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 both weekly and daily realized volatility of the S&P 500 Index over different market regimes (bull and bear market periods). In the sections below, we first explain our choice of explanatory variables including our data sources and the frequency of each feature. Second, 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. Third, we round up our observations from EDA and apply feature engineering to prepare the data to be used in machine learning algorithms. Fourth, we show the results from our preliminary linear models. Finally, we discuss steps on how we plan to avoid over (under) fitting, use non-linear models for volatility prediction and further model development.solute returns - this implies that financial time series incorporate memory of the history.

Selection of explanatory variables

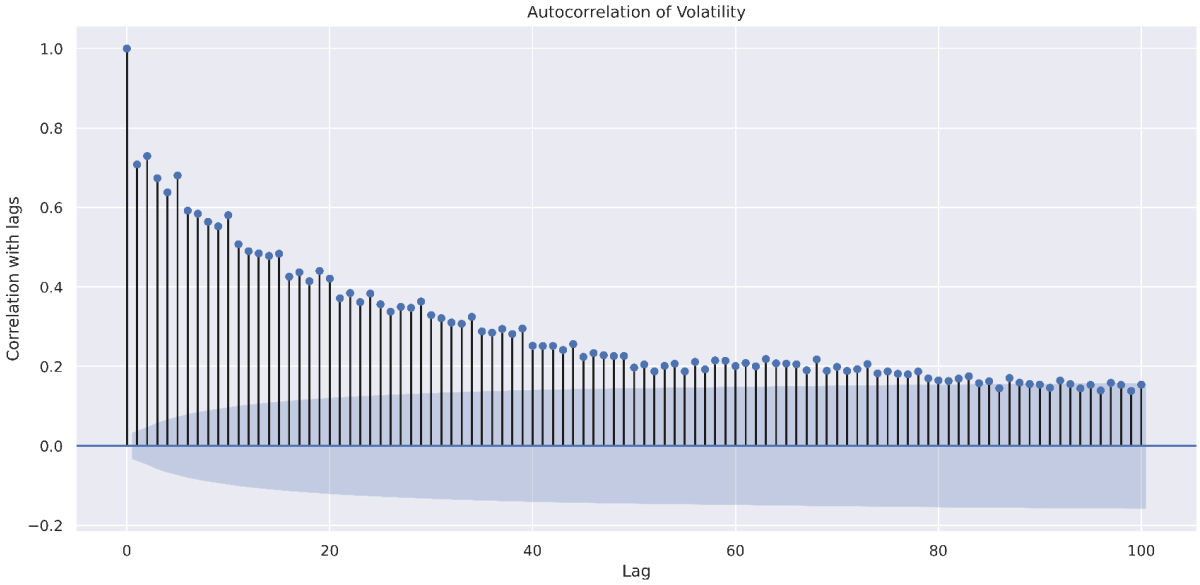
For the prediction of volatility, we have used the 1 minute tick data for S&P 500 from Kaggle. We first compute daily returns from 1 minute tick data by compounding minute returns. This is then used to compute the daily standard deviation in the S&P 500 price returns — this gives us the daily volatility 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
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 semi- conductor chip crisis where IBM has ordered less semiconductor chips from its supplier in Taiwan due to which all tech companies are getting impacted.
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.
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.
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 (details in EDA).
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.
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
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.

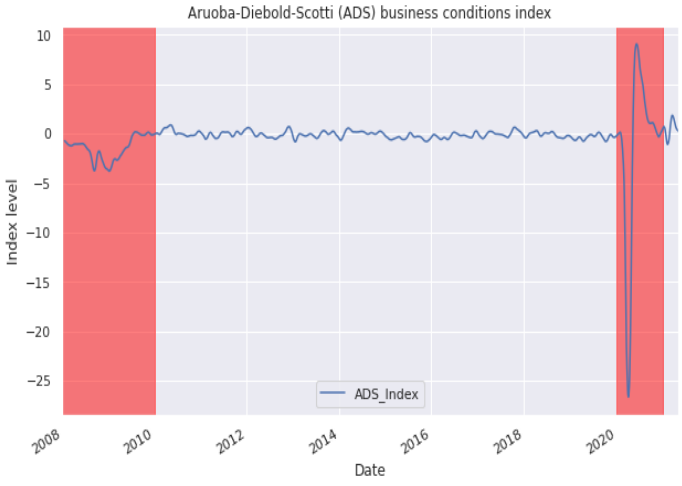
Exploratory Data Analysis

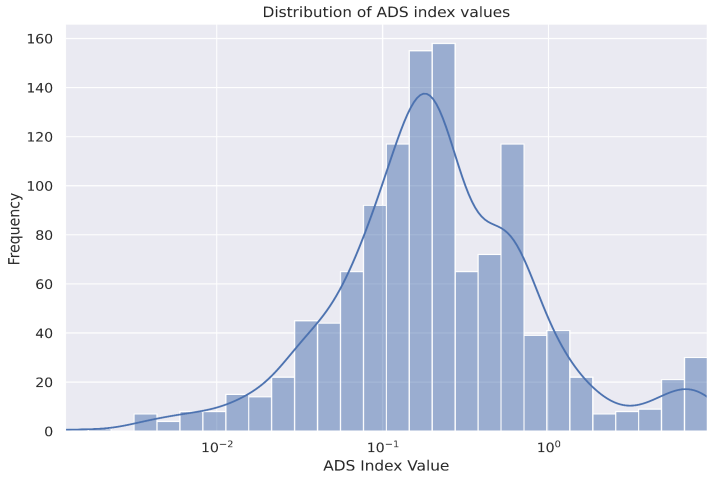
The following section covers the data analysis done for every feature and how we dealt with missing/corrupted data.

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.

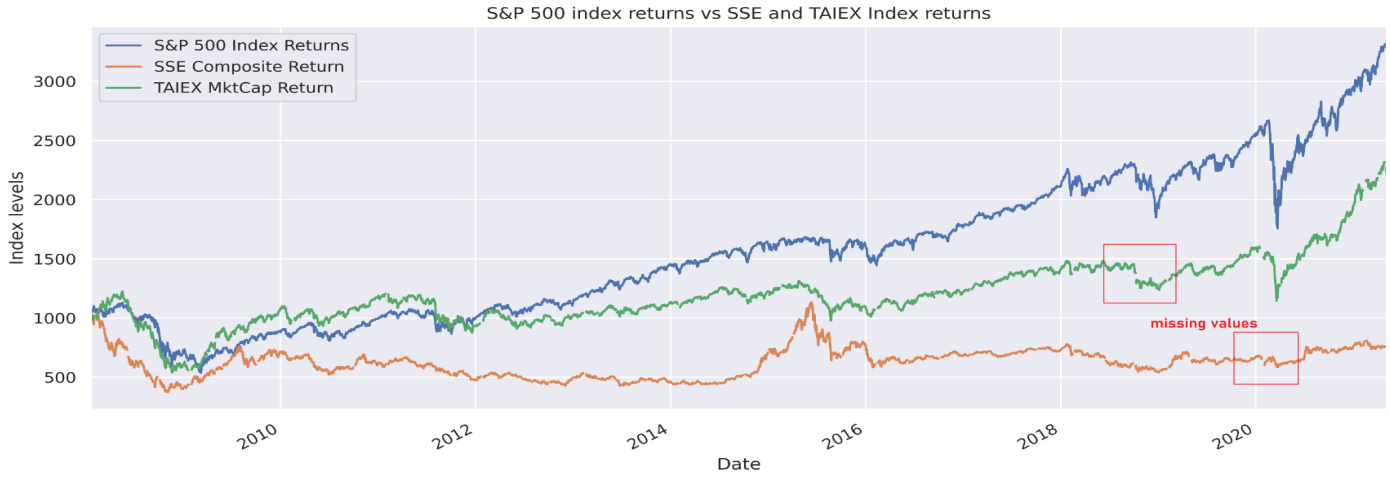


Trend and distribution of ADS Business Condition Index - In the line plot below, we can see in highlight the crisis periods when ADS index levels crashed dramatically (2008 financial crisis and 2020 covid crisis) and the market was very volatile. The distribution of index values shows heavy tails.

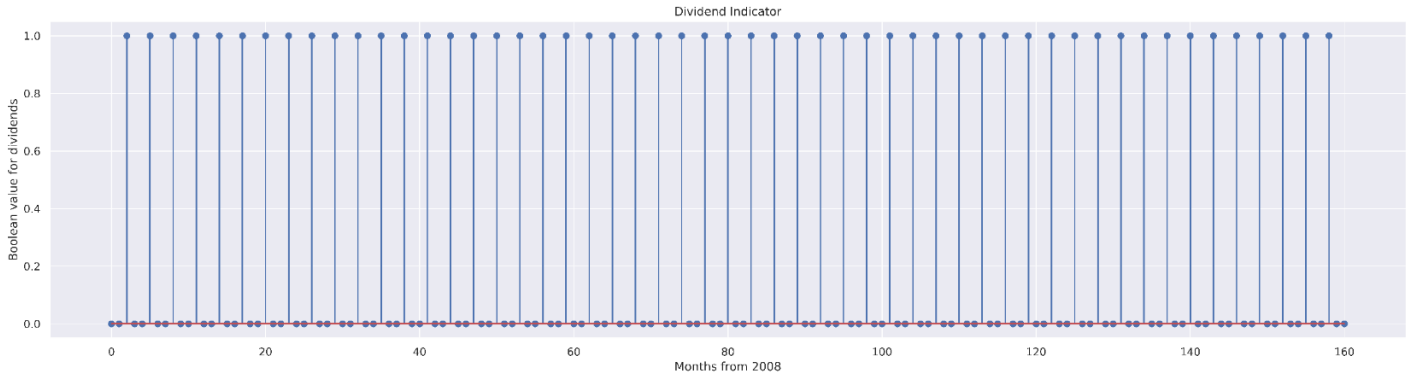




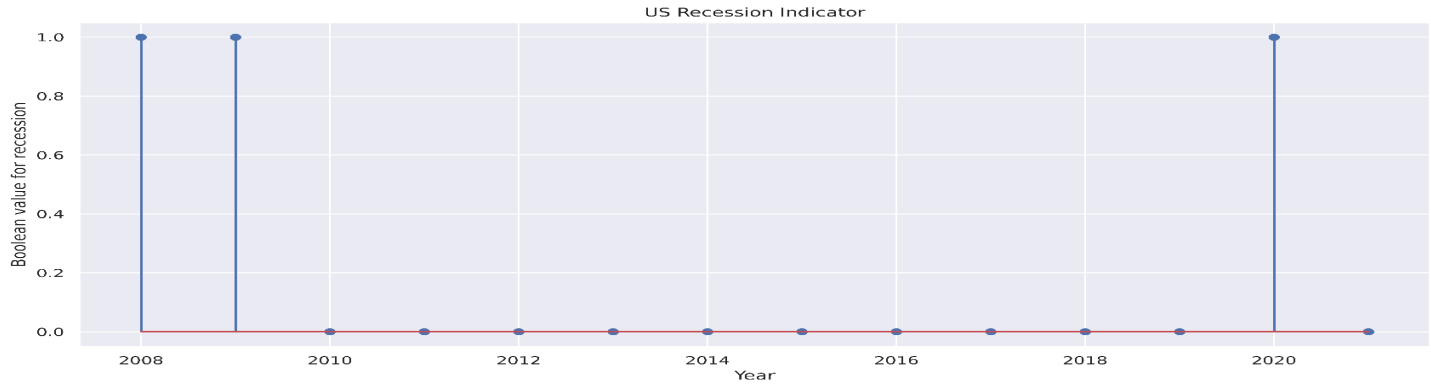
Comparison of China and Taiwan Index with S&P 500 index - In the line plot below, we can see how all indices dropped during the 2008 financial crisis. The S&P 500 index and Taiwan index dropped during the 2020 Covid crisis also. However, the China index was relatively stable then. Also, as we see from 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.



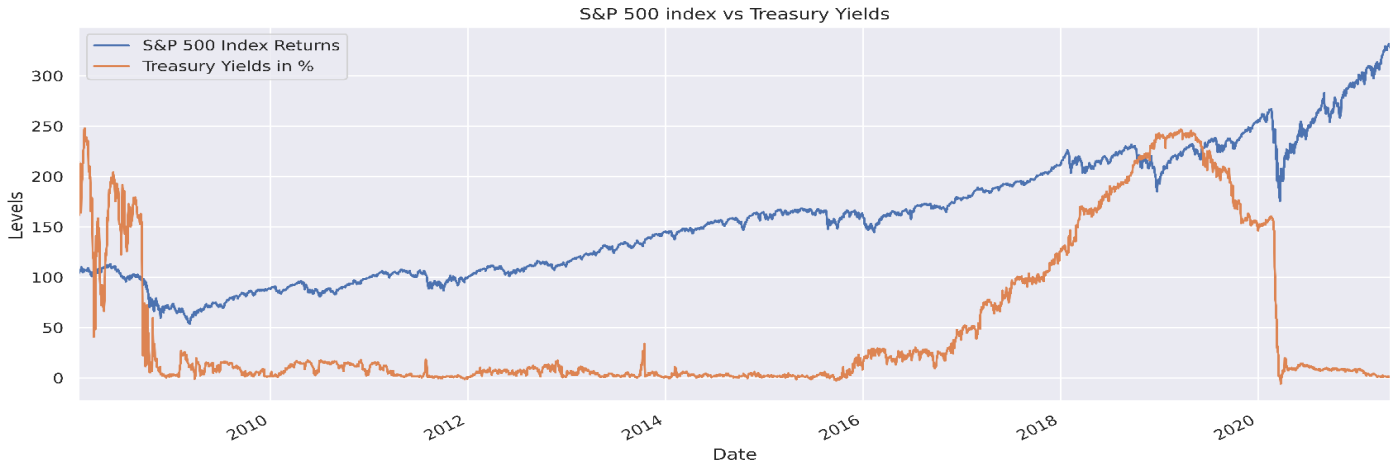
US Dividend Indicator - We have taken the dividend history of S&P ETF. This was then transformed to a boolean time series where the value was 1 for the date of the dividend. This is then plotted as a stem plot for boolean values.



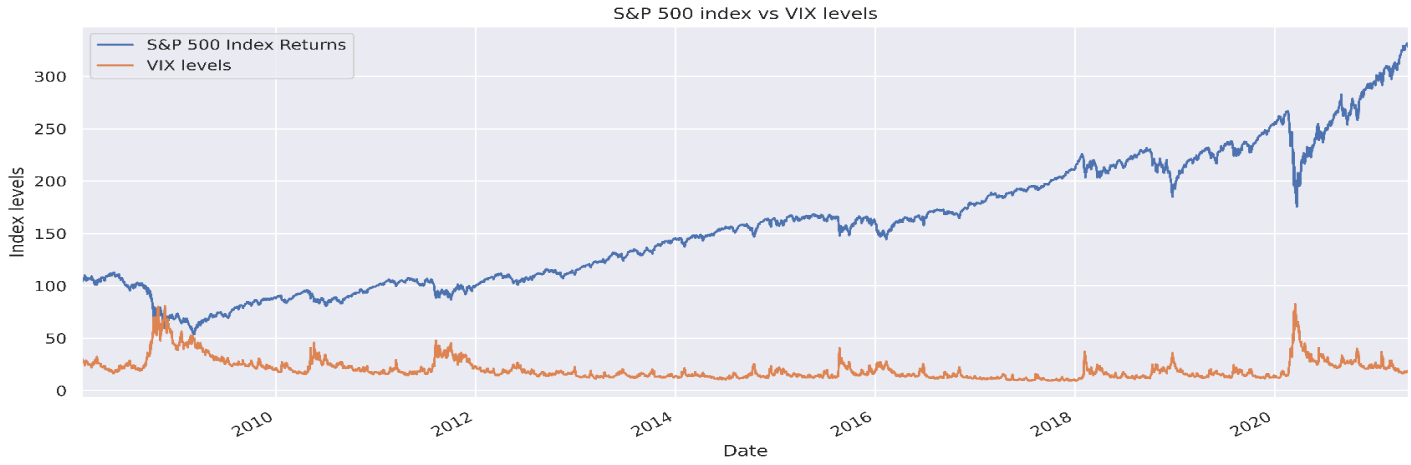
US Recession Indicator - This was a daily data depicting recession in the US economy through boolean values (1 depicting recession and 0 depicting no recession). The missing values on federal holidays and/or stock market holidays are filled with 0. To analyse the data, we have taken the maximum of these 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 the US economy now.



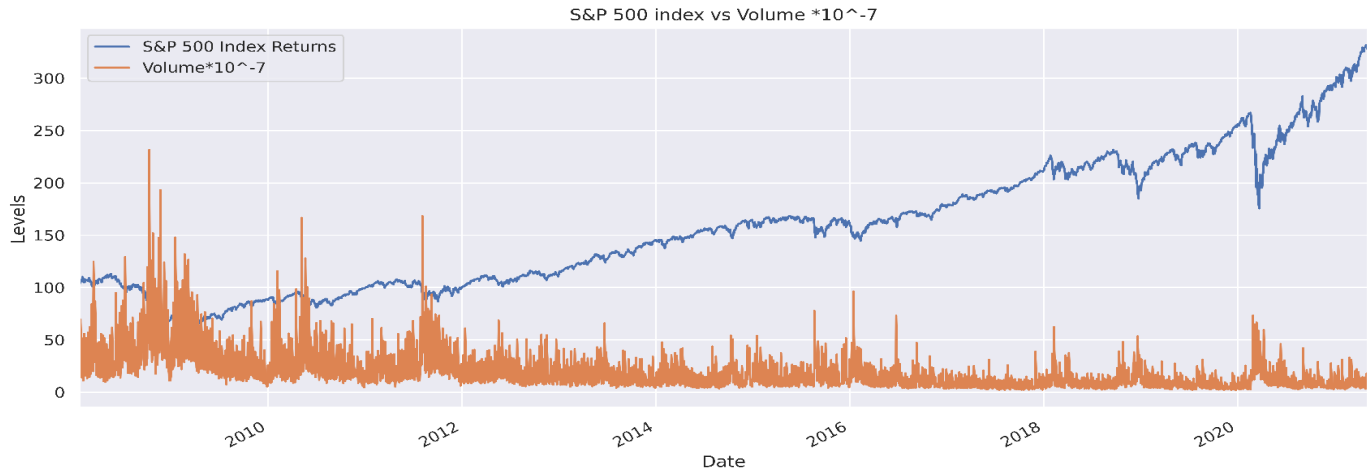
Relation between treasury yields and S&P 500 returns - Treasury yields have a direct inverse relationship with S&P 500 returns. As the treasury yields increase the returns decrease and this is depicted from the line chart as well. Hence the economic intuition of the relationship of treasury yields with volatility stands corrected. 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 the last day.



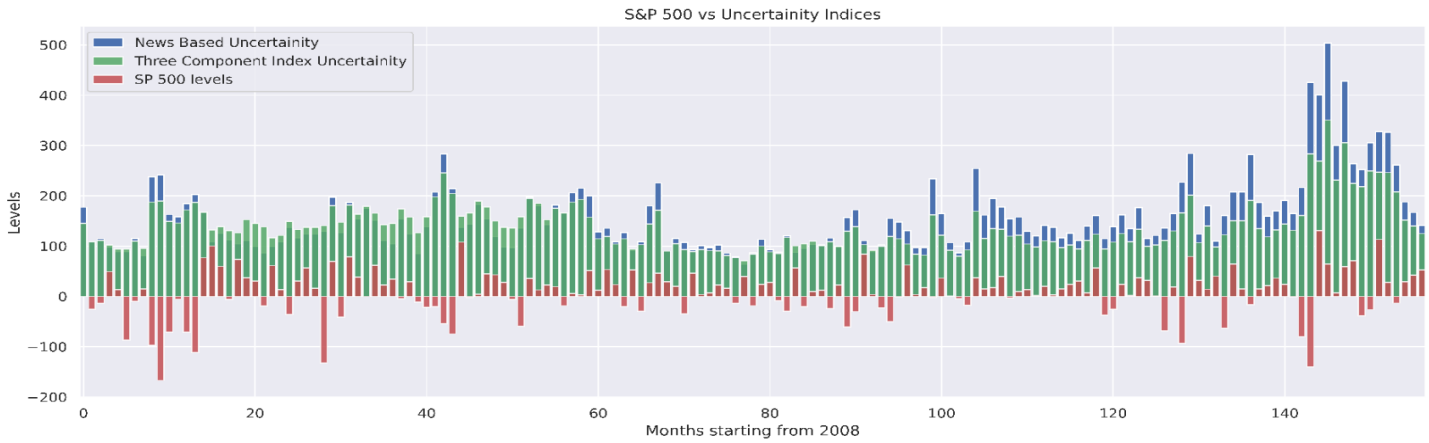
Relation between S&P 500 returns and changes in VIX index levels - The direct relation between the S&P 500 index returns and the VIX levels are seen from the line chart above. This clearly depicts the relation between S&P 500 and the VIX (volatility index levels)- the close relationship with the daily volatility will also apply to the daily standard deviation and S&P 500 levels.



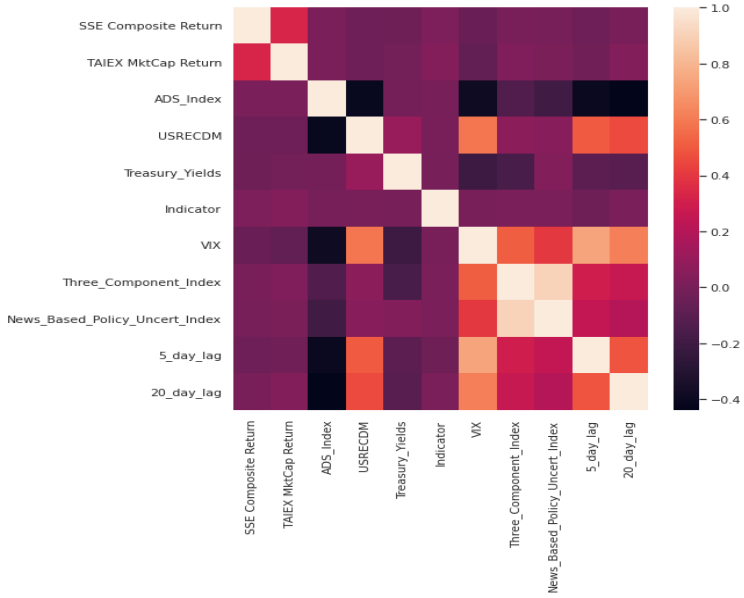
Relation between S&P 500 and Volume - As the volume of trading increases in S&P 500 levels, we see sudden drops in S&P 500 levels which means the volatility will increase. This goes with the economic intuition as well - as the supply and demand dynamics change the price levels of S&P changes and hence 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 eda, 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, 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. So we just keep VIX and drop all other features correlated with it to reduce multicollinearity.



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 started on with 2070834 rows since we started with 1 min tick data for S&P 500 price levels from 2008 to 2020. The transformation and feature engineering is described in the next section.

Date	daily_vol	SSE Composite Return	TAIEX MktCap Return	ADS_Index	USRECDM	Treasury_Yields	Indicator	VIX	Three_Component_Index	News_Based_Policy_Uncert_Index	5_day_lag	20_day_lag
2008-02-20	0.000933	0.452281	0.467351	-0.364483	1.0	2.368111	0.0	24.275731	107.849452	108.169588	-0.007342	-0.006692
2008-02-21	0.000638	0.464410	0.507796	-0.373953	1.0	2.412111	0.0	24.995732	107.849452	108.169588	-0.007314	-0.006089
2008-02-22	0.000387	0.438479	0.486361	-0.382433	1.0	2.341111	0.0	23.935730	107.849452	108.169588	-0.007376	-0.007087
2008-02-25	0.000716	0.432462	0.505446	-0.403023	1.0	2.379111	0.0	22.905732	107.849452	108.169588	-0.007708	-0.007441
2008-02-26	0.000581	0.484022	0.486121	-0.409693	1.0	2.292111	0.0	21.775731	107.849452	108.169588	-0.007217	-0.007254

Data Summary and Feature Engineering

Feature	Feature engineering
CBOE volatility (VIX) index	• These features are normalised by using a scaling factor of (X - Min) / (Max - Min) • 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)
China (SSE) and 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 simply impute missing values by 0
Ex-dividend dates	

Preliminary 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 with an AutoML feature selection package as well as 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. The regression summary is provided below:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.706
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	577.6
Date:	Mon, 01 Nov 2021	Prob (F-statistic):	0.00
Time:	23:30:16	Log-Likelihood:	517.68
No. Observations:	2661	AIC:	-1011.
Df Residuals:	2649	BIC:	-940.7
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8586	0.268	6.937	0.000	1.333	2.384
x1	-1.1389	0.259	-4.405	0.000	-1.646	-0.632
x2	-2.0542	0.364	-5.640	0.000	-2.768	-1.340
x3	0.0088	0.009	1.022	0.307	-0.008	0.026
x4	-0.0235	0.023	-1.002	0.316	-0.069	0.022
x5	0.0121	0.009	1.362	0.173	-0.005	0.030
x6	-0.1129	0.032	-3.534	0.000	-0.176	-0.050
x7	0.0306	0.001	35.253	0.000	0.029	0.032
x8	-0.0019	0.000	-8.074	0.000	-0.002	-0.001
x9	0.0015	0.000	8.937	0.000	0.001	0.002
x10	129.0415	16.785	7.688	0.000	96.129	161.954
x11	-68.9510	14.999	-4.597	0.000	-98.361	-39.541

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.	8.55e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.55e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Initial Model

OLS Regression Results

Dep. Variable:	y	R-squared:	0.696
Model:	OLS	Adj. R-squared:	0.695
Method:	Least Squares	F-statistic:	866.9
Date:	Mon, 01 Nov 2021	Prob (F-statistic):	0.00
Time:	23:30:17	Log-Likelihood:	473.43
No. Observations:	2661	AIC:	-930.9
Df Residuals:	2653	BIC:	-883.8
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.8070	0.963	-8.105	0.000	-9.696	-5.918
x1	-1.1537	0.262	-4.396	0.000	-1.668	-0.639
x2	-2.3035	0.369	-6.246	0.000	-3.027	-1.580
x3	0.0289	0.007	3.937	0.000	0.015	0.043
x4	-0.1086	0.032	-3.344	0.001	-0.172	-0.045
x5	0.0283	0.001	38.427	0.000	0.027	0.030
x6	149.9467	16.848	8.900	0.000	116.910	182.984
x7	-74.8262	15.036	-4.977	0.000	-104.309	-45.343

Omnibus:	1612.052	Durbin-Watson:	1.709
Prob(Omnibus):	0.000	Jarque-Bera (JB):	41201.363
Skew:	2.402	Prob(JB):	0.00
Kurtosis:	21.669	Cond. No.	9.49e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.49e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Revised OLS model

The revised model has less multicollinearity and features with significant p statistics, although our model's R^2 dropped slightly. The Jarque-Beta (JB) test result shows that our data is not Gaussian — which confirms our assumption that financial data (especially volatility) is not normally distributed.

To reduce multicollinearity in the model, we introduced three regularized models — Ridge, Lasso and ElasticNet. The penalty term in ElasticNet is simply a combination of the Ridge and Lasso penalties:

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

Surprisingly, Lasso regression (and ElasticNet) sets most of the coefficients to zero barring the coefficients for treasury yields which increase drastically- this is not suitable. After running these regression models, we present our model effectiveness scores using MSE and AIC attributes:

	OLS	Ridge	Lasso	Elastic
Train MSE	0.041020	45.891484	0.063275	0.058221
Test MSE	0.091811	47.160537	0.149484	0.141572
AIC	-1574.425902	2582.469270	-1249.776372	-1285.994017

The MSEs 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 do not increase significantly from train MSEs, meaning that our linear model is not over-fitted and is quite generalised. We have however, removed some of the features from our linear model to remove multicollinearity. These features will be re-visited again for non linear models which we plan to run in the next half of the semester.

Discussion and next steps for the semester

Discussion on current implementation: We have cleaned the data, run exploratory data analysis, applied feature engineering and looked at linear models discussed in class — linear regression, ridge & lasso regressions and elastic net regularization (which combines the best of lasso and ridge regularization parameters). We looked at multicollinearity between our features in our model — and have dropped some of our correlated features for linear models. When implementing the non linear model we will consider these features again.

Plan to avoid over/under fitting: We see that there is some overfitting in our linear models that we tried to overcome by using regularization parameters — further, we plan to mitigate this using log transformations of the features as well as looking into combining features (eg, take ratio of two features as a single independent variable) and running cross validation to select the best set of features. Since our time series data has different market regimes, we can also look at analyzing rolling regressions (by regimes) and check the stability of our model coefficients — this may help in effectively removing over/underfitting of the models. Time permitting, we may also look at PCA for selection of features for linear models.

Testing effectiveness of models: For our current models, we have implemented MSE and AIC (Akaike information criterion is an estimator of prediction error and thereby relative quality of models) for our linear models. In the future, we plan to add feature importance (for nonlinear models) as a means to check the features with best predictive power. We still need to explore other measures of accuracy and model effectiveness in order to decide on which model performs best.

Next steps in model development and what remains to be done: In the next phase of the project, we will implement non linear models which will include tree based regressions (ensemble methods like boosting, bagging and ControlBurn) to improve accuracy of the model predictions. 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. Time permitting, we may also implement a simple neural network which may have better predictive accuracy.