Portfolio Forecasting Using Neural Networks (LSTMS) & Optimization

Akhil Sanjay Potdar

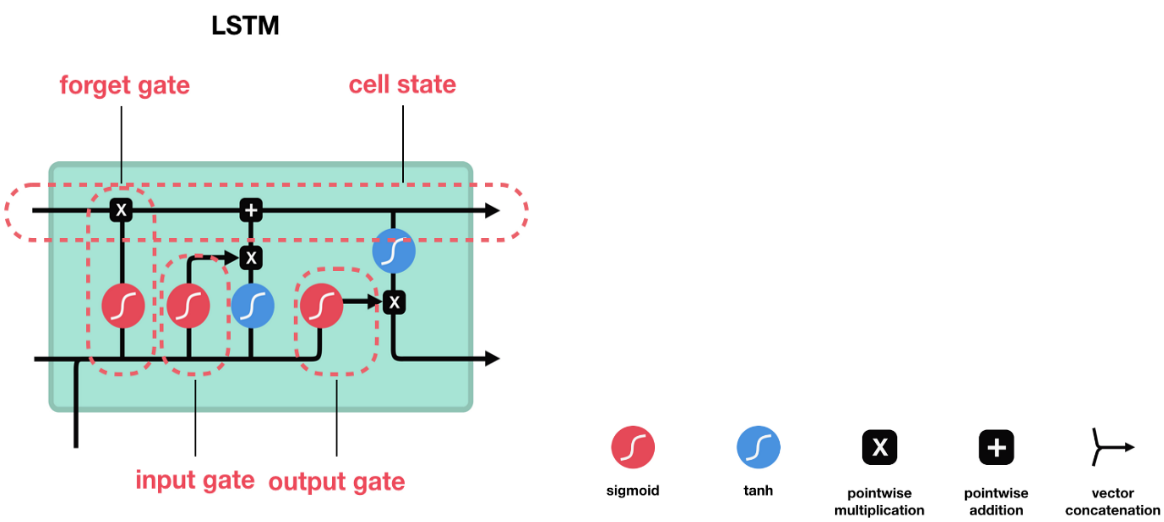
UNIVERSITY OF OKLAHOMA Norman , OK

Contents

Introduction

In this project we attempt to perform portfolio price prediction using neural networks (LSTM) along with optimization using Monte Carlo and Scipy optimizers. This project aims to be tool for an amateur stock trader to use in his/her trading or investment activities. A diversified portfolio of 5 stocks 2 tech companies (Alphabet, Tesla), 2 mutual funds (Vanguard 500, SAP 500) and 1 non-traditional fund (water stocks i.e American Water Works) is used to demonstrate the use of these tools. We will be checking the forecasts of these stocks to identify trends and take calculative decisions whether to invest further based on various factors such as risk, growth, stability and so forth.

This project is divided in two parts. Part 1 focuses of stock price prediction using Neural Networks and Part 2 focuses of portfolio optimization. We use LSTMs for stock forecasting. Long Short-Term Memory (LSTMS) is a neural network that has data prior layers as a 'memory' which allows the model to find the relationships between in intra, inter data and output. They are widely used for sequence prediction problems and have proven to be extremely effective for time series prediction. They can predict an arbitrary number of steps into the future. Their ability to store past information that is important and forget the information that is not is crucial for its success. An LSTM module (or cell) has 5 essential components which allows it to model both long-term and short-term data.



Contents	Description
Cell State	Represents the internal memory of the cell which stores both short term memory and long-term memories
Hidden State	This is output state information calculated w.r.t. current input, previous hidden state and current cell input which you eventually use to predict the future stock market prices
Input Gate	Decides how much information from current input flows to the cell state
Forget Gate	Decides how much information from the current input and the previous cell state flows into the current cell state
Output Gate	Decides how much information from the current cell state flows into the hidden state, so that if needed LSTM can only pick the long-term memories or short-term memories and long-term memories

Portfolio Management is the process of maximizing the return on a portfolio. It involves analysis of different assets, their strengths and weaknesses before deciding about which equities are to be held in a portfolio for balancing the risk and drawing maximum returns. This makes portfolio management a difficult process. However, using optimization techniques such as CAPM and calculation of variability, sharpie ratio and annual returns, one can take calculative risks in portfolio investment. In Part 2, we deal with portfolio analysis using correlation and optimization procedures.

Dataset

Stock price data was from Yahoo Finance. 5 stocks 2 tech companies (Alphabet, Tesla), 2 mutual funds (Vanguard 500, SAP 500) and 1 water stocks i.e American Water Works. 5 yrs data from 2014 till date is taken for analysis. There are multiple variables in the dataset – date, open, high, low, close and Volume

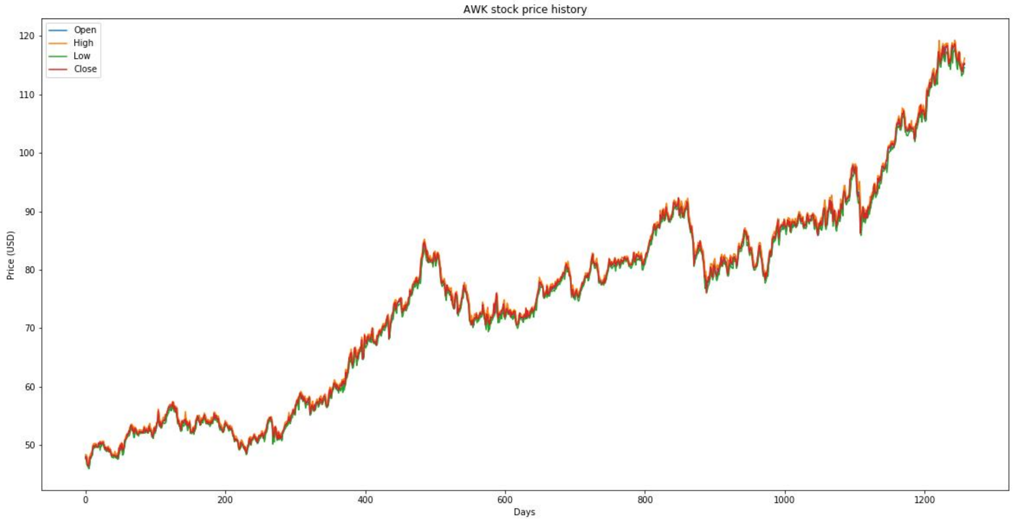
- The columns Open and Close represent the starting and final price at which the stock is traded on a day.
- High, Low represent the maximum, minimum of the share for the day.
- Volume is the number of shares bought or sold in the day

Part 1: Stock Price Prediction using Neural Networks (LSTMS)

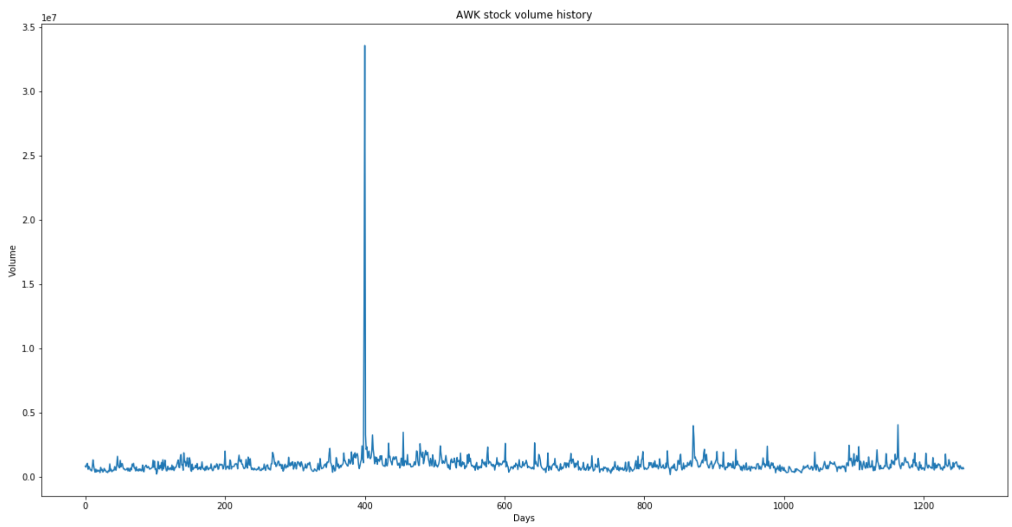
Procedure & Results

The results are only shown for one stock of the 5 for simplicity.

1. Data Visualization and Exploration:



The profit or loss calculation is usually determined by the closing price of a stock for the day, and we will be using that for our prediction. The surge in volumn can be acounted for when anaylsis of sentiments is taking place.



2. Training and Test = 80:20 ratio for Train:Test division of data was done. Test data here indicates a validation set as this is to see if our model perfectly validates are data. We will validate for the last approx. 8 months of the dataset (252 days).

3. Normalizing Data using MinMaxScalar : Normalizing data helps the algorithm in converging i.e. to find local/ global minimum efficiently MinMaxScaler from Sci-kit Learn was used

4. Converting data to time-series and supervised learning problem.
Time Steps define how many units back in time you want your network to see. In our case we will be using 60 as time step i.e. we will retain information of prior 2 months of data to predict next day's price. We will also sort the dataset in ascending order and then

create a separate dataset so that any new feature created does not affect the original data. Transformation into numpy arrays for neural network fit.

5. Creating neural net model (Keras): LSTM was used for this task, a variation of Recurrent Neural Network. 2 LSTM layers finalized with a Dense layer were used.

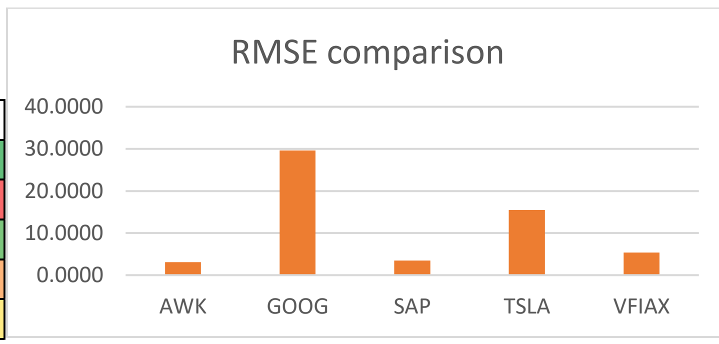
```
# create and fit the LSTM network
model = Sequential()
model.add(LSTM(units=50, return_sequences=True, input_shape=(x_train.shape[1],1)))
model.add(LSTM(units=50))
model.add(Dense(1))
```

6. Training, predicting and visualizing the result

```
model.compile(loss='mean_squared_error', optimizer='adam')
model.fit(x_train, y_train, epochs=1, batch_size=1, verbose=1)
```

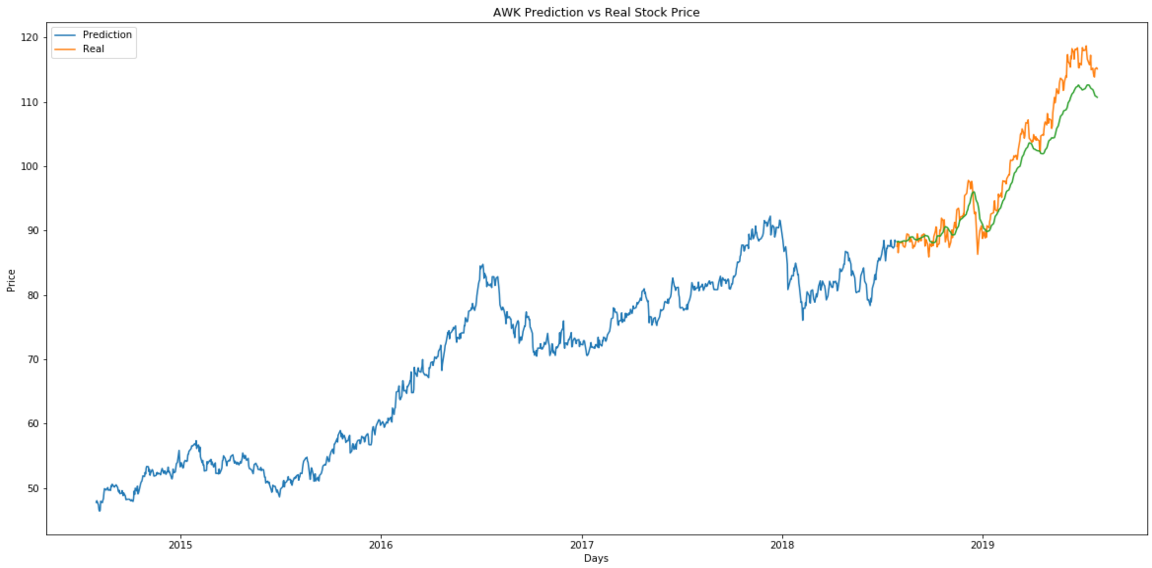
Tuning of hyper-parameters was not required but may be necessary if higher fluctuations are seen in the stocks. The metric being mean squared error through Adam's optimizer helps us get the RMSE values for the prediction. This indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. RMSE is an absolute measure of fit. Lower RMSE value is better.

Company	RMSE
AWK	3.0994
GOOG	29.5894
SAP	3.4691
TSLA	15.5090
VFIAX	5.3367

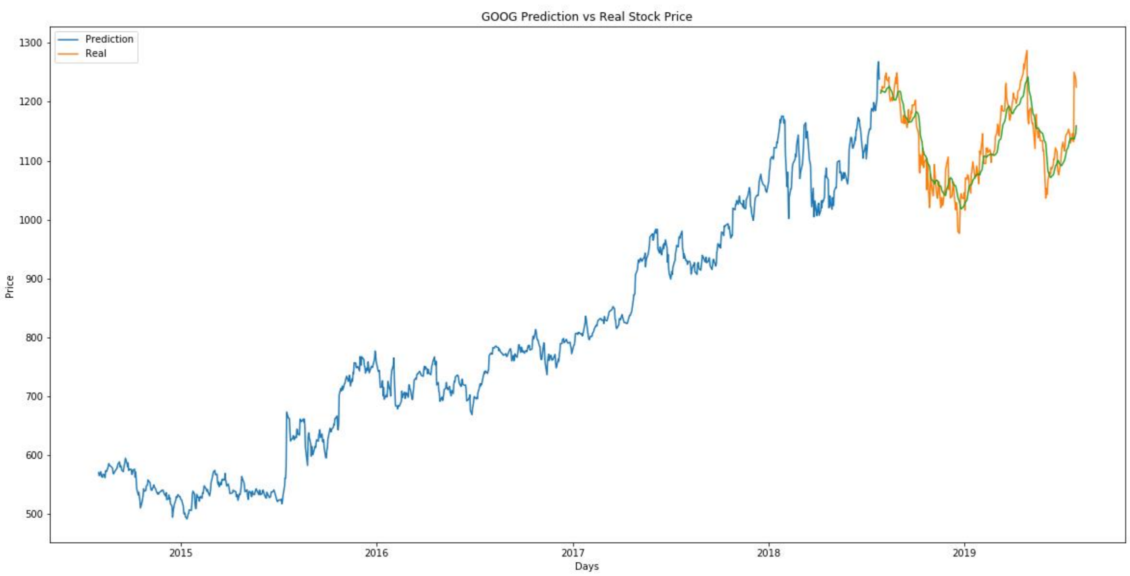


For a more intuitive understanding a plot of the predicted values along with the actual values. American Water works (AWK) has a forecast with highest accuracy as seen in colour grade.

Least RMSE – Americal Water Works

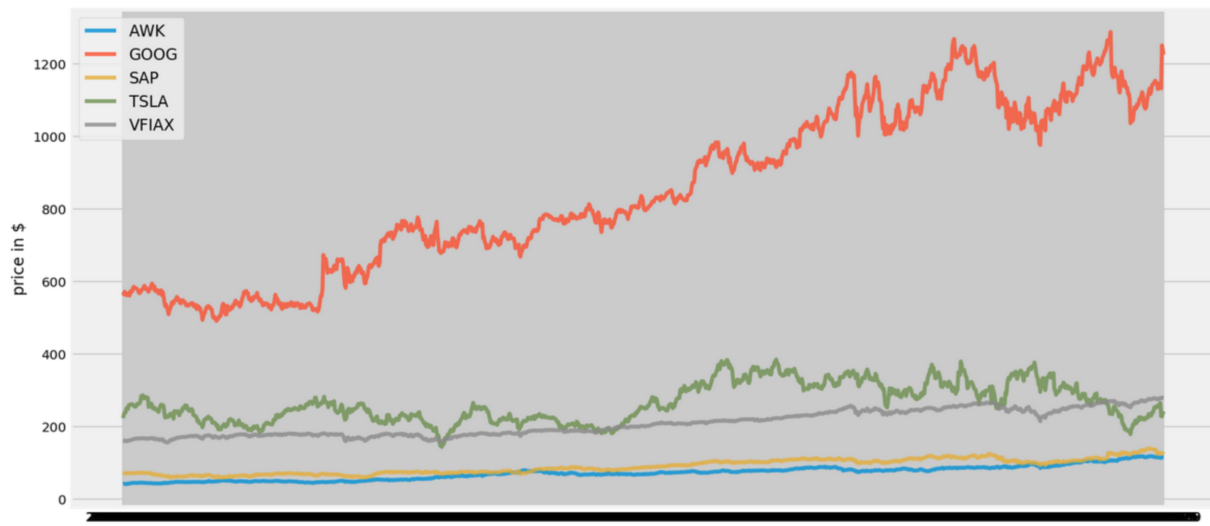


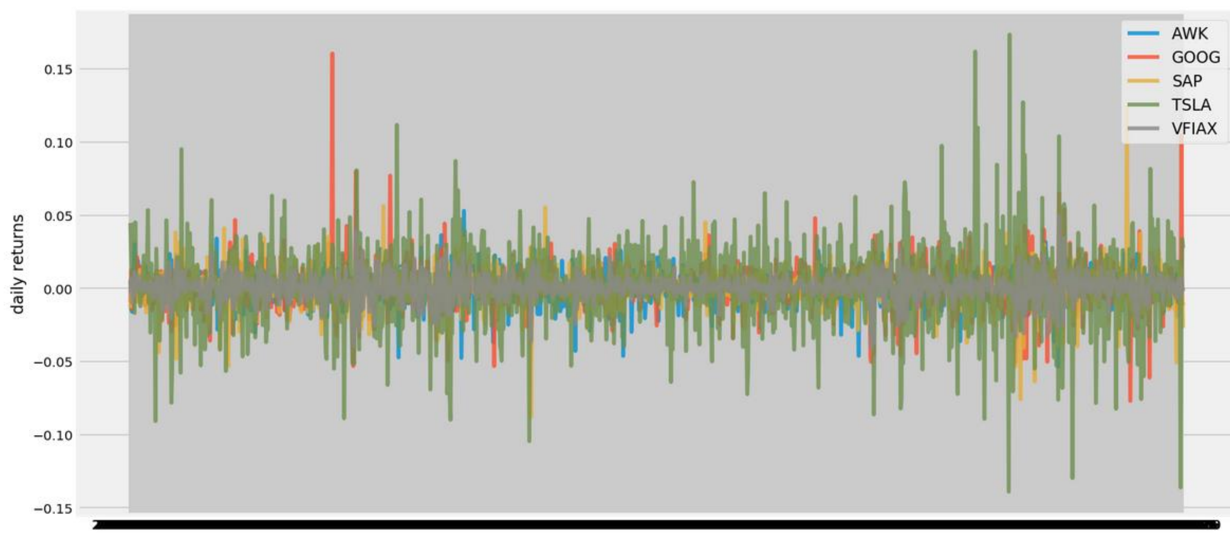
Highest RMSE – Google



Part 2: Portfolio Analysis

Data Trends of Portfolio





Correlation Analysis.

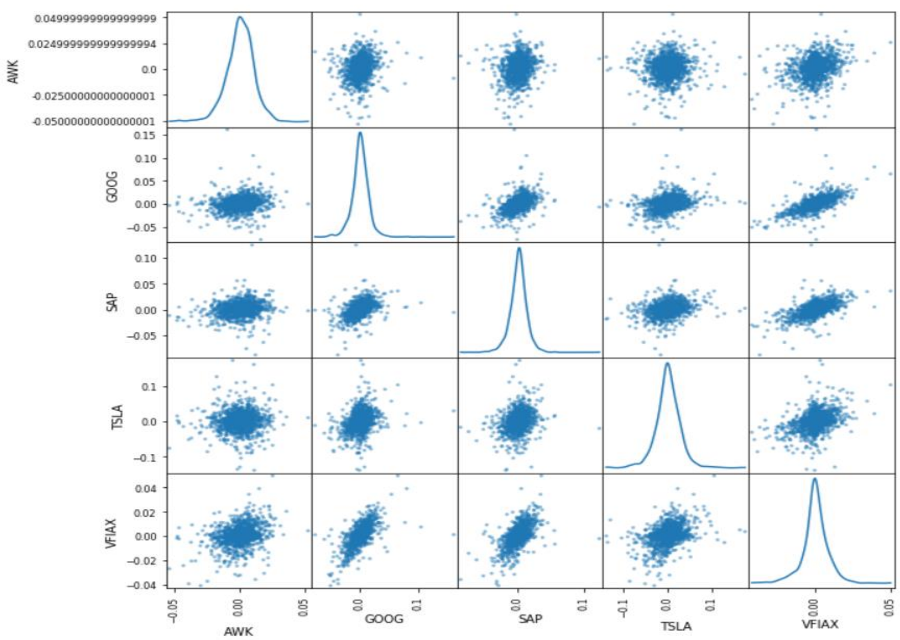
This is a primary stock exploration technique to check company performance with one another and their correlations.

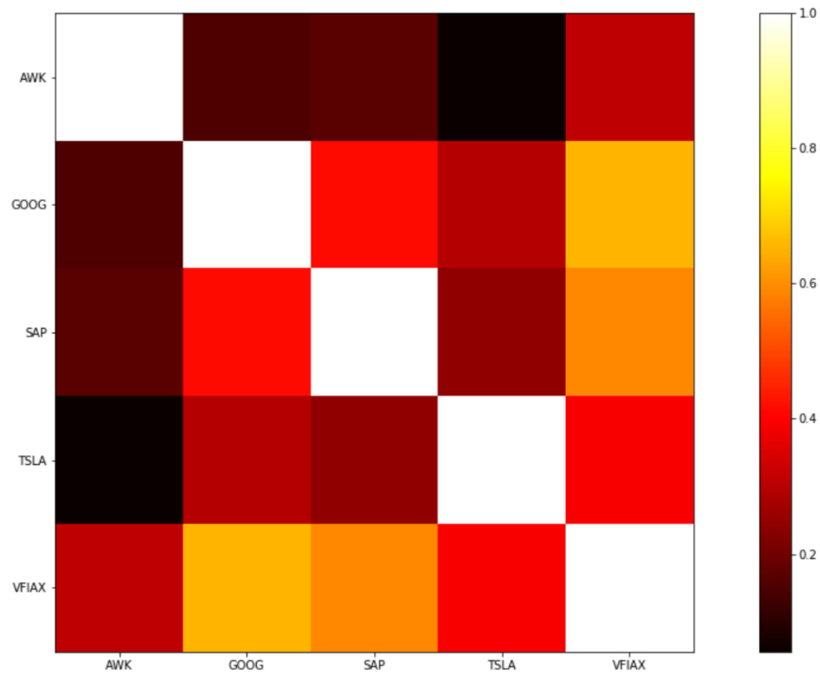
We do this by using **Pandas** dataframe.**pct_change()** function which calculates the percentage change between the current and a prior element. This function by default calculates the percentage change from the immediately previous row.

	AWK	GOOG	SAP	TSLA	VFIAX
AWK	1.000000	0.152723	0.168185	0.055258	0.311942
GOOG	0.152723	1.000000	0.415938	0.296482	0.655016
SAP	0.168185	0.415938	1.000000	0.245221	0.591497
TSLA	0.055258	0.296482	0.245221	1.000000	0.391165
VFIAX	0.311942	0.655016	0.591497	0.391165	1.000000

Through a scatter matrix with all the competitor's data we find the KDE of each stock. Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. KDE will determine if your chart is more normally distributed leaning to the left, center or right. It specifies the high probability for the returns in the long run to be

	Returns
Left	Negative
Center	0
Right	Positive

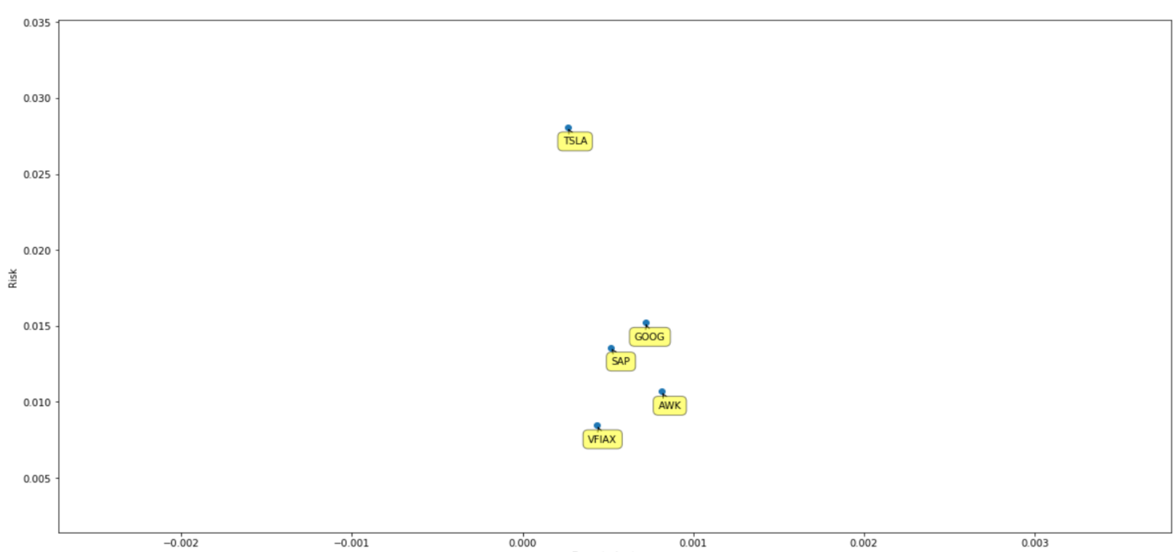




To prove the correlations, we will use heat maps to visualize the correlation ranges among the competing stocks. Lighter colors are directly proportional to higher correlation. As such through the above graph we see that most of the distributions among stocks which approximately positive correlations, except for combinations such as [TSLA, AWK] & [AWK, GOOG] both of which while not negative have almost no correlation. While not to indicate causality, this shows trends of competing stocks and we will later link and check it with the optimized portfolio.

Optimization

Stocks Returns rate and Risk



The chart of risk and return comparisons for competing stocks helps us identify naively which stocks we should buy and sell. We should aim at minimizing the risk and maximize returns. TSLA having the highest risk can be considered for selling/not buying while the rest can consider buying/holding. This shows us the risk-return tolerance in the portfolio.

Methodology

Portfolio weights that would match the "optimal" means the portfolio with the highest Sharpe ratio, also known as the "mean-variance optimal" portfolio. Aim is to optimise a portfolio of stocks based on minimising different cost functions (i.e. the negative Sharpe ratio, the variance and the Value at Risk). Among the various portfolio optimization methods, the following methods are shown:

1. Monte Carlo, "brute force" style optimization:
2. Scipy's "optimize" function for "minimizing (or maximizing) objective functions, possibly subject to constraints"

We try to discover the optimal weights by creating many random portfolios, all with varying randomly weights, calculating and recording the Sharpe ratio of each of these and then finally extracting the details corresponding to the result with the highest value. We will always experience some discrepancies to replicate the exact weights we are searching for and we can never get it correct but rather just a ball point.

$$\boldsymbol{Sharpe\ ratio} = \frac{\bar{\boldsymbol{r}}_{\boldsymbol{p}} - \boldsymbol{r}_{\boldsymbol{f}}}{\boldsymbol{\sigma}_{\boldsymbol{p}}}$$

$$\bar{r}_p = expected\ return\ of\ the\ portfolio\ or\ investment$$
$$r_f = risk\ free\ rate$$
$$\sigma_p = standard\ deviation\ of\ portfolio\ returns$$

Steps:

1. Calculate the mean returns and co-variance matrix of our list of stocks. Set risk-free rate is required for the calculation of the Sharpe ratio that should be provided as an annualised rate.
2. Calculate the annualised return, annualised standard deviation and annualised Sharpe ratio of a portfolio through a function

The portfolio of 5 stocks was augmented with 100,000 simulated portfolios to produce our results. These results are then be plotted and both the "optimal" portfolio with the highest recorded Sharpe ratio and the "minimum variance portfolio" are displayed. The "minimum variance portfolio is the portfolio with the lowest recorded variance (standard deviation or "volatility")

Results:

The stock weightings of the portfolios, along with the annualised return, annualised standard deviation and annualised Sharpe ratio. These are shown below firstly for the maximum Sharpe portfolio, and then for the minimum variance portfolio in the Monte Carlo Method while the Scipy Optimize function only works by minimizing negative sharpie ratio and doesn't return the ration values.

Method	Portfolio Type	AWK	GOOG	SAP	TSLA	VFIAX	ret	stdev	sharpe
Monte Carlo	Max Sharpie	0.678225	0.142191	0.03301	0.003693	0.14288	0.185189	0.137841	1.343495
	Min Var	0.392604	0.009397	0.012022	0.013198	0.57278	0.147822	0.120829	1.223391
Scipy Optimize	Max Sharpie	0.66	0.2	0.09	0	0.05			

American Water Works has the highest weight in the optimized portfolio which lends to its credibility as against that of TSLA.

Conclusions.

As seen through the results above, the user of these tools gets a fair picture of his/her investments in the stock market. Through the above examples we see that a non-traditional stock i.e water stock (AWK) has a higher weight in the optimized portfolio and that the Vangaurd 500 mutual funds can also be a good investment choice. While the results state that investment in Tesla can be a bad in the long run, one must also consider sentiment analysis which has not been included in this project. Furthermore, the LSTM models can also help a client analyze his portfolio for trading practices, however with caution as even till date, no machine learning model can correctly predict all the price movements all the time. With the other tools at disposal the trader can also include this for further validity into his practices.

As such the prediction and optimization of a portfolio of 5 stocks helps the user get insights to take calculative decisions whether to invest further based on various factors such as risk, growth, stability. Having had said that, one must always follow his instincts upon reading of the various information available and only then take a stand.

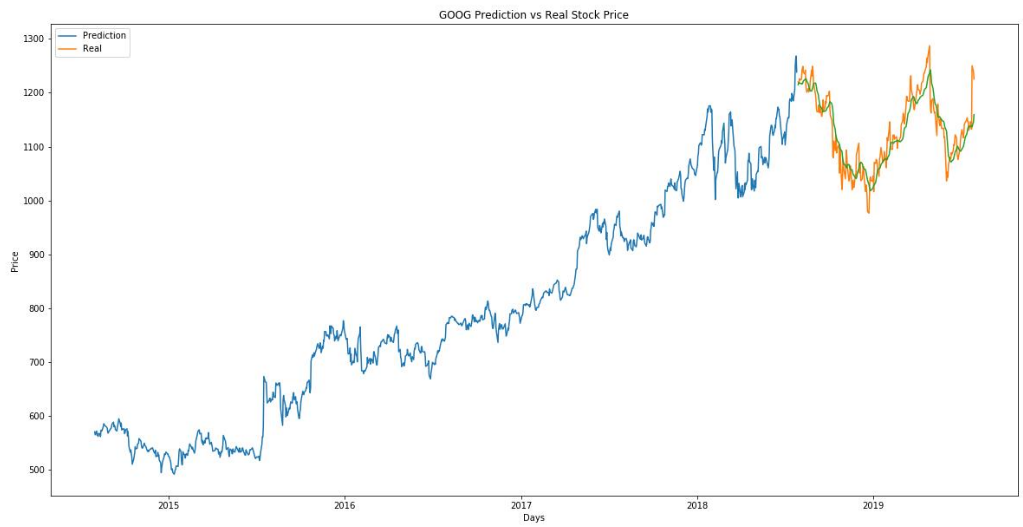
References

1. https://towardsdatascience.com/predicting-stock-price-with-lstm-13af86a74944
2. https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/
3. https://medium.com/sfu-big-data/predicting-stable-portfolios-using-machine-learning-f2e27d6dbbec
4. https://pythonforfinance.net/2019/07/02/investment-portfolio-optimisation-with-python-revisited/
5. https://www.datacamp.com/community/tutorials/lstm-python-stock-market?source=post_page---------------------------
6. https://towardsdatascience.com/in-12-minutes-stocks-analysis-with-pandas-and-scikit-learn-a8d8a7b50ee7
7. https://towardsdatascience.com/efficient-frontier-portfolio-optimisation-in-python-e7844051e7f

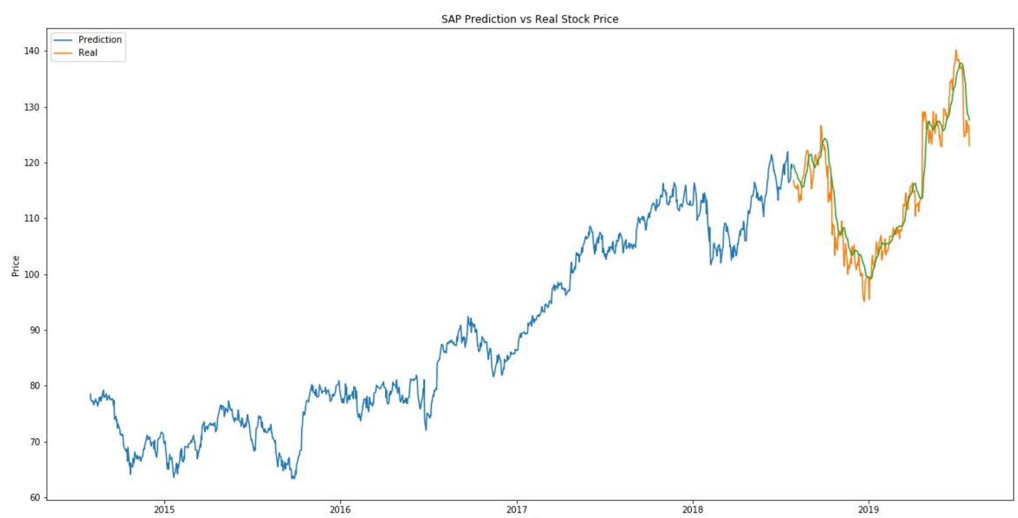
Appendix

Prediction vs Actual

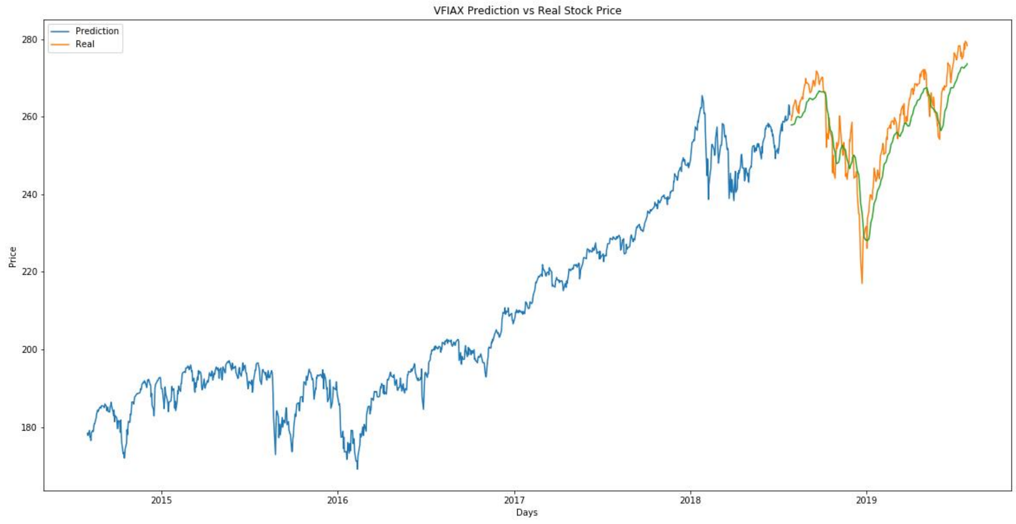
1. Google



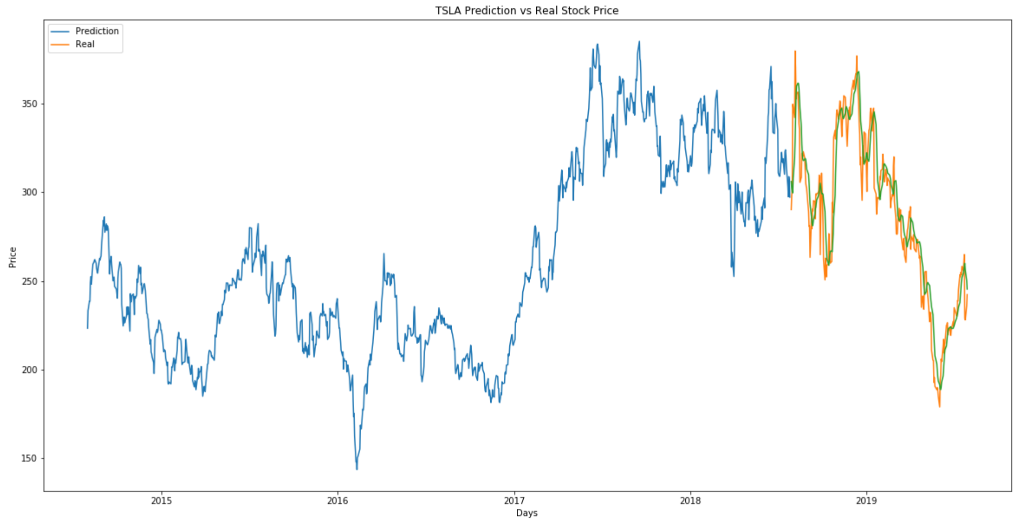
2. SAP500



3. VFIAX

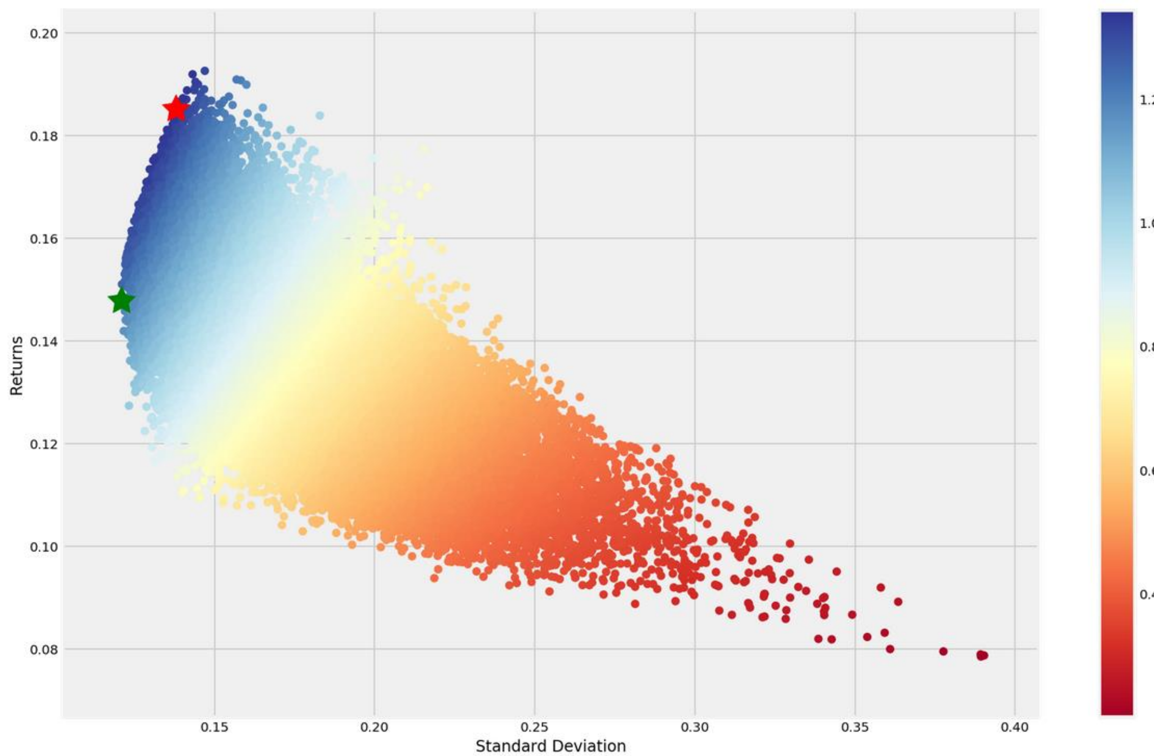


4. TSLA



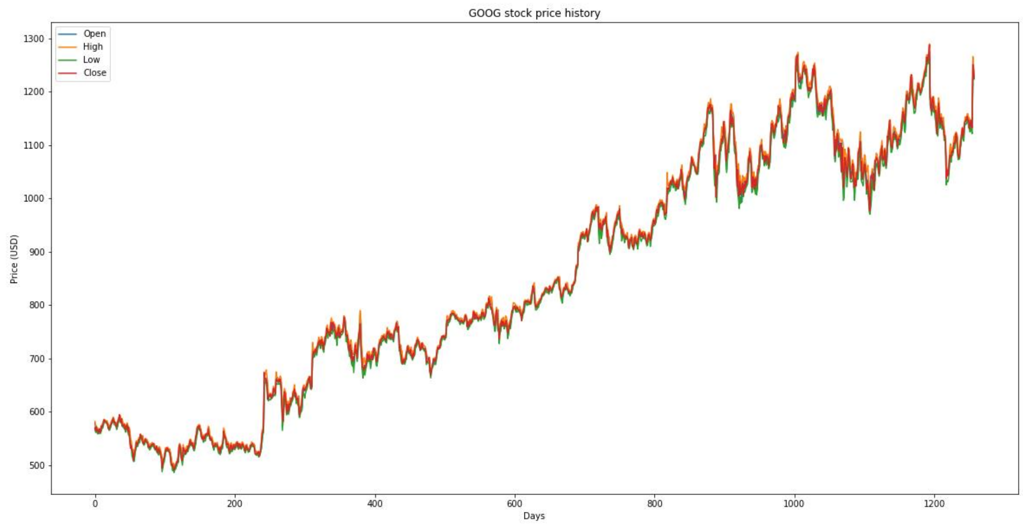
EFFICIENT FRONTIER

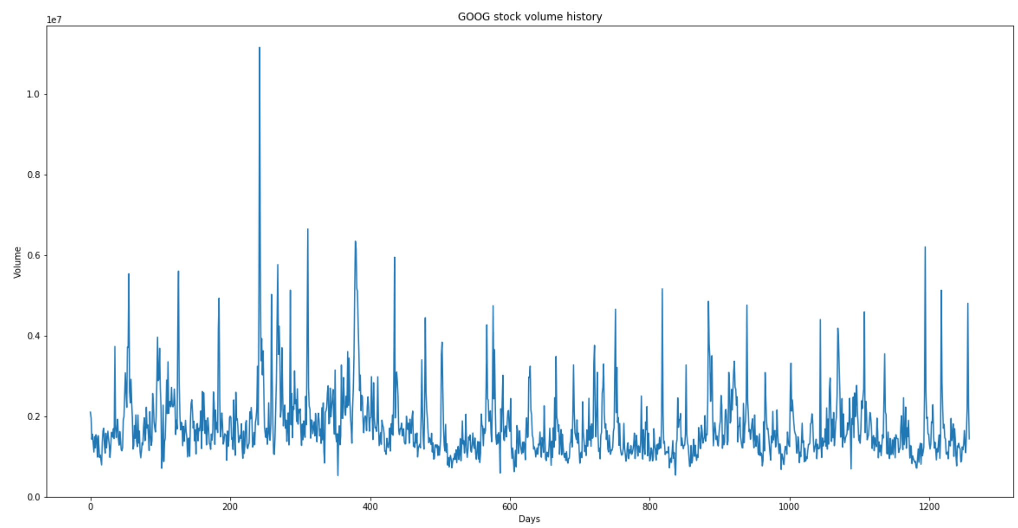
1. Monte Carlo Model



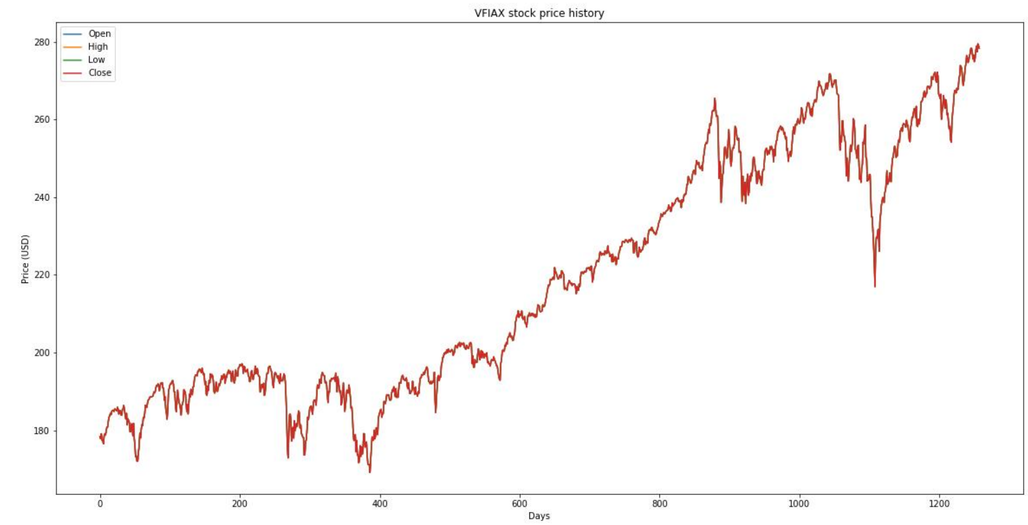
VISUALIZATION

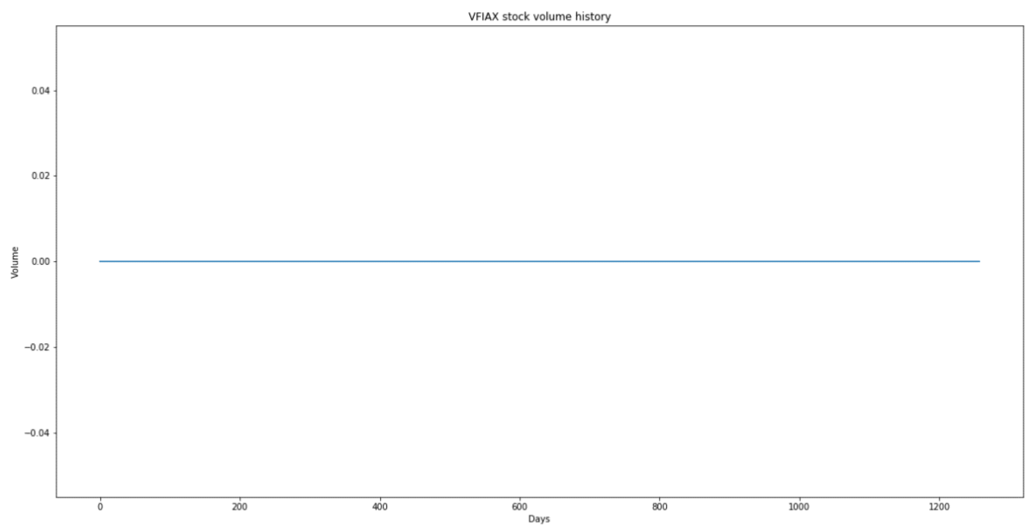
1. GOOG


GOOG stock price history
Open
High
Low
Close
Price (USD)
Days

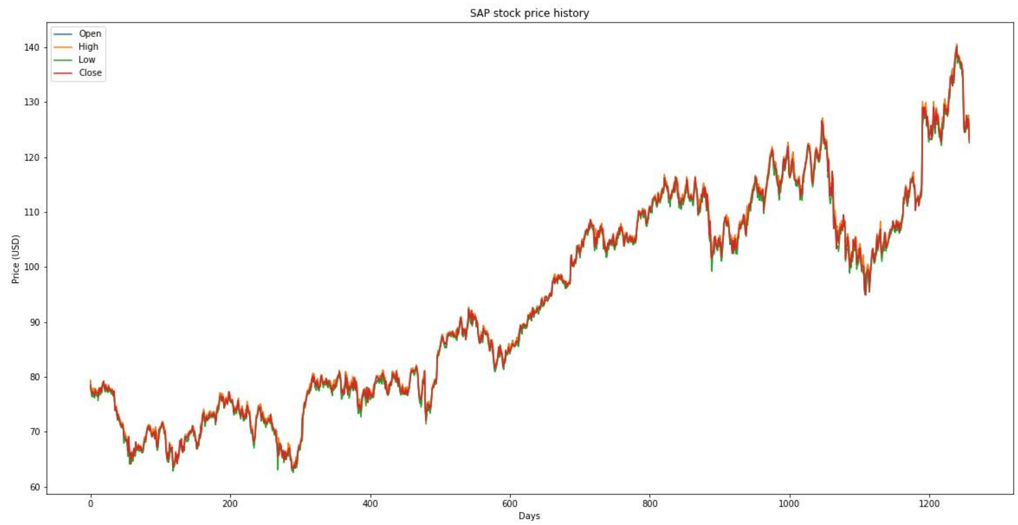

GOOG stock volume history
1e7
Volume
Days

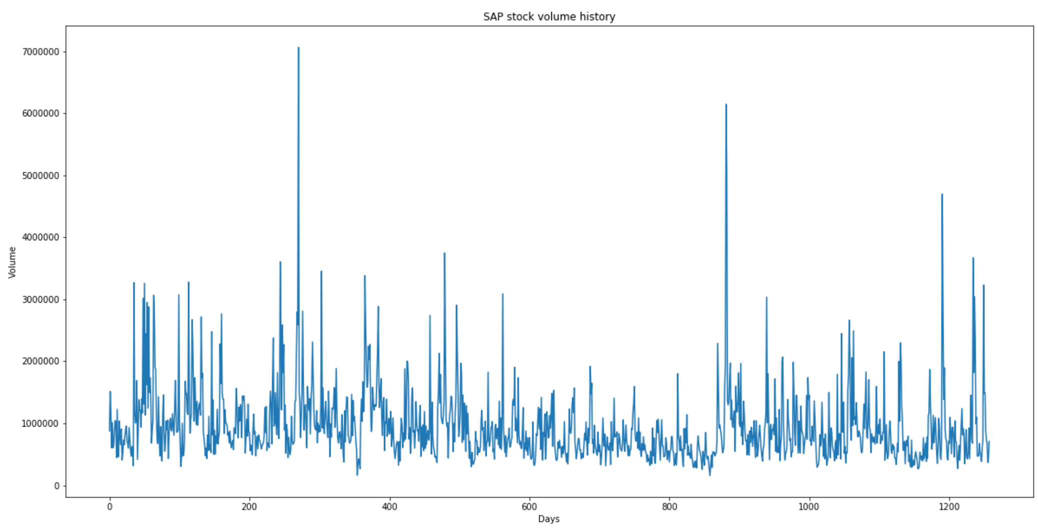
2. VFIX


VFIAX stock price history
Open
High
Low
Close
Price (USD)
Days

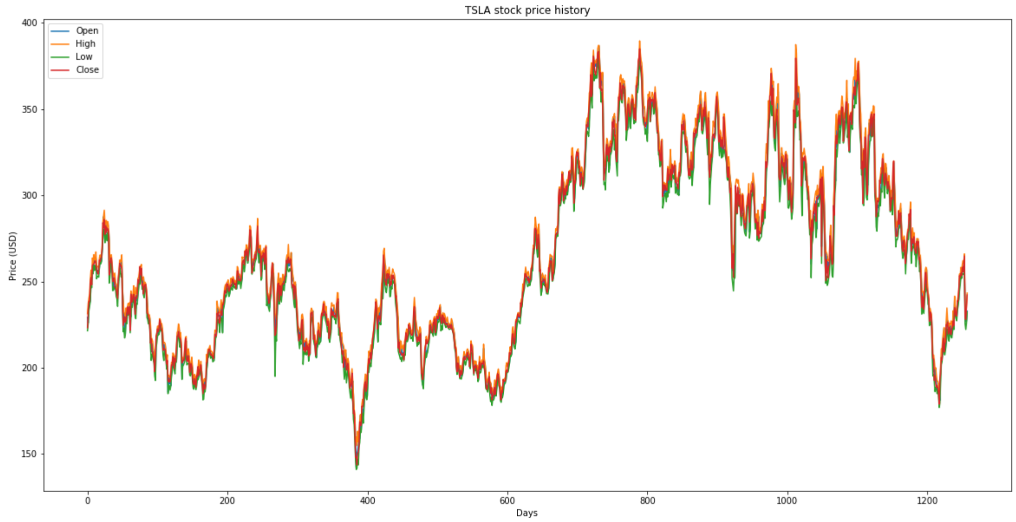

VFIAX stock volume history
Volume
Days

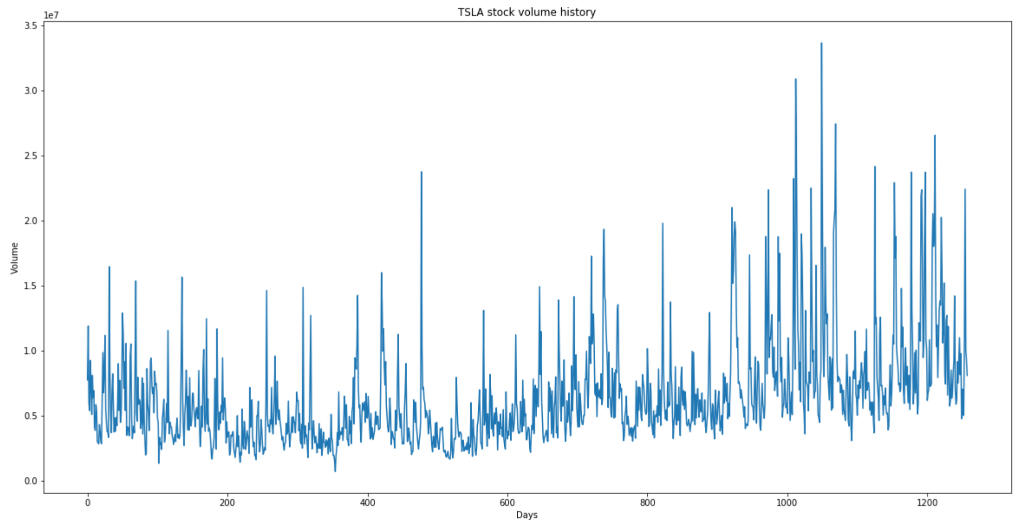
3. SAP


SAP stock price history
Open
High
Low
Close
Price (USD)
Days
140
130
120
110
100
90
80
70
60
0
200
400
600
800
1000
1200


SAP stock volume history
Volume
Days
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
200
400
600
800
1000
1200

4. TSLA


TSLA stock price history
Open
High
Low
Close
Price (USD)
Days


TSLA stock volume history
1e7
Volume
Days