Prepared By:
Shaishav Shah

Introduction to Healthcare Stock (Procter & Gamble Co)

Machine Learning & Python in Finance

Why P&G?

2ND LARGEST MARKET CAP IN HEALTHCARE WITH $314 BILLION

DIVIDEND YIELD PER SHARE: $2.39

PRICE-EARNINGS RATIO: 83.13

52 WEEK HIGH: $127 (ACHIEVED YESTERDAY)

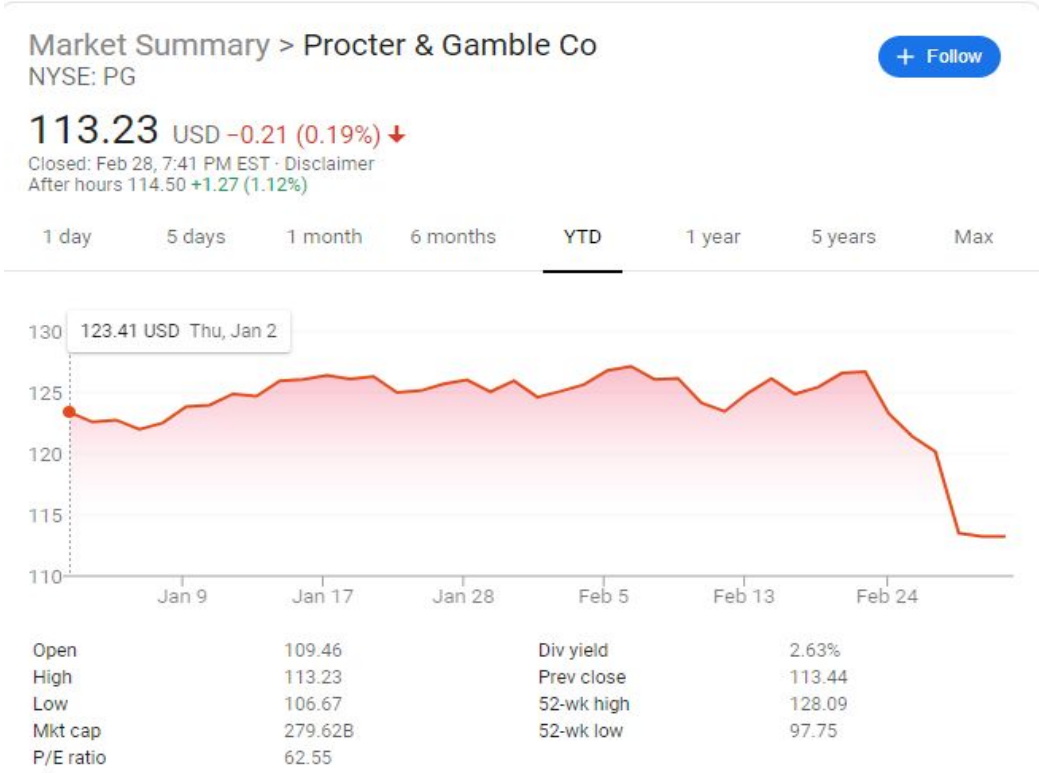
P&G (39.47%) OUTPERFORMED S&P 500 (24.19%) IN A YEAR

QUARTERLY EARNING GRAPH

91% ANALYST RATING FOR BUY AND HOLD

PG Stock- Past Return Index

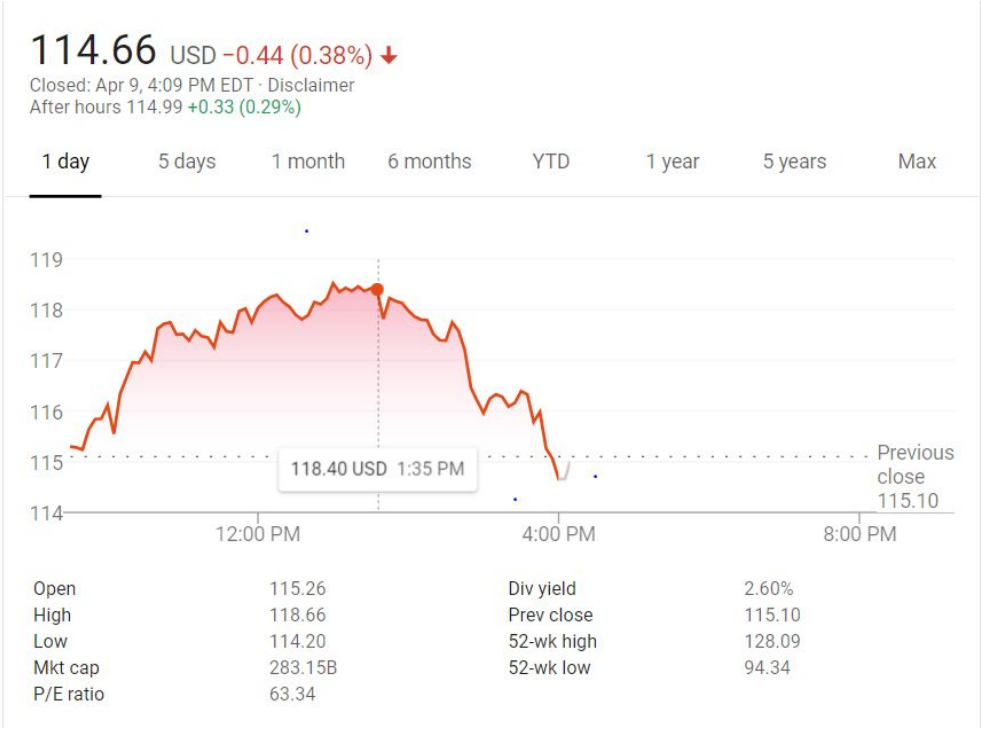
- Year to date stock
- Current P/E ratio higher than SP500-HC
- High Revenue generating company regardless of current financial strain



PG Stock Present Return Index

- Even with the current COVID-19 situation we can see that the company was expected to post earnings of $1.24 per share and it actually produced earnings of $1.37 per share, delivering a surprise of 10.48%.

- Even with the drop in the price index by 38% the P&G market is still stable and is considered as top 10 company in Healthcare sector



DATA

To begin with we are considering P&G stock prices for at most 20 years, ranging from 2000 to 2019

Data Definition:

- Date: Date of trading day
- High: Highest Price of Stock on that day
- Low: Lowest Price of Stock on that day
- Open: Opening Price of Stock on that day
- Close: Closing Price of Stock on that day
- Volume: Number of shares or contracts traded in a security or an entire market during a given period of time
- Adj Close: Adjusted closing price amends a stock's closing price to accurately reflect that stock's value after accounting for any corporate actions

S&P-500 and S&P-500 HealthCare Sector

S&P 500- The S&P 500, or just the S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States. It is one of the most commonly followed equity indices, and many consider it to be one of the best representations of the U.S. stock market.

To see how P&G stock prices correlate with S&P 500 Health Care Sector data

- HC-Open- Cumulative opening price of all HealthCare stocks on that day.
- HC-Close- Cumulative closing price of all HealthCare stocks on that day.
- HC- High- Cumulative high price of all the HealthCare stocks on that day.
- HC-lOW- Cumulative low price of all the HealthCare stocks on that day.

To see how P&G Stock prices correlate with S&P 500 Data

- S&P Open- Cumulative opening price of all S&P 500 stocks on that day.
- S&P Close - Cumulative closing price of all S&P 500 stocks on that day.
- S&P High- Cumulative high price of all the S&P 500 stocks on that day.
- S&P Low- Cumulative low price of all the S&P 500 stocks on that day.

Fama French Factor Model

Fama French Model is an asset pricing model that expands on the capital asset pricing model (CAPM) by adding size risk and value risk factors to the market risk factor in CAPM.

- **Mkt-RF**: MKTRF (or Rm-Rf) is the excess return on the market. It is calculated as the value-weight return on all NYSE, AMEX, and NASDAQ stocks (from CRSP) minus the one-month Treasury bill rate (from Ibbotson Associates)
- **SMB**: Small Minus Big (S M B) is the average return on the small portfolios minus the average return on the big portfolios. It is basically the portfolio returns
- **HML**: High Minus Low (H M L) is the average return on the value portfolios minus the average return on the growth portfolios
- **RF**: This is just Risk Free interest you can get(Like treasure bills or say basic checking account returns as an example)
- **RMW**: Difference between the returns of firms with robust (high) and weak (low) operating profitability.(Probability factor)

ADS Index

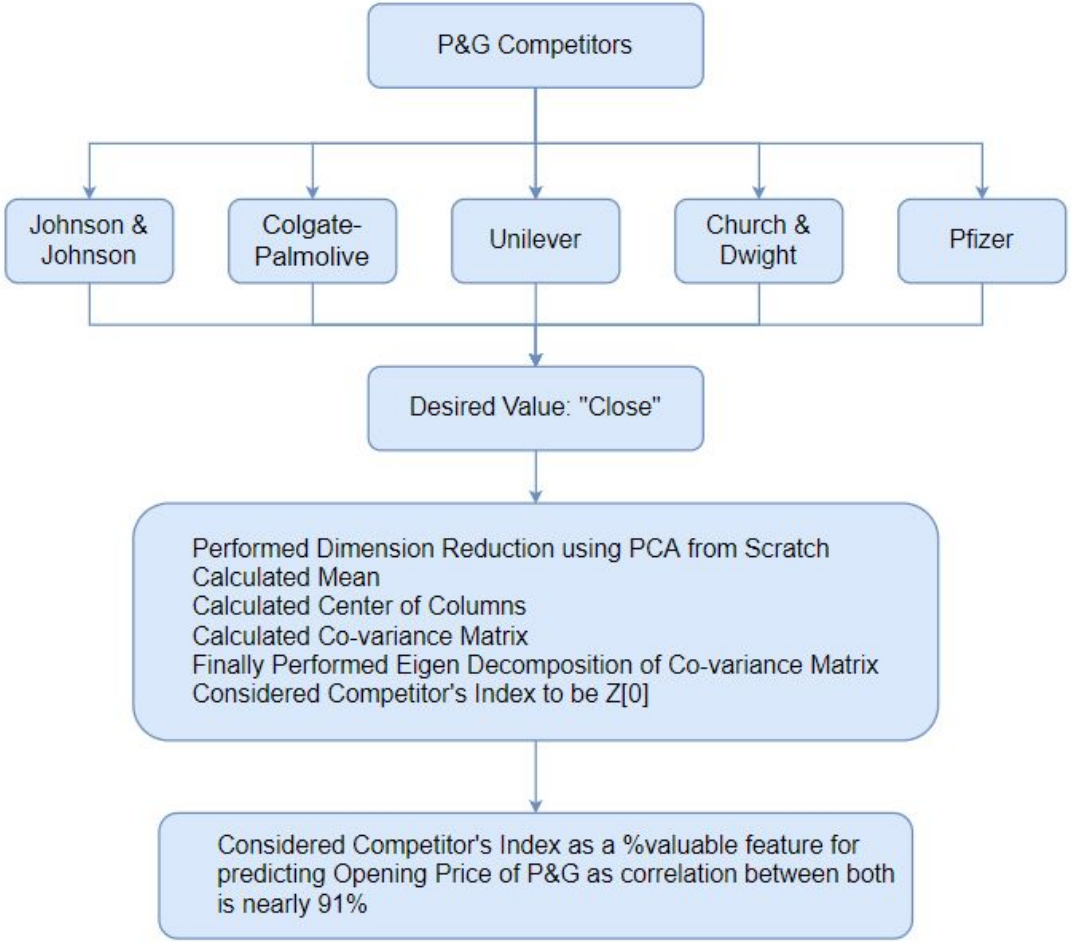
- ADS-Index-The Aruoba-Diebold-Scotti business conditions index is designed to track real business conditions at high observation frequency. The average value of the ADS index is zero.

- Progressively bigger positive values indicate progressively better-than-average conditions, whereas progressively more negative values indicate progressively worse-than-average conditions. The ADS index may be used to compare business conditions at different times.

P&G Competitors Data

We considered these companies because firstly they belong to the same business sector i.e HealthCare and are mainly into consumer goods

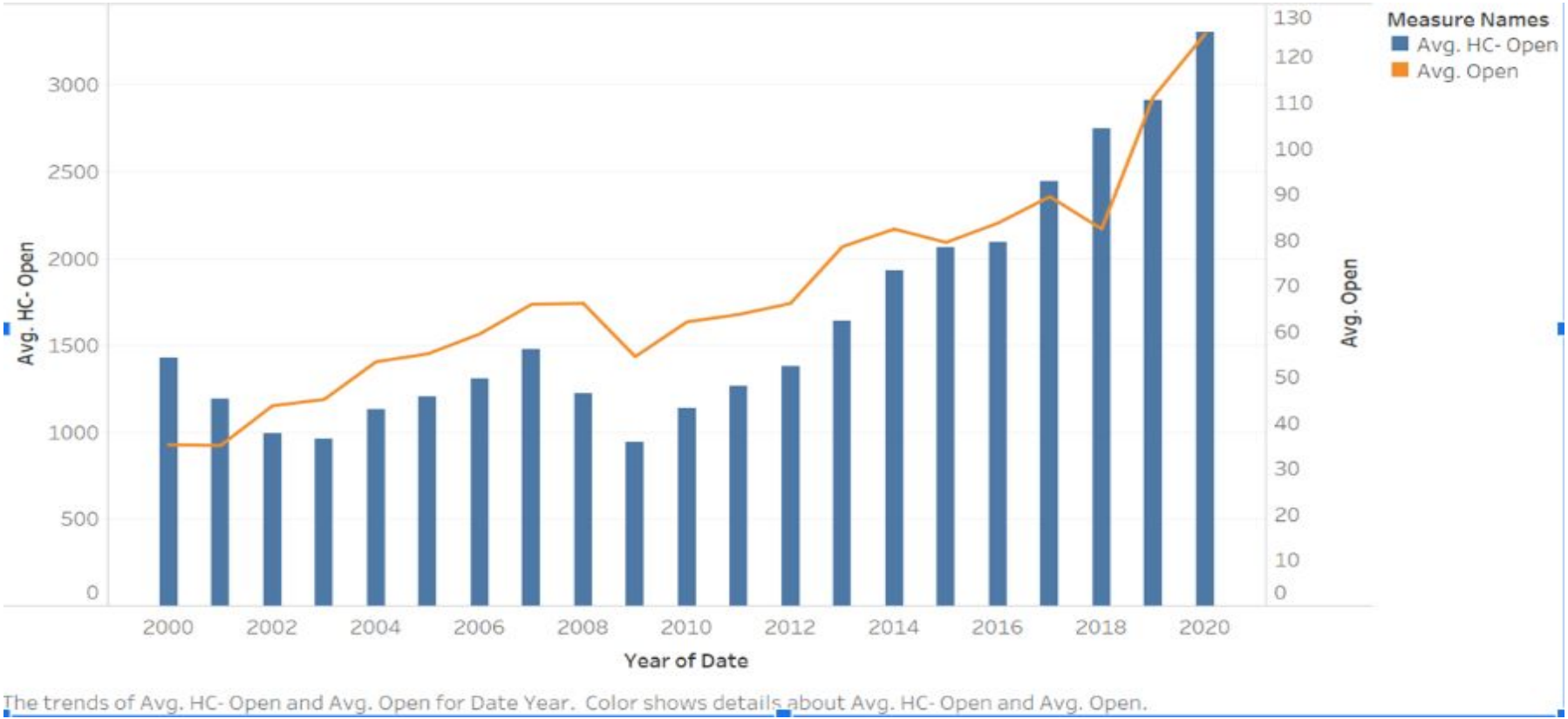
- Procter & Gamble is a very well-known consumer products company, owning major brands like Crest, Gillette, Pampers, and Tide.
- Major competitors for P&G include Colgate-Palmolive, Church and Dwight, and Unilever.
- Nearly two-thirds of P&G's revenues are generated from developed markets, while Unilever and Johnson & Johnson gets the majority of its revenues from faster-growing emerging markets.
- CL is much smaller than PG, though — Colgate has amassed $62 billion in market capitalization vs. $220 billion for Procter & Gamble. However, I think the smaller size is beneficial to Colgate for many of the same reasons large size hurts Procter & Gamble: oversight and control of the business by management.
- Procter & Gamble (PG +0.5%) is in talks with Pfizer (PFE +0.3%) over a purchase of the company's consumer products business or a joint venture arrangement as both belong to the same market domain


P&G Competitors
Johnson & Johnson
Colgate-Palmolive
Unilever
Church & Dwight
Pfizer
Desired Value: "Close"
Performed Dimension Reduction using PCA from Scratch
Calculated Mean
Calculated Center of Columns
Calculated Co-variance Matrix
Finally Performed Eigen Decomposition of Co-variance Matrix
Considered Competitor's Index to be Z[0]
Considered Competitor's Index as a %valuable feature for predicting Opening Price of P&G as correlation between both is nearly 91%

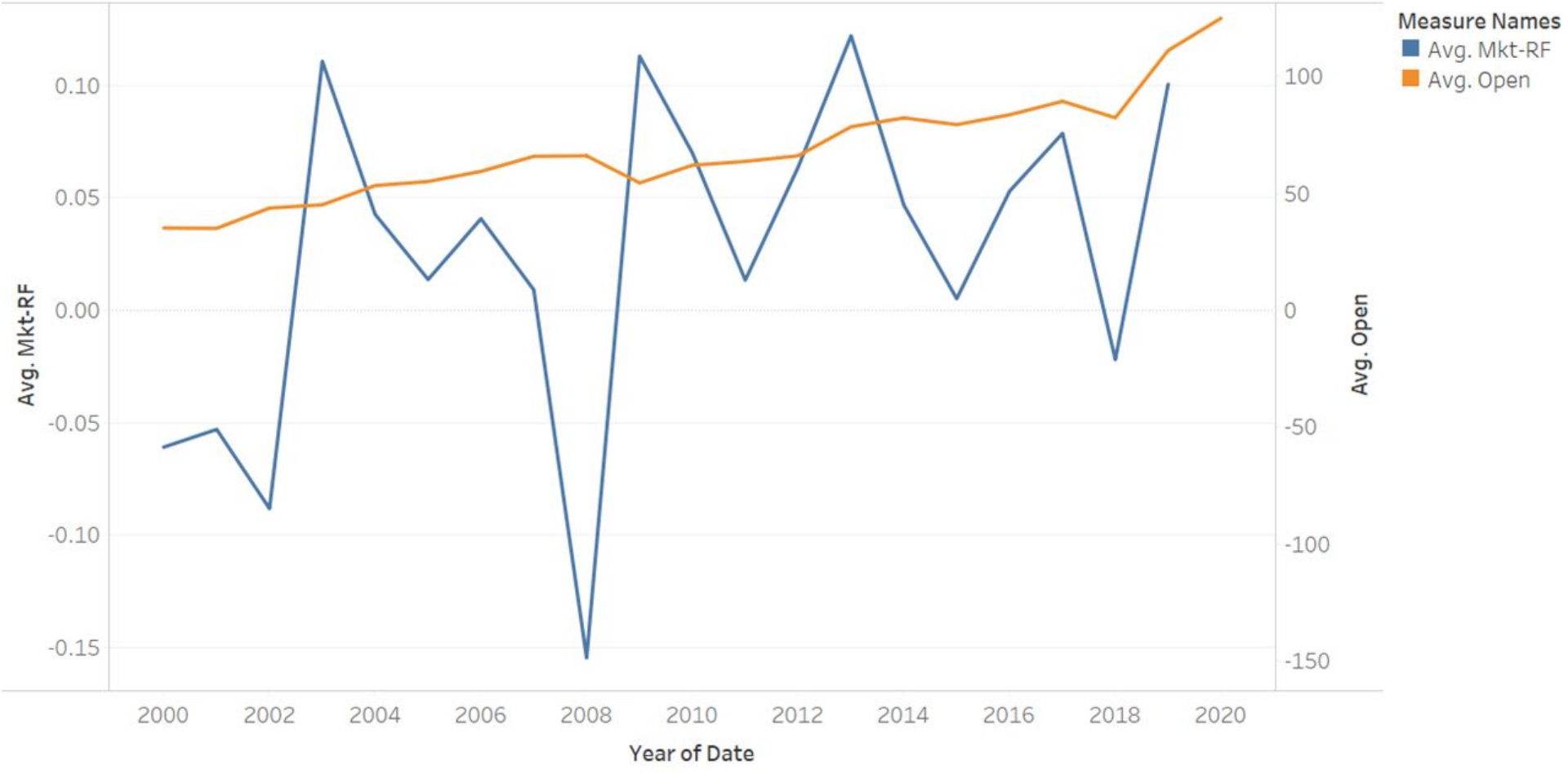

PF Open and Competitors_Index with Time
150
100
50
0
-50
-100
1/3/2000
1/3/2001
1/3/2002
1/3/2003
1/3/2004
1/3/2005
1/3/2006
1/3/2007
1/3/2008
1/3/2009
1/3/2010
1/3/2011
1/3/2012
1/3/2013
1/3/2014
1/3/2015
1/3/2016
1/3/2017
1/3/2018
1/3/2019
Open
Competitors_Index

EXPLORATORY DATA ANALYSIS

Graph Representing AVG Open of S&p 500 with AVG Open of P&G Stock



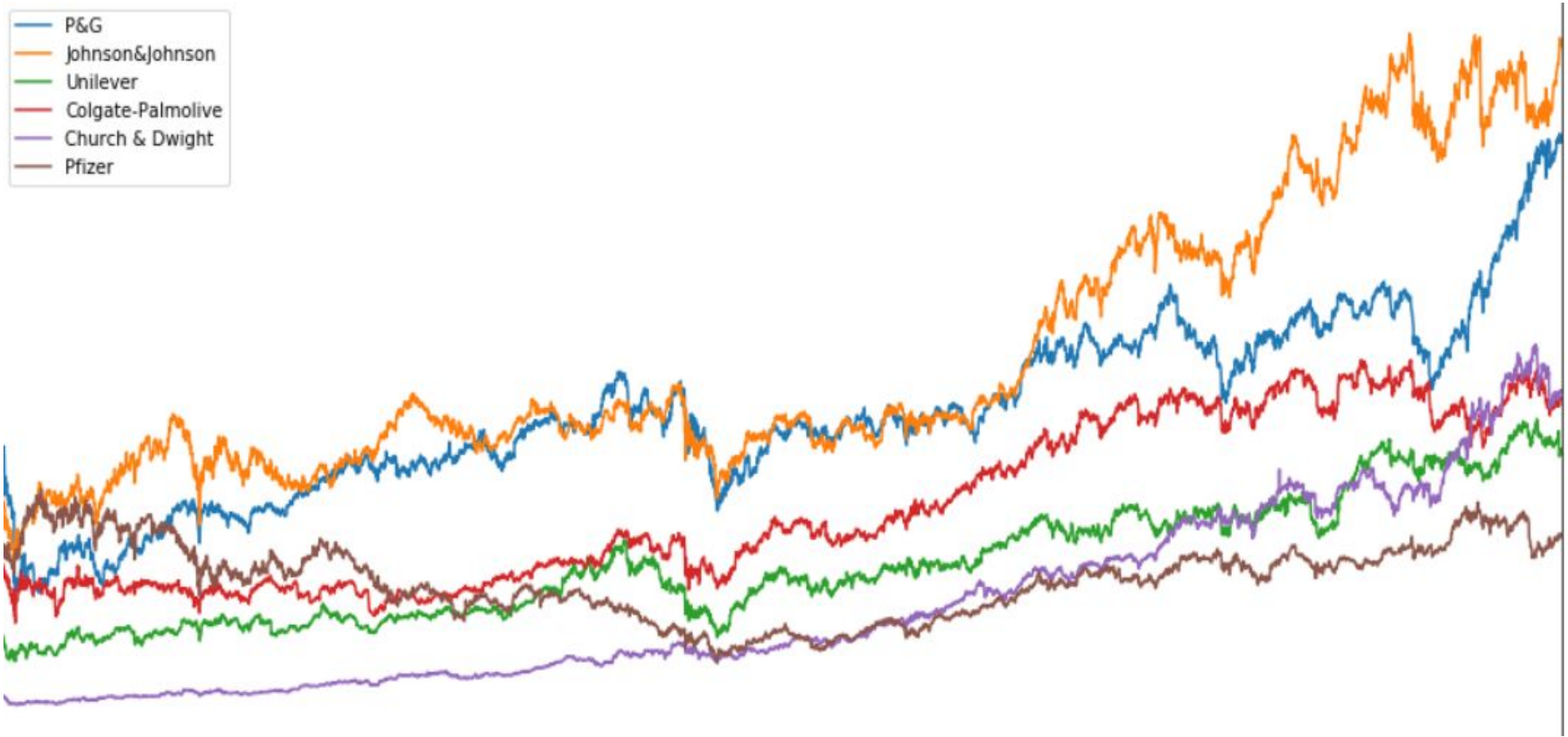
Max Correlation Between Mkt-Rf and P&G Open



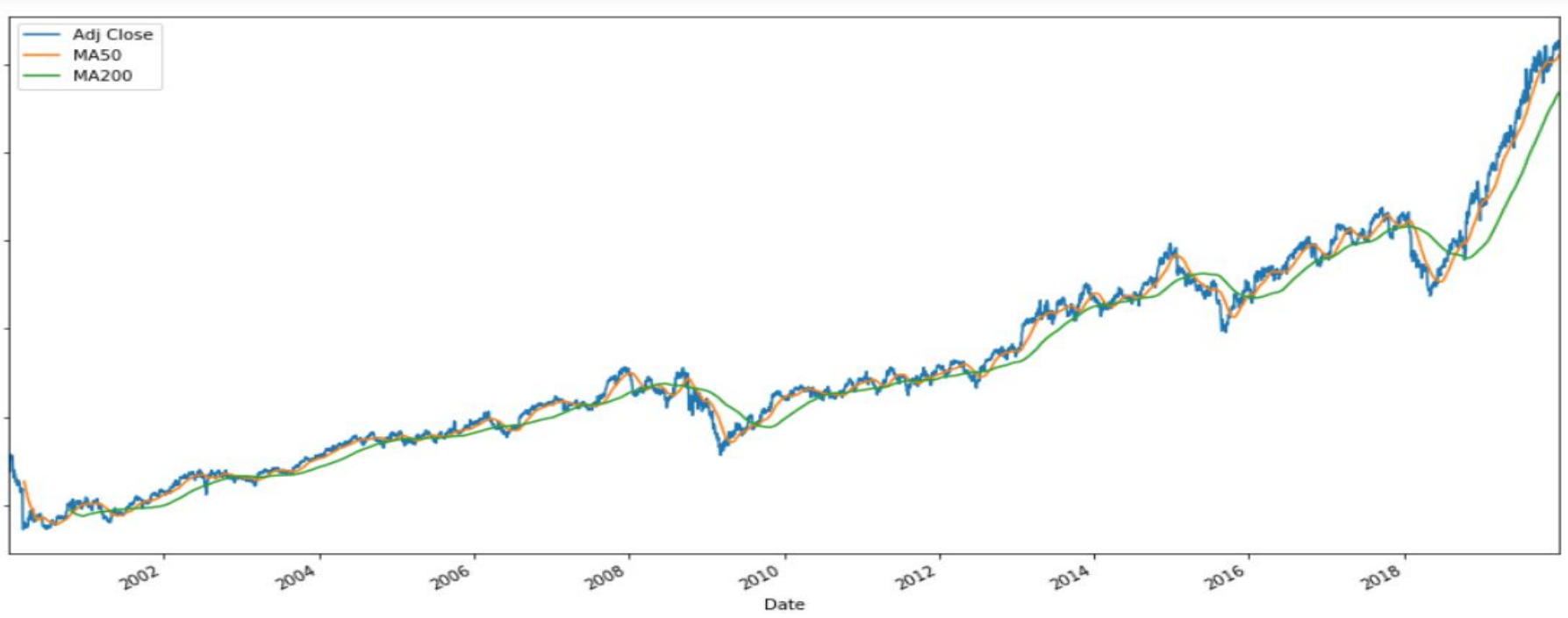
Least Correlation between SMB & p&G Open

The trends of Avg. Open and Avg. RMW for Date Year. Color shows details about Avg. Open and Avg. RMW.

Graph Representing the Opening Price of P&G with their Competitors

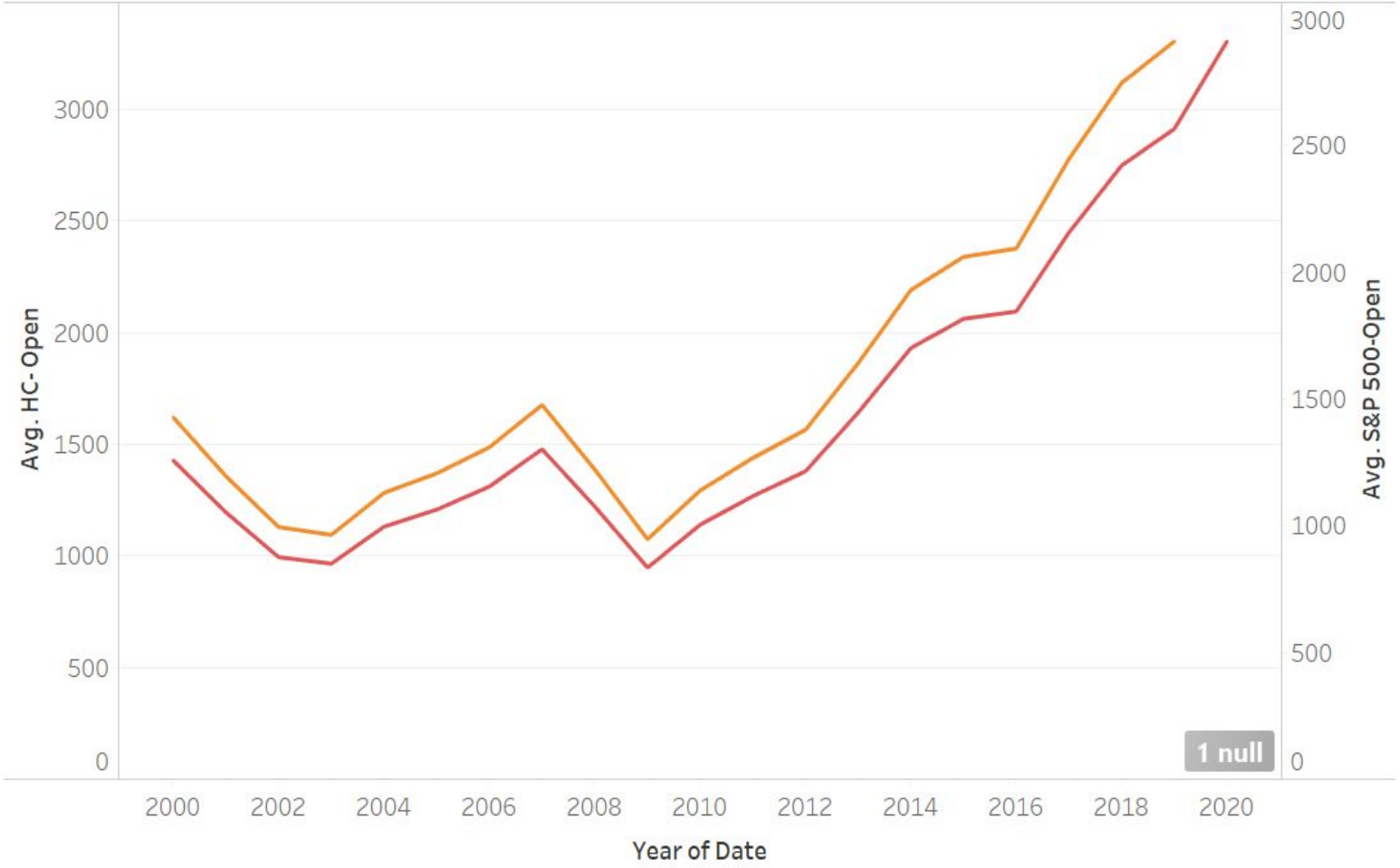


Moving average- between the rolling period of 50 and 200 days



CandleStick Visualization

Avg Open of S&P 500 HealthCare Sector with AVG Open of S&P 500



Measure Names

- Avg. HC- Open
- Avg. S&P 500-Open

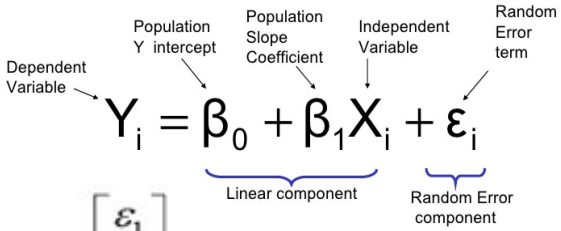
Machine Learning Models and Strategies

- Linear Regression
- Linear Regression with PCA
- Random Forest Algorithm
- Time Series Forecasting ARIMA model
- Pairs Trading Algorithm
- Garch Model
- Long Short Term Memory (RNN)

Linear Regression

Definition: Linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables

Equations:

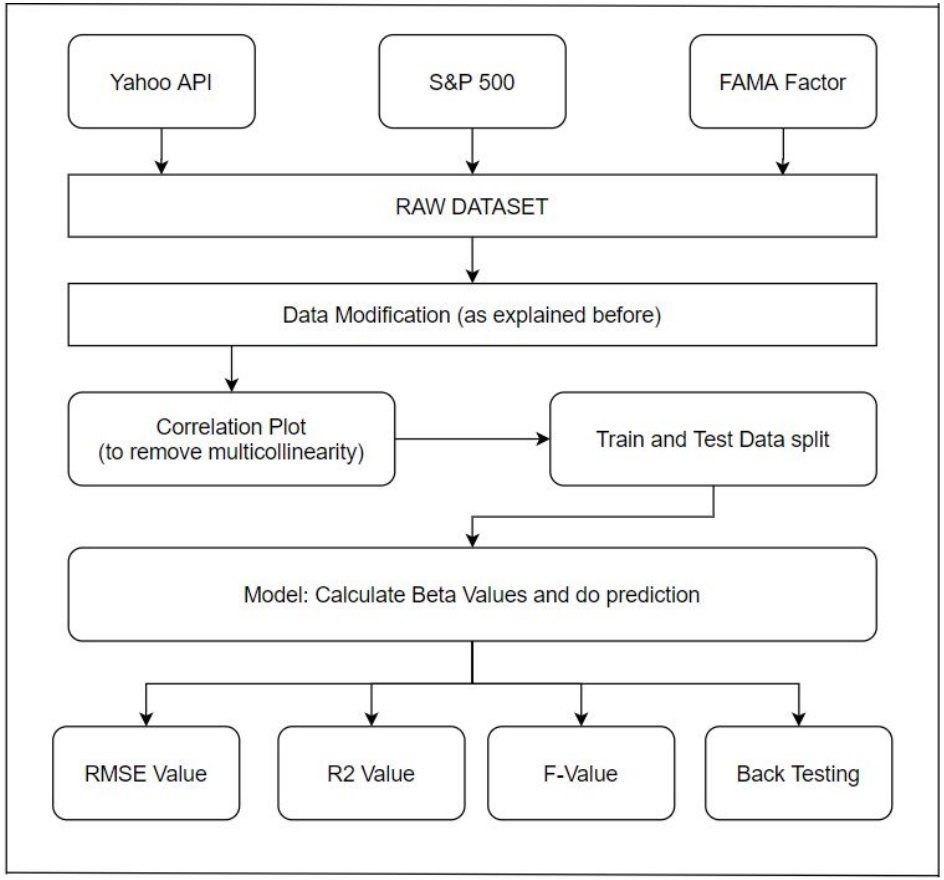


$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

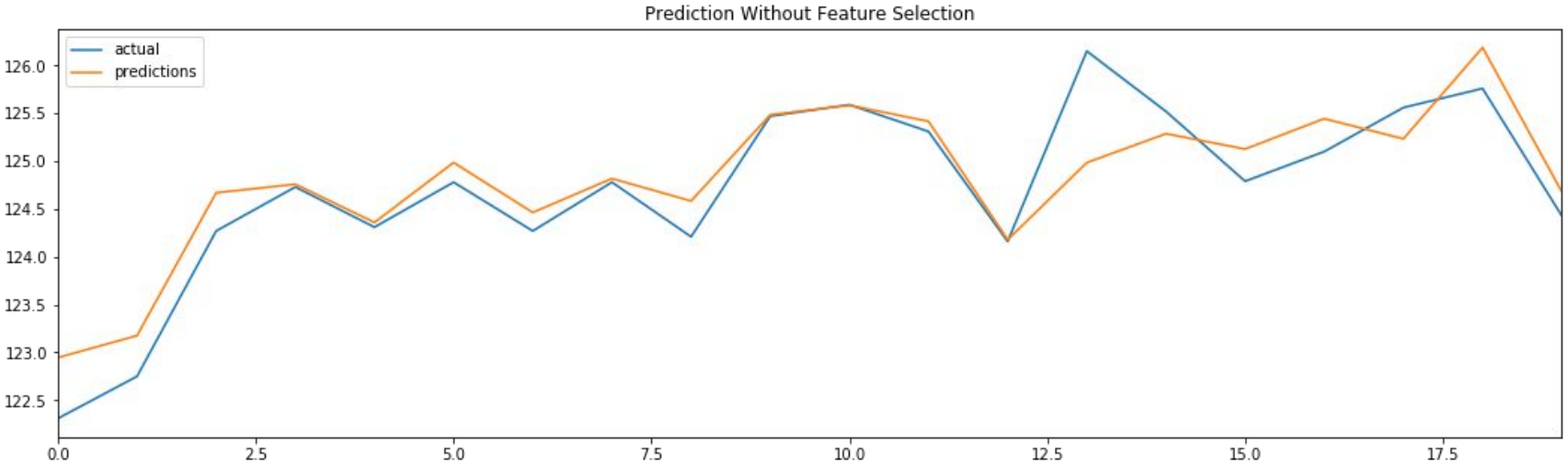
$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

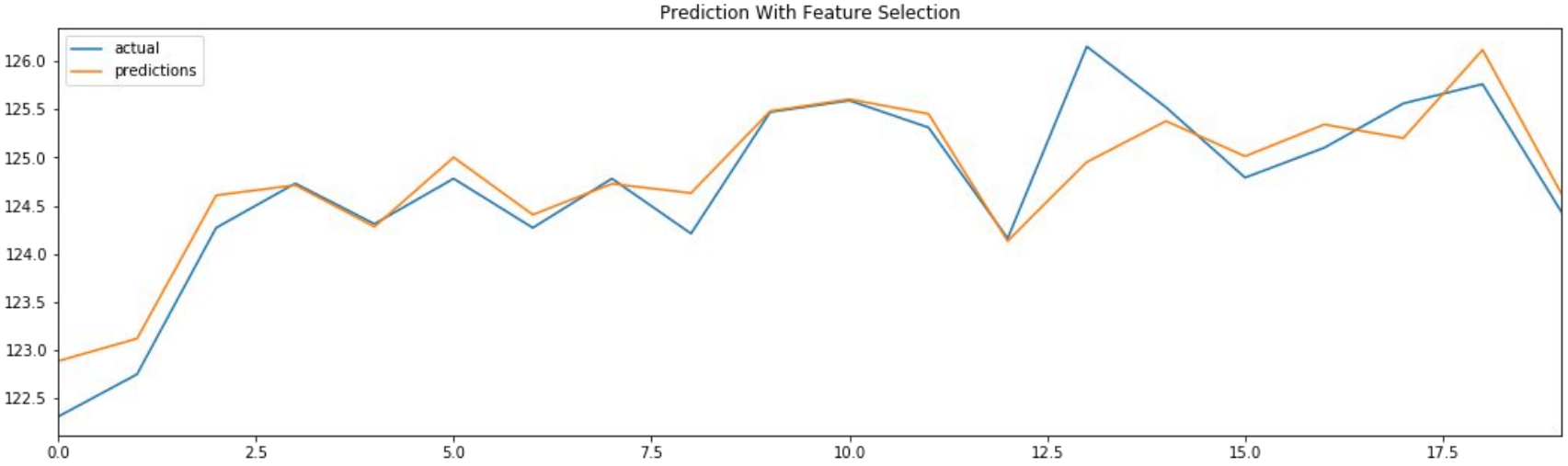
$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$



Assumptions: Multiple linear regression analysis makes several key assumptions:

1. There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.
2. Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
3. No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
4. Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.


Prediction Without Feature Selection
actual
predictions
126.0
125.5
125.0
124.5
124.0
123.5
123.0
122.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Prediction With Feature Selection
actual
predictions
126.0
125.5
125.0
124.5
124.0
123.5
123.0
122.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

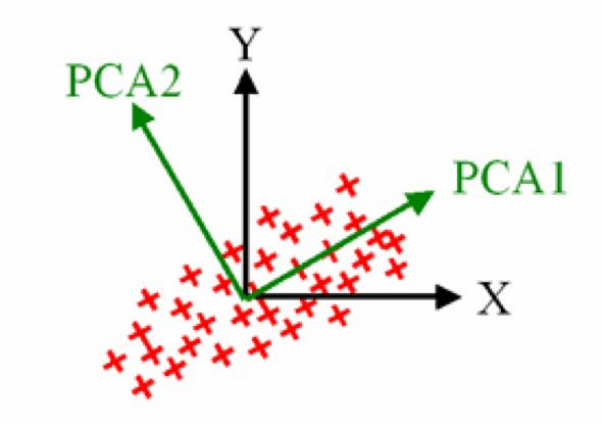
Results and Outputs

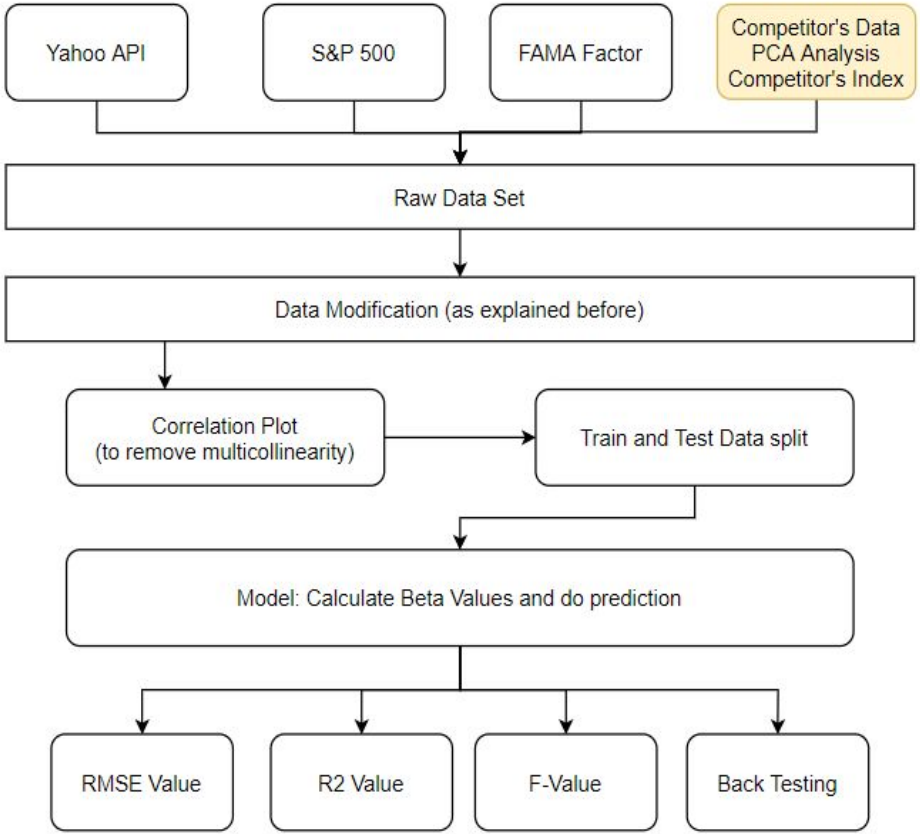
Accuracy	Linear Regression Model (w/o feature Selection)	Linear Regression Model (with feature Selection)
RMSE Value	0.3858	0.3680
R2 Value	0.8270	0.8426
F Value	3.466	16.212

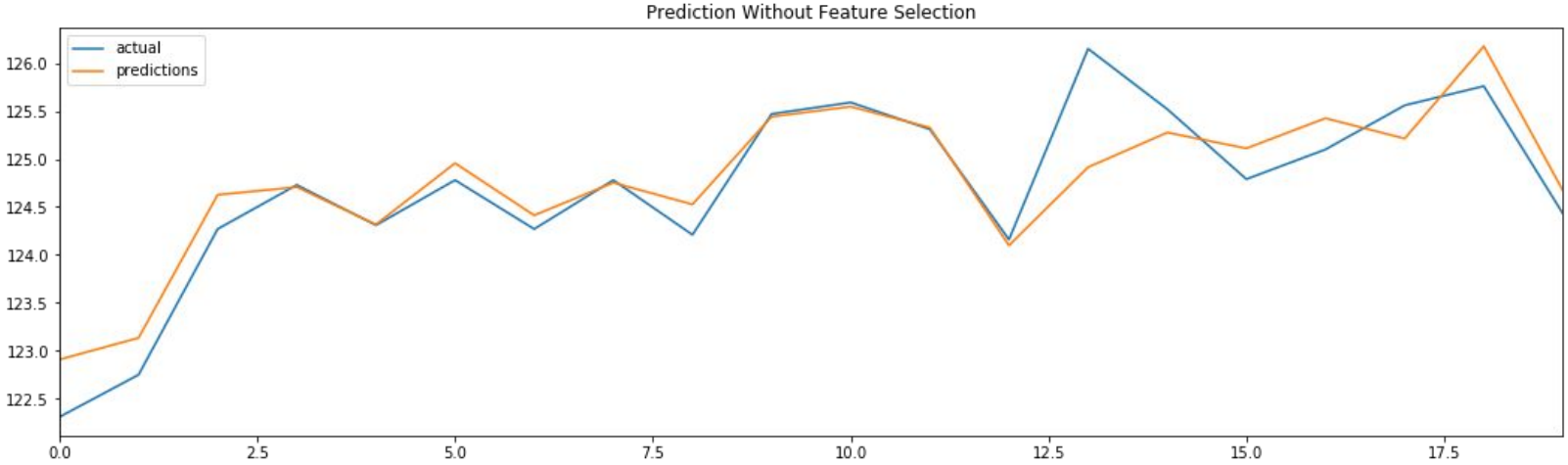
Sharpe Ratio	1.17
Treynor Ratio	1.83
Profit (%)	0.5792%
Hit Rate	1.0

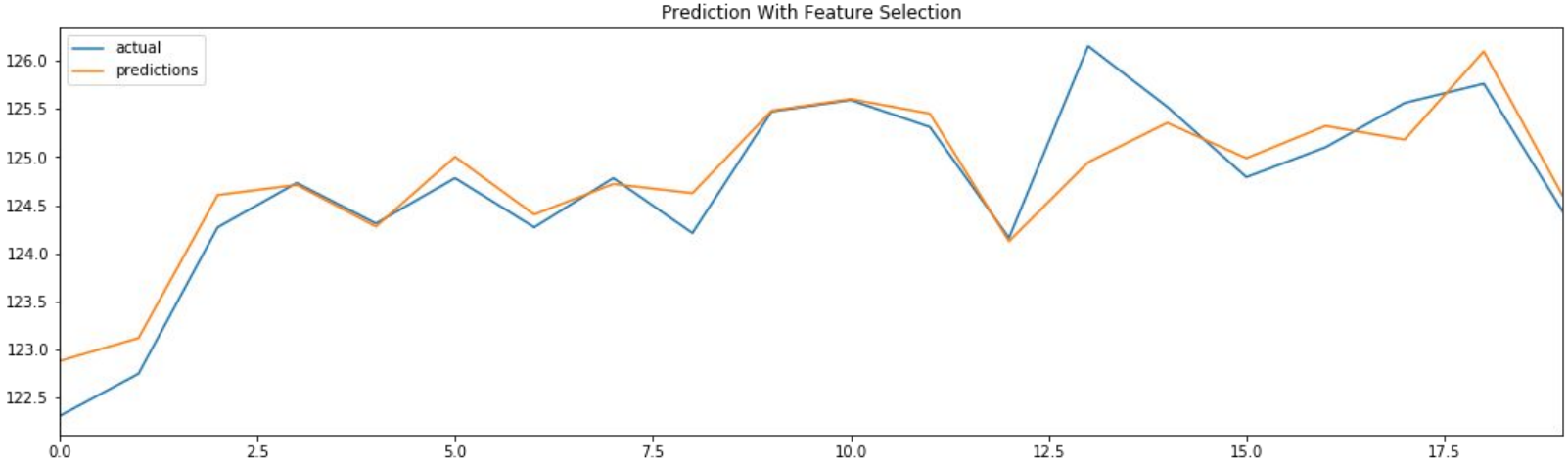
Linear Regression With PCA

PCA: Principal component analysis (PCA) is a technique used for identification of a smaller number of uncorrelated variables known as principal components from a larger set of data. PCA is a standard technique for visualizing high dimensional data and for data pre-processing. PCA reduces the dimensionality (the number of variables) of a data set by maintaining as much variance as possible.






Prediction Without Feature Selection
actual
predictions
126.0
125.5
125.0
124.5
124.0
123.5
123.0
122.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Prediction With Feature Selection
actual
predictions
126.0
125.5
125.0
124.5
124.0
123.5
123.0
122.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

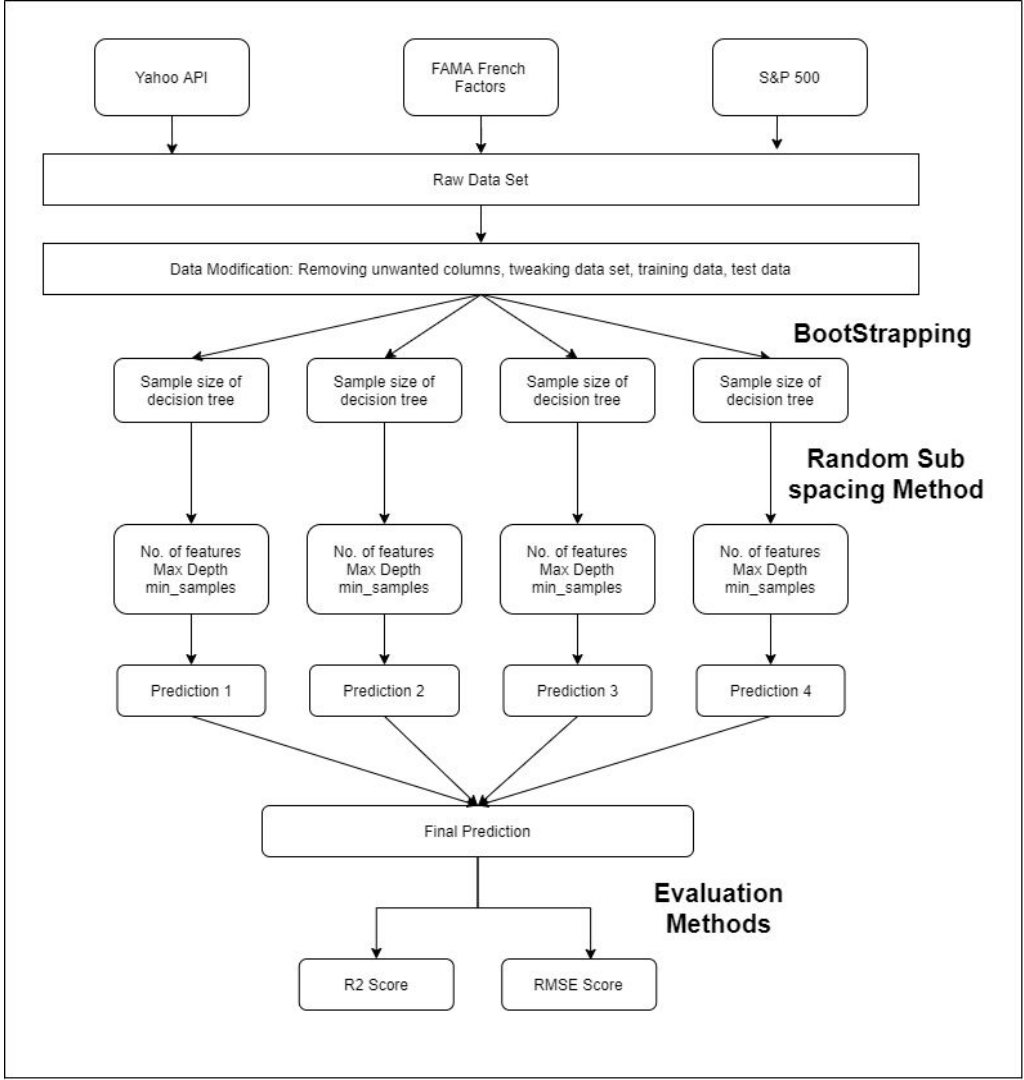
Results and Outputs

Accuracy	Linear Regression Model (w/o feature Selection)	Linear Regression Model (with feature Selection)
RMSE Value	0.3834	0.3669
R2 Value	0.8292	0.8436
F Value	2.843	12.060

Sharpe Ratio	1.16
Treynor Ratio	1.84
Profit (%)	0.5764%
Hit Rate	1.0

Random Forest Algorithm

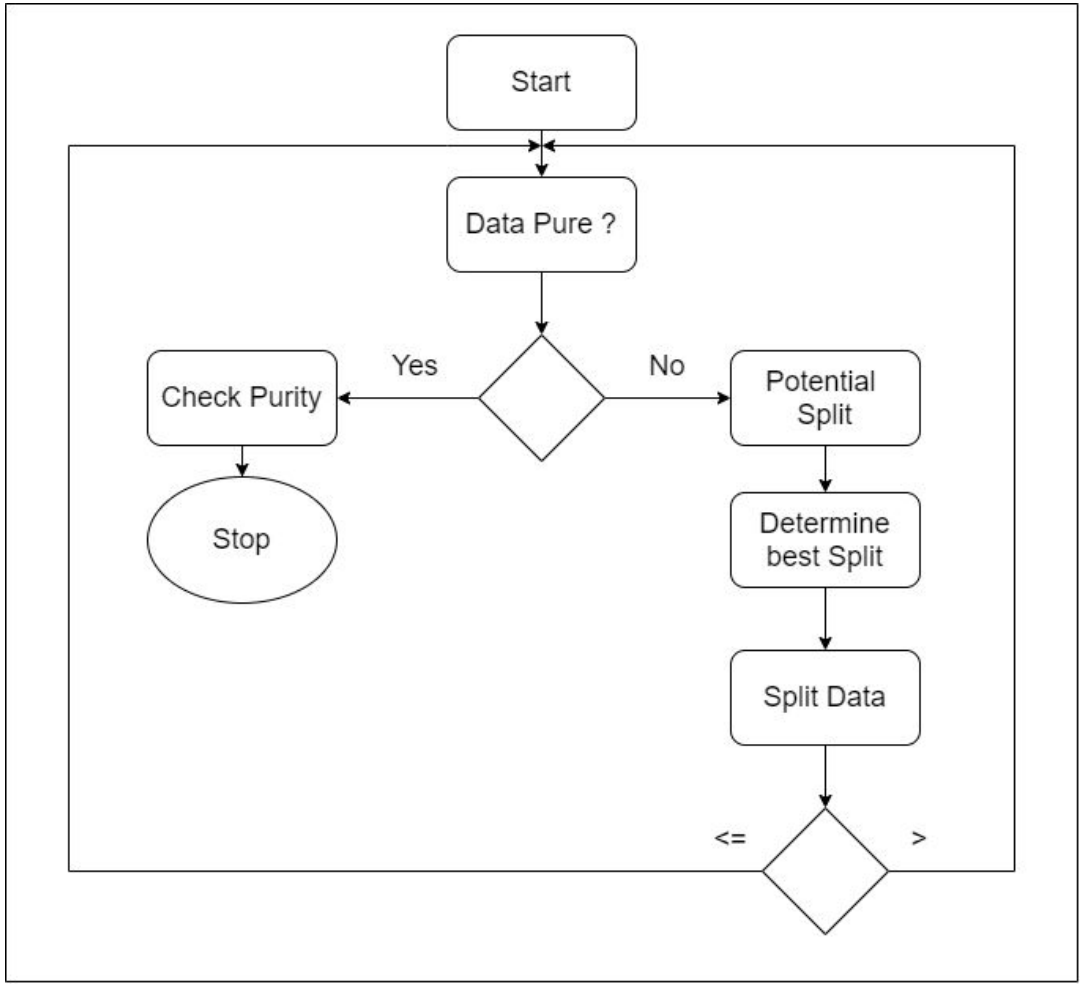
Definition: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. We have focussed on Random Forest Regressor for this Algorithm.



Decision Tree Logic

Helper Function:

- Check Purity
- Create Leaf
- Potential Split
- Split Data
- Determine Best Split
- Determine Column Type



Results and Outputs

Accuracy	Random Forest Model
RMSE Value	1.015
R2 Value	0.7984

Sharpe Ratio	1.36
Treynor Ratio	0.8
Profit (%)	0.6148 %
Hit Rate	0.94

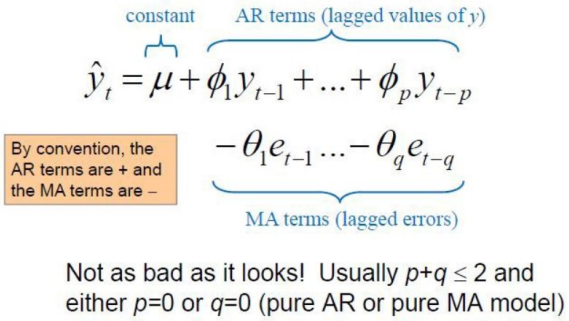
Time Series Forecasting Using ARIMA Model

Time Series: A time series is a series of data points indexed (or listed or graphed) in time order.

Time Series Forecasting: Time series forecasting is a technique for the prediction of events through a sequence of time.

ARIMA Model: ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Equations:



Not as bad as it looks! Usually $p+q \le 2$ and either $p=0$ or $q=0$ (pure AR or pure MA model)

Undifferencing the forecast

The differencing (if any) must be *reversed* to obtain a forecast for the original series:

If $d = 0$: $\hat{Y}_t = \hat{y}_t$

If $d = 1$: $\hat{Y}_t = \hat{y}_t + Y_{t-1}$

If $d = 2$: $\hat{Y}_t = \hat{y}_t + 2Y_{t-1} - Y_{t-2}$

Basic Methodology

1. Data Loading
2. Data Modification as per requirement
3. Time Series Format (Univariate Series)
4. Check for Stationary of the Series
5. Estimating & Eliminating Trend
    a. Log Transformation
    b. Moving Average
    c. Exponentially Weighted MA
6. Eliminating Trend and Seasonality
    a. First Order Differencing
    b. Decomposing
7. Forecasting a Time Series: Partial auto correlation function (PACF) and Auto correlation function (ACF) (Estimating the parameter p and q in ARIMA model)
8. Grid Search to find the best Arima Model Order (p,d,q)
9. Model Prediction
10. Accuracy Test
11. Visualisation Plot
12. Back Testing

Results and Outputs

Accuracy	ARIMA Model
RMSE Value	0.8165
R2 Value	0.2256

Sharpe Ratio	0.05
Treynor Ratio	0.03
Profit (%)	0.288%
Hit Rate	1.0

Pairs Trading Algorithm

Underlying Principle:

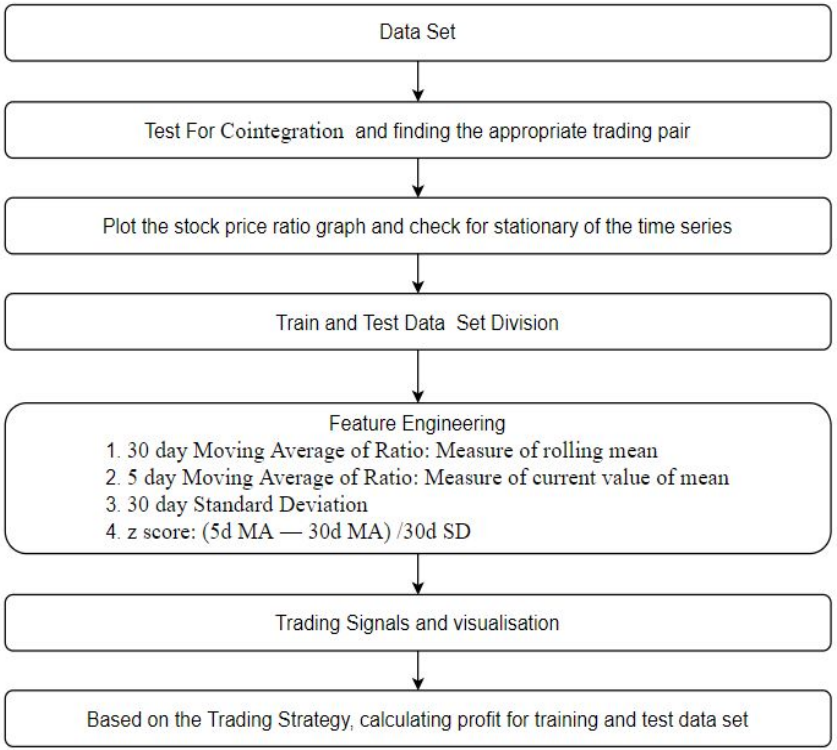
Let's say you have a pair of securities X and Y that have some underlying economic link, for example two companies that manufacture the same product like Pepsi and Coca Cola. You expect the ratio or difference in prices (also called the ***spread***) of these two to remain constant with time. However, from time to time, there might be a divergence in the spread between these two pairs caused by temporary supply/demand changes, large buy/sell orders for one security, reaction for important news about one of the companies etc. In this scenario, one stock moves up while the other moves down relative to each other. If you expect this divergence to revert back to normal with time, you can make a pairs trade.

When there is a temporary divergence, the pairs trade would be to sell the ***outperforming*** stock (the stock that moved up)and to buy the ***underperforming*** stock (the stock that moved down). You are making a bet that the *spread* between the two stocks would eventually converge by either the ***outperforming*** stock moving back down or the ***underperforming*** stock moving back up or both — your trade will make money in all of these scenarios. If both the stocks move up or move down together without changing the spread between them, you don't make or lose any money.

Basic Methodology

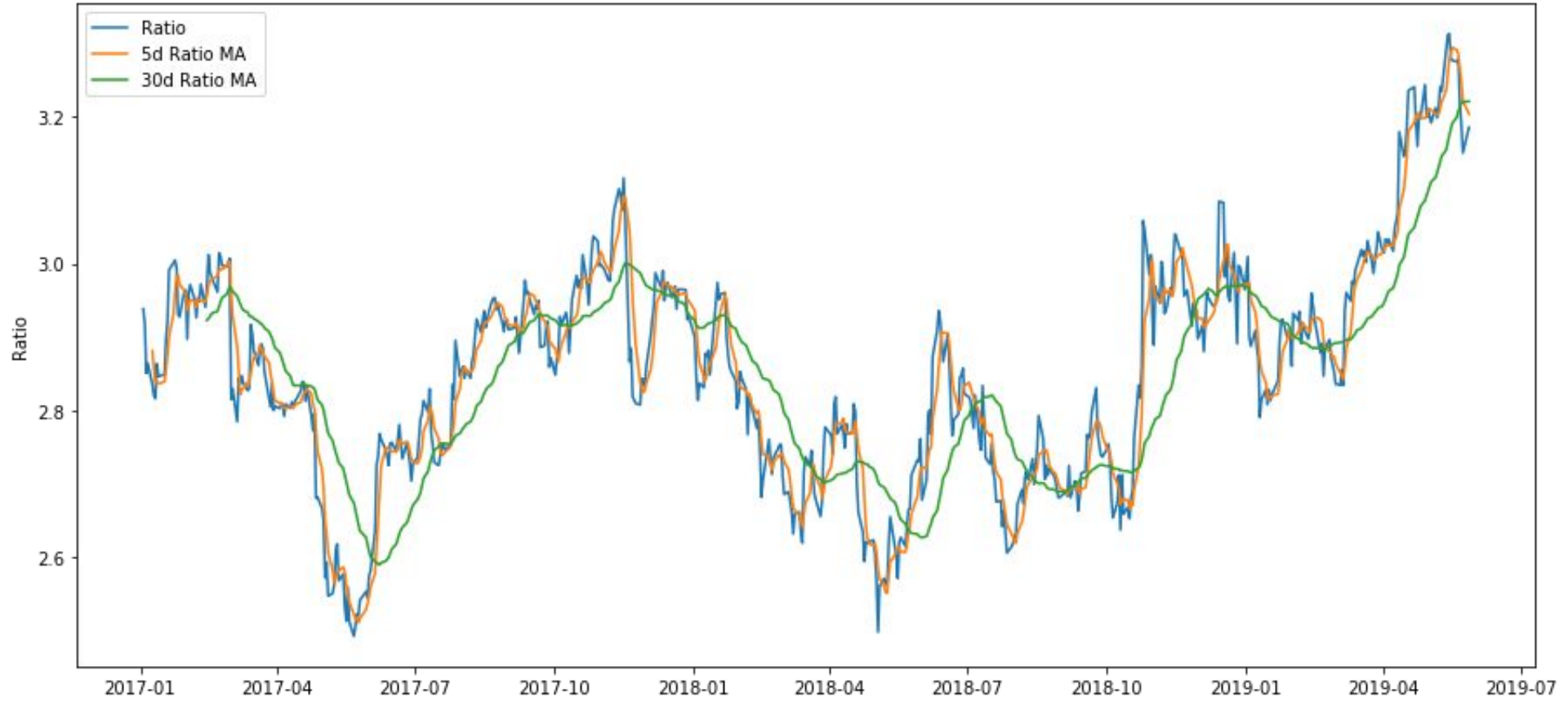
1. Getting data for all the stocks for which pairs trading needs to be checked
2. Test for Cointegration pairs
3. Finding the most appropriate pair
4. Finding the ratio and check for stationary of the series
5. Train and Test Data Set
6. Feature Engineering:
   a. 30 day Moving Average of Ratio: Measure of rolling mean
   b. 5 day Moving Average of Ratio: Measure of current value of mean
   c. 30 day Standard Deviation
   d. z score: (5d MA — 30d MA) /30d SD
7. Finding the trading signals
8. Calculation of profit based on trading strategy for accuracy

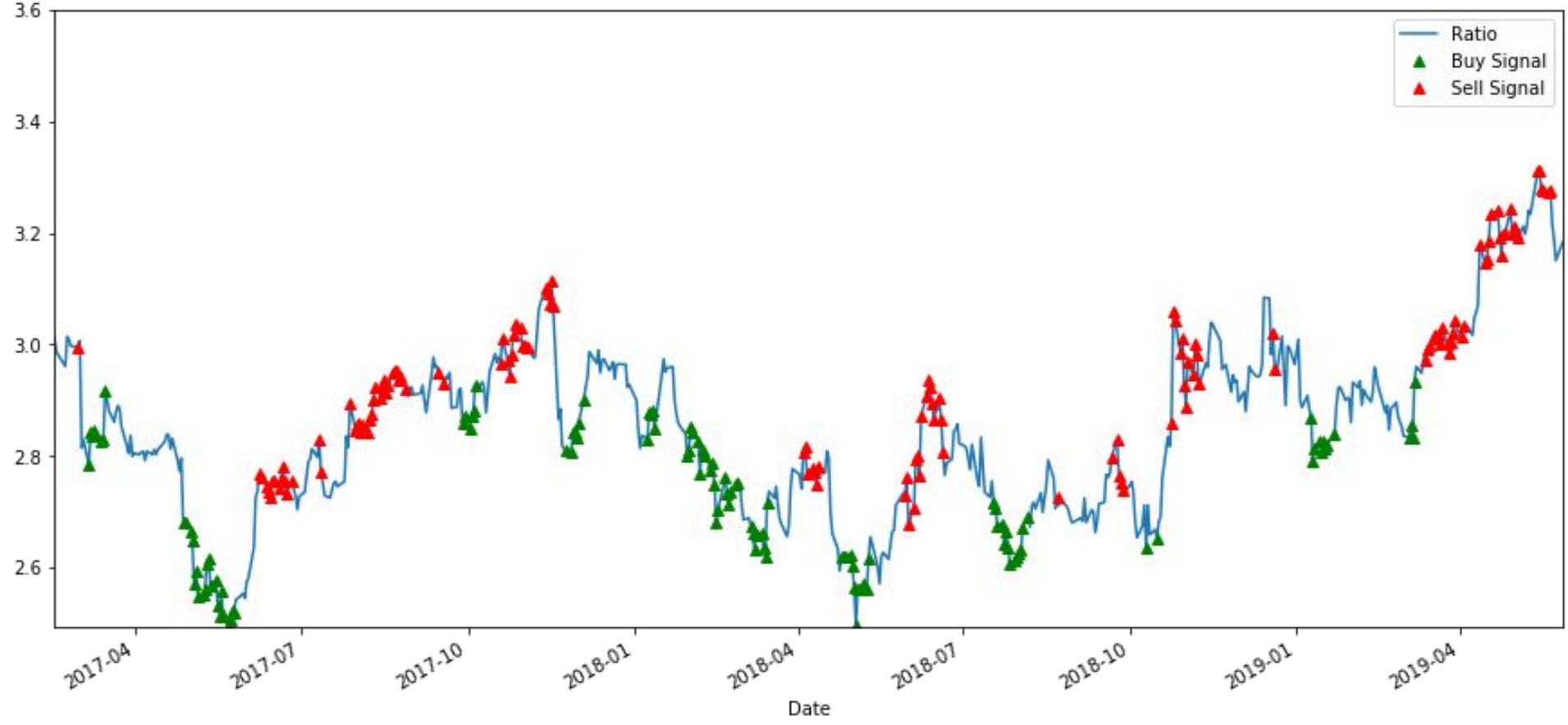
Assumption: Stock Price Ratio is considered to be normally distributed in order to calculate Z-score

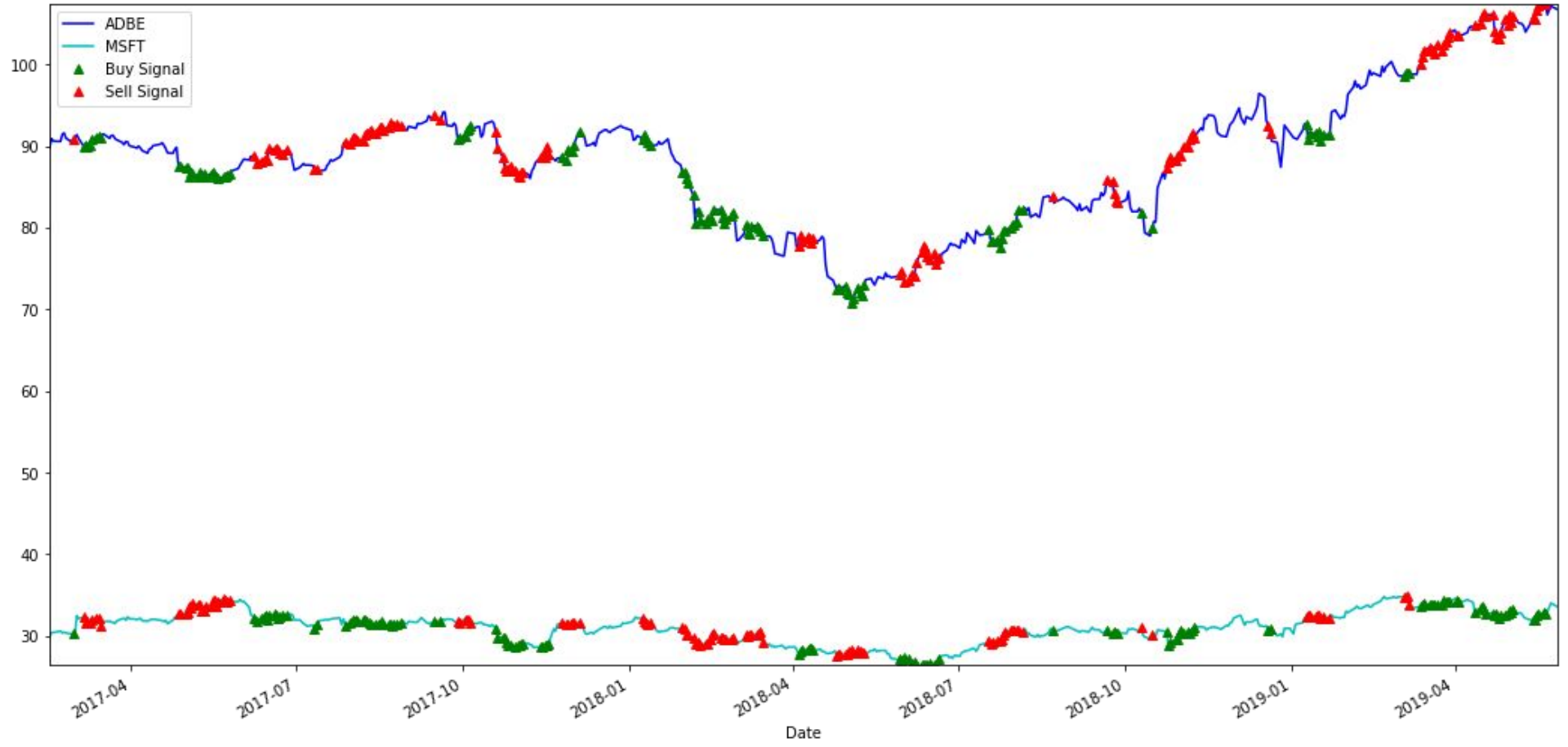


Result of the Algorithm:

Profit from training dataset: 135.85 USD
Profit from testing dataset: 111.05 USD


Ratio
5d Ratio MA
30d Ratio MA
Ratio
3.2
3.0
2.8
2.6
2017-01
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
2018-10
2019-01
2019-04
2019-07


Ratio
Buy Signal
Sell Signal
3.6
3.4
3.2
3.0
2.8
2.6
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
2018-10
2019-01
2019-04
Date


ADBE
MSFT
Buy Signal
Sell Signal
100
90
80
70
60
50
40
30
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
2018-10
2019-01
2019-04
Date

Garch Model

Definition:

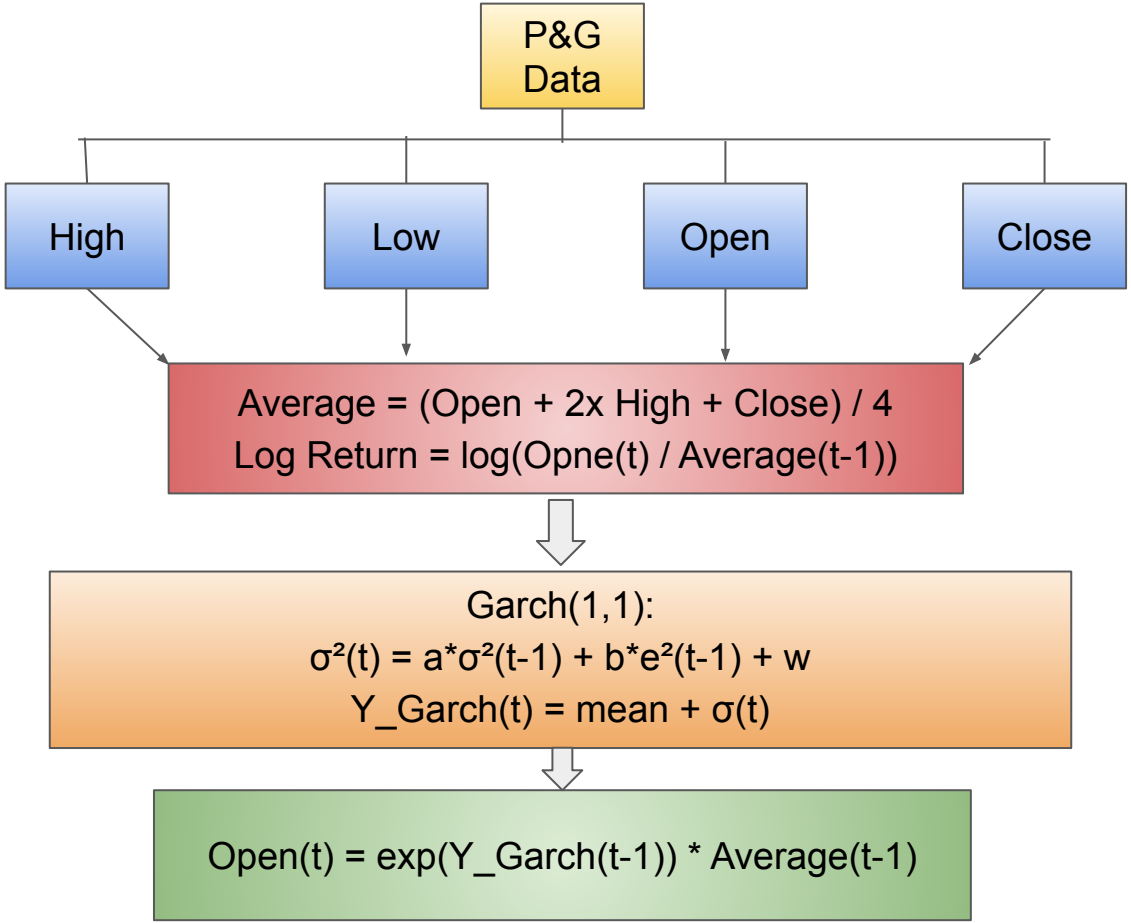
Garch(p,q):Generalized Autoregressive Conditionally Heteroskedastic Models

It is a statistical model used in analyzing time-series data where the variance error is believed to be serially auto correlated.

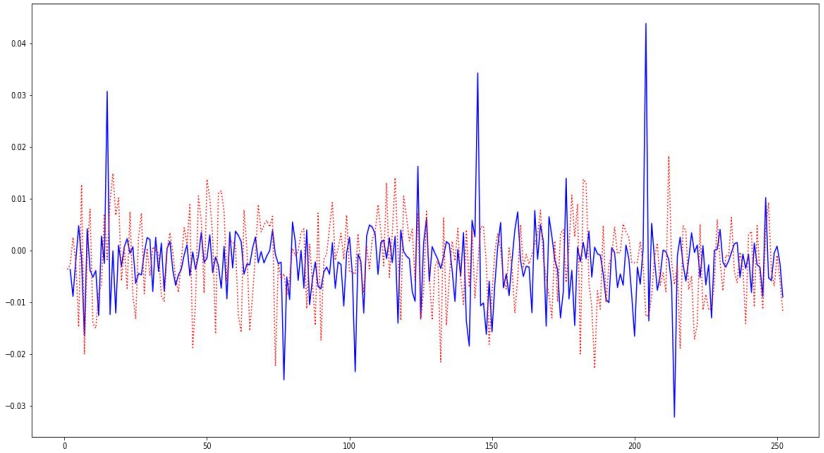
The GARCH(1,1) model is:
$\sigma^2(t) = a*\sigma^2(t-1) + b*e^2(t-1) + w$

Assumptions:

GARCH models assume that the variance of the error term follows an autoregressive moving average process.



Plot for Garch Model Output



Log return (blue)
Garch Estimation (red)

Candlestick and Line Chart of Predicted Open

Results and Outputs

Statistical Measure:

R_square value:	**0.678**
Root Mean Square Error:	**5.857**

Backtesting:

Sharpe Ratio	**0.955**
Treynor Ratio	**-0.692**
Profit (%)	**0.437**
Hit Rate	**1.0**

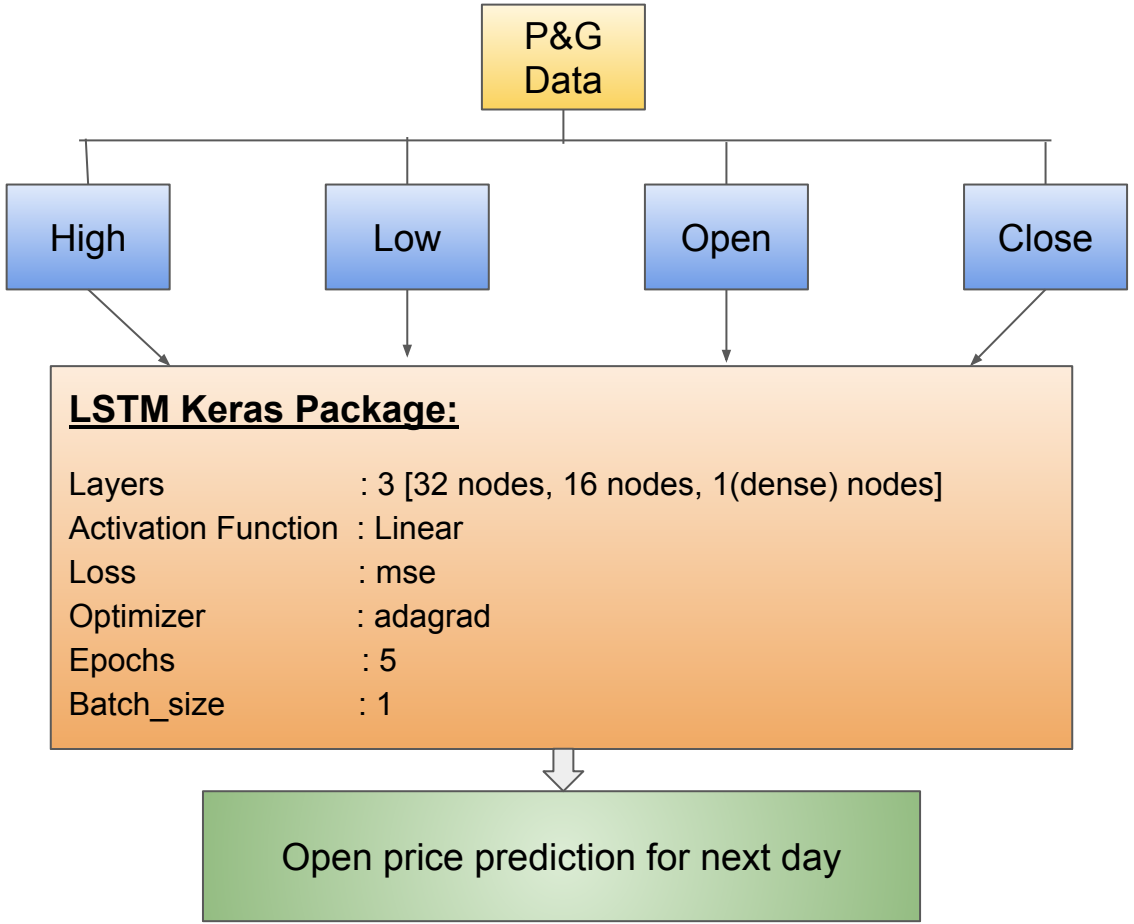
Long Short Term Memory

Definition:

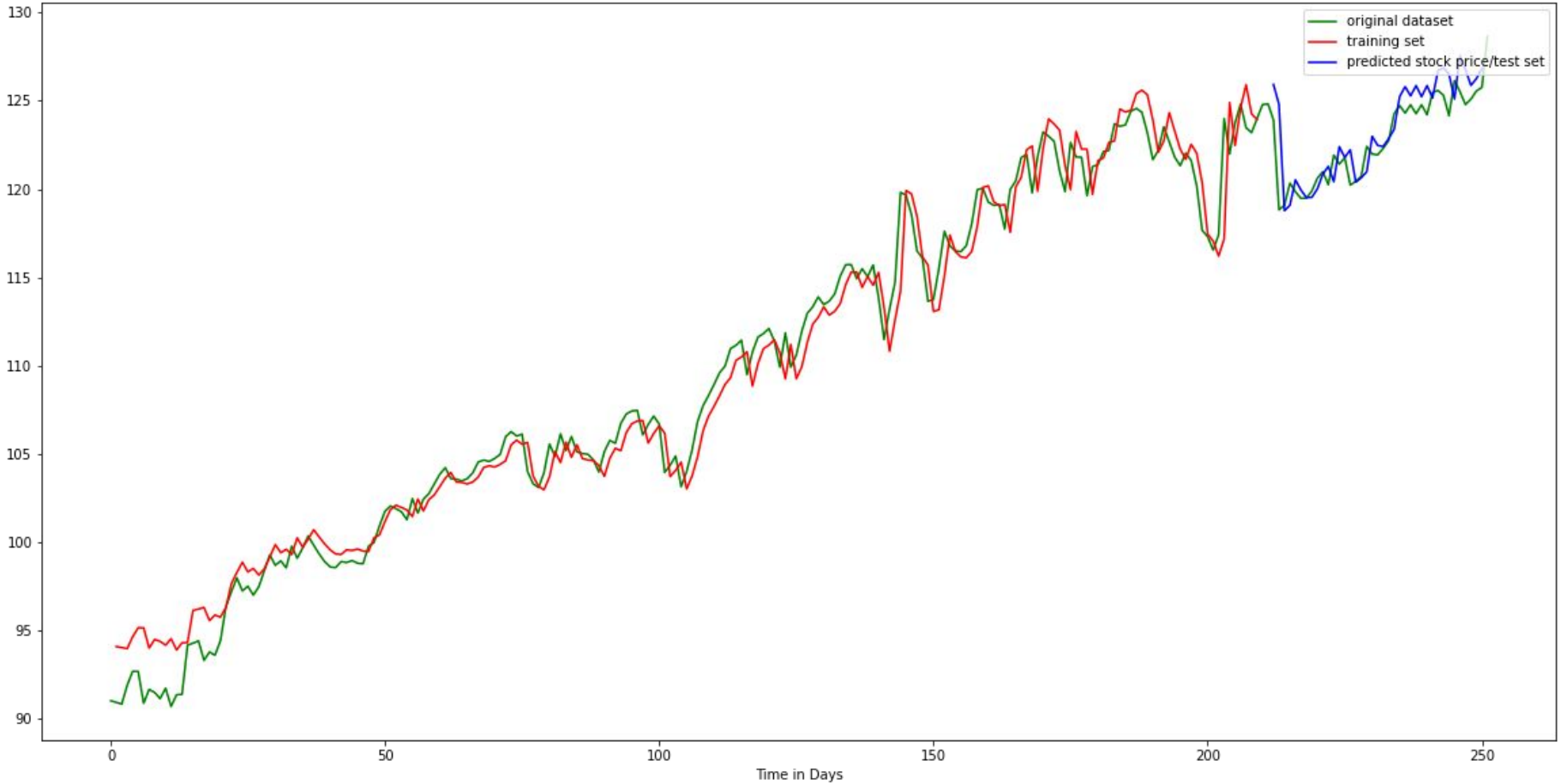
Long short-term memory is an artificial recurrent neural network architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points, but also entire sequences of data.

Assumptions:

LSTM assumes that the state at current time step depends on previous time step. This assumption constraints the time dependency modeling capability.



Plot for LSTM model Output



Results and Outputs

Statistical Measure:

R_square value:	**0.57**
Root Mean Square Error:	**1.47**

Backtesting:

Sharpe Ratio	**1.27**
Treynor Ratio	**-2.92**
Profit (%)	**0.49**
Hit Rate	**0.91**

Statistical Comparison of Models

	LR	LR - CI	RF	ARIMA	Garch	LSTM
RMSE	**0.36**	**0.37**	**0.97**	**0.82**	**5.86**	**1.47**
R_square	**0.84**	**0.84**	**0.82**	**0.22**	**0.68**	**0.57**

Backtesting Development

Backtesting is the process of applying a trading strategy or analytical method to historical data to see how accurately the strategy or method would have predicted actual results.

Backtest includes:

- Sharpe Ratio
- Treynor Ratio
- Average Profit
- Hit Rate

- **Sharpe Ratio:** It is measures the performance of an investment (e.g., a security or portfolio) compared to a risk-free asset, after adjusting for its risk.

$$S_a = \frac{E[R_a - R_b]}{\sigma_a} = \frac{E[R_a - R_b]}{\sqrt{\mathrm{var}[R_a - R_b]}},$$

here , Ra = Profit for P&G ;
Rb = Risk free rate (S&P 500 benchmark)

Sa > 1.0 :: considered acceptable to good by investors
Sa > 2.0 :: Very Good
Sa > 3.0 :: Excellent
Sa < 1.0 :: Sub-optimal
Sa < 0.0 :: Loss w.r.t. Benchmark

- **Average Profit:** It is average profit gained in percentage per trade.

- **Treynor Ratio:** It is a performance metric for determining how much excess return was generated for each unit of risk taken on by a portfolio.

$$T = \frac{r_i - r_f}{\beta_i}$$

here , ri = Profit for P&G ;

rf = Risk free rate (S&P 500 reference benchmark)

Bi = portfolio i's beta $\beta = \frac{\mathrm{Cov}(r_a, r_b)}{\mathrm{Var}(r_b)}$,

T > 0 :: A positive ratio indicates that the investment has added value in relation to its risk.
T < 0 :: A negative ratio indicates that the investment has performed worse than a risk free instrument.

- **Hit Rate:** It is defined as the percentage of the observations (in-sample) that is correctly predicted by the model.

Financial Comparison of Models using Backtesting

	LR	LR - CI	RF	ARIMA	Garch	LSTM
Sharpe	**1.17**	**2.11**	**0.55**	**0.05**	**0.95**	**1.27**
Treynor	**1.83**	**1.03**	**0.78**	**0.03**	**-0.69**	**-2.92**
Profit (BPS)	**58**	**63**	**28**	**29**	**44**	**49**
Hit Rate (%)	**100%**	**100%**	**95%**	**100%**	**100%**	**91%**

On a Final Note,

- P&G is one of the few American businesses that has been deemed essential and can remain open. The Pennsylvania factory was told to stay up and running to continue to produce paper products and diapers. California issued a similar edict. Demand for paper products produced by P&G is so great that the company reopened an idle Georgia plant to meet the insatiable demand. The Procter & Gamble Company (PG) could offer safe haven against the pandemic storm in coming weeks, so PG stock should be an essential part of your portfolio.
- We can offer different trading strategies based on regression, pairs trading, random forest and neural nets. Our quantitative strategies and machine learning model are developed from scratch so we have total control over result. Sum up, these can optimized portfolios.
- Naive investment in P&G leads to average 4 bps of loss per trade. Our strategic investment technique can gain average 63 bps profit per trade with impressive success ratio and risk adjustment.