# CAN WE PREDICT EQUITY INDEX VOLATILITY USING ML?

## BACKGROUND

Financial markets have very complex structures that are highly nonlinear (stochastic) with a low signal-to-noise ratio — this makes predicting asset returns (holy grail) difficult. However, it is well established in literature that volatility is easier to predict than returns — and this would be the basis of our project. Numerous factors tend to suggest that volatility is inherently more forecastable, for example, slow decay of autocorrelation in absolute returns.

The aim of our project is to predict/forecast daily (and potentially weekly) realized volatility of the S&P 500 Index over different market regimes (bull and bear market periods). In the sections below, we explain our choice of explanatory variables including our data sources and the frequency of each feature. Then, we run exploratory data analysis on each of the feature variables and analyze the different distributions of the data — including comments on how these can affect volatility. Then, we round up our observations from EDA and apply feature engineering to prepare the data to be used in machine learning algorithms. Next, we present a baseline model for volatility prediction and later share the results from linear models and non-linear models, compared with the baseline model. Finally, we discuss the fairness of the models, its destruction power  and the future improvements that we see in the model.

## SELECTION OF EXPLANATORY VARIABLES

For the prediction of volatility, we have used the 1 minute tick data for S&P 500 from Kaggle*. We first compute daily returns from 1 minute tick data by compounding minute returns. This is then used to compute the daily standard deviation in the S&P 500 price returns — this gives us the daily volatility for the S&P 500 Index from 2008-2020. As detailed in our project proposal, the data covers different market regimes including the financial crisis, sovereign debt crisis, flash crashes as well as the recent Covid crisis in 2020. Based on academic literature, we have considered the following macroeconomic variables, market variables and lagged volatility terms in order to predict volatility.

| Feature | Frequency | Rationale |
|---|---|---|
| Volatility lag variables | Derived data | There are periods when we observe volatility clusters - which mean that past volatile periods may help predict next period volatility. We check the correlation using ACF to include lags |
| CBOE volatility (VIX) index* | Daily | The VIX Index, also known as the "Fear Index" is a gauge of the next month's implied volatility based on the options market. This feature has high predictive power for the actual S&P 500 volatility. |
| China (SSE) and Taiwan (TAIPEX) index returns* | Daily | Majority of companies (in S&P 500) use semiconductor chips developed in Taiwan and have manufacturing units in China.  For example: the recent semiconductor chip crisis where IBM has ordered less semiconductor chips from its supplier in Taiwan due to which all tech companies are getting impacted. |
| US 1-month T-bill rate* | Daily | The US Treasury rates have a huge impact on the real interest rates which impacts the volatility. As the rates rise, investors move to bonds and the markets will crash leading to increase in volatility. |
| Recession Indicator* | Daily | Technically, recession is defined as a decline in GDP over two successive quarters. This boolean indicator helps us identify business cycles over time which in turn affects the stock market volatility. |
| Ex-dividend dates* | Quarterly | Boolean indicator which represents 1 for dates on which dividends are announced for the S&P 500 ETFs. When dividends are announced, the stock rally and hence the volatility of the stock increases. |
| ADS Business Conditions index* | Daily | The index is designed to track real business conditions in the US which has real implications on the stock market performance and volatility. As the real business conditions worsens, the stock market will crash causing the volatility to rise. |

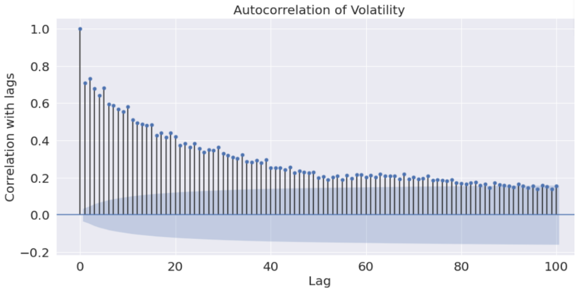
| | | |
|---|---|---|
| Economic Policy Uncertainty Index (EPU)* | Monthly | EPU denotes the contribution of government policy makers to the uncertainty regarding fiscal, regulatory, or monetary policy. Higher economic policy uncertainty leads to increases in stock volatility. Consists of two features - one economic data based and one news sentiment based. |

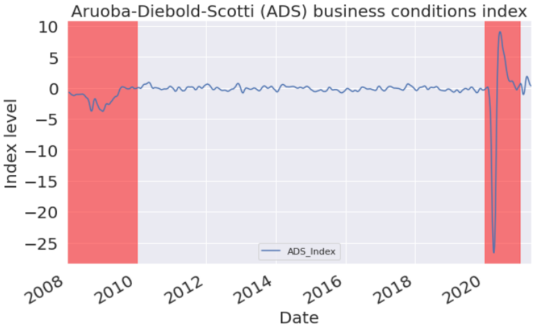
*(* [Link inserted here] Note: Please download the PDF from Github to access links - Github PDF renderer does not allow you to click the link.)*

## EXPLORATORY DATA ANALYSIS

The following section covers the data analysis done for every feature and how we dealt with missing/corrupted data.

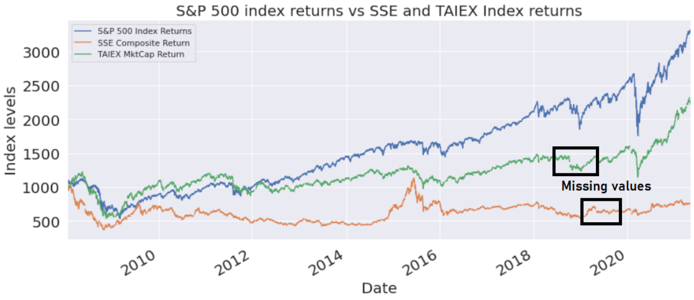
**Autocorrelation with lagged volatility variables -** Based on correlation values from the autocorrelation plot shown below, we choose weekly (5-days) and monthly (20-days) lagged volatility variables as additional input features. However, note that the monthly lagged variable has a slightly weak correlation.





**Trend and distribution of ADS Business Condition Index -** In the line plot below, we can see in highlight the crisis periods when ADS index levels crashed dramatically (2008 financial crisis and 2020 covid crisis) and the market was very volatile. The distribution of index values shows heavy tails.

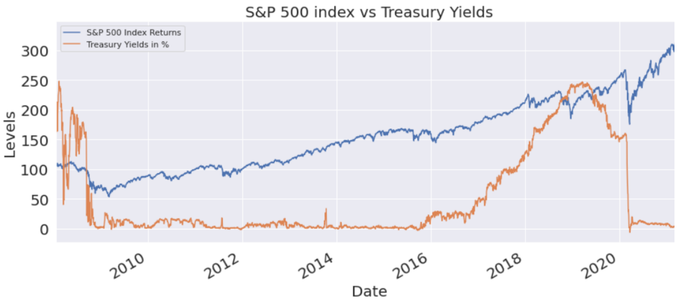
**Comparison of China and Taiwan Index with S&P 500 index -** In the line plot below, we can see how all indices dropped during the 2008 financial crisis. The S&P 500 index and Taiwan index dropped during the 2020 Covid crisis also. However, the China index was relatively stable then. Also, as we see from the graph, there are gaps in the middle, which is justified because the trading days in US are different from China and Taiwan. These are filled using the most recent previous value available in the data because in the absence of price levels we are assuming that the return is 0 so the price level remains the same as the last day.



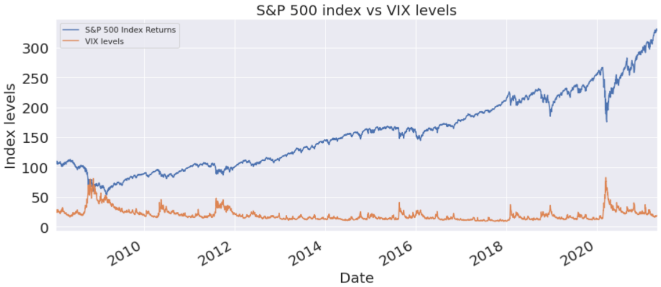
**US Dividend Indicator -** We have taken the dividend history of S&P ETF. This was then transformed to a boolean time series where the value was 1 for the date of the dividend.

**US Recession Indicator -** This was a daily data depicting recession in the US economy through boolean values (1 depicting recession and 0 depicting no recession). The missing values on federal holidays and/or stock market holidays are filled with 0. To analyse the data, we have taken the maximum of these boolean values for each year and tried to understand the years of recession of the US economy. The stem chart clearly depicts the years of recession for US economy now.

**Relation between treasury yields and S&P 500 returns** - Treasury yields have a direct inverse relationship with S&P 500 returns. As the treasury yields increase, the returns decrease and this is depicted

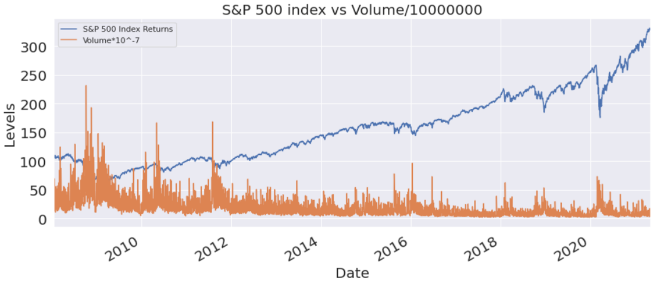


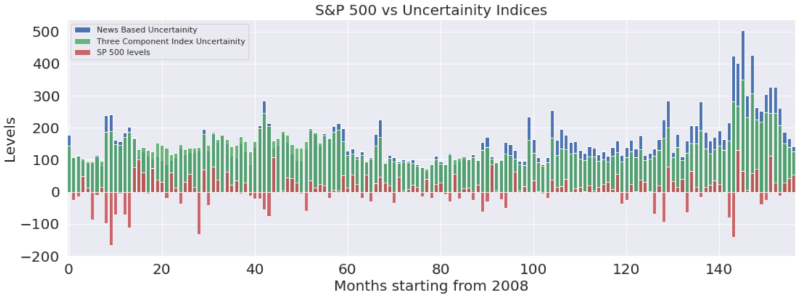
from the line chart as well. Hence the economic intuition of the relationship of treasury yields with volatility stands correct. The treasury yields data and the S&P 500 returns are using daily data and the missing daily levels (probably in case of Fed holidays) of treasury yields are filled using previous day's close price since in the absence of prices we will trade the yields on the price levels of last day.



**Relation between S&P 500 returns and changes in VIX index levels -** The direct relation between the S&P 500 index returns and the VIX (volatility index levels) levels are seen from the line chart. There exists a close negative relationship between the VIX index levels and S&P 500 return levels.

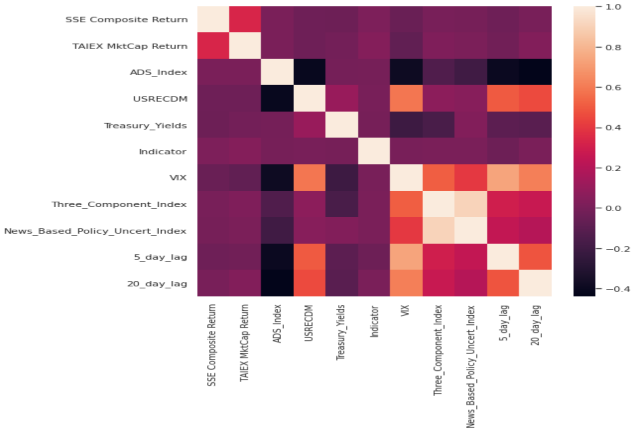
**Relation between S&P 500 and Volume -** As the volume of trading increases in S&P 500 levels, we see sudden drops in S&P 500 levels which means the volatility will increase. This goes with the economic intuition as well - as the supply and demand dynamics change the price levels of S&P changes and hence the volatility.



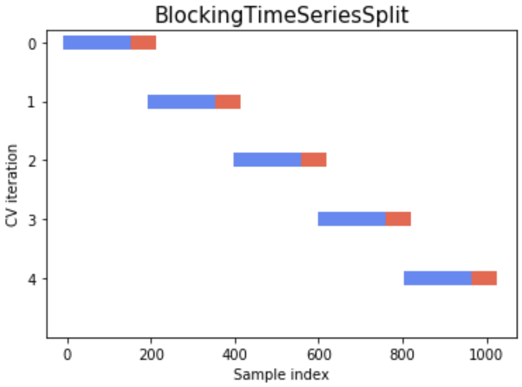


**Relation between Uncertainty indices and S&P 500 -** From the bar plot, we can see that as the new based uncertainty increases/decreases the cumulative returns of S&P 500 which will increase the volatility of the S&P 500. The Uncertainty Indices were all monthly levels so for exploratory data analysis, we compounded the daily returns of S&P 500 to monthly to analyse the impact of the uncertainty indices on S&P 500 returns.

**Correlation between different input features** - Plotting the heat map (as shown on bottom right), we observe that the Economic Policy Uncertainty Index (both economic data based three component index and news based uncertainty index) are significantly positively correlated with VIX while ADS Business Condition Index is significantly negatively correlated with VIX. Hence, we drop these highly correlated features to reduce multicollinearity in our linear models.



**How many features and examples are present? -** We have used 11 input features and have divided the dataset into 80% for training purposes and 20% for the testing purposes. For a time series dataset, we need to carefully divide our dataset. We cannot simply pick a random subset of data for training and testing. We took the first 80% of the rows for testing and the last 20% of the rows (recent years) constitute our test dataset. We have not explicitly created a validation split because OLS model has no hyperparameters to tune via cross-validation and for all the remaining linear as well as non-linear models, we have used inbuilt



cross-validation functions that identify the best hyperparameters by splitting data on its own. Going by the academic literature, we used blocked cross-validation (example shown on left) to prevent leakage from future data to the model; otherwise, the model will observe future patterns to forecast and try to memorize them. We started with 2070834 rows since we started with 1 min tick data for S&P 500 prices from 2008 to 2020.

## DATA SUMMARY AND FEATURE ENGINEERING

| Feature | Preliminary feature engineering | Final feature engineering |
|---|---|---|
| CBOE volatility (VIX) index | These features are normalised by using the following scaling factor: (X - Min) / (Max - Min)<br>For these features, we fill missing values with the previous day's value (which is the best estimator in case the prices are not available) | These features are standardised by using the following scaling factor: (X - Mean) / (Standard Deviation)<br>Here, the Mean and the Standard Deviation are that of the train data<br>No changes made to how we fill missing values |
| China (SSE), Taiwan (TAIPEX) index returns | | |
| ADS Business Conditions index | | |
| US 1-month T-bill rate | | |
| Economic Policy Uncertainty Index (EPU) | | |
| Volatility lag variables | | |
| NBER based Recession Indicator | For these boolean features, we impute missing values by 0 | |
| Ex-dividend dates | | |

We have decided to change the way we scale the training and test data set from our mid term report (hence the change in the summaries from the mid term report). In our mid term report, we had normalised the entire dataset without segregating it first into training and test data. If we take the maximum and minimum of the whole dataset we will be introducing future information into the training explanatory variables (i.e. the maximum and minimum).

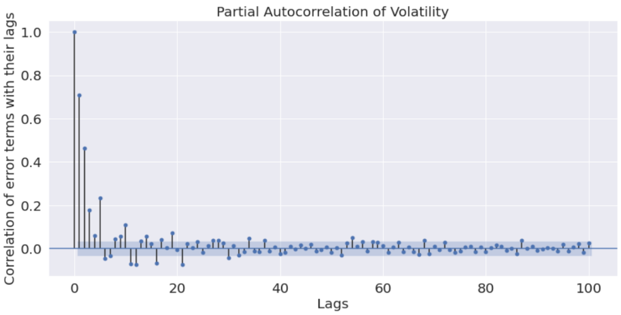
Therefore, we should perform feature scaling over the training data, then perform standardisation on testing data, but this time using the mean and variance of training data only. In this way, we can test and evaluate whether our model can generalize well to new, completely unseen data points.

## BASELINE GARCH MODEL

$$X_t = e_t\sigma_t$$

$$\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \ldots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2.$$

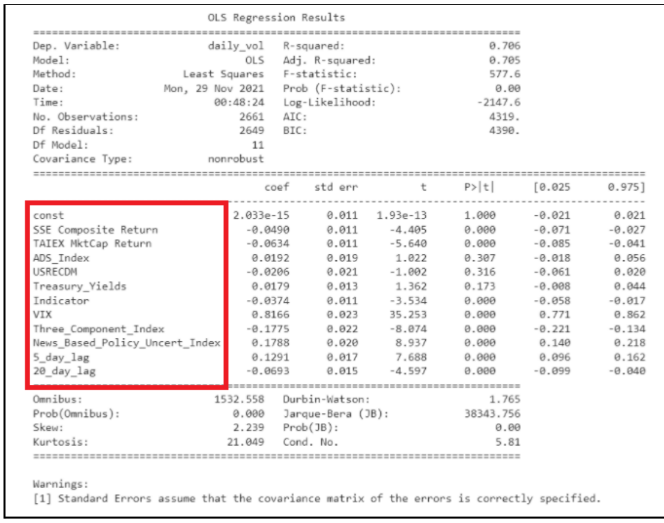
In the finance industry, the norm is to use the GARCH model (shown on right) for prediction of volatility. We have thus compared the results of our models with the GARCH model, treating it as the baseline model for our purposes. There are lots of academic studies which suggest the use of TGARCH and EGARCH model voicing the reason that financial time series data is asymmetric (a greater impact by negative news than its counterpart is observed by exhibiting statistically significant asymmetric effects at one percent level) but GARCH model is still in use in the industry and also the symmetric GARCH model outperformed all the other varieties of GARCH models in forecasting the volatility of S&P 500 Index.



We have thus considered the GARCH model for comparing the model performance of our linear and non linear models. The GARCH (p,q) model is modelled as shown above, where p is the number of lagged return terms being considered and q is the number of lagged error variance terms being considered. For our model, we have used the GARCH (2,1) model. The parameters p and q for the GARCH model have been found using the Autocorrelation Function (ACF chart in the EDA analysis section above) as well as the Partial Autocorrelation Function (PACF)- shown on the left.

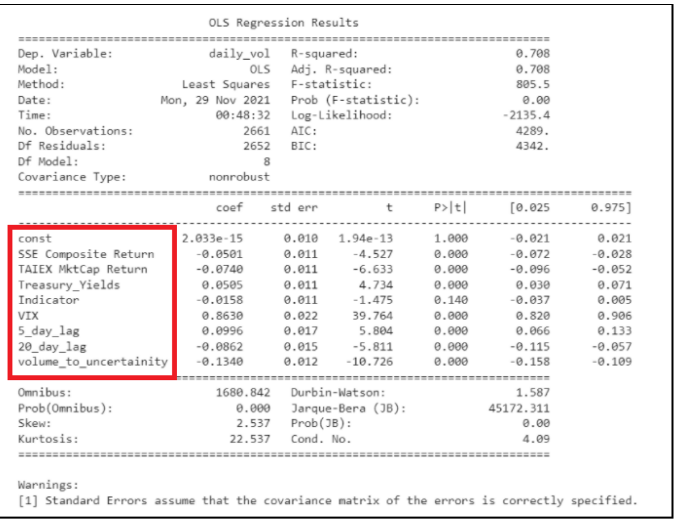
## LINEAR PREDICTION MODELS

For the initial model development, we first consider running an Ordinary Least Squares regression including all features. The result is not optimal as expected since we have many correlated variables (presence of multicollinearity). We cross-validated our intuition by running a heatmap of correlations between our features (in the EDA section) and ran the new model, dropping some of the correlated features. The results turn out to be much better. We also transformed one of the features to deal with multicollinearity. The feature was made using uncertainty indices as well as the volume of the S&P 500 contracts traded every minute on the exchanges. We took the ratio of total volume of all the contracts traded in a day to the sum of the news-based policy uncertainty index as well as the three component uncertainty index. The initial as well as final regression summary is provided below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              daily_vol   R-squared:                       0.706
Model:                            OLS   Adj. R-squared:                  0.705
Method:                 Least Squares   F-statistic:                     577.6
Date:                Mon, 29 Nov 2021   Prob (F-statistic):               0.00
Time:                        00:48:24   Log-Likelihood:                -2147.6
No. Observations:                2661   AIC:                             4319.
Df Residuals:                    2649   BIC:                             4390.
Df Model:                          11
Covariance Type:            nonrobust
=================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                         2.033e-15      0.011   1.93e-13      1.000      -0.021       0.021
SSE Composite Return            -0.0490      0.011     -4.405      0.000      -0.071      -0.027
TAIEX MktCap Return             -0.0634      0.011     -5.640      0.000      -0.085      -0.041
ADS_Index                        0.0192      0.019      1.022      0.307      -0.018       0.056
USRECDM                         -0.0206      0.021     -1.002      0.316      -0.061       0.020
Treasury_Yields                  0.0179      0.013      1.362      0.173      -0.008       0.044
Indicator                       -0.0374      0.011     -3.534      0.000      -0.058      -0.017
VIX                              0.8166      0.023     35.253      0.000       0.771       0.862
Three_Component_Index           -0.1775      0.022     -8.074      0.000      -0.221      -0.134
News_Based_Policy_Uncert_Index   0.1788      0.020      8.937      0.000       0.140       0.218
5_day_lag                        0.1291      0.017      7.688      0.000       0.096       0.162
20_day_lag                      -0.0693      0.015     -4.597      0.000      -0.099      -0.040
==============================================================================
Omnibus:                     1532.558   Durbin-Watson:                   1.765
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            38343.756
Skew:                           2.239   Prob(JB):                         0.00
Kurtosis:                      21.049   Cond. No.                         5.81
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Initial Model



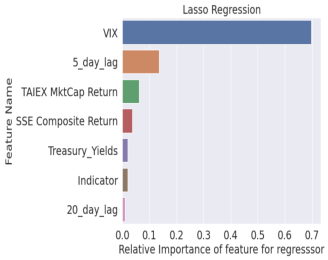
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              daily_vol   R-squared:                       0.708
Model:                            OLS   Adj. R-squared:                  0.708
Method:                 Least Squares   F-statistic:                     805.5
Date:                Mon, 29 Nov 2021   Prob (F-statistic):               0.00
Time:                        00:48:32   Log-Likelihood:                -2135.4
No. Observations:                2661   AIC:                             4289.
Df Residuals:                    2652   BIC:                             4342.
Df Model:                           8
Covariance Type:            nonrobust
===========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                   2.033e-15      0.010   1.94e-13      1.000      -0.021       0.021
SSE Composite Return      -0.0501      0.011     -4.527      0.000      -0.072      -0.028
TAIEX MktCap Return       -0.0740      0.011     -6.633      0.000      -0.096      -0.052
Treasury_Yields            0.0505      0.011      4.734      0.000       0.030       0.071
Indicator                 -0.0158      0.011     -1.475      0.140      -0.037       0.005
VIX                        0.8630      0.022     39.764      0.000       0.820       0.906
5_day_lag                  0.0996      0.017      5.804      0.000       0.066       0.133
20_day_lag                -0.0862      0.015     -5.811      0.000      -0.115      -0.057
volume_to_uncertainity    -0.1340      0.012    -10.726      0.000      -0.158      -0.109
==============================================================================
Omnibus:                     1680.842   Durbin-Watson:                   1.587
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            45172.311
Skew:                           2.537   Prob(JB):                         0.00
Kurtosis:                      22.537   Cond. No.                         4.09
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

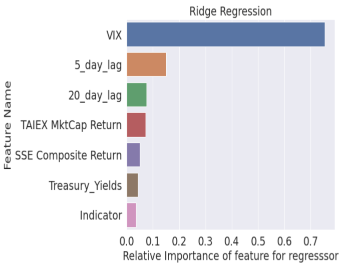
Revised OLS model

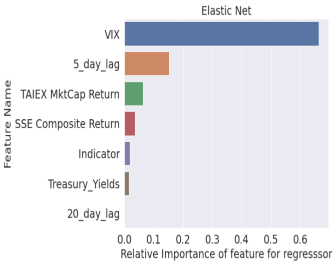
The revised model has less multicollinearity and features with significant p statistics. Our model's  adjusted R2 has also increased slightly in the process. The Jarque-Beta (JB) test result shows that our data is not symmetrically distributed around the normal, which confirms our assumption that financial data (especially volatility) is not perfectly normally distributed.

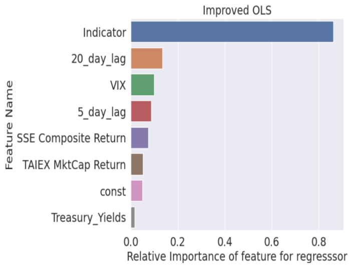
To reduce more multicollinearity in the simple linear model, we also introduced three regularized models — Ridge, Lasso and ElasticNet. The penalty term in ElasticNet is simply a combination of the Ridge and Lasso penalties (as shown on the right). Unsurprisingly, Lasso and ElasticNet set most of the coefficients to zero barring the coefficients for treasury yields which increase drastically and hence this is not suitable. Ridge gives the best performance. Following are the importance assigned to each feature by each linear model. As expected, all the models give maximum importance to VIX, because the VIX is an indicator of the next month's implied volatility and it has very high predictive power for the actual S&P 500 volatility.

$$\phi(\beta;\lambda,\alpha)=\lambda\left(\alpha\sum_{i=1}^{J}\beta_i^2+(1-\alpha)\sum_{i=1}^{J}|\beta_i|\right)$$





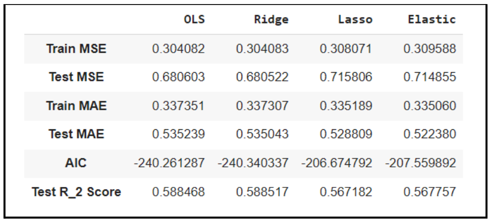




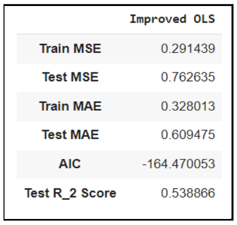
After running these regression models, we present our model effectiveness scores using different evaluation metrics on the test dataset.

The MSEs and MAEs of all our linear models are quite low and the AIC's (Akaike information criterion is an estimator of prediction error and thereby relative quality of models) are negative, which is an indicator of the good quality of the model. The test MSEs for the linear models as well as the train MSEs are quite low and do not change a lot, meaning that our linear model is not over-fitted and is quite generalised. Same is the case with MAEs where the test and training MAEs are quite low. It is interesting to see how the train MSEs are smaller than the train MAEs but the test MAEs are smaller than the train MSEs.

We have however, removed some of the features from our linear model to remove multicollinearity. These features were again revisited for non-linear models. For the model where feature transformation was done, even though the train MSEs and MAEs perform better than the ordinary least squares regression, ridge and lasso regressions the test MSEs and MAEs are larger for the feature transformed model which shows that the model can not generalise well. The AIC score has also been reduced for the featured transformed model. The R2 score for the feature transformed model, however, is better than the least squares model and the ridge and lasso regression models. When compared to the GARCH model which we have used as baseline model, our model has better predictive power than GARCH (train MSE 1.304) model.

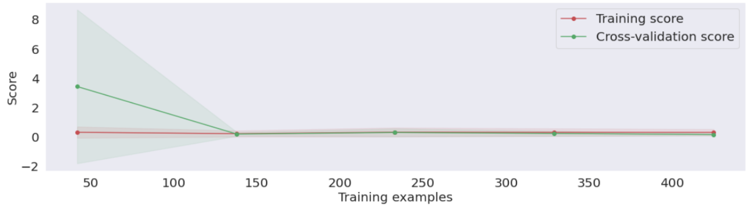


| | OLS | Ridge | Lasso | Elastic |
|---|---|---|---|---|
| Train MSE | 0.304082 | 0.304083 | 0.308071 | 0.309588 |
| Test MSE | 0.680603 | 0.680522 | 0.715806 | 0.714855 |
| Train MAE | 0.337351 | 0.337307 | 0.335189 | 0.335060 |
| Test MAE | 0.535239 | 0.535043 | 0.528809 | 0.522380 |
| AIC | -240.261287 | -240.340337 | -206.674792 | -207.559892 |
| Test R_2 Score | 0.588468 | 0.588517 | 0.567182 | 0.567757 |

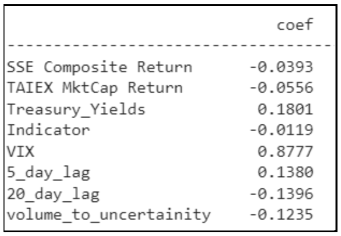


| | Improved OLS |
|---|---|
| Train MSE | 0.291439 |
| Test MSE | 0.762635 |
| Train MAE | 0.328013 |
| Test MAE | 0.609475 |
| AIC | -164.470053 |
| Test R_2 Score | 0.538866 |

Of all the linear models, Ridge regression performs the best according to MSE, AIC as well as R_2 score and its learning curve is as shown below.

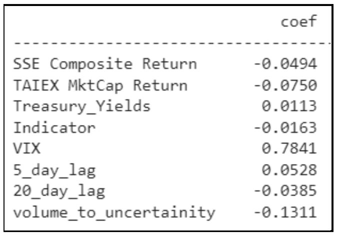


**Comparison of coefficients of the linear model for different market regimes -** Since we have considered the time frame from 2008 to 2021, the time series includes various crashes as well as rallies. We have tried to segregate the very well known financial crisis of 2008-2009, the covid crisis of 2020-2021 as well as the rally period in between and compared how the coefficients of the linear model changed between the various crashes and rallies in the market. We have also compared the measure of goodness of fit i.e. adjusted R-squared for the three time periods.



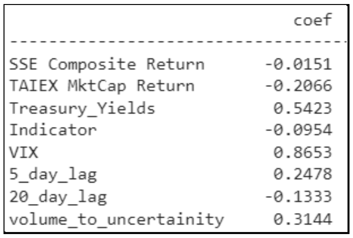
```
                                 coef
-----------------------------------
SSE Composite Return          -0.0393
TAIEX MktCap Return           -0.0556
Treasury_Yields                0.1801
Indicator                     -0.0119
VIX                            0.8777
5_day_lag                      0.1380
20_day_lag                    -0.1396
volume_to_uncertainity        -0.1235
```

**Regime 1 (Jan 2008 to Dec 2010)**
**Adj. R-squared (uncentered) : 0.744**



```
                                 coef
-----------------------------------
SSE Composite Return          -0.0494
TAIEX MktCap Return           -0.0750
Treasury_Yields                0.0113
Indicator                     -0.0163
VIX                            0.7841
5_day_lag                      0.0528
20_day_lag                    -0.0385
volume_to_uncertainity        -0.1311
```

**Regime 2 (Jan 2011 to Dec 2019)**
**Adj. R-squared (uncentered) : 0.642**



```
                                 coef
-----------------------------------
SSE Composite Return          -0.0151
TAIEX MktCap Return           -0.2066
Treasury_Yields                0.5423
Indicator                     -0.0954
VIX                            0.8653
5_day_lag                      0.2478
20_day_lag                    -0.1333
volume_to_uncertainity         0.3144
```

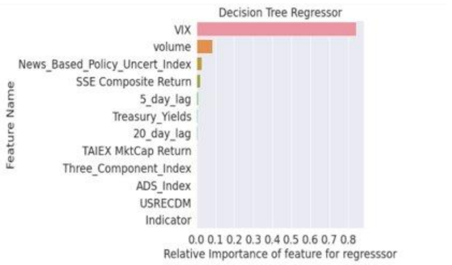
**Regime 3 (Jan 2020 to May 2021)**
**Adj. R-squared (uncentered) : 0.708**

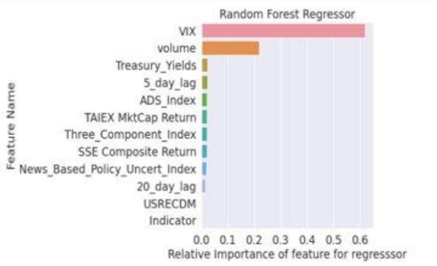
As is visible from the tables, we see that for all the three regimes the volatility in the market is negatively related to the returns of Chinese and Taiwan Indices. For treasury yields, we see a strong correlation with volatility during the crisis time period of 2008-2009 and 2020-2021 which makes sense as well since the yields are increased by the Fed during times of high volatility which coincides with crisis time periods. For all the three time regimes we see that the strongest predictor of realised volatility is VIX (implied volatility) which checks out with various academic studies that have been done in this domain. It is also observed that the 5 day lag in realised volatility has a higher coefficient during crisis times of 2008-2009 and 2020-2021. This is also observed in the market — in case of crisis times we see that there are sustained periods of high volatility and hence higher dependence on immediate past returns. Same is the case for 20 day lag as well but the 20 day lags are negatively correlated with the realised volatility. The transformed feature of volume to uncertainty however performs differently during Covid crisis as compared to the other durations.

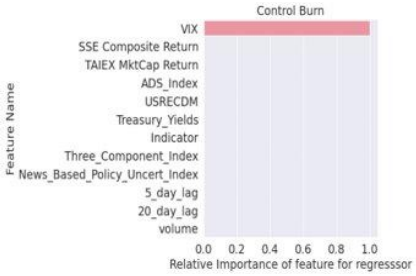
## NON LINEAR PREDICTION MODELS

Next, we will show the result of non-linear model implementation. One of the advantages of non-linear models is that we don't need to manually correct for multicollinearity and it is easy to prevent overfitting of the data by hyperparameter tuning. For the non-linear models, we also added cross validation sets for hyperparameter tuning since we cannot use regular k-fold cross validation for time series analysis.

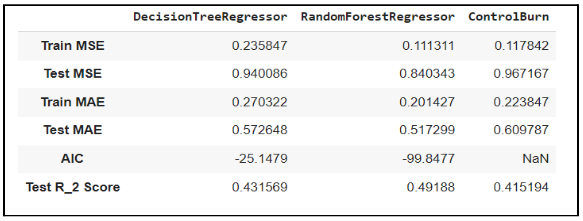
We implemented three non-linear models from class - DecisionTreeRegressor, RandomForestRegressor, and ControBurn. For the first two models, we made use of random search cross validation in order to find the best hyperparameters. For Random Forest, the focus is limited to just a random subset of input features (and not a subset of data points, hence no issues even for time series) for finding the best split point, so it is able to decorrelate the trees. RandomizedSearchCV using a blocking time series split cross validation selects the best hyperparameter combination for each model which saves time on manually tuning the model. The reason for choosing Random Grid Search rather than the standard Grid Search is that while it's possible that RandomizedSearchCV will not find as accurate of a result as GridSearchCV, it surprisingly picks the best result more often than not and in a fraction of the time it takes what GridSearchCV would have taken. Given the same resources, Randomized Search can even outperform Grid Search. Generally GridSearchCV works best with small datasets (<5000 samples) and lots of computational resources. With large data sets like the high dimensional data that we have (60,000 x 29), GridSearchCV will greatly slow down computation time and be very costly. Following are the importance assigned to each feature by each non-linear model. Same as before, all models give maximum importance to VIX.





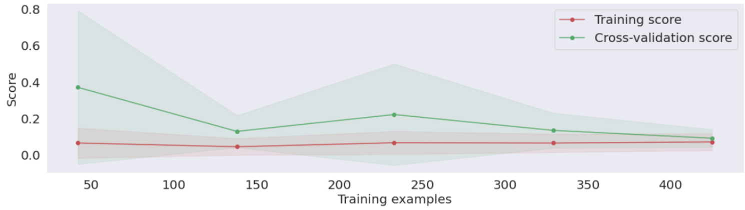


As expected, we saw that the Random Forest performed much better (on the test set) compared to the Decision Trees. This is because random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than a single decision tree since they can handle overfitting and show less variance. As we saw in the class, the non-linear models fit the data very well (better than linear regression) but they do not generalize well. The test MSE for all three non-linear models is greater than that for the linear models (OLS, Ridge, Lasso, ElasticNet and the well engineered model with selective features).



| | DecisionTreeRegressor | RandomForestRegressor | ControlBurn |
|---|---|---|---|
| Train MSE | 0.235847 | 0.111311 | 0.117842 |
| Test MSE | 0.940086 | 0.840343 | 0.967167 |
| Train MAE | 0.270322 | 0.201427 | 0.223847 |
| Test MAE | 0.572648 | 0.517299 | 0.609787 |
| AIC | -25.1479 | -99.8477 | NaN |
| Test R_2 Score | 0.431569 | 0.49188 | 0.415194 |

The test MAE on the other hand is smaller than the linear models for Random Forest Regressor but more for Decision Tree Regressor as well as Control Burn. The test R_2 score however is better for all non-linear models than linear models. The AICs for non linear models are less negative than the linear models.
It is also known from academic literature that the non-linear models which are related to the time component (lag component in our case) are not able to generalise well. In such cases, linear regression or neural networks should be used. When compared to the GARCH model which we have used as the baseline model, our nonlinear model has better predictive



power than the GARCH (train MSE 1.304) model. Of all the non-linear models, Random Forest performs the best according to MSE, MAE as well as AIC and its learning curve is on the right hand side.

## FAIRNESS OF MODELS

For our models, we have only used the macroeconomic variables as well as returns from publicly available indices as the independent variables. We have not used any variable which we think might introduce any discrimination or bias. Hence, we do not believe that our algorithm will bias information or will try to avoid unintended negative consequences. In terms of legal requirements as well since we are using the publicly available data, we do not believe that any legal requirements would be violated by our model. Additionally, the output of our model is not feeding into the financial markets so that it won't have a market impact.

**What is the harm of false positives and false negatives? -** The impact of false positives or false negatives will be that the future prediction of reality volatility will not be accurate and hence the trading strategy using our model will get disadvantaged because of it. If we use this model directly in production for making investment decisions, we may suffer losses due to it.

**How can errors change the data distribution in the future? -** Since our model uses lagged terms of realised volatility, the errors in the prediction will lead to errors in the future as well but only for the trading strategies employing our model.

However, none of these have any adverse impact on model fairness. Thus, we believe that fairness of the models is not relevant to our realised volatility prediction model.

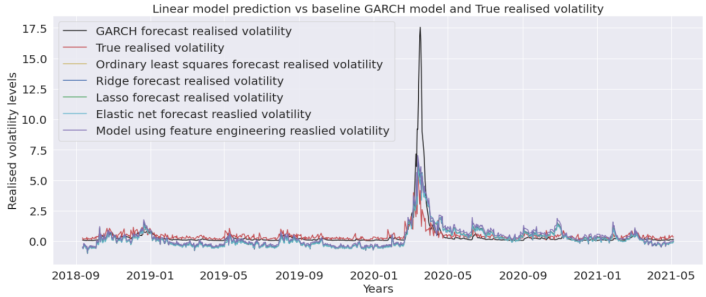
## WEAPON OF MATH DESTRUCTION

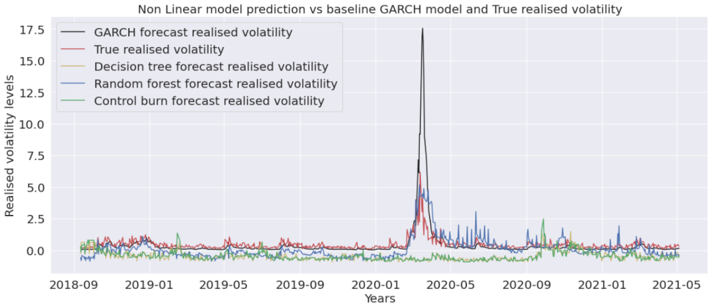
We do not believe our project has created a Weapon of Math Destruction since our models do not fit the three main criteria of a *Weapon of Math Destruction:*

1. Our model has easily measurable outcomes. The goal of our model is not ambiguous and we are able to directly measure our results.
2. The predictions given by our model do not have negative or unintended consequences. Our predictions do not affect the wellbeing of others or the health of any of the companies we performed analysis on.
3. Our models do not create feedback loops and therefore we are not worried about creating a feedback loop.

## SUMMARY AND CONCLUSION

In this project, we have tried to come up with a model that predicts volatility on the basis of the market conditions (deriving its value from macroeconomic values) as well as from its own past values (momentum). The data that we analysed for our model development includes 10+ years of data which captures the evolution of the market. It was also observed that the predictive power of our linear models was better than the non linear models. For comparison of errors between the different models, different error estimates have also been compared and the best performance is observed under ridge regression (under linear models). Surprisingly, we have been able to come up with a model which has better predictive power than the widely used GARCH models. In fact all our models, be it linear or nonlinear, have better predictive power than the GARCH model. Even in outlier situations, the predicted values are much closer to the true realized volatility as opposed to the GARCH model predictions which shoot through the





roof. For comparison purposes, we have also plotted the predictions of our models with the baseline GARCH model as well as the true realised volatility that was observed. Having said this, our model is still in the model development stage and we should make sure that the developed model makes sound financial sense.

## FUTURE IMPROVEMENTS

For future improvements on the preliminary models that we've constructed, we would like to implement the best performing models on weekly volatility as well since that might be more predictable than the daily volatility, which can fluctuate a lot and generally need additional manual intervention/decision making. The models we have implemented have shown a fair amount of predictive power, however, we can make use of more powerful/complex models in order to have a more accurate model. We can extend our project to use stacked ensemble methods (popular as competition winning algorithms) and deep neural networks to improve the prediction accuracy. Ensemble methods are commonly used to boost predictive accuracy by combining the predictions of multiple machine learning models. The traditional wisdom has been to combine so-called "weak" learners. However, a more modern approach is to create an ensemble of a well-chosen collection of strong yet diverse models.

We can also work towards choosing different models (models with different coefficients) in case of different market regimes - we can essentially switch models based on bull/bear markets or market bubble signals dynamically to ensure high predictive accuracy. This will ensure that the model is dynamically chosen depending on the market conditions enabling the trading strategies using our model to make better decisions. The Log Periodic Power Law Singularity (LPPLS) Model from Sornette et al. (python implementation here) can help in analyzing market bubble behaviour to model signals to switch models. However, this is a notoriously difficult task (even for institutional funds) and still in the nascent research stage.

As part of future improvements, we should also be recalibrating the model periodically to make sure that a black swan event if happens or if a situation which has not been captured in the past during the model development stage arises, the model is agile enough to readjust so that the future predictions from the model are not impacted severely.

## REFERENCES

**Research papers we referenced:**

- Bollerslev, T., 1986, "Generalized autoregressive conditional heteroscedasticity," Journal of Econometrics, 31(3), 307
- Breiman, L., 1996, "Bagging predictors," Machine Learning, 24(2), 123-140.
- Breiman, L., 2001, "Random forests," Machine Learning, 45(1), 5-32.
- Breiman, L., and A. Cutler, 2004, "Random Forests,"
- Kim Christensen, Mathias Siggaard and Bezirgen Veliyev, "A machine learning approach to volatility forecasting", CREATES Research Paper 2021-03

**Links we used:**

- https://medium.datadriveninvestor.com/why-wont-time-series-data-and-random-forests-work-very-well-together-3c9f7b271631
- https://machinelearningmastery.com/develop-arch-and-garch-models-for-time-series-forecasting-in-python/
- https://machinelearningmastery.com/findings-comparing-classical-and-machine-learning-methods-for-time-series-forecasting/
- https://medium.datadriveninvestor.com/why-wont-time-series-data-and-random-forests-work-very-well-together-3c9f7b271631
- https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/
- https://stackoverflow.com/questions/49444262/normalize-data-before-or-after-split-of-training-and-testing-data
- https://www.sciencedirect.com/science/article/pii/S1042443110000430