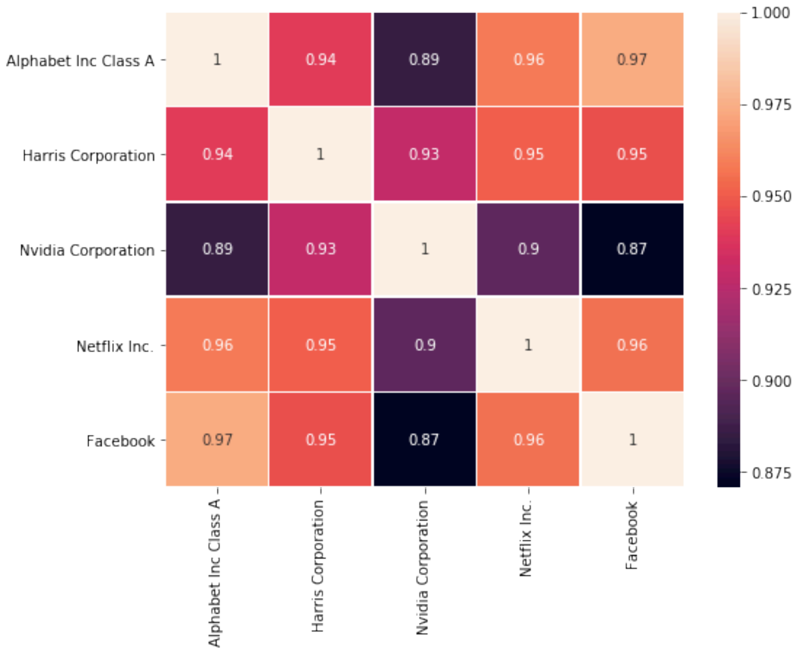


Figure 11: Correlation between stocks of Information Technology sector

Compared to all other sectors, stocks of Information Technology sector are found to be highly positively correlated with each other. The strongest positive correlation between Facebook and Netflix could be attributed to the fact that they both relate to social media.

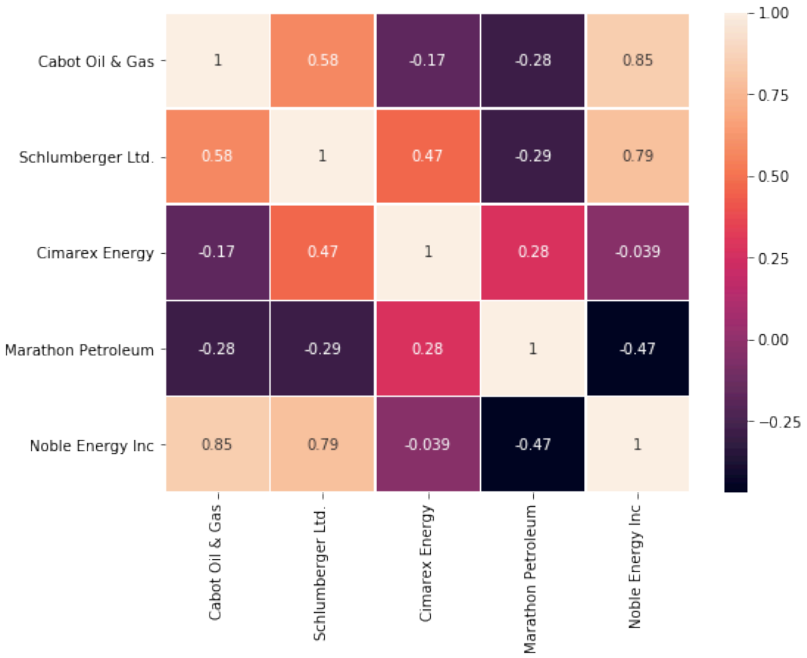


Figure 12: Correlation between stocks of Energy sector

A low correlation between the stocks of Energy sector could be attributed to the fact that they all are competitors in the same field of petroleum and natural gas. All other sectors contain stocks from different sub-sectors within each sector.

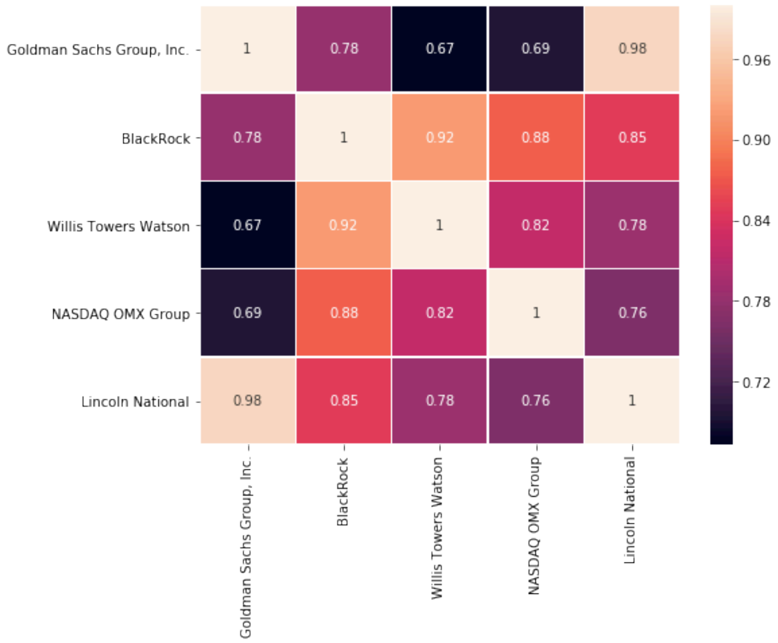


Figure 13: Correlation between stocks of Financials sector

The stocks of Financials sector have a strong positive correlation between each other. Goldman Sachs and Lincoln National, the two firms that deal in investment management, have the strongest positive correlation with each other.

II. Correlation between the sectors and stocks across sectors can be seen from the correlation matrix as given below:

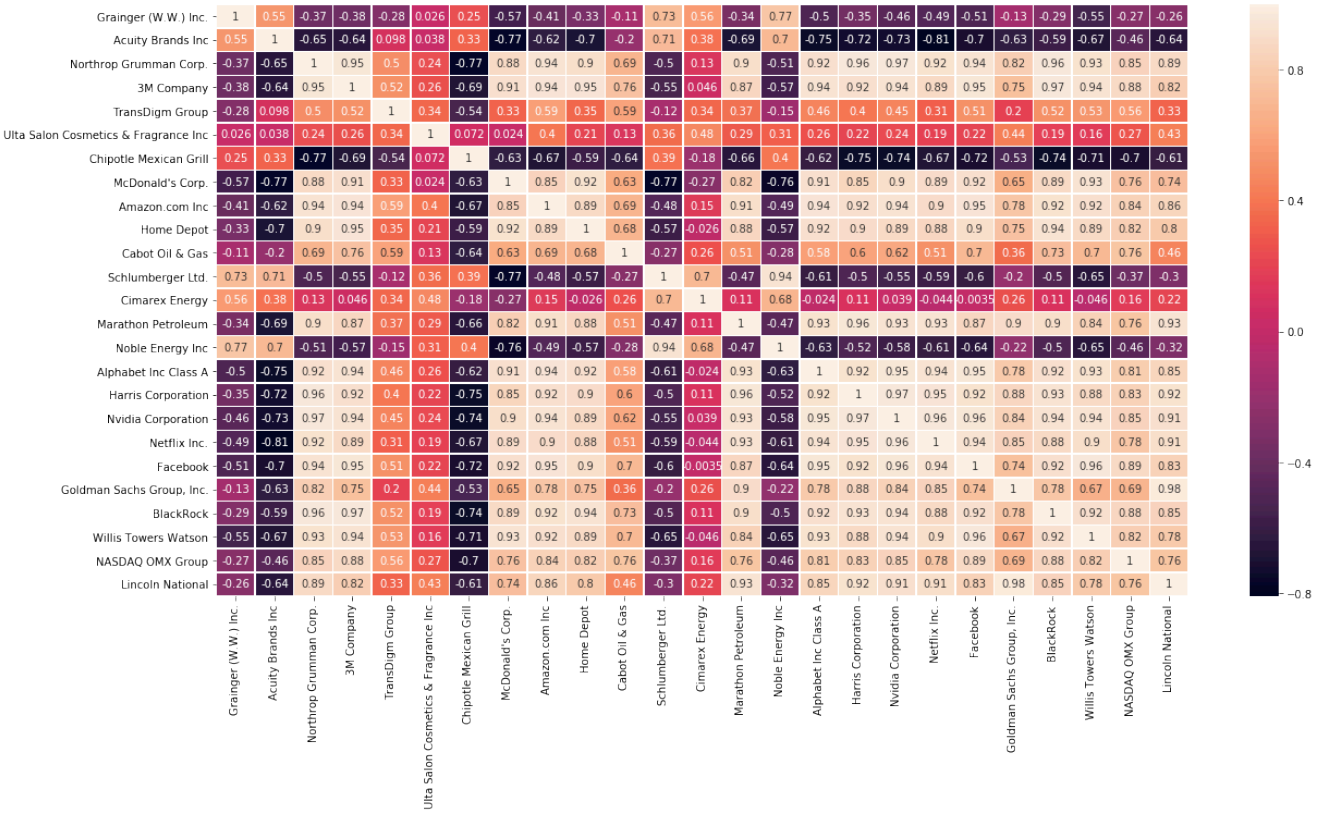
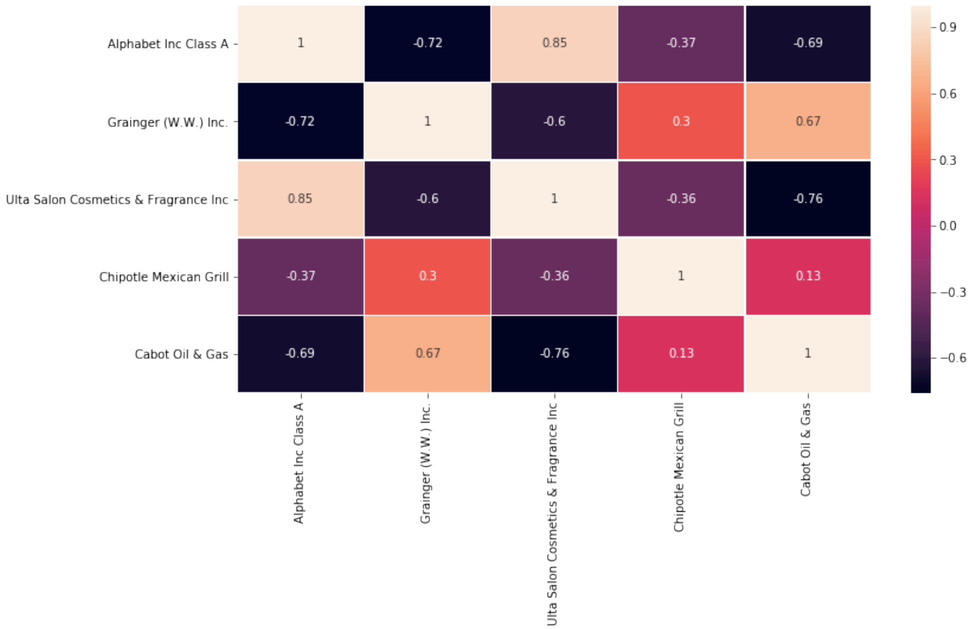


Figure 14: Correlation between sectors and stocks across sectors

The bottom right part seems to have a strong correlation with each other. This part represents Financials and Information Technology (IT) sectors. Thus, it could be said that Financials and IT sectors are highly positively correlated with each other. This could be interpreted by considering that in the modern times, a lot of Financial firms collaborate with IT firms to carry out their day-to-day business activities in a more efficient manner.

Also, three of the five stocks of Consumer Discretionary sector, viz. McDonald's, Amazon.com and Home Depot are seen to have a strong positive correlation with Financials and Information Technology sectors. Emphasizing the strongest correlation between Amazon.com with IT sector could be attributed to the fact that Amazon.com deals in cloud computing as well.

III. Correlation between the top 5 performing stocks can be seen from the following correlation matrix:



	Alphabet Inc Class A	Grainger (W.W.) Inc.	Ulta Salon Cosmetics & Fragrance Inc	Chipotle Mexican Grill	Cabot Oil & Gas
Alphabet Inc Class A	1	-0.72	0.85	-0.37	-0.69
Grainger (W.W.) Inc.	-0.72	1	-0.6	0.3	0.67
Ulta Salon Cosmetics & Fragrance Inc	0.85	-0.6	1	-0.36	-0.76
Chipotle Mexican Grill	-0.37	0.3	-0.36	1	0.13
Cabot Oil & Gas	-0.69	0.67	-0.76	0.13	1

Figure 15: Correlation between the top 5 performing stocks

The top 5 performing stocks have low to negative correlation with each other. This helps in forming a diverse portfolio. Diverse portfolio helps to reduce risk and maximize returns, as per Modern Portfolio Theory.

2.2.3 Cumulative return

The cumulative return of the 5 selected stocks and S&P index over the period from February 2013 to December 2017 is as shown below:

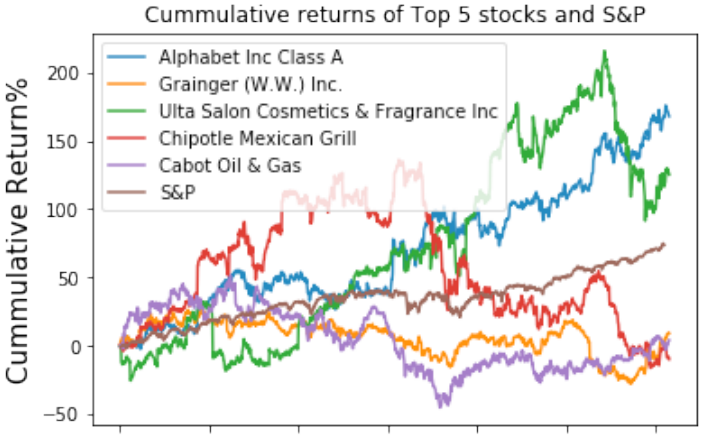


Figure 16: Cumulative returns

It can be seen that cumulative returns of the 5 selected stocks are highly unstable, whereas, that of S&P index has increased gradually and have been more stable over the five years.
It is observed that at the end of five years, Alphabet Inc. gives the highest return, whereas, Chipotle Mexican Grill gives the lowest return. The strong decline in return of Chipotle towards the end of 2015 can be attributed to E. coli virus outbreak.

2.3 Using Machine Learning for Modern Portfolio Theory

To use ML to pick the best portfolio, features and targets were generated. Exponentially weighted moving averages (ewma) of prices were taken as features and portfolios with highest Sharpe ratio were taken as targets. The predicted monthly return of selected portfolio was compared with the actual monthly return of S&P index for the last 5 months from August 2017 to December 2017.

It was found that monthly returns of S&P index have been stable during these 5 months, whereas, returns of selected portfolio fluctuated a lot in this period. In November 2017, the monthly returns of selected portfolio surpassed the monthly return of S&P index.

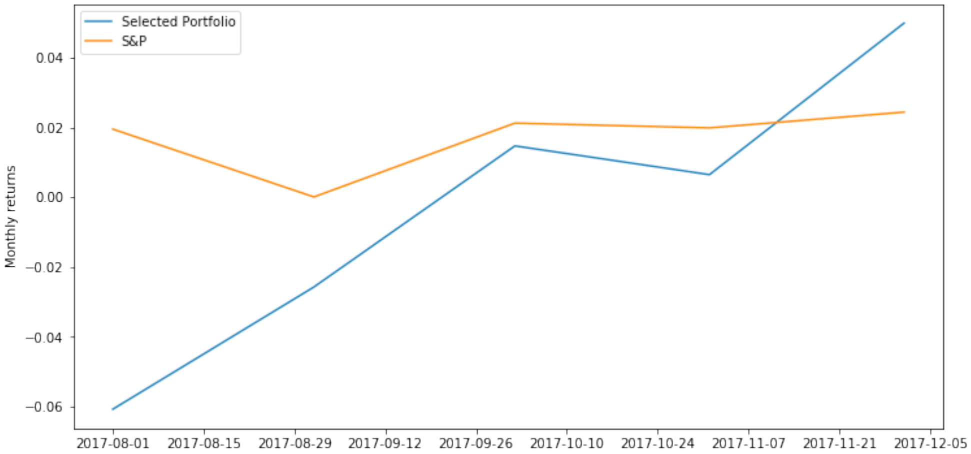


Figure 17: Monthly returns of Selected portfolio and S&P index

In order to calculate the absolute return on investing in the selected portfolio and in the S&P index, an investment of GBP 1000 was considered. It was seen that at the end of five months, S&P index paid a positive return of 8.77% and the portfolio paid a negative return of 0.07%.

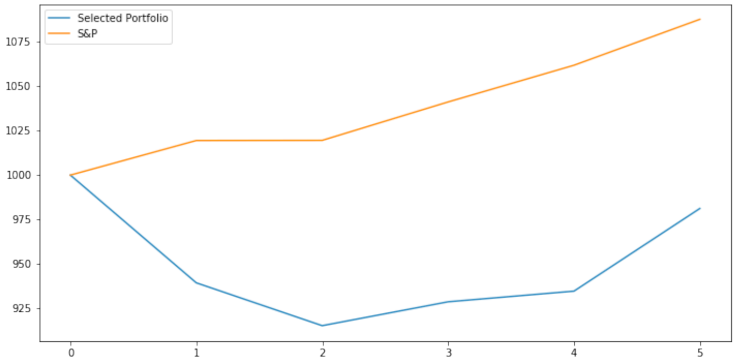


Figure 18: Return on Selected portfolio and S&P index

3. Limitations and Scope

The accuracy of the model predicting the portfolio return is low. This could be improved by using more advanced algorithms such as Neural Networks and Deep Learning.

4. Public HTML Link

The Public HTML link to my work is:
https://smcse.city.ac.uk/student/aczd064/INM430_Computational_Notebook_Aishwarya_Chandra_Shekhar.html

5. References

[1] (n.d.). Retrieved from https://www.investopedia.com/terms/m/modernportfoliotheory.asp
[2] BRUDER, B. (2017). *MACHINE LEARNING AND ASSET MANAGEMENT.* LYXOR.
[3] Link to the referred data sets : https://www.kaggle.com/camnugent/sandp500#all_stocks_5yr.csv , https://www.kaggle.com/dgawlik/nyse, and https://www.kaggle.com/adityarajuladevi/sp-index-historical-data