Tze Fung Lung, Jim

Data 698 - Final presentation

Topic: **Portfolio optimization and Machine learning with visualization analysis for S&P 500**

Introduction

- We are looking at the S&P 500, an index of the largest US companies. The S&P 500 is an American stock market index based on the market capitalization of 500 large companies having common stock listed on the NYSE, NASDAQ Exchange.

- I loaded all 500 dataset in S&P 500 for analysis by using portfolio optimization to get the possible several stocks with higher return and lower risk. And using the machine learning predict the investment trend for S&P 500 index.

Methodology

- This project are separated parts of analysis from data exploration, visualization, correlation and monthly return for data extraction by mathematical programming, portfolio optimization and machine learning .

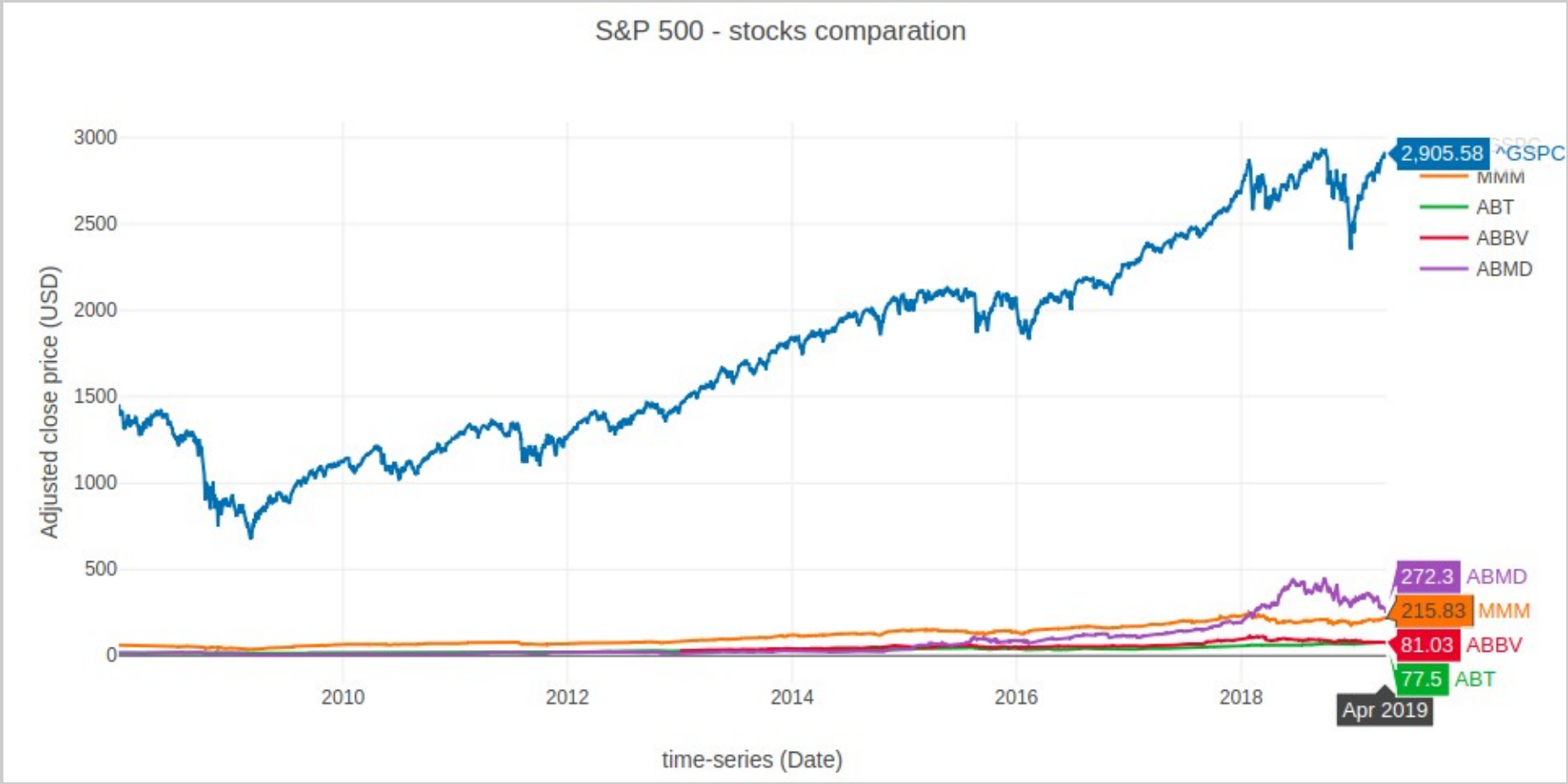
- 1. Data Exploration
- 2. Higher monthly return
- 3. Portfolio Optimization
- 4. Machine Learning

Data Exploration

- We considered if it plot all 500 stock shares, it can't appear a meaningful comparation graph. Initially, we view the first 5 column including S&P 500 index in the time series from 2008 to 2019 (2019-04-18).

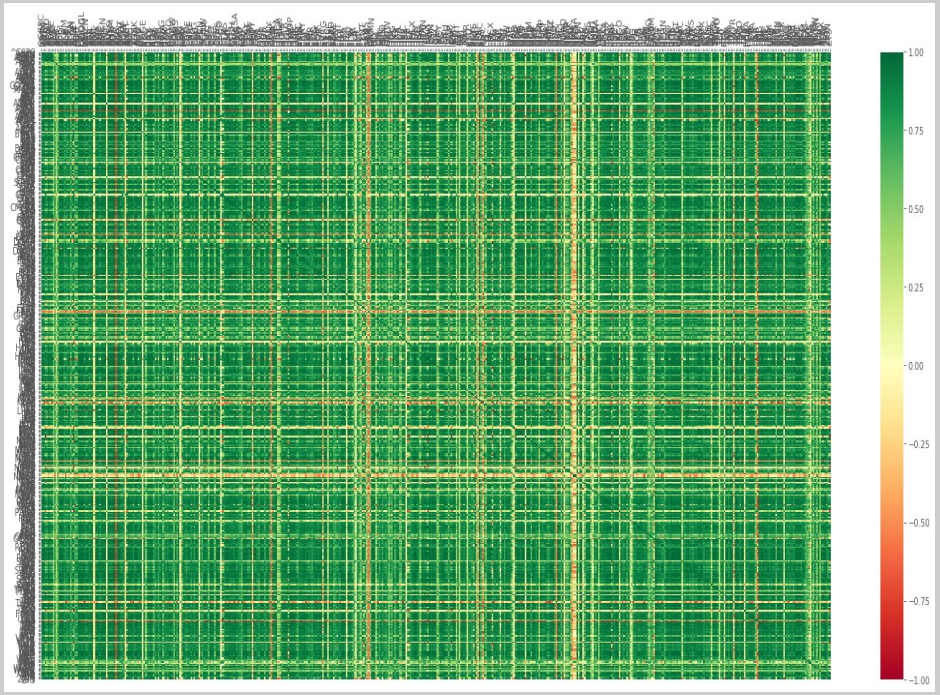
- We will further see the top 10 monthly return, top 10 for higher correlation with S&P 500 index, and top 10 investing stocks strategic from the portfolio optimization in our data time series.

Data Exploration



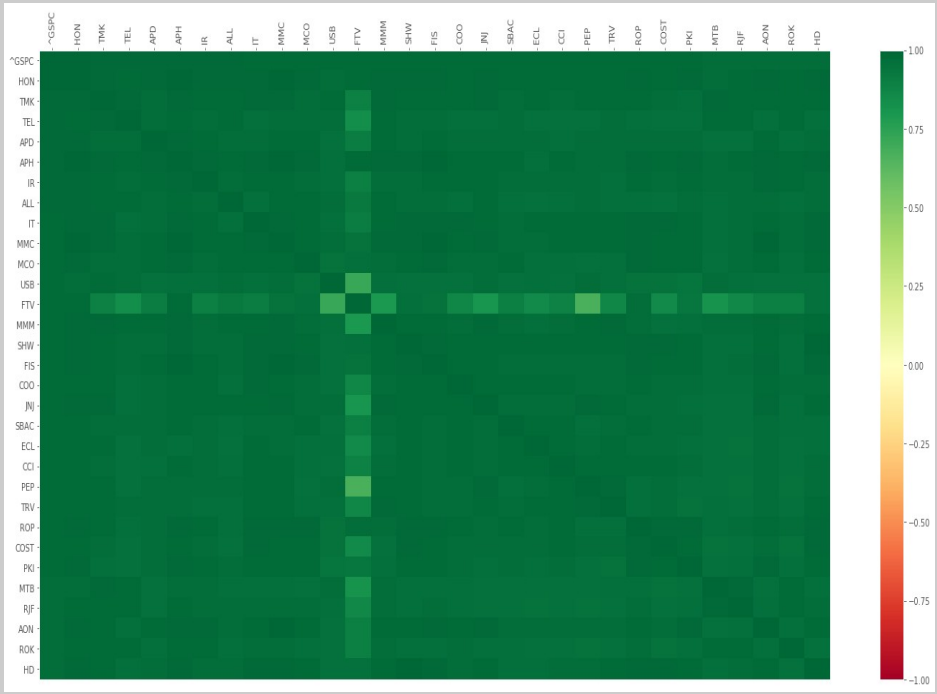
Correlation for all 500 stocks

Most people think that they will get higher return when the main index going up, so we can calculate the correlation to check which stocks share are strong relationship and low relationship.



Highest Correlation (Sort out top 30 from all 500 stocks share)

- to find out the top 30 stocks which are higher correlation with S&P 500 index from 500 stocks share.
- Last part of this project will conduct the prediction of machine learning for S&P 500 index

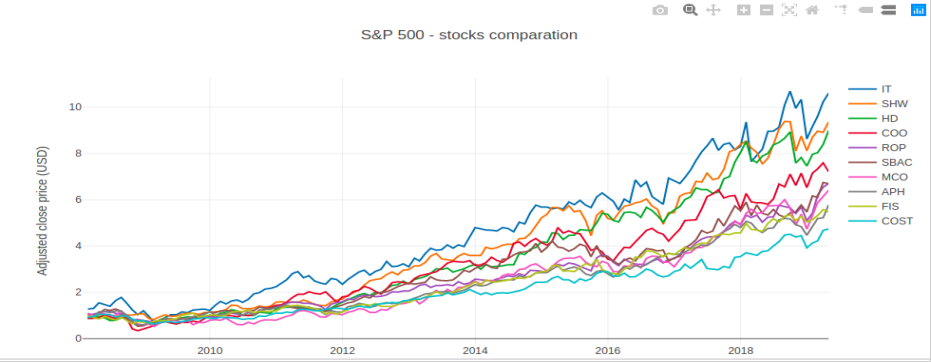


Highest monthly return (Sort out top 10 from 30 stocks share)

- Apart from the risk consideration, the percentage of increasing return can be 9.8 times over 10 year if just only invest "IT" tickers of stock.

Date	IT	SHW	HD	COO	ROP	SBAC	MCO	APH	FIS	COST
2018-07-31	9.132164	9.055837	8.472125	6.675123	5.756390	5.344478	5.750779	5.129259	5.036642	4.187018
2018-08-31	10.098449	9.379355	8.655667	6.554177	5.689082	5.242485	5.998048	5.187952	5.282784	4.474156
2018-09-30	10.687795	9.371737	8.930724	7.101767	5.647898	5.424856	5.633489	5.169493	5.342351	4.507741
2018-10-31	9.947404	8.100652	7.582601	6.619007	5.401586	5.476866	4.901674	4.920970	5.098915	4.387793
2018-11-30	10.329737	8.748771	7.820522	7.144816	5.682071	5.768659	5.375525	4.835197	5.287492	4.449202
2018-12-31	8.620364	8.117271	7.451876	6.521378	5.088832	5.467409	4.732435	4.466833	5.038099	3.918829
2019-01-31	9.163183	8.696162	7.959740	7.143824	5.417782	6.164471	5.356600	4.847247	5.135373	4.128901
2019-02-28	9.595415	8.960154	8.029566	7.329108	6.190303	6.097940	5.867519	5.180798	5.313217	4.219332
2019-03-31	10.227916	8.908651	8.384287	7.589993	6.540701	6.742992	6.137644	5.219292	5.574535	4.670700
2019-04-30	10.600135	9.357900	8.985942	7.206867	6.697916	6.693009	6.407770	5.779133	5.465607	4.741492

Using 10 years cumulative return for selecting higher return, it may be some problem, like it does not reflect the trend of industry, because new industry may have large return only these few years, it may use only 3 years cumulative return, it would be more realistic. But now in research level, we keep using 10 years sum of return.

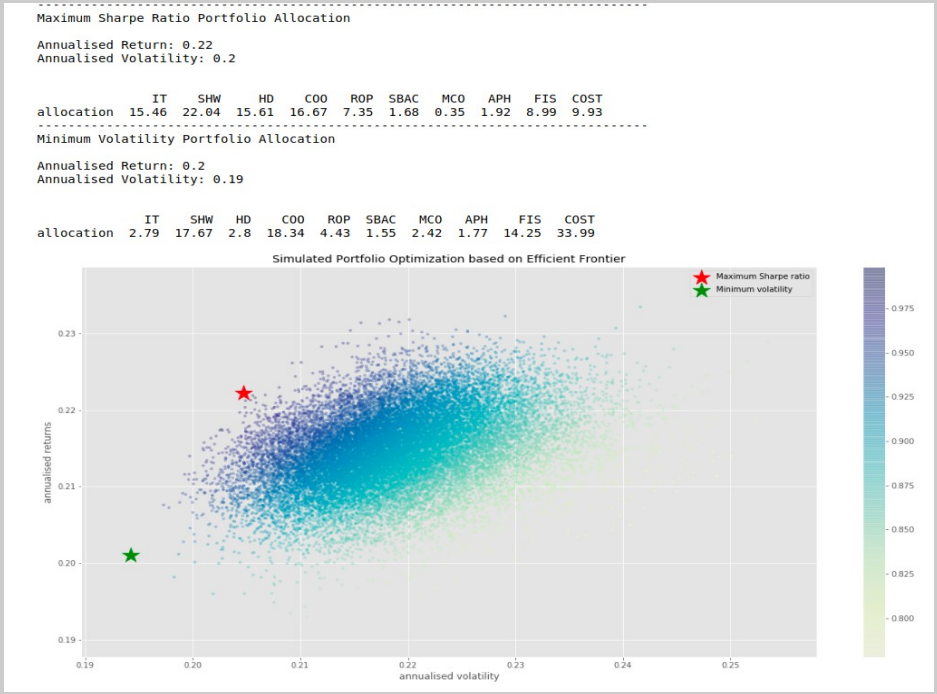


Highest monthly return (Sort out top 10 from 30 stocks share)

- Gartner Inc
- Sherwin-Williams
- Home Depot
- The Cooper Companies
- Roper Technologies
- SBA Communications
- Moody's Corp
- Amphenol Corp
- Fidelity National Information Services
- Costco Wholesale Corp.

Portfolio Optimization - Random Portfolios

- In order to choose from 10 number of stocks share from the result of last part which are higher monthly return and lower risk as our final medium or long term investment strategic.



Portfolio Optimization - Random Portfolios

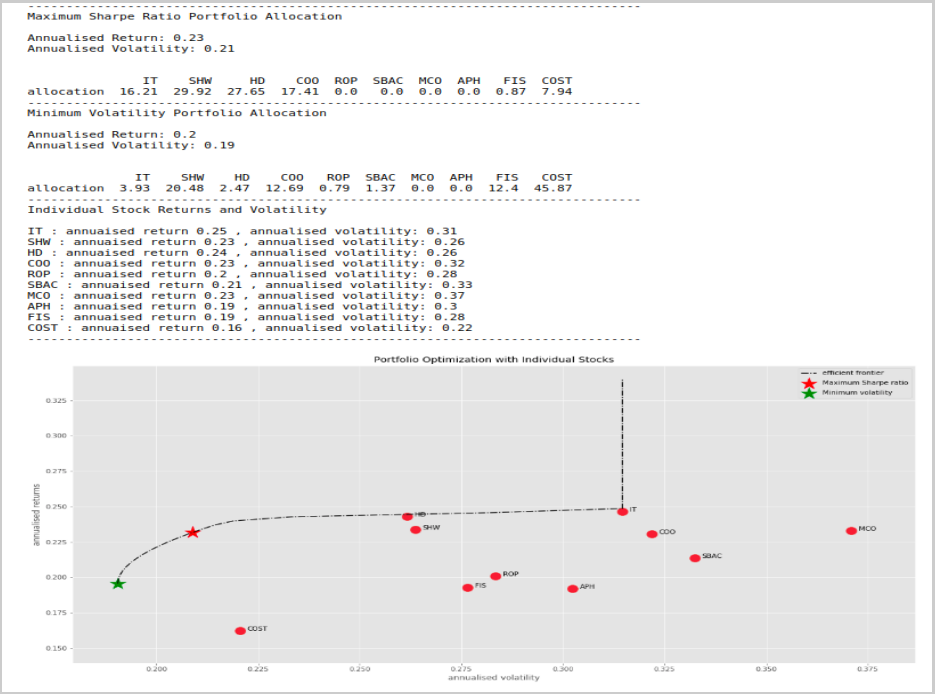
- For minimum risk portfolio, we can see around 30% of our budget is allocated to "Cost" - Costco. If you take another look at the daily return plot from earlier, the Costco is the least volatile among these stocks, so allocating a large percentage to Costco for minimum risk portfolio makes sense.

- In this scenario, we are allocating a significant portion to "SHW" - Sherwin-Williams and "IT" - Gartner Inc, which are quite volatile stocks from the previous plot of daily returns. And "Cost" - Costco which had around 30% in the case of minimum risk portfolio, has only 10% budget allocated to it.

Portfolio Optimization - Efficient Frontier

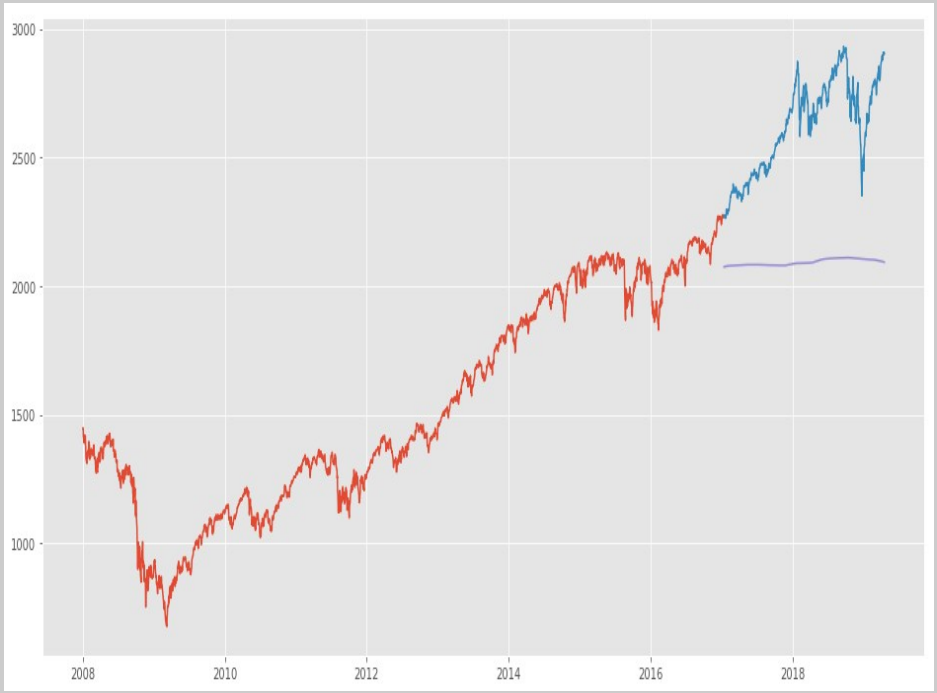
- Stocks with the least risk is COST at around 0.22, but the return is only around 0.16. If we will to take slightly more risk around 0.225 return, we should consider to choose HD and SHW rather than IT with higher risk with portfolio optimisation.

```
--------------------------------------------------------------------------------
Maximum Sharpe Ratio Portfolio Allocation

Annualised Return: 0.23
Annualised Volatility: 0.21

                IT    SHW     HD    COO  ROP  SBAC  MCO  APH   FIS  COST
allocation  16.21  29.92  27.65  17.41  0.0   0.0  0.0  0.0  0.87  7.94
--------------------------------------------------------------------------------
Minimum Volatility Portfolio Allocation

Annualised Return: 0.2
Annualised Volatility: 0.19

                IT    SHW    HD    COO   ROP  SBAC  MCO  APH   FIS   COST
allocation  3.93  20.48  2.47  12.69  0.79  1.37  0.0  0.0  12.4  45.87
--------------------------------------------------------------------------------
Individual Stock Returns and Volatility

IT : annuaised return 0.25 , annualised volatility: 0.31
SHW : annuaised return 0.23 , annualised volatility: 0.26
HD : annuaised return 0.24 , annualised volatility: 0.26
COO : annuaised return 0.23 , annualised volatility: 0.32
ROP : annuaised return 0.2 , annualised volatility: 0.28
SBAC : annuaised return 0.21 , annualised volatility: 0.33
MCO : annuaised return 0.23 , annualised volatility: 0.37
APH : annuaised return 0.19 , annualised volatility: 0.3
FIS : annuaised return 0.19 , annualised volatility: 0.28
COST : annuaised return 0.16 , annualised volatility: 0.22
--------------------------------------------------------------------------------
```



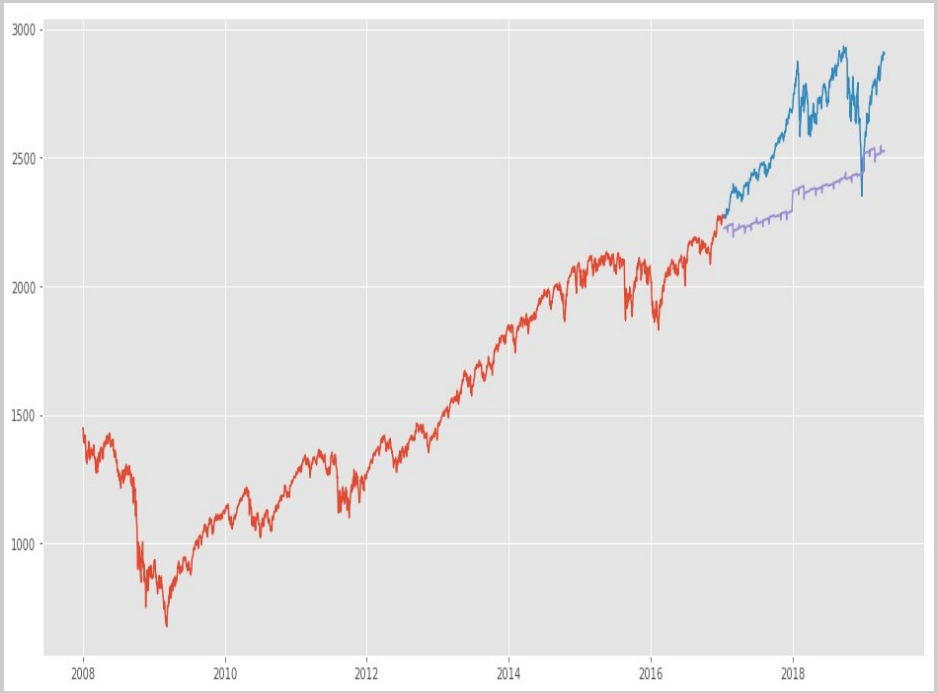
Machine Learning - Moving Average

- The root mean square error is 556.7202, prediction is 2092.7282
- the results are not very promising, the predicted values are not the same range as the observed values in the train set
- (there is not obvious an increasing trend and then a slow decrease)



Machine Learning - Linear Regression

- The root mean square error is 296.1361, prediction is 2525.14814
- The RMSE value is lower than the previous technique, which clearly shows that linear regression has performed better.

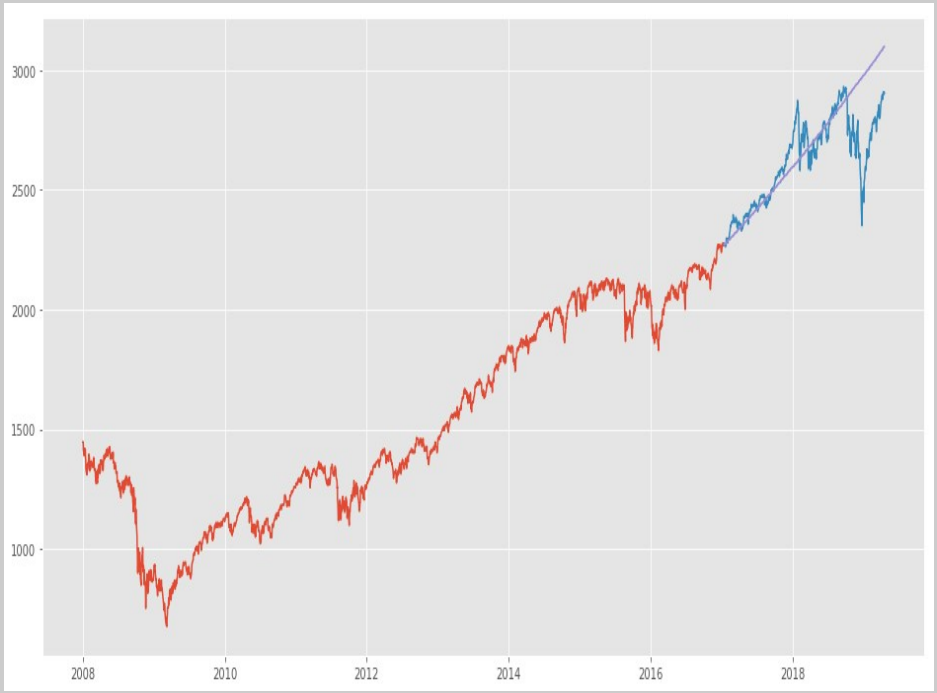


Machine Learning - k-Nearest Neighbours

- The root mean square error is 1142.9905, prediction is 2266.327
- The RMSE value of KNN is higher than the linear regression model and the plot shows the same pattern
- kNN also identified a raising trend from January 2019 since that has been the pattern for the past years

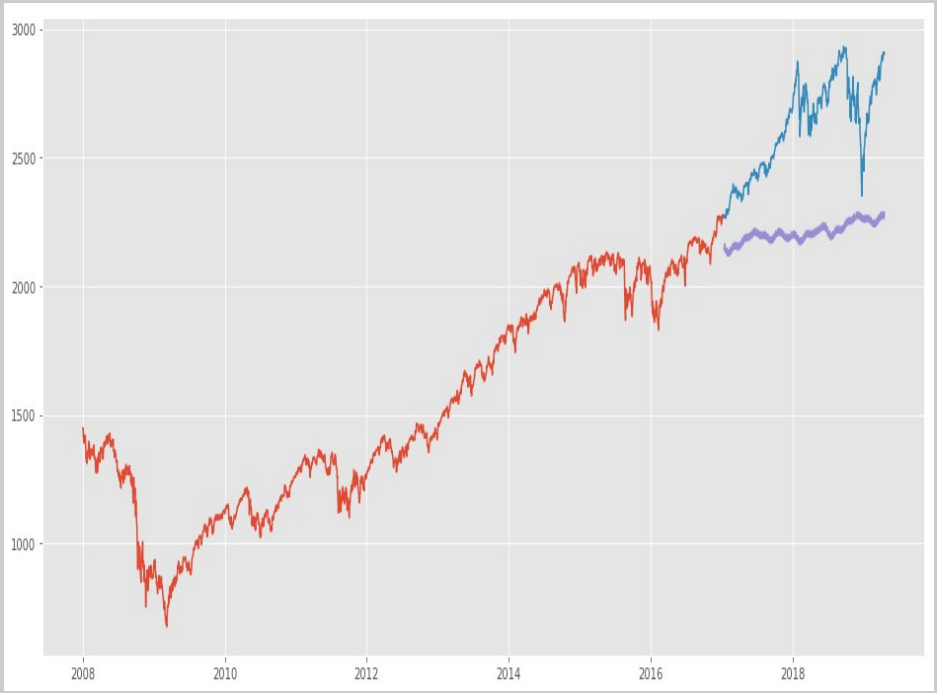
Machine Learning - Auto ARIMA

- The root mean square error is 149.4734, prediction is 3098.4159
- Although the predictions using this technique are far better than that of the previously implemented machine learning models, these predictions are still not close to the real values.



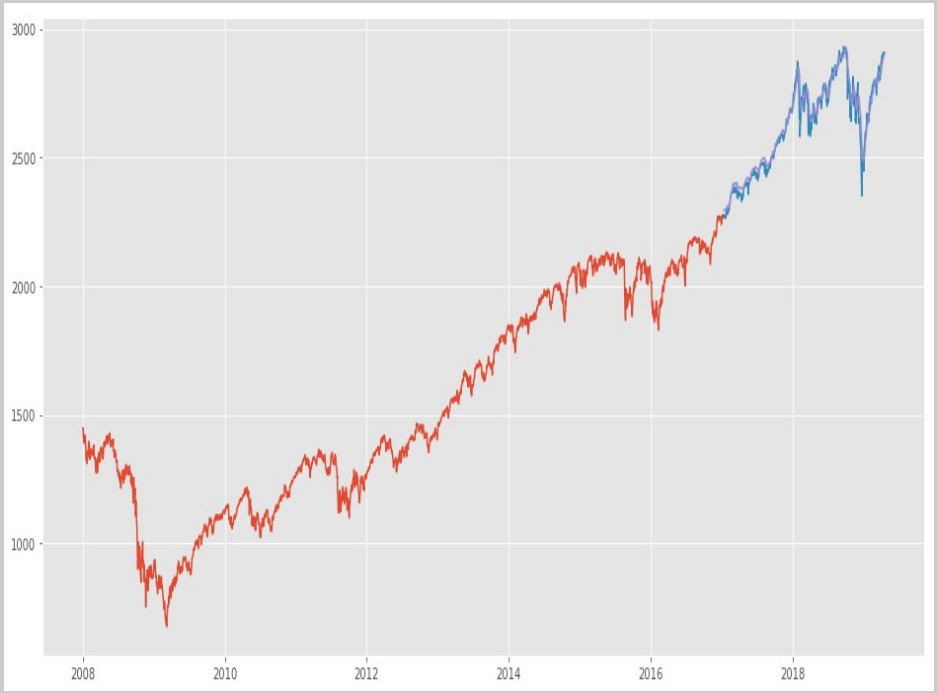
Machine Learning - Prophet

- The root mean square error is 446.4384233932304, prediction is 2285.3302
- Forecasting techniques like ARIMA and Prophet would not show good results for this particular problem.



Machine Learning - Long Short Term Memory (LSTM)

- The root mean square error is 36.8895, prediction is 2899.4375
- LSTMs are widely used for sequence prediction problems and have proven to be extremely effective.
- the model seem to predict well with real data



Model Comparsion

- The S&P 500 index is 2905.030029 on 2019-04-18
- the most lowest rmse of model is by using of Long Short Term Memory method, prediction value is 2899.4375 As we mentioned as last part, we should also consider the news about the company and other factors like demonetization or merger/demerger of the companies.

	RMSE	Prediction
ma	556.720265	2092.728205
linear	296.136189	2525.148140
knn	1142.990569	2266.327772
arima	149.473499	3098.415958
prophet	446.438423	2285.330298
lstm	36.889548	2899.437500

```
The last S&P 500 index is 2905.030029
```