GR5293-FinalReport

1. **Business Understanding:**

Volatility Index (VIX) was created by the Chicago Board Options Exchange (CBOE) and it's the real-time market index that represents the market’s expectation of 30-day forward-looking volatility. Derived from the price inputs of S&P 500 index options, it provides a measure of market risk and investors’ sentiments. It’s also known by other names like “Fear Gauge” or “Fear Index”. Investors, research analysts, and portfolio managers look to VIX values as a way to measure market risk, fear, and stress before they take investment decisions. Volatility value, investors' fear and the VIX index values move up when the market is falling. The reverse is true when the market advances – the index values, fear and volatility decline.

VIX index has paved the way for using volatility as a tradable asset, although through derivative products. CBOE launched the first VIX-based exchange-traded futures contract in March 2004, which was followed by the launch of VIX options in February 2006. Such VIX-linked instruments allow pure volatility exposure and have created a new asset class altogether.  Like all indexes, one cannot buy the VIX directly. Instead, investors can take positions in VIX through futures or options contracts, or through VIX-based exchange-traded products (ETP). It is a measure of the level of implied volatility, not historical or statistical volatility, of a wide range of options, based on the S&P 500.

VIX index values are calculated using the CBOE-traded standard SPX options (that expire on the third Friday of each month) and using the weekly SPX options (that expire on all other Fridays). Only those SPX options are considered whose expiry period lies within 23 days and 37 days.

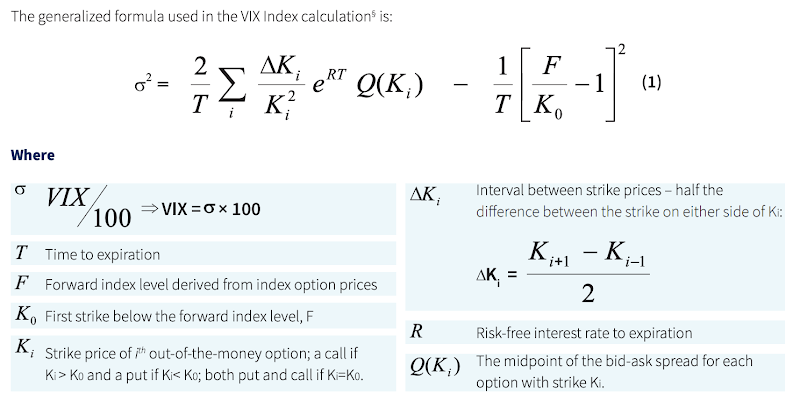


Figure 1: Formula for VIX Index Calculation

Whether the CBOE volatility index is set to rise can be an important question for traders and investors focused on short term investments.  Usually, a higher VIX reflects increased fear of investor and a lower VIX indicates complacency. Historically, this pattern in the relationship between the VIX and the behavior of the investors’ sentiment and the stock market, has repeated itself in bull and bear cycles. During periods of market turmoil, the VIX spikes higher, largely reflecting the panic in the market. During bullish periods, there is less fear on the contrary. For example, what happened last week could present us with a clear indication of how the S&P 500 index interacts with VIX. The rally Friday afternoon on May 10 pulled back the major indexes from their worst weekly declines since late December as the U.S.-China trade dispute escalated. The Dow and S&P 500 index each fell more than 2% as investors sold trade-sensitive shares in a broad sell-off that extended the market’s slide into a second week. Technology stocks led the way lower, with digital storage companies and chipmakers among the big decliners. Heavy equipment makers Deere and Caterpillar drove losses in the industrial sector. Within the S&P 500 index, information technology shares performed worst. Industrials and materials shares were also weak as investors worried about rising trade barriers. Volatility, as measured by the CBOE VIX, spiked to its highest level since late January.

Inspired by the significance of VIX as a leading indicator of the S&P 500, we want to see how well do Natural Language Processing (NLP) sentiment scores with Machine Learning Techniques (MLT) predict the future VIX percent change compared to the linear VIX forecasting model. Sentiment Analysis is also known as Opinion Mining, which is a field within NLP that builds systems that try to identify and extract opinions with text. With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics or any topic that people can express opinions about. For our target VIX in the project, we believe there exists a strong correlation between past US event sentiments and the VIX. The data we have gathered are from Google Bigquery which contains all past world events data with a Goldstein value, people involved and overall tone, etc. We selected out only the events that happened or heavily involved in the United States. As for features, we only selected the Goldstein value, overall tone and the number of articles. The number of articles decides the weight of the particular event while the other two values determine the sentiment of the particular event.  In our project, we want to use event data to predict the percent change in VIX and whether VIX is decreasing or increasing.

We first use the linear regression model as our baseline model to predict the percent change in VIX and get the prediction results with MSE 0.051 and R square 0.0188. Then we try to use machine learning models to see if we could outperform the baseline model. The main machine learning models in our project are: Gradient Boosting regression Tree, logistic regression, and Gradient Boosting Tree (classification). We mainly have two kinds of models: a quantitative model and a qualitative model. For the quantitative model, our response is the percent change of VIX. We use the quantitative model to predict the exact percent change in VIX. The models here are linear regression and Gradient Boosting Regression Tree, and the bases are  R-squared and MSE. For the qualitative model, our response is 1 if today‘s VIX is greater than or equal to yesterday’s and response is 0 if today’s VIX is smaller than yesterday’s.  We use the qualitative model to see if the VIX is increasing or decreasing. The models here are logistic regression, Support Vector Machine (SVM) and boosted decision tree. And the bases are accuracy and ROC-AUC. By building quantitative and qualitative models, we can fully explore how the event data affect the predictions of percent change in VIX. When we fit the model, we get better MSE and R-squared values compared to the baseline.

**2. Data Understanding**

The reason we want to predict the percent change of VIX is that, it can provide us with valuable information.  The VIX index measures the expectation of stock market volatility over the next 30 days implied by S&P 500 index options.The percentage changes in VIX index, driven by the increases and decreases in value of the securities it comprises, which are of most value to the market. It reflects the likely range of possible index levels the market “expects” in a month. We often use trading strategies requiring assumptions in changes in volatility relative to the market. And the  percent change of VIX can provide us with the information to gain insights from the market, and make better investment. The VIX data was extracted from Yahoo Finance ([https://finance.yahoo.com](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://finance.yahoo.com%26amp;sa%3DD%26amp;ust%3D1604003880658000%26amp;usg%3DAOvVaw3HV2s0qwU5CBTunQtWWydo&sa=D&ust=1604003880700000&usg=AOvVaw2hSOtecHbYnccobvdDoLq9)), and we calculated its percent change to better serve for the following model building process.

We use the Google Bigquery Database to collect features data ([https://bigquery.cloud.google.com/ welcome/ecbm4040-ta](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://bigquery.cloud.google.com/welcome/ecbm4040-ta%26amp;sa%3DD%26amp;ust%3D1604003880658000%26amp;usg%3DAOvVaw3MurtvQXLvPRBN6HpeM-wT&sa=D&ust=1604003880701000&usg=AOvVaw3I8j7gwTy4Z6ZghV-1GyYa)). Google BigQuery is a serverless, highly-scalable, and cost-effective cloud data warehouse with an in-memory BI Engine and [machine learning](https://www.google.com/url?q=https://www.google.com/url?q%3Dhttps://cloud.google.com/bigquery/%2523bigqueryml%26amp;sa%3DD%26amp;ust%3D1604003880659000%26amp;usg%3DAOvVaw1JWwtuQ1wG2BeU-0N2xhlt&sa=D&ust=1604003880701000&usg=AOvVaw2q-HLHyvD6SJ9YwylUNzS-) built in. The dataset includes all the events from 2015 to 2019 with very specific information, but for the purpose of this project, we will only extract the Date, GoldsteinScale, Average Tone, Number of articles(this is for the weight of the event) and QuadClass.  We selected out only the events that happened or heavily involved the United States. We use SQL to extract event data that happened in the US between Jan 1 2015 and Feb 2019. Quadclass was initially a categorical feature, but we average the amount of each quadclass category happening in a day to be a new feature and use one hot encode quad class into 4  binary categorical features. Also, we aggregate each feature by daily average. Lastly, we add daily median and 10%, 25%, 75%, 90% quantile of GoldsteinScale, NumArticles, and AvgTone to the forecasting model so as to develop a thorough model.

In the project, we used both numerical and categorical features to predict the percent change of VIX.  Numerical data have meaning as a measurement. Moreover, by sorting collected data into groups or categories, categorical data represents characteristics. These two types of data can help us select the influencing factors more thoroughly. The detailed descriptions of the predictors are as follows:

**GoldsteinScale:**It ranges from -10 to 10 and captures the potential impact of that type of event will have on the stability of a country, which is based on the type of event, not the specifies of the actual event record being recorded. VIX provides estimate of expected market volatility. Usually, a lower VIX generally corresponds to stable, stress-free periods in the markets. Similarly, a higher VIX indicates a volatile market. The ‘GoldsteinScale’ feature can reflect the stability of the country, and therefore helps to predict the percent change of VIX.

**Avg Tone:**It’s the average “tone” of all documents  containing one or more mentions of this event during the 15 minute update in which it was first seen. The score ranges from -100 (extremely negative)  to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This can be used as a method of filtering the “context” of events as a subtle measure of the importance of an event and as a proxy for the “impact” of that event. For example, a riot event with a slightly negative average tone is likely to have been a minor occurrence, whereas if it had an extremely negative average tone, it suggests a far more serious occurrence,  which might cause market turbulence and further affect the percent change of VIX.

**NumArticles:**It’s the total number of source documents containing one or more mentions of this event  during the 15 minute update  in which it was first seen. It can be used as a method of assessing the “importance” of an event: the more discussion of that event, the more likely it it to be significant. A frequently mentioned topic would be more likely to reflect people’s sentiment towards a certain event, and sway the opinions of decision makers (i.e.: investors). The total universe of source documents varies over time, so it’s recommended that this field be normalized by the average or other measure of the universe of events during the time period of interest.

**QuadClass:** This field specifies this primary classification for the event type. The numeric codes in this field map to the Quad Classes as follows: Quad 1 = Verbal Cooperation, Quad 2 = Material Cooperation, Quad 3 = Verbal Conflict, Quad 4 = Material Conflict. We encode QuadClass into four binary categorical features. What is more, QuadClass was initially a categorical feature. We averaged the amount of each quadclass category happening in a day to be a new feature

After knowing about the descriptions of the predictors, let’s see the correlation coefficients between different variables. From the figure, we can notice that there exists a positive relationship among the percent change of VIX and ‘Quad 4’ (Material Conflict) and ‘Close price’. Meanwhile, there is a negative relationship among the percent change of VIX and ‘Goldstein Scale’, ‘Number of Articles’,  ‘Average Tone of documents’, ‘Quad 1’ (Verbal Cooperation), ‘Quad 2’ (Material Cooperation), ‘Quad 3’(Verbal Conflict), and the ‘Average Source’. From the heat map below of Pearson correlation of features, we can analysis the relationship among features in a more clearly way.

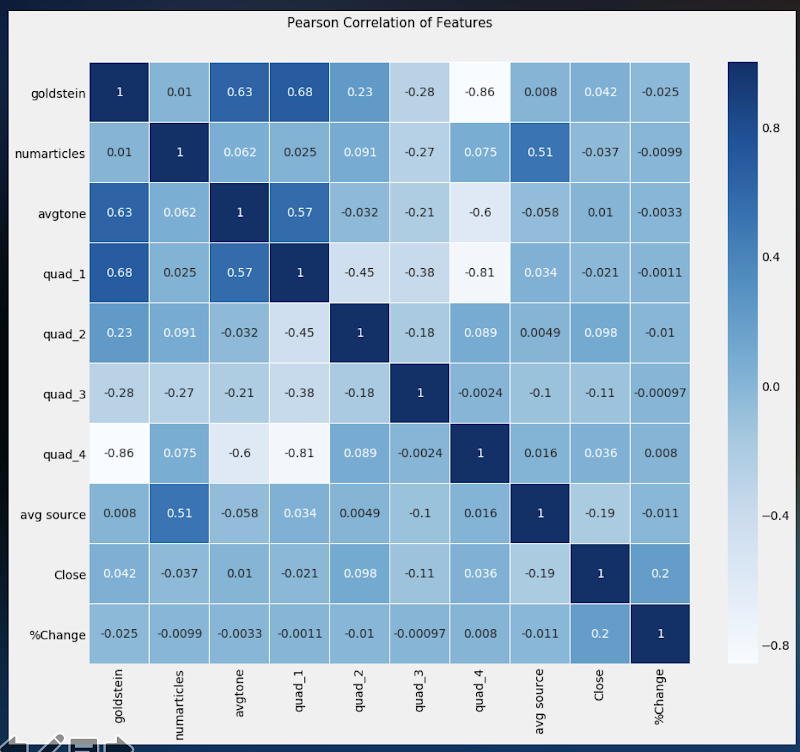


Figure 2. Correlation of variables

Considering data manipulation, we used severals methods to dealt with it. There exist missing variables in our dataset.Firstly, we compare the missing and non-missing cases on variables where information is not missing. We added plausibility to our results (or detect potential biases) by comparing sample members on variables that are not missing. Secondly, we dropped some ‘NA’ variables. Since it is no great loss if the NA variables had little effect on the percent change of VIX, for those missing variables, a substantial proportion of cases lack data, we simply chose to drop the variables. Thirdly, we tried to use several ways to substitute the missing variables by using the mean of its previous variable and its next variable. We also changed the discrete variable (for instance:quad class variable) into dummy variables when doing the regression and prediction part. What is more, we grouped the data by daily average, and added daily median and 10%, 25%, 75%, 90% quantile of the influencing features (Goldstein Scale, NumArticles, and AvgTone) to the forecasting model and  develop a thorough model, by depicting the movement of VIX index and make further prediction.

Furthermore, We find that there exists signaling effect of the sentiment to the change of VIX. Therefore, we using lag to manipulate the date, and better depict the lag effect. Response lag is the time it takes or responding to certain economic signals. Time lags may be days, weeks or months between the response data and the features selected, depending on the nature and significance of the features. Thus, we implemented lag operator which shifts the time series in model training and predicting results. With this time series methods, we lagged the data with different orders and calculated cross-validation scores of all candidate models with different lag orders. We select the lag We compare the signaling effect by different days. After calculating average MSE, we found that using the data of sentiment shifted 26 days before VIX have a lower MSE value, and therefore have a good predicting power.

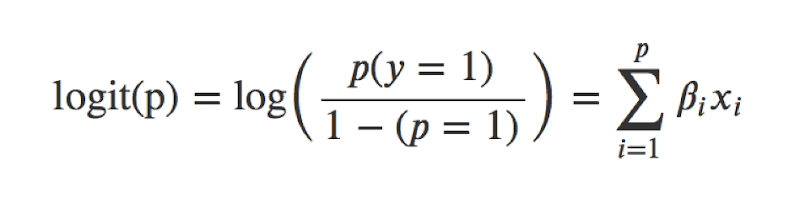
**3. Model Understanding**

3.1 Qualitative Model

As for our qualitative model, we aim to predict whether the daily VIX will go up or down, which is a binary classification problem. We used three machine learning techniques, namely Logistic Regression, Support Vector Machine and Gradient Boosting Decision Tree. For each method, we not only learned the algorithm and the methodology behind, but also tuned the hyper parameters according to the metric of classification accuracy. We then compare the classification accuracy as well as the (AUC-ROC) Area Under the Receiver Operating Characteristics for the models.

1. Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Mathematically, logistic regression estimates a multiple linear regression function defined as:

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using it’s underlying logistic function.These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function. The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier. 

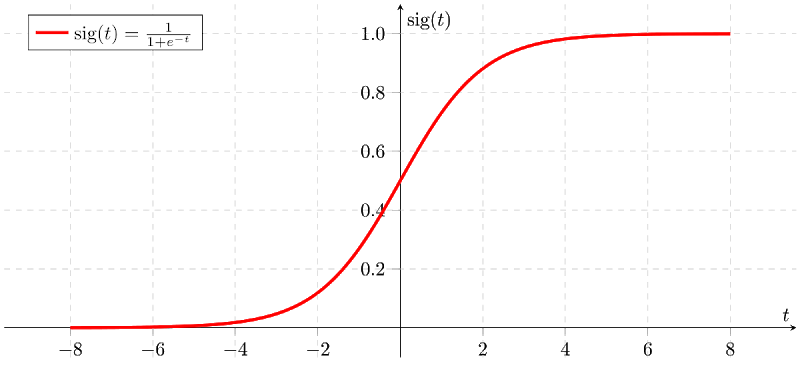
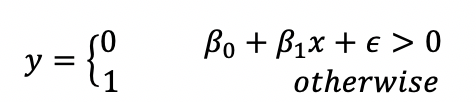
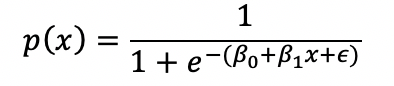


Figure 3: Sigmoid Function

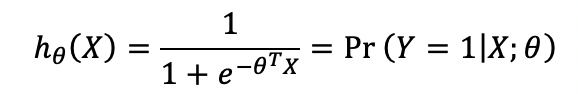
Logistic regression assumes the data follows Bernoulli distribution, with probability p to get the value 1 and probability (1 – p) to get 0, which makes logistic regression especially suitable in our case, since we will look for the direction of direction of percent change of VIX, which is a binary variable (either positive or negative). It can be simply understood that we aim to find the b parameters that best fit the formula:



And we can also make some calculation and get the formula for its probability density function:



We can also rewrite the function as:



Using the likelihood function above, we can calculate the estimator vector θ with maximum likelihood estimation. After the model is fitted, we could use the logistic model to get the outcome (either 1 or 0).

For the Logistic Regression, there is no hyperparameters we need to tune. Since comparing to the other two models SVM and Gradient Boosting Tree, the result of Logistic Regression not as good as other two models, we do not put further emphasis on this model. The logistic model is cheap and fast in calculation, however, its classification accuracy isn’t very precise. Also, it always got a low score in cross validation even sometimes it generates a surprisingly low sample error.

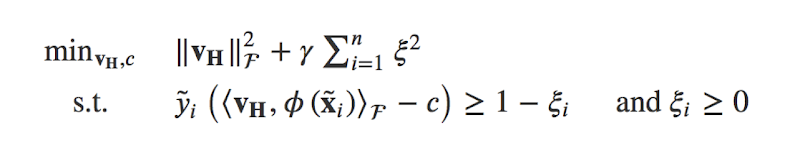
2. Support Vector Machine

The support vector machine (SVM) is one of the most famous and widely used kernel-based algorithms that deal with binary classifiers. For example, we are given a labeled training data set (supervised learning), after utilize some machine learning method, the algorithm outputs an optimal hyperplane which categorizes new examples.

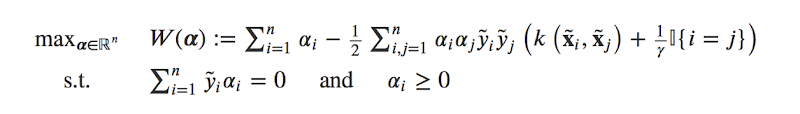
A Support Vector Machine is a discriminative classifier formally defined by a separating hyperplane. The objective of SVM is to find a hyperplane which maximize the margin separating the two categories of data. It is a classification method of the supervised learning. The changing of different kernel functions makes SVM not only work for linear problems, but also for nonlinear problems. However, it requires correct scaling of data, and it is sensitive to parameters.

In this project, our goal is to maximize the distance between the line and the support vectors, which turns the whole problem into an optimization problem. We choose the sigmoid function as our kernel function since it has the best performance among other kernel functions. Besides the kernel functions, there are also other parameters we have to choose in order to achieve a precise prediction. In our research we employ the Support Vector Machine Cross Validation (SVMCV) model, which uses the grid search function from ‘sklearn’ package to help adjust the parameters, and the SVMCV proves to outperform the simple SVM model in most situations.

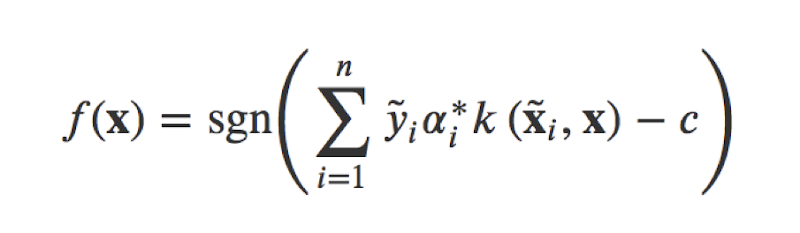
The following formula formally describes SVM.



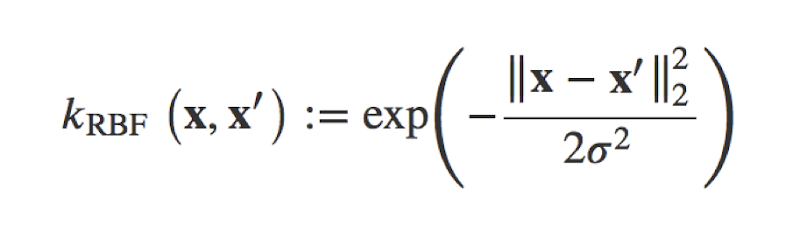
Where  s the support vectors, is the penalty term of the misclassified object, is a transforming function of the data to a higher dimensional space. The variables ξi are slack variables. To solve the problem implicitly, we can instead solve a dual optimization problem:



The classifier is:



Where k is the kernel of the SVM. There a many choices of the kernels, such as the linear kernel, the Gaussian kernel, etc. In this paper, we choose the Gaussian kernel:



As for the SVM, there are two key parameters  and .  is the parameter for the penalty of the misclassification sample. The larger , the larger penalty the model shall make to the misclassified sample, which means  the margin of the Support Vector are smaller, vice versa.  controls the variance of the Gaussian kernel, which is a transformation of the original data to a higher dimension. The larger , the data will get more close to each other so that it could be harder to separate them.

We use the 10-fold cross validation to find the best parameters. It turns out that a combination of  = 0.2 and   = 10 is the best one, which makes the prediction accuracy to 0.503.  It turns out that SVM does not do a decent job to predict the VIX since the best parameters of the SVM only gives a slightly increase in the prediction accuracy.

3. Gradient Boosting Decision Tree

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models.  It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

It is common that decision trees are used as the weak learner in gradient boosting. Specifically regression trees are used that output real values for splits and whose output can be added together, allowing subsequent models outputs to be added and “correct” the residuals in the predictions. Trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss.

It is common to constrain the weak learners in specific ways, such as a maximum number of layers, nodes, splits or leaf nodes.This is to ensure that the learners remain weak, but can still be constructed in a greedy manner. In our model, we try to analyze different combinations of the number of layers and leaf nodes to get the best predictive results.

Because of these characteristics of Gradient boosting, we can easily conclude that the gradient boosting is especially effective when dealing with non-linear relationships. And from the previous correlation heatmap, it is evident that there exists little linear relationships between the independent variables against the dependent variable. As a result, in order to make a predictive model on non-linear variables, we select gradient boosting technique to discover any non-linear relationships between the variables.

It turns out that the Gradient Boosting Tree method has the best performance among the models for the qualitative analysis.It is among the most powerful and widely used model for supervised learning. Similarly to other tree-based models, the algorithm works well without scaling and on a mixture of binary and continuous features, which fits our project theoretically.

Therefore, we will make a great emphasis on the gradient boosting tree.

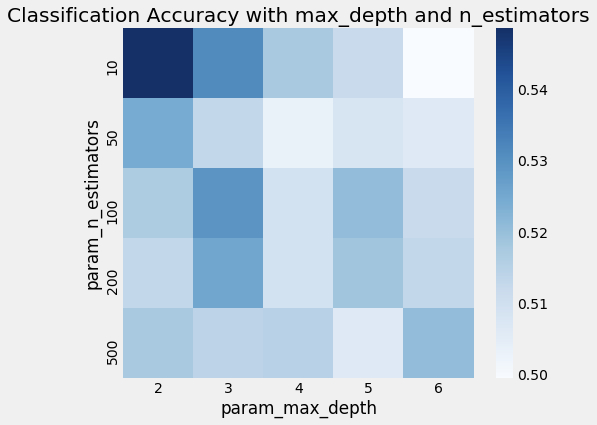


Figure 4 Cross validation of parameters

Figure 4 shows a heatmap of different combination of the hyperparameters. We find that the best ones are number of trees equal to 10 and max depth equals to 2, and the classification accuracy is over 0.54. Since our total number of feature is only 9, the hypothetical best max depths should be . And it turns out that 3 max depth does also a decent job in our model.



Figure 5: Feature  Importance

We get the feature importance plot for the Gradient Boosting Tree, which is shown in Fig 3-2. The best feature is Num Sources, followed by avgtone and quand\_2. It should be noted that the feature importance is not worked as the correlation. We can not get the direction or the symbol of the variable, because the weak learner of the Gradient Boosting Tree is based on decision tree, and the feature importance is calculated based on which variable will produce the maximum  information gain.

3.2 Quantitative Model

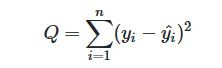
As for our quantitative model, we aim to predict change percent (return) of the daily VIX’s, which is a regression problem.. We used two machine learning techniques, namely Linear Regression and Gradient Boosting Regression Tree. For each method, we not only learned the algorithm and the methodology behind, but also tuned the hyper parameters according to the metric of classification accuracy. We then use the Mean square errors (MSE) and R-square to measure the performance of models .

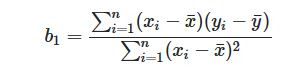
1. Linear Regression

linear regression is used to model relationship between continuous variables. Xs are the predictor, explanatory, or independent variable. Y is the response, outcome, or dependent variable. There are some assumptions for this model. The mean function is linear and the variance function is constant. Residuals are statistically independent, have uniform variance, are normally distributed; the epsilon ϵϵ term is assumed to be a random variable that has a mean of 0 and normally distributed. The error terms are also assumed to be independent.

For the algorithm of linear regression, we want to find the best fit line:

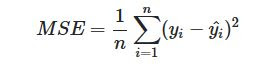


We can achieve that by using least squares estimation, which minimizes the sum of the squared predicted errors. We need to find the values of  b0 and b1 that minimize:

such that

And

The cost function of linear regression is



Linear regression is an extremely simple method. It is very easy to use, understand, and explain.; the best fit line is the line with minimum error from all the points, it has high efficiency. However, linear regression only models relationships between dependent and independent variables that are linear. It assumes there is a straight-line relationship between them which is incorrect sometimes.Also, it is very sensitive to the anomalies/outliers in the data if the number of the parameters are greater than the samples, then the model starts to model noise rather than relationship

2. Gradient Boosting Regression Tree

Gradient Boosting Regression Tree is similar to the Gradient Boosting Decision Tree from before, but this time we would use Regression model instead the classification model to build up the models and test their performances. The main idea of boosting is to add new models to the ensemble sequentially. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful “committee” that are often hard to beat with other algorithms.

The cost function of Gradient Boosting Regression Trees on the model, could be square loss or exponential loss. For any loss function, we can derive a gradient boosting algorithm. Absolute loss and Huber loss are more robust to outliers than square loss.

Gradient Boosted Methods generally have 3 parameters to train shrinkage parameter, depth of tree, number of trees. Now each of these parameters should be tuned to get a good fit. And you cannot just take maximum value of ntree in this case as GBM can overfit fit higher number of trees. But if you are able to use correct tuning parameters, they generally give somewhat better results than Random Forests. Also, it often provides predictive accuracy that cannot be beat and it can optimize on different loss functions and provides several hyperparameter options that make the function fit very flexible. In addition, it can handle missing data which imputatino is not required

However, GBMs will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting. Must use cross-validation to neutralize and computationally expensive. Also, the high flexibility results in many parameters that interact and influence heavily the behavior of the approach. Furthermore, the poor of interpretation could be an issue.

**4. Model Performance**

After fitting different models, it is important to have appropriate metrics to measure the performance of the models. We will introduce the metrics one by one and why we have selected such metrics in evaluating the validity of the models in terms of both mathematical and business perspective.

4.1 Quantitative model

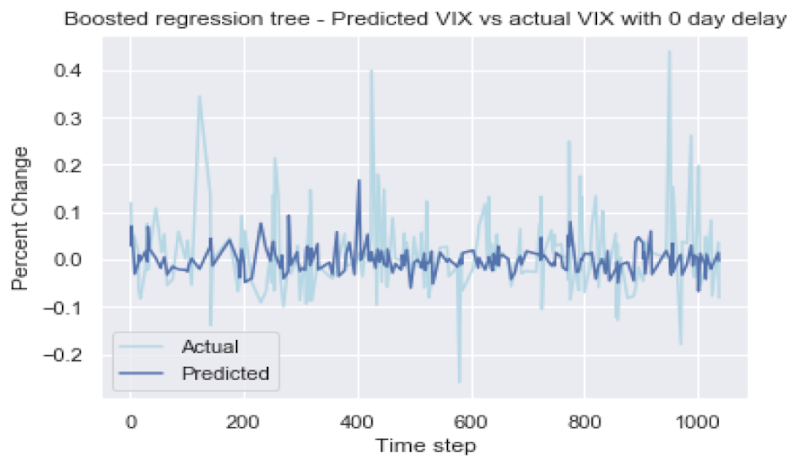
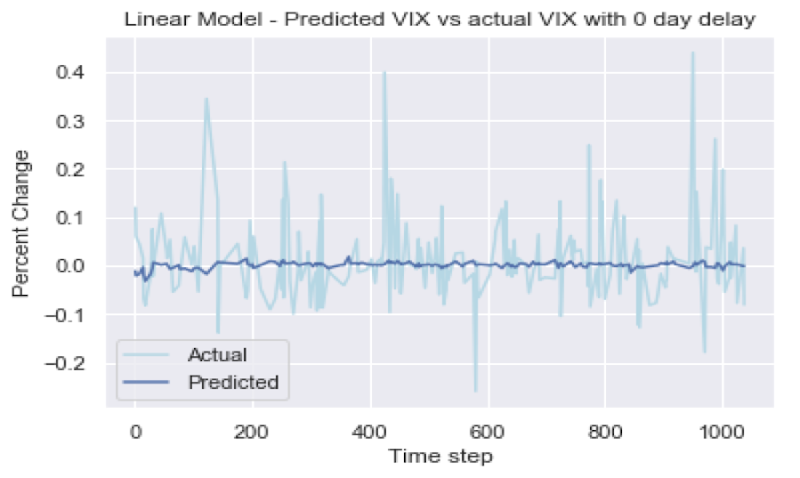
The first metric we will introduce is mean squared error (MSE). In statistics, the MSE is an estimator that measures the average of the squares of the errors. It acts as a risk function, corresponding to the expected value of the squared error loss. MSE is a simple but useful way to examine our quantitative models because it measures the deviance of our predicted results in comparison to the actual results. We generally deem a model with a low MSE to be better compared to a model with a high MSE. Moreover, it acts as a loss function for the linear regression to minimize as a baseline performance metric. However, MSE is sometimes not the best performance metrics due to the characteristic of heavily weighting outliers. When we minimize MSE, we are also discouraging extreme predictions because that loss is amplified if not correctly predicted. Additionally, MSE is largely dependent on the scale of the data. In our case, a small MSE could be rather misleading due to the percentage scale of our dependent variable VIX percent changes.

The second metric we use for quantitative models is the R-squared. R-squared is a special and simple metric to interpret since it does not depend on the scale of the data. Unlike the MSE, the scale of the data has no direct impact on R-squared. Hence, we can directly interpret and determine the fit of a model. By definition, the R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. As a result, we generally aim for a model with a high R-squared value. Also, note that there are cases where the computational definition of R-squared can yield negative values, depending on the occasion. In cases where negative values arise, the mean of the data provides a better fit to the outcomes than do the fitted function predicted values. It is a rare occurrence but it is important to keep in mind when conducting our performance check. From a business perspective, we can easily compare the R-squared between different models predicting the same continuous variable. It provides a direct and efficient way to select out the model that best explains the dependent variable.

On top of deciding the metrics we be looking at, it is also essential to select a method to validate the scoring metrics. To select the best model that have good generalization across different training and testing datasets, we use K-fold cross validation. It is a statistical method of evaluating models’ performance and reducing overfitting problems. The training data may be capricious sometimes, in which train the models with one dataset might not reflect the actual performance of the models. Therefore, we need to split the training set into multiple subsets using k-fold Cross Validation.

We use one of the most commonly used versions of cross validation, k-fold cross validation, to select the best models from all the candidate models. The data is firstly partitioned into k parts in equal size, called folds. Next, the first model is trained using the first fold as test set and remaining （k-1) folds are used as training set. The model is run using the training data but the accuracy is evaluated on test data. The process is repeated in all folds of data. In our project,  we eventually decide to implement a 10-fold cross validation technique. K-fold cross validation procedure generally have a lower bias than other validation methods. It is a good way to see how a model performs out-of-sample after the training is complete in-sample. In our case, we split the datasets randomly into 10 groups. Then for each unique group, we will treat the group as a test set and the remaining groups as a training data set. Next, we will fit a model using the training data set to verify on the test set by checking the scoring metrics previously mentioned. Eventually, we take the mean of the scoring metrics to examine the average performance in each sub-group. This method results in less biased or less optimistic estimate of the model skill than other methods. Or in simpler terms, it prevents us from mistakenly selecting out an over-fitted model that performs extremely well in-sample.

Using a base linear regression, we receive a baseline MSE value of .0511 and a baseline R-squared value of .0188. Clearly, the linear regression does not provide the best performance with a low explained variance in the dependent variable. However, after selecting out the best hyper-parameters for the boosted regression tree using grid search method like discussed earlier, we notice that the MSE is significantly lower at .0369 and the R-squared is also significantly lower at .0013. This result was rather disappointing since we were hoping for a better fit that would verify our early hypothesis that the sentiment scores would be correlated with the percent changes in VIX. On a closer exploration, we discover something extremely interesting when we decided to plot the predicted values against the actual values. As shown in the two graphs, we notice that the linear model shows a very flat line for the predicted values compared to the actual values. This occurrence is likely caused by the nature of linear model which emphasizes on minimizing the MSE. As a result, the model acts extremely conservatively to make the MSE as low as possible. However, this nature is unhelpful in a business perspective because the linear model fails to provide any practical information in the real world. On the contrary to the linear model, the boosted regression tree actively predicts the ups and downs, and we suspect that there is an early ‘signaling effect’ with this model since the extreme value predictions always seem to appear slightly earlier than when the actual extreme values occurs. To investigate further on this signaling hypothesis, we decide to shift the dependent variables forward so that we are really using the independent variables to predict the variables in the futures.

Figure 6 Signalling effect 

To determine the best time period of shift, we run a validation across the data with different time shifts ranging from 2 to 30 trading days. We receive the best MSE values when the data is shifted 7, 15, and 26 days and the best R-squared values when the data is shifted 3,8, 26 days. We conclude that after shifting the data by 26 days, there is an improvement in the performance metrics. Additionally, by plotting the graph shown below, we can see that the actual extreme conditions are roughly predicted after the shifts. From a business perspective, we can conclude that there likely exists a signaling effect in our features when predicting the percent changes in VIX. The sentiment scores signals the occurrences of extreme swings in VIX 26 days before the swings actually occur. This result was a big surprise to us and it reaffirmed our initial idea of conducting qualitative analysis to see if the qualitative models will give out more impressive results.

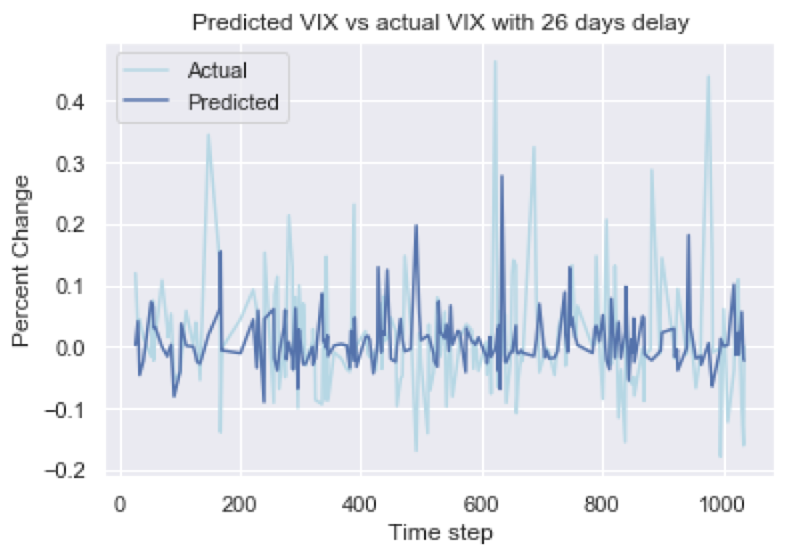


Figure 7 Predicted VIX vs actual VIX

4.2 Qualitative Model

We use two metrics to compare the classification models: AUC and accuracy to compare the three models. Accuracy is a direct and simple way to examine a classification model performance. Generally, a well-performing model is able to have prediction accuracy of higher than the baseline sample distributions. In our case, we need to achieve above 53% accuracy to conclude the model to have better performance than random guessing. While accuracy is an important measure to consider, we decide to include AUC-ROC as another metric to review the performance of the models.

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability (ability to be separated). It tells how much model is capable of distinguishing between classes. The higher the AUC, the better the model predicts the responses. In our case, the higher the AUC, the better the model is at distinguishing between the positive and negative changes in VIX. From the graph below, we see that the ROC curve is mostly above the baseline in the middle for this particular case of boosted tree classification. This graph is indicative of a well-performing model and it is an especially useful metric on top of our simple metric of accuracy. An excellent model has AUC near to the 1 which means it has good measure of separability and a poor model has AUC near to the 0 which means it has almost no ability to separate the classes. Additionally, a model is performing better than base distribution if it has AUC greater than 0.5, so it is a simple method to determine if a model outperforming random guessing. The graph below represents the ROC curve of a unique case using the gradient boosting tree method. The overall area under the curve is greater than .5 and the curve is on the left side of the dotted red line majority of the time. Hence, we are able to conceive this model to be performing better than random guessing. Viewing the ROC curve gives the advantage of seeing the tradeoff between sensitivity and specificity for all thresholds rather than one that was chosen by the model itself.

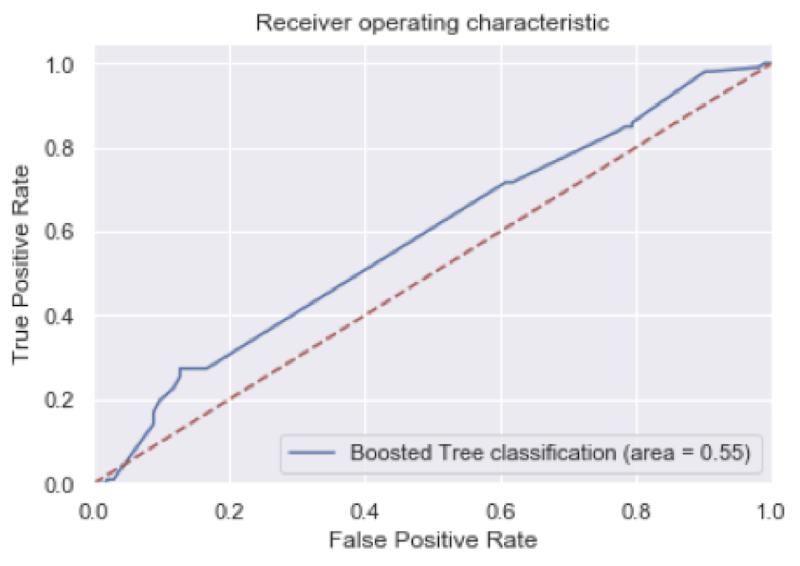


Figure 8: ROC curve of the model

Using the same 10-fold cross validation method mentioned from earlier in quantitative analysis, we compare the three model performances at predicting the positive and negative changes in VIX on a daily basis. The result was an overwhelming victory for the gradient boosting tree model. As shown in Table 3-1, the gradient boosting tree model is the only model that predicts at a higher accuracy than the base distribution. Moreover, the gradient boosting method also has highest ROC-AUC scores, showing a stronger predicting power than the other two models. This is no real surprise to us since we expected the gradient boosting model to have the best performance. From the early exploration results, we expected little to no linear relationships between the independent variables and the dependent variable which would mean that the relationship are non-linear if there exists any. And gradient boosting does very well in this occasion because it is a robust out of the box classifier that can learn complex non-linear decision boundaries via boosting and reinforcement learning. The logistic regression model performs slightly better than the SVM model in terms of both ROC-AUC and accuracy. This was slightly surprising to us since SVM models are good at learning complex non-linear decision boundaries in a high dimensional space. It was also noting that the logistic regression method took the lowest training time and the SVM took the highest training time for the fitting process. In terms of elegancy, the SVM model was also not able to compete against the other two models.



Table 1: Comparison of different models

Despite having a higher than baseline performance for the gradient boosting model, we are looking for additional techniques to improve the performance. By conducting additional data manipulation to the raw data sets such as taking quantile values at different percent, we were able to improve the overall accuracy by an additional 2% to 57% and ROC-AUC to roughly 0.52. In addition to simply improving model, we inspect further to analyze the behaviors and the reason for each feature to have an impact or correlation to our dependent variable. We suspect that the average number of sources of events to have highest feature importance due to the fact that higher frequency of events could lead to people panic and thus increasing the implied volatility in the market.



Figure 9 Performance of different models

In conclusion, the best model at forecasting values of VIX is the gradient boosting regression tree model on a 26-day shifted dataset which beats the baseline linear model by a small margin. While the best model at predicting the directions of VIX is the gradient boosting classification tree model which significantly beats the baseline of the sample distribution. It is possible to develop statistical arbitrage opportunities based on these models with more supporting indicators. For example, the previously daily changes in NASDAQ index or other type of macro indicators could possibly improve the model performance. In a real world application based on these two models behaviors, we propose an arbitrage strategy where we use the quantitative model to discover any extreme values 26 days later when we will avoid trading within 5 days of the predicted occurrence of extreme values. And besides these days the portfolio will simply consist of an ETN based on the VIX, for which we will open long position or short positions based on the classification model predictions. Following this strategy, we can safely avoid extreme conditions to minimize risks and trade at a small margin to take advantage of the classification model.