University of Connecticut

MS in Business Analytics and Project Management

OPIM 5512

Data Science with Python

Zombie Companies:

Do Winning Stocks Get Sick with COVID-19?

Team 9:

Sidhdesh Gupta

Supeng He

Rachan Vamsi

Phillip Zaprzalka

Table Of Contents

Executive Summary

The idea that a company can operate without growth and fail to eliminate or even accumulate larger and larger debts can lead to a condition widely considered being a Zombie Company. There is not a commonly agreed-upon definition of such state but general findings can be distilled to a company that does not show growth, whilst its revenue is only enough to pay minimal operation costs and payment of interest on its accumulated debts. This data is not easily attained as it is closely held by companies out of the publicly accessible online resources. During the time of COVID-19, all stocks were impacted some excelled while others took incredible hits. There was general recovery at different rates creating winning and losing sectors. These can be used to build a baseline of performance during the COVID-19 period from January to June 2020. The construction of a model to compare the performance of zombie stocks to the greater baseline performance of the stock market can offer further insights into possible definitions of zombie stocks.

Problem Statement

For some time there has been an increase in corporate debt from taking advantage of low-interest loans or increased investment in businesses by others who can attain the low-interest debt. Companies have leveraged significantly in bets to spur growth, purchase new equipment in modernizing, or to pay down older debt burdens with restructuring. Those are all positive uses of new debt.

Alternatively, Zombie companies have been rising from the negative side of this debt. Companies have pursued increasing their debt load with the knowledge or at times unknowingly being unable to reduce that load over time. They have assumed a debt with interest whose payments exceed their ability to pay it back in a timely or responsible period. When the payments of interest only are met

along with basic operating costs to result in breakeven or dwindling cash reserves, this would be the formation of a Zombie Company.

During this time of analysis, the world is responding to the effects of COVID-19 as they change daily personal and professional behaviors. Regulations imposed by the government and recommendations from the health industry continue to offer guidance to remain socially distanced. All of this compounded often negatively affects company functions. This in turn will cause companies to take on debt to retain talent, reduce their overhead by sending people to work from home or to close their doors and hold their breath operationally till the situation changes.

Some companies have fared well which were already designed for modern remote work conditions or those that specialize in less contact distribution of their goods and services. COVID-19 has shaped how many companies are able to perform for their public shareholders. The zombie companies which were in a state of idle unchanging survival, now find themselves shaken and disturbed to repay on those debts. Struggling as they were they are now likely to fail in these uncertain times.

It is the goal of this analysis to answer the question as to how to define and identify zombie companies. And in accomplishing this also establish a baseline of normal company performance during times of COVID-19, all through the use of US stock market performance of publicly traded companies.

Methodology

Analysis Plan

As stated above the initial analysis was expected to focus on Zombie Companies with a Baseline being established through analysis of common stock performance. This will therefore be broken up into two different processes but will become integrated within the modeling process. One for zombie companies and one for the COVD-19 stock baseline.

Zombie Companies

First, to find zombie companies to perform the analysis we will need to find or establish a list of companies fitting this description. Unfortunately, there is no definitive unified source with a list per se of these companies. There are common discussions on forums and articles written of companies that have characteristics that analysts and journalists characterize by this name though. The greatest anticipated challenge was finding/establishing defining metrics of “zombie companies”.

DEFINE: To define the zombie companies we must either subscribe to a commonly understood definition or create one. We did an assessment of online articles and forums to come to a common core metric which is used to define the operating dead of a zombie company. The definitions available are not exact and can vary. We then compared these sources of definitions to establish our assumed distilled definition. A zombie company is one in which the debt burden held by the company does not decline and likely rises, only being maintained through payment of accumulating interest requirements and regular operating expenses covered by incoming revenue. Characteristically there is no revenue growth and often an amassing, growing debt.

IDENTIFY: To identify the zombie stocks of the COVID-19 period of interest, which ascribe to the

definition, we can scan stock data sources for those metrics. Unfortunately, the public release of interest payments, frequency, and scale cannot be attained. With a lack of professional credentials, it is not currently possible to acquire this information efficiently.

EVALUATE: To evaluate zombie companies during the time of COVID-19, first establish a baseline of the historical performance of zombies compared to the market with the use of indices such as the S&P 500, DowJones, or NASDAQ indices.

MODEL: Use various models often focusing on time series models due to the nature of stock price data. Then perform model comparisons for best fit.

ADDITIONAL FACTORS: To add a sentiment analysis into the model will increase the dimensionality and greater resolution to the further definition and future identification of zombie companies. This requires data acquisition from Twitter or other online forums in the form of raw data. Then scale the weighting of sentiment to show a trend of support behind various stocks.

RESULTS: To resolve which companies which have been defined as zombies are maintaining performance along with industry peers versus those which are doomed to fail and operate outside of normalcy. This would differentiate which zombie companies live on and which should get buried.

COVID-19 Baseline

DEFINE: To define the baseline stocks to be used, both indices as well as individual stocks were chosen. There was additionally a choice to use the Electronically Traded Funds (ETF) for various trading indexes as these have more automated and rules-based stock inclusions. Five ETF Indexes were chosen from three ‘winning’ sectors and three ‘losing’ sectors as commonly found between the time window of peak COVID-19 for both winning and losing sectors (Kastner 2020).

IDENTIFY: To identify each sector for inclusion, a simple calculation of the losses tallied in early COVID to mid-march added to the two-month recovery by sectors till mid-May (see Figure 1). This well frames a set of sectors to pursue in establishing strong companies and weak companies from the ETFs as our baseline.

EVALUATE: To evaluate the stocks for a baseline, those sectors would need to be refined and sorted delivering a consolidated dataset. The consensus of five ETFs within each sector would be used to compare their top 10 holdings. Like index funds, they should offer a general consensus of top 3 stocks in the sector when rank averaged from the publicly disclosed top ten holdings, by net asset percent holdings, for each ETF. This provides us an effective reduction in target stocks from which to build a baseline for winning and losing sectors both.

MODEL: To model a baseline the data will be recent in scope from 01-01-2017 to 12-31-2019 (Train/Validate) will that predictive model show the stocks to be predicted higher or lower than current levels within 2020. The ranking will be in accordance to Outperform, Market Perform, or Underperform. Then comparisons from baseline predictions to the actuals allow for an understanding of how the general market has fared in the COVID period 01-01-2017 to 06-01-2020.

RESULTS: The results will show Pre-COVID performance, COVID Crash, and COVID Recovery. Then conclusions can be made as to which stocks are therefore the ‘best of the best’ COVID Recovery stocks to watch or buy.

Furthermore, the plan of analysis can have outcomes of not only a baseline during a time of global pandemic for possible use in the future but also how would companies perform without the grand impact of this magnitude. If COVID-19 did not happen, where would the stock be today, would be defined by the variance 2017-2019 vs. 2020. How did the COVID-19 winning sector perform

compared to predicted versus the losing sector?

Data Description

We extracted data from thirty ETF indices and filtered it down to six lists of three stock ticker values. The eighteen stock ticker symbols were then used to extract data from Yahoo Finance Database from 1 January 2017 till the current date. This is a dynamic function of the model currently set to pull to the most current date of available information into the database. The datasets contain records of recent stock price changes and the volume of stocks sold.

The variables in the datasets are:

1. **Dates**: These dates are used as indices for the data.
2. **High Price**: The Highest price of Stock recorded at the end of the daily trading session.
3. **Low Price:** The Lowest price of Stock recorded at the end of the daily trading session.
4. **Open Price:** The Open price of Stock recorded at the beginning of the daily trading session.
5. **Close Price**: The Closing price of Stock recorded at the end of the daily trading session.
6. **Adjusted Close Price**: Adjusted closing price amends a stock's closing price to accurately reflect that stock's value after accounting for any corporate actions
7. **Volume:** Volume is the amount of an asset or security that changes hands over some period of time, often over the course of a day.

The datasets used one main directory and six sector subdirectories. In each of these subdirectories, are further subdirectories of the top three stock ticker symbols to place the extracted results respectively. The code and the results were mounted in Google Drive ensuring the user to have access to the data anywhere with internet access in any device.

Data Preprocessing

We imputed data using forward fill and backward fill because the time series model needed data to be continuous at a daily rate. This made the data flexible for adding a variable for future scope and ensuring there are no missing values. We ensured there are no outliers accounted for as the stock data can be random. Missing a price would make our model biased over a certain range of prices

where the company was more active. So, we have used normalization to make all the prices in the range of 0 to 1.This ensured that we have accounted for every price in our model. We also have plotted density plots, scatter plots, box plots, and line plots to visualize data for 18 companies.

Data Exploration

We noticed that all the price variables were following the same trend.and were having the same confidence interval even after data preprocessing. After we made the correlation matrix, we noticed that the price variables had a correlation value of 1 making them the directly derived variable. We assess that the model would have less complexity if we added price variables with Adjusted Close Price as the target. We added the Volume of the stocks as a predictor as the correlation value was relatable and data was spread out and random. This also added some business relevance in showing the volatility of the stock or whether the stock was more active. We made pair plots, correlation heat maps, and generated a pandas GUI profile for all the datasets to allow the user to explore data and different correlation matrices. We have also calculated the mean, median, maximum prices, minimum prices, and quantiles to make it easier for the user to access the stock data. The present value if the user invests on 1 January 2020 after 1 year if the stock has no greater changes, Net Present Value of the Stock at 12% discount rate, Internal rate of return for the stock prices, Return of Investment of the stock and whether the company is a potential zombie company. This allows the user to analyze top companies and know about updated stock information daily.

Data Modeling

The models built, explored, and compared for performance are listed below:

1. Linear Regression: Linear regression quantifies the relationship between one or more predictor variables and one outcome variable.

2. Lasso Regression: Lasso regression is a type of linear regression that uses shrinkage with the help of loss function and a penalty function.

3. Elastic Net Regression: The Elastic Net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods.

4. Decision Tree: Decision tree builds regression models in the form of a tree structure.

5. KNN Model: K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure.

6. SVR Model: Support Vector Regression (SVR) applies the idea of SVM to predict real values rather than a class.SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model.

7. Sentiment Analysis: Neural networks are a set of algorithms, modeled loosely after the human brain, that is designed to recognize patterns. The sentiment was recorded and the variables were inputted to a neural network model.

8. Random Forest Regressor: Random Forest Regressor is an ensemble model that uses a bunch of decision trees to get the prediction.

9. **Gradient Boosting Regressor (2nd Best):** Gradient Boosting Regressor is a type of ensemble model which uses many decision trees to predict the target variable.

10. Extra Trees Regressor: Extra Trees Regressor is an ensemble model that uses many trees to make a prediction.

11. AdaBoost Regressor: AdaBoost Regressor is an ensemble model that uses decision trees to make a prediction.

12. **ARIMA Model (Best Performance):** ARIMA model or Autoregressive integrated moving average model is a type of time series model.

13. MA Model: MA model or Moving Average model is a type of time series model.

The time series models were predicted by variable Adjusted Close Price as Target. The other models were modeled and fitted using variable Volume as predictor and variable Adjusted Close Price as Target. All these models were run on the 18 different datasets and predictions were plotted for different models with different datasets.ten of thirteen models were under fitted modeling and were depicting a lackluster performance after modeling. The models were analyzed by getting the lowest Root Mean Square Error and determining it as the best model. The residuals were also extracted for different models and the spread of the residuals was found for different models to allow the user to select a more accurate model (see Figure 2).

For instance, as seen in figure 2, figure 3 and figure 4, the residual spread of Amazon Stock Prediction models, we can see that MA model has a tight range of the range of error and has lowest error value even though it doesn't fit the model as well as the best model Gradient Boosting Regressor Model prediction. The user might think to get the model with the least error value and select the MA model instead of a Gradient Boosting Regressor Model. This allows the user to interpolate the numbers to show in the results making it possible for the user to select data with minimum knowledge of data analytics and have a choice to select different models for prediction. This whole model selection process has gone through the generalization of the process by making it into a box plot.

Best Model - ARIMA Time Series

Our best performance model so far is the time series. To find out how the pandemic had impacted the stock market, we aim to choose two companies from different industries to make comparisons. The first target company is Visa. Visa is a multinational financial corporation, its main product is offering electronic funds transfers service throughout the world. Another company is Hertz, Hertz is also an international company, the main product is offering a car rental service.

The time range is the same as other models, the date range is 3 years. One difference is that the partition for Time Series is in sequential order, so we set the first 70% as our training, and 30% as test data.

Use Visa company as an example, from the stationarity plot we can tell it is not stationary thus we applied a second difference here. In the process of determining the p and q, the PACF plot suggests lag 1 and 2 are all significant spikes, so the range of q should be 1 or 2. Autocorrelation plot shows that somewhere starting at 2 the lines are within the critical boundaries, therefore we decide the P is 2. In the end, the best order we can generate is (2,2,1). *(Stationary plot, PACF, and ACF plots can be found in the Appendix)*

By looking at the prediction for Visa in 2019, you may wonder why forecasts align with true values so well. Since we used time-cross validation here that yields the prediction fits so well.

To dig more about how the pandemic had impacted the companies, we compare the stock price trend in the year 2020 for Visa and Hertz. The left one is Visa, the right one is Hertz. Given a 90% confidence interval, starting from late February, we can see both companies had been influenced negatively by the COVID-19. After the financial support from the U.S federal government, Visa started to raise from late March, however, Hertz didn't. We can intuitively think that the quarantine policy has posed a significant negative impact on the travel industry.

Next Best Model - Gradient Boosting Regressor

Gradient boosting Regressor Model builds a model in a stage-wise fashion and generalizes them by allowing optimization of an arbitrary differentiable loss function. All the ensemble models from Scikit learn Package was chosen after observing the randomness of the prediction of the decision tree. Even though the explainability of this model is lower, we wanted to make use of the randomness in the prediction to predict even if we were using one predictor variable, Volume to predict our target, Adjusted Close Price. It was evident we made a good decision as the Gradient Boosting Regressor model was showing lower RMSE value than other models in some company stock price predictions. If we notice the above prediction graph of Amazon Stock in Figure 2, we notice the accurate trend prediction from March 2020 to April 2020 when there was more COVID activity in the US. This shows the model is performing great in some time interval. If the same COVID situation reoccurs, then this model might be reliable for the user to predict the stock price for some companies.

The interest to find retention percentage of the existing students created the scope to remodel the dataset to perform a churn model analysis. Below are the steps involved in creating this model.

Results

The models produced were not able to accurately predict the performance of zombie companies due to the inability to access certain data elements. The models are able to perform a weighted analysis with introduced sentiment data. As well they will be able to continuously update for live tracking and up to date analysis as the market changes. The results achieved are most markedly in the area of model design and structure for future usefulness when specific variables, sentiment data, and higher quality data sources can be accessed. The reliability of the outputs received from Yahoo Finance resulted in non-convergent correlations.

To be successful next time we would pursue a larger volume of data spanning a longer timeline. This would give a higher resolution for market changes and normalization of seasonality. The pursuit of a higher quality data source than Yahoo finance would increase our ability to use more variables in the model which are currently being left out. Those variables being left out have irregular relation to the prior day open sales. If we had it how would that change our results?

Conclusions and Recommendations

This analysis did find a distilled definition of zombie companies to be those which do not grow and carry a debt burden, where revenue covers only interest and operating costs. Our evaluation was not able to subsequently test and validate this with data and the supporting modeling constructed due to limited data accessibility for interest payment information being close hold by corporations. In the future with specific data access to Debt and Interest Payments compared with Revenue**,** we can achieve this. Towards the goal of predicting Zombie Stock prices, the models are not conclusive due to the above constraints. The model is designed to input these values at a future time when attained. In light of not having conclusive results through a defined and targeted modeling outcome, the models did produce a possibility for a proxy of identification for prospective zombie companies. This binary is based on falling prices and failure to return to pre COVID-19 levels behavior.

When comparing a ‘Zombie Reference’ (Hertz) from ‘normal’ stock behaviors, Visa, for example, the COVID stock Sector data capture and comparison allowed us to establish a solid baseline for ‘normal’ behavior.

The best performing models were the time series ARIMA and Gradient Boosting Regressor are the top performers. These results were conclusive using the RMSE for comparison.

In the scope of future work, the models are configured for real-time monitoring and identification of targets of focus. Sentiment analysis data sources once identified will offer to weight the quality of Zombie identification outputs. Additionally seeking to improve variables and time complexity of the algorithm. This will speed up the inefficiencies in the coding. More variables like alternative data, Company Accounting Details, Sentiment, etc should be attained. Including a larger collection of companies or representation from each sector can offer a more well-rounded representation of

market baseline performance for any large scale events which can impact other industries in the future. Making a package to generate results as the report in GitHub would be advisable to reduce processing time and increase intercommunication and crowdsourcing of coding content.

References

Kastner, David. “Schwab Sector Views: Changes Are Coming”. Charles Schwab. June 18, 2020. https://www.schwab.com/resource-center/insights/content/sector-views

Korstanje, Joos. “Text mining for Dummies: Sentiment Analysis with Python”. TowardsDataScience.com. March 8, 2020. https://towardsdatascience.com/text-mining-for-dummies-text-classification-with-python-98e47c3a9deb

Appendix

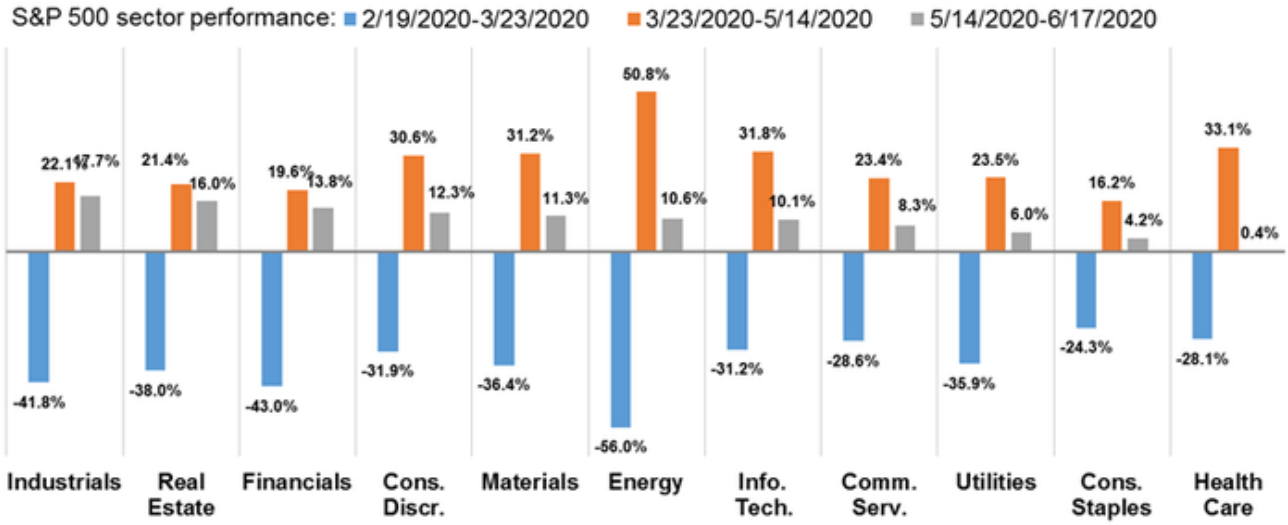

Figure 1: List of Stock Sector Performance in the time of COVID-19

Figure 2:Best model for Amazon Stock prediction using RMSE is Gradient Boosting Regressor Model

Figure 3:MA Model for Amazon Stock

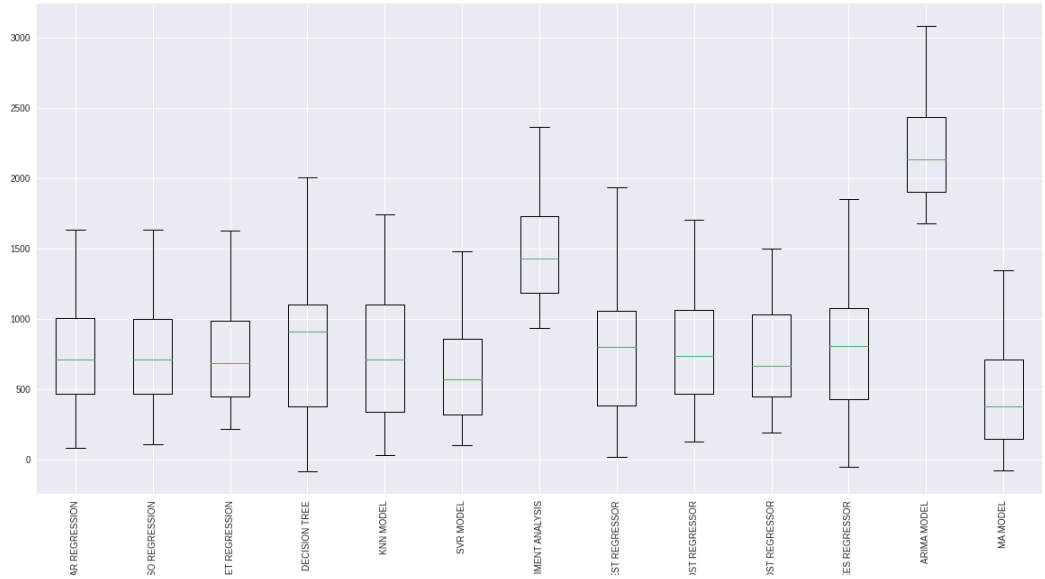

Figure 4:Residual Comparison for Amazon Stock

Figure 5:Visa Prediction Plot in the year 2019

Figure 6: Visa 2020 prediction

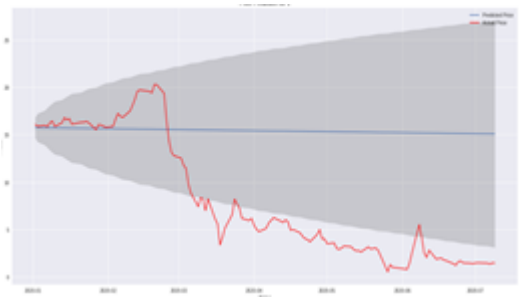

Figure 7: Hertz 2020 prediction

Figure 8: Time Series- Stationarity plot

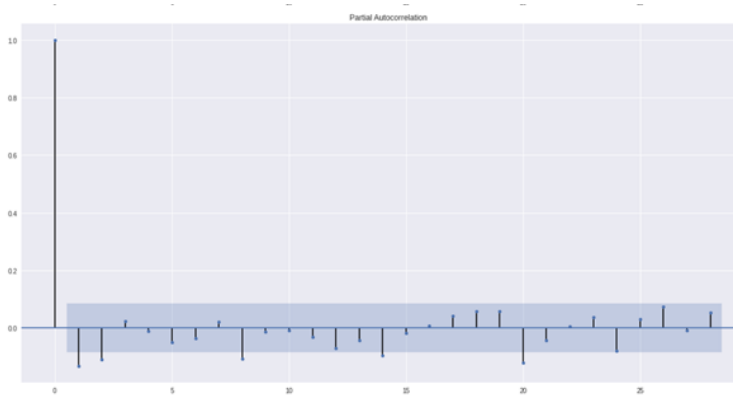

Figure 9: Time Series- PACF plot

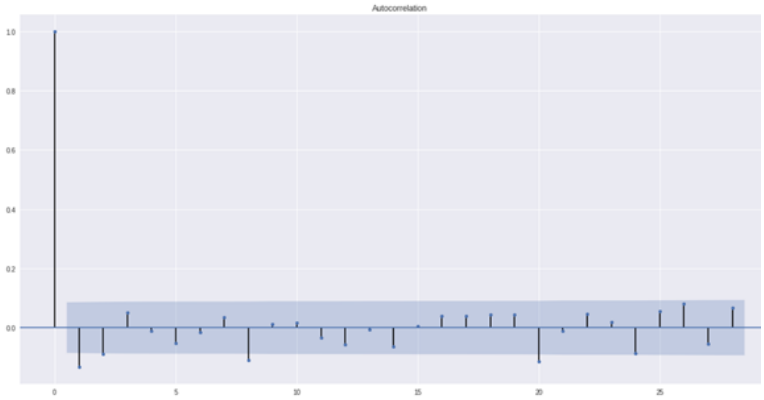

Figure 10: Time Series- ACF plot

The packages we used are below:

```
from google.colab import drive
drive.mount('/content/drive')
import numpy as np
import pandas as pd
pd.core.common.is_list_like = pd.api.types.is_list_like
import matplotlib.pyplot as plt
import pandas_datareader as pdr
import datetime
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
import os
import shutil
% matplotlib inline
```

```
import fix_yahoo_finance as fyf
from pandas_datareader import data as pdr
fyf.pdr_override()
#for setting figure size
from matplotlib.pylab import rcParams

from pandas_profiling import ProfileReport
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.linear_model import Lasso
from sklearn.linear_model import ElasticNet
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.ensemble import AdaBoostRegressor
from statsmodels.graphics.tsaplots import plot_acf  #autocorrelation
from statsmodels.tsa.stattools import adfuller as ADF  # stable test
from statsmodels.graphics.tsaplots import plot_pacf    #pacf
from statsmodels.stats.diagnostic import acorr_ljungbox    #white noise
from statsmodels.tsa.arima_model import ARIMA
from statsmodels.tsa.stattools import adfuller
from sklearn.metrics import mean_squared_error
from math import sqrt
!pip install pyramid-arima
from pyramid.arima import auto_arima
import math
from statsmodels.tsa.arima_model import ARIMA
%pip install numpy-financial
import numpy_financial as npf
```