Data Mining: A Stock Strength Indicator With Predictive Power

Drake Morey

Department of Statistics, Columbia University

STATUN3106: Applied Data Mining

Professor Wayne Lee

May 7, 2022

Data Mining: A Stock Strength Indicator With Predictive Power:

Introduction:

Who doesn't love making more money? This report will attempt to explore using large-cap and sector ETFs (exchange traded funds) and an individual stock (of which is included in said sector ETFs) to synthesize a stock strength indicator for that particular individual stock. Additionally, these ETFs will serve as the platform to predict if an individual stock's price performance will be upward or downward for a single day period.

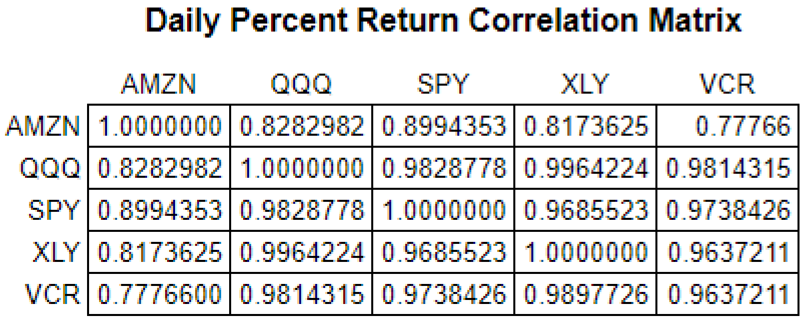
Audience and potential value for that audience:

The creation of a stock strength indicator that has predictive value would be useful to a large pool of individuals and institutions. Namely, these individuals would include persons that invest money into the stock market (although, the indicator may be more valuable to active traders - as in stock market day traders - that buy/sell stock positions on a daily basis). Additionally, institutions such as hedge funds, investment management corporations, or quantitative trading divisions would find this stock strength indicator useful. For the scope of this report, we will narrow our audience focus to quantitative trading divisions within hedge funds - as these institutions would benefit greatly from the ability to increase returns for investors in their funds and they trade frequently enough to employ the indicator. Therefore, this audience would have value realized through the dual ability to increase investor returns (increased reward) and decreased uncertainty (such as decreasing volatility of funds). Increased investor returns due to utilizing the stock strength indicator would come from the quant division being able to increase leverage when the stock strength indicator is implying a higher likelihood of positive return days and decrease leverage when the stock strength indicator is implying a higher likelihood of negative return days. On the other hand, the quant division could reduce fund uncertainty (through managing fund volatility) by selling out individual positions that are indicated to have a high likelihood of negative return days; thus reducing the drawdown of said fund.

Datasets:

For this report and analysis, various datasets were used. Datasets for stock market data (individual stocks and ETFs) were acquired from Yahoo Finance via the utilization of the package called “quantmod” in R. Yahoo Finance receives its historical data on stocks from Commodity Systems. The data from Yahoo Finance via quantmod provide us with dates, daily open prices, daily high prices, daily low prices, daily close prices, daily volume, and adjusted prices (adjusted for stock splits). Data was pulled for Amazon (AMZN), Nasdaq 100 (QQQ), S&P 500 (SPY), Consumer Discretionary Select Sector SPDR Fund (XLY), and Vanguard Consumer Discretionary (VCR). Additionally, a dataset of historical put/call ratios on the S&P 500 index (SPX) was acquired via CBOE - which is the entity that compiles and owns the data for SPX put/call ratios. The CBOE data contains dates, daily put/call ratios, put volume, call volume, and total volume. The CBOE data on SPX put/call ratios was constrained temporally and provided data from 7/6/2010 to 10/4/2019, therefore the previously mentioned stock market data was pulled only to include that date range.

Individual stock and ETF data were used together because the ETFs (all of which contain the target individual stock) are used to produce insights into the individual stock, we can see that the individual stock and ETF data are related because their daily returns are correlated - daily returns calculated by the formula: (((stock close price / stock open price) - 1)*100) (see correlation matrix below):



Daily Percent Return Correlation Matrix

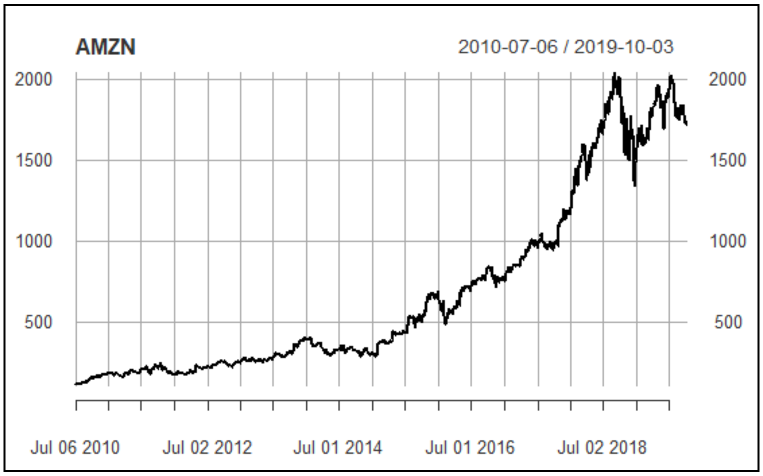
	AMZN	QQQ	SPY	XLY	VCR
AMZN	1.0000000	0.8282982	0.8994353	0.8173625	0.77766
QQQ	0.8282982	1.0000000	0.9828778	0.9964224	0.9814315
SPY	0.8994353	0.9828778	1.0000000	0.9685523	0.9738426
XLY	0.8173625	0.9964224	0.9685523	1.0000000	0.9637211
VCR	0.7776600	0.9814315	0.9738426	0.9897726	0.9637211

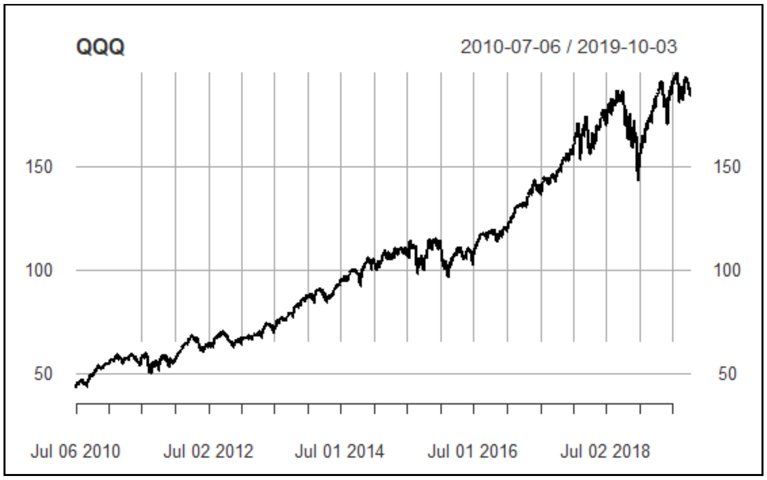
Furthermore, from the correlation matrix above, we can see that AMZN has the highest daily return correlation with the S&P 500 (SPY), and thus the SPX put/call ratio data was used (instead of another index). Additionally, a Pearson correlation test was performed to examine the relationship between AMZN daily returns and the SPX put/call ratios; there was a significant correlation w/ p-value < .0001 (slight negative correlation, r = -0.119). Therefore, there was some indication that the SPX put/call ratios could complement our existing ETF data in producing the stock strength indicator and producing predictive insights.

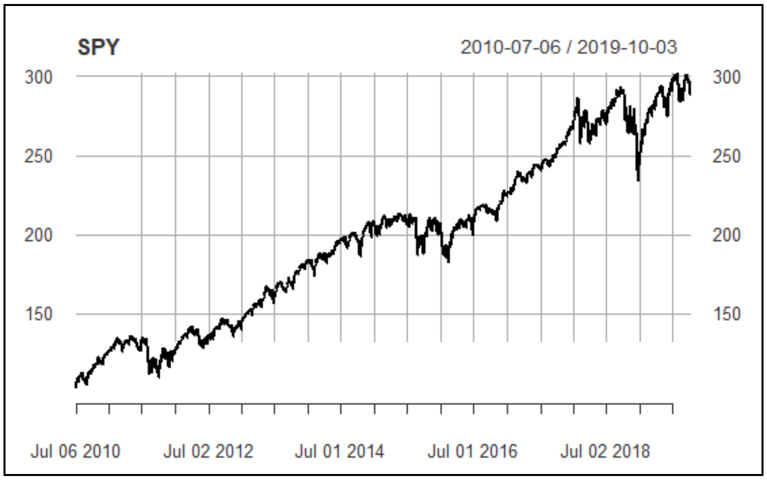
Exploratory data analysis:

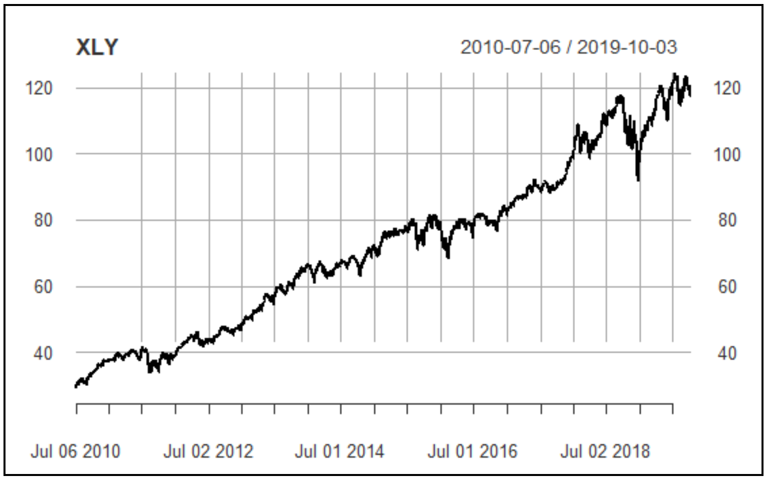
For this portion of the report, we will perform some introductory exploratory data analysis on the previously mentioned datasets to determine the completeness and quality of the data:

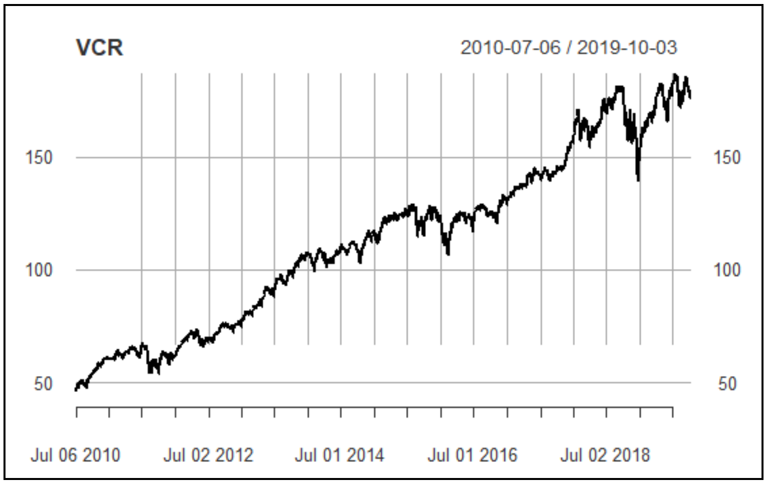
Historical Price Plots:





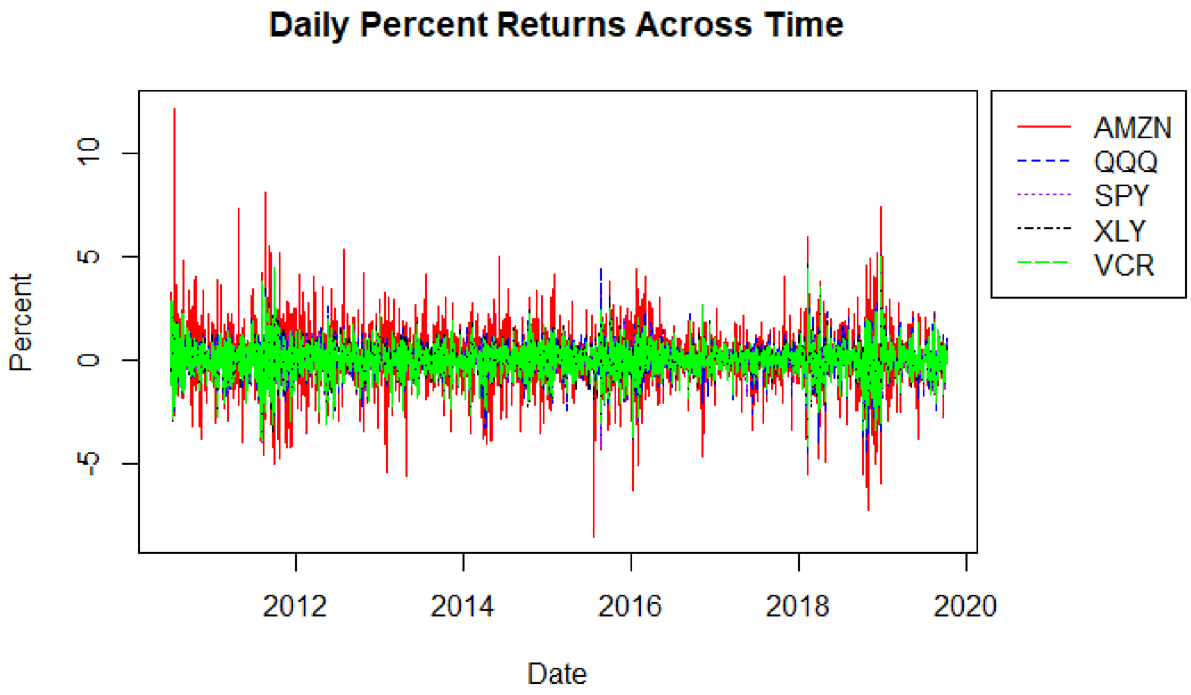


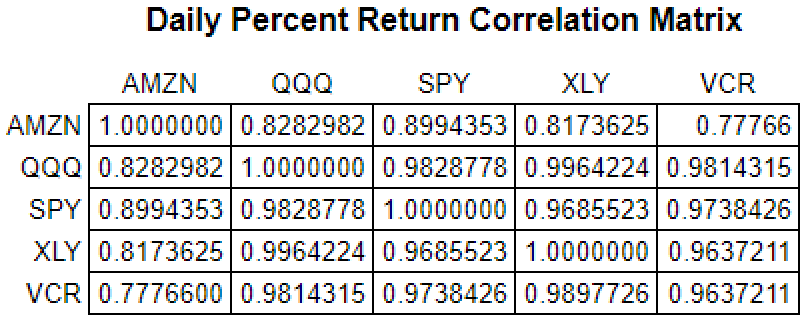




From the above historical price plots (using the closing prices) we can see that the individual stock (AMZN) and all of the ETFs follow a similar upward trend.

Daily Returns of Individual Stock and ETFs Graph & Correlation Matrix:



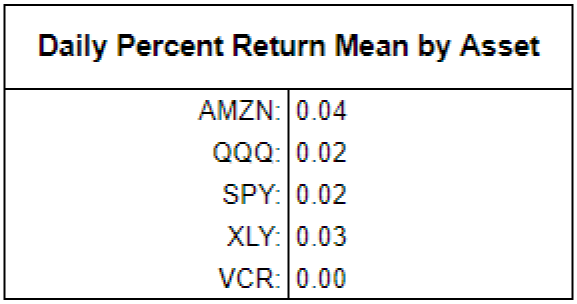


Daily Percent Return Correlation Matrix

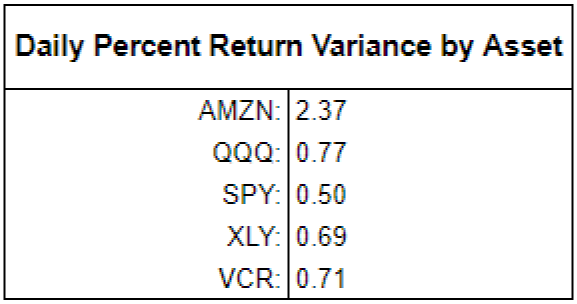
	AMZN	QQQ	SPY	XLY	VCR
AMZN	1.0000000	0.8282982	0.8994353	0.8173625	0.77766
QQQ	0.8282982	1.0000000	0.9828778	0.9964224	0.9814315
SPY	0.8994353	0.9828778	1.0000000	0.9685523	0.9738426
XLY	0.8173625	0.9964224	0.9685523	1.0000000	0.9637211
VCR	0.7776600	0.9814315	0.9738426	0.9897726	0.9637211

From the graph above we can see the overlaid daily returns of the selected stock and ETFs for this report. AMZN appears to have a wider variance in daily returns (see red line), but there does appear to be a correlation between the AMZN and the various other ETF when it comes to large up/down daily returns (see spikes around 2012 & 2016). Additionally, the correlation matrix above provides evidence for similar movements in daily percent returns via the high correlation between AMZN and the other ETFs.

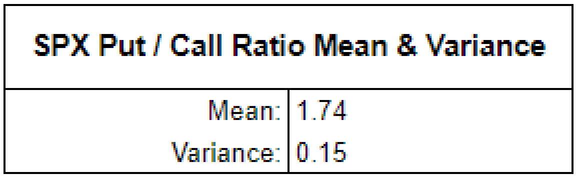
Summary Statistics of Daily Returns by Asset & SPX Put/Call Ratio:



Daily Percent Return Mean by Asset	
AMZN:	0.04
QQQ:	0.02
SPY:	0.02
XLY:	0.03
VCR:	0.00



Daily Percent Return Variance by Asset	
AMZN:	2.37
QQQ:	0.77
SPY:	0.50
XLY:	0.69
VCR:	0.71



SPX Put / Call Ratio Mean & Variance	
Mean:	1.74
Variance:	0.15

From these summary statistics of the daily returns, we can see that all assets had positive mean daily returns, with AMZN showing the highest. On the other hand, when taking a look at the variance of the daily returns, we can see that AMZN has the highest variance in daily returns among all the assets. Furthermore, the SPX put/call ratio had a mean of 1.74 and a variance of 0.15.

Feature Creation:

To create the stock strength indicator and produce predictive insights, two features were engineered (to eliminate data seasonality) from the individual stock (AMZN) and ETF (QQQ, SPY, XLY, VCR) data. The SPX put/call ratio is also technically a feature that is produced from data on puts and calls on the S&P 500 index, but CBOE has already completed this feature creation and thus the ratio is our third feature. Firstly, a joint MACD and signal line was created using the closing data from the asset's datasets - see MACD / signal formula below:

$$MACD_p = EMA_{12}(p) - EMA_{26}(p)$$

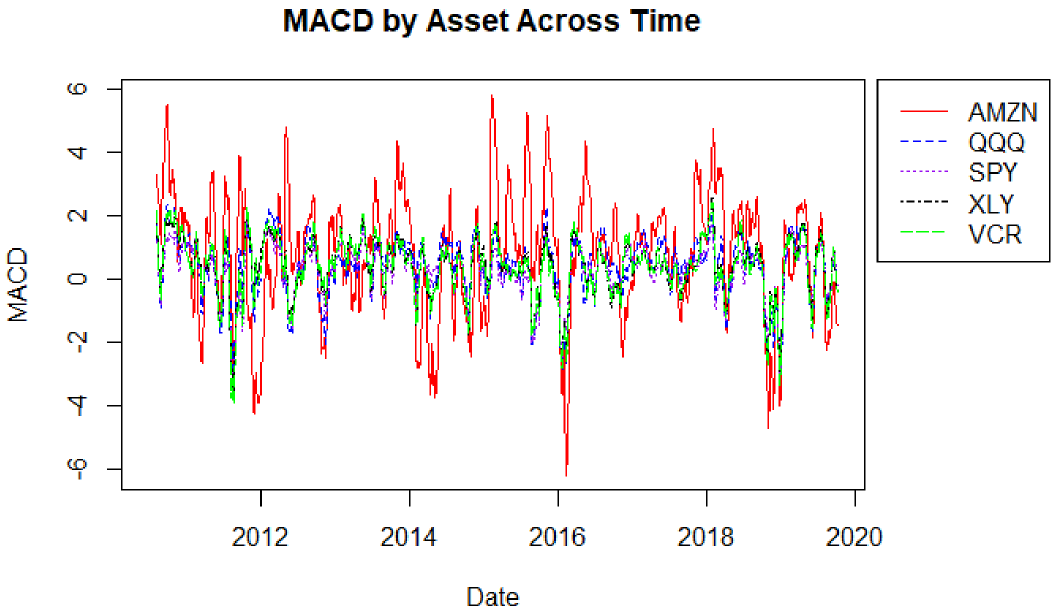
$$S_{MACD} = EMA_9(MACD)$$

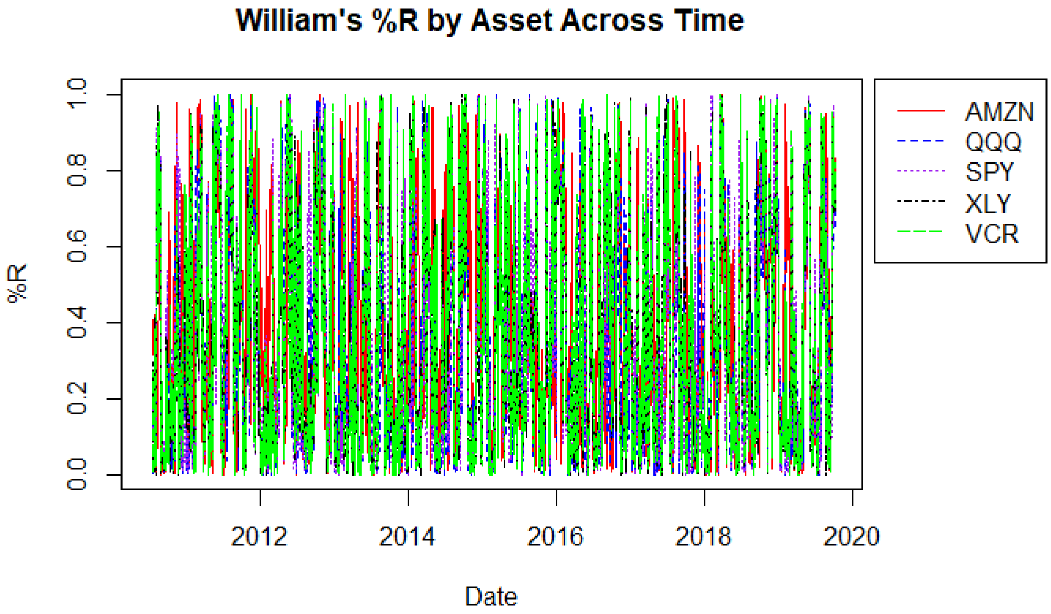
Where a 12 and 26-day period EMA (exponential moving average) of the closing price was calculated and used to produce the MACD and signal outputs.

For the second feature, William's %R was engineered from the asset's datasets. A typical 14 day period was used for the calculations, and HLC (high, low, and close) data was used from the asset's datasets - see %R formula below:

$$\left[\frac{\text{highest high (14 periods)} - \text{most recent close}}{\text{highest high (14 periods)} - \text{lowest low (14 periods)}}\right] \times (-100)$$

Therefore, after the feature extraction, we had combined MACD and signal outputs for each ETF and the individual stock, William's %R outputs for each ETF and the individual stock, and SPX put / call ratios. MACD and William's %R were chosen for feature creation because they are commonly used technical indicators within the day trading community, and have been previously shown to have predictive power on their own. See below for overlaid time series graphs of the two engineered features for each ETF and the individual stock:

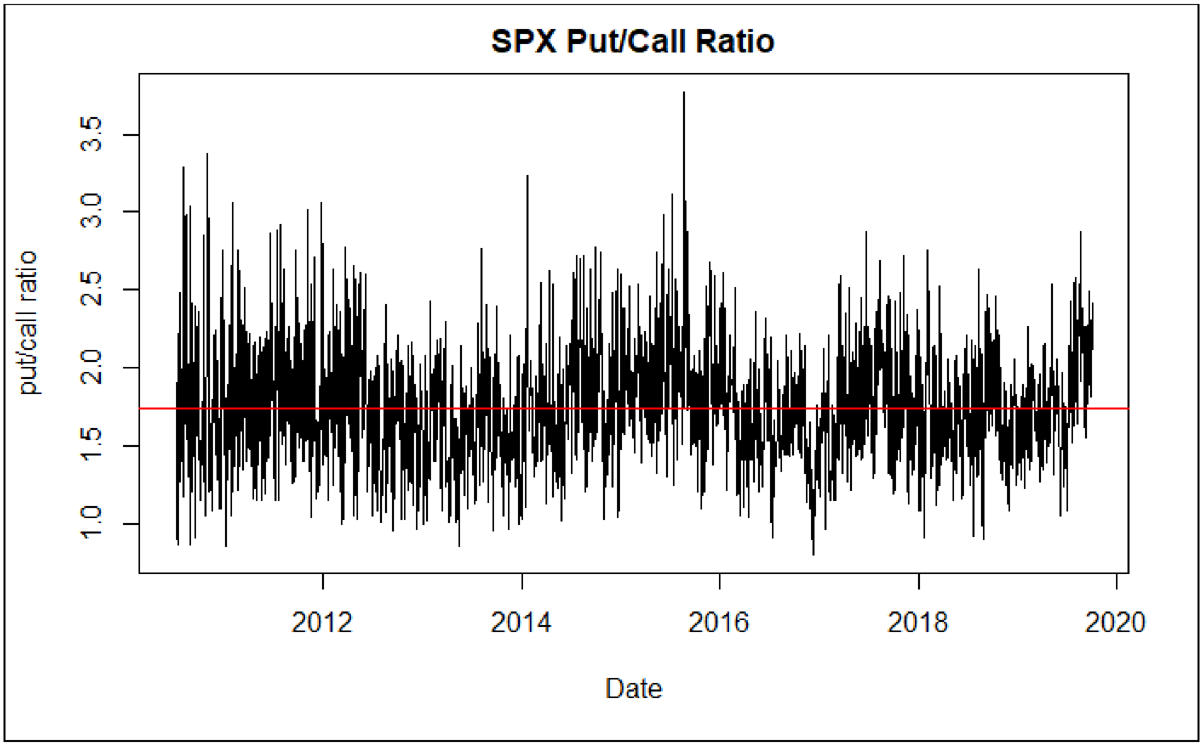




From the graph above for each feature, it is illustrated that both the MACD and %R feature acts as oscillators; where William's %R oscillates between 0 and 1, and MACD oscillates between a range with baseline 0. Additionally, from the graphs, we can gather that the MACD is less noisy and provides longer term oscillations, whereas the %R oscillates rapidly and thus produces a lot of noise in the graph above. Therefore, the longer term trend from MACD and the faster oscillations from William's %R should help complement each other, giving us exposure to each end of the spectrum.

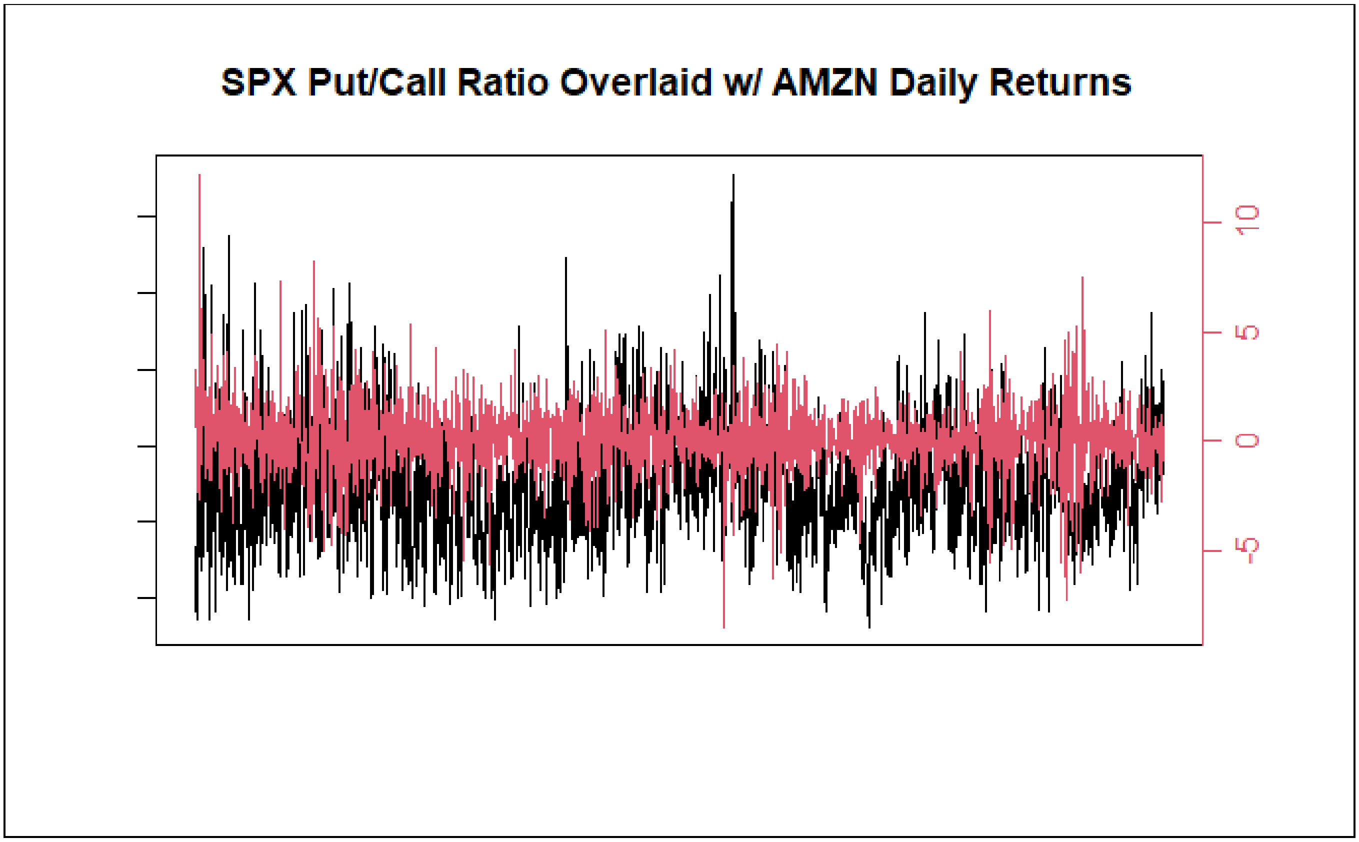
The SPX put/call ratio feature is engineered by producing the proportion of total put and call options volume in a single day. CBOE, in their dataset that is being utilized for this analysis, has already created the put/call ratio. See below for a time series graph of the put/call ratio on SPX:

SPX Put/Call Ratio Across Time:



From the graph above, the redline indicates the mean put/call ratio. Additionally, we can see that during periods of low stock market returns, the put/call ratio rises to its upper bounds. The graph below depicts SPX put/call ratios with AMZN daily returns overlaid, which help illustrate how the SPX put/call ratio interacts with daily returns of stocks:

SPX Put/Call Ratio Plotted w/ AMZN Daily Returns:

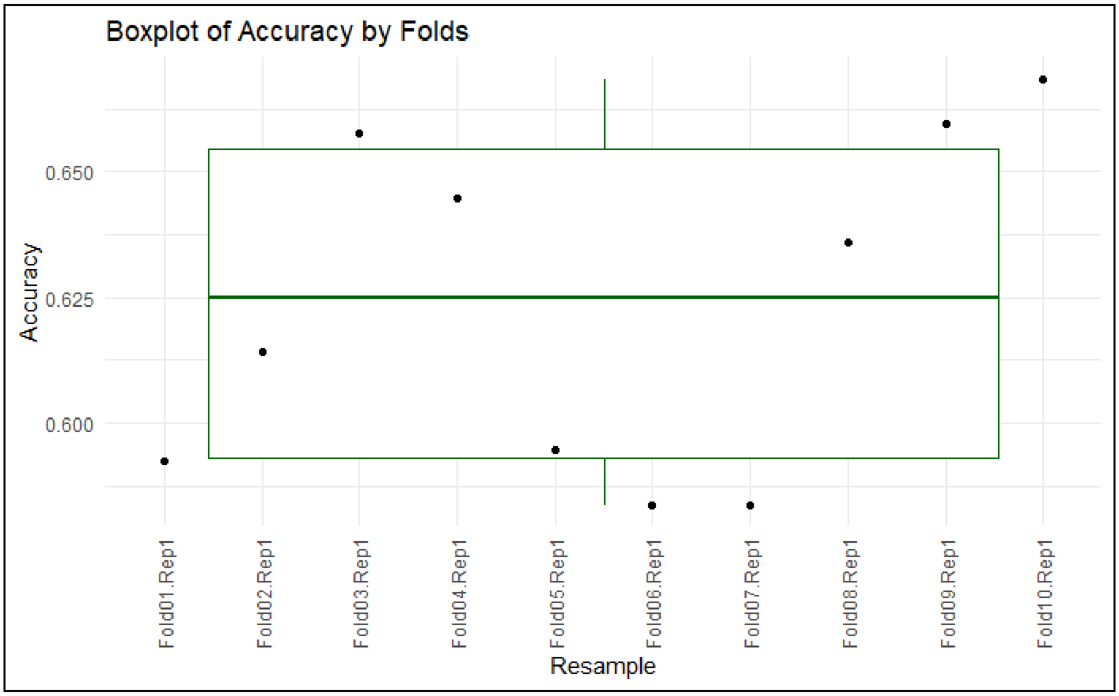


Algorithm, Predictive Power, and Model Performance/Evaluation:

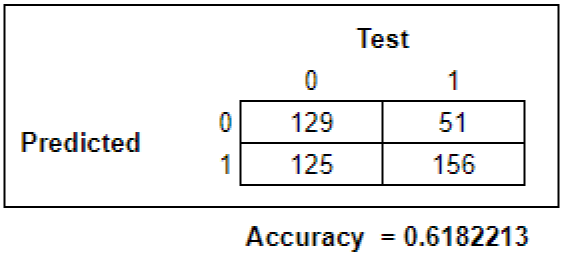
For this analysis/report, a binary logistic regression will be performed (with stepwise AIC optimization). First, the percent daily returns of AMZN were turned into a binary classifier - where, if the daily percent change was above 0, a classifier of 1 was used; but if the percent change was 0 or below, a classifier of 0 was used. Next, data was split into 80/20 train and test sets (80% of the parameter/feature dataset was used for training the model and 20% was used to test the model via predictions). Then, a logistic regression was performed, alongside a stepwise variable selection to reduce AIC - via the stepwise variable selection, AIC was reduced from 2383.849 to 2381.334, indicating a more frugal model. Additionally, a chi-square test was performed between the two models (base logistic regression and stepwise variable selected model), and with a p-value of 0.4759, there was no significant difference between the base and stepwise model - meaning that even though the stepwise model used fewer parameters, it fit to the data just as well as the base model.

The model was then applied to the test data and predicted responses were rounded - meaning if the predicted response was < 0.5, it was treated as a 0 (or a down daily return), and if the predicted response was > 0.5, it was treated as a 1 (or a up daily return). Next, the accuracy was computed for the predicted responses from the model via a confusion matrix - see below. Additionally, a K-fold cross validation was performed on the model with 10 folds, and a mean accuracy of 0.61 was achieved - see boxplot of accuracy by fold:

Boxplot of Various Accuracies From K-Fold Cross Validation:



Confusion Matrix & Accuracy - Stepwise Model (no cutoff optimization):

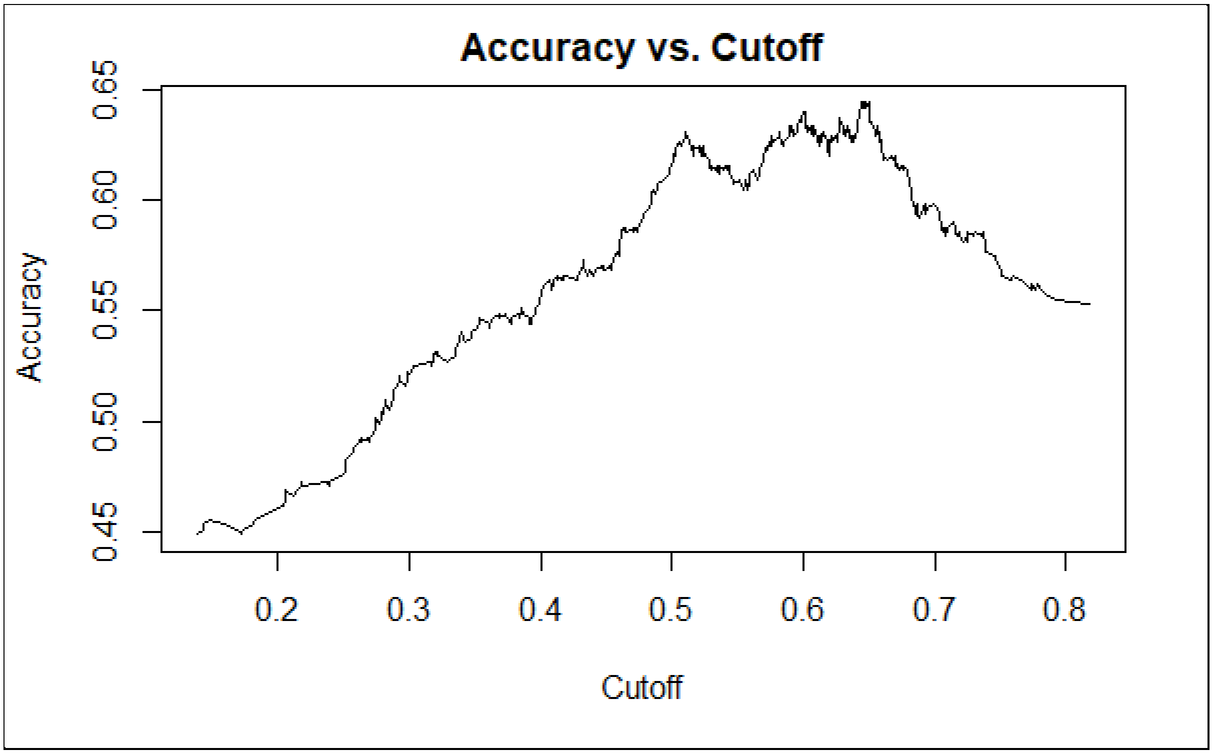


		Test	
		0	1
Predicted	0	129	51
	1	125	156

Accuracy = 0.6182213

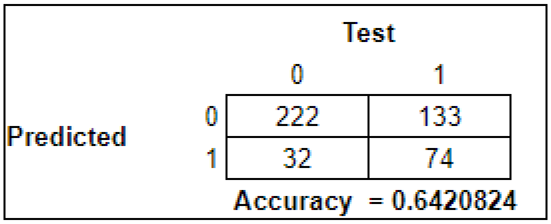
From the boxplot above, we can see that there are no outliers, the accuracies have a fairly small variability, and the lowest accuracy was .58. Additionally, the stepwise logistic regression had a accuracy of 0.62, but this accuracy was achieved using a arbitrary cutoff for up/down days (0.5), therefore to improve the accuracy of the model, we will plot varying cutoffs of the stepwise logistic regression model to determine which cutoff provides our model with the highest accuracy - see accuracy and cutoff graph below:

Various Cutoffs w/ Accuracies Plotted:



From these calculations and graph, the highest accuracy achievable by the model is 0.64, using a cutoff at 0.65. When reproducing the predictions with the new cutoff, the following confusion matrix is achieved:

Confusion Matrix & Accuracy - Stepwise Model (w/ cutoff optimization):



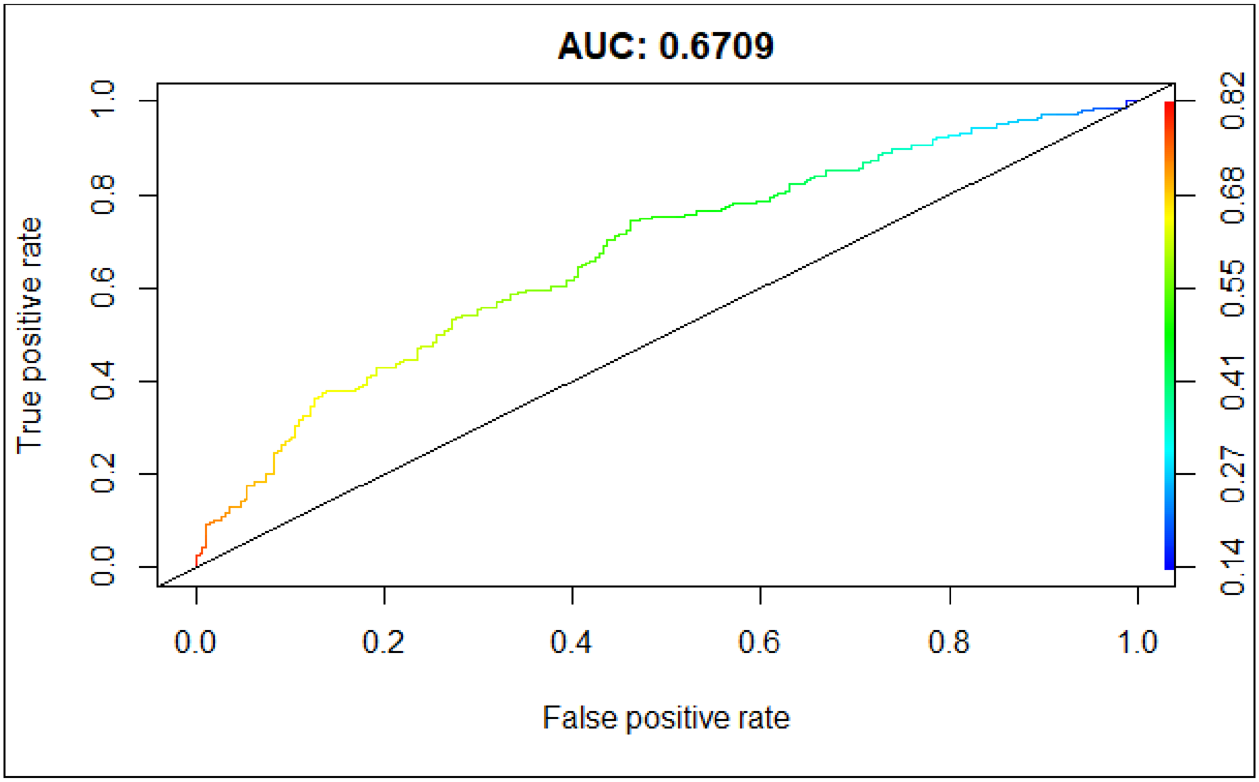
		Test	
		0	1
Predicted	0	222	133
	1	32	74

Accuracy = 0.6420824

With the optimal model produced (achieved via reducing AIC with stepwise selection, and utilizing the optimal cutoff value), we can now evaluate the model's quality by computing its recall, precision, and F1 score. The precision will allow us to determine the accuracy of the positive predictions, the recall will allow us to determine the percent of positive which were correctly identified, and the F1 score will allow us to compute a combined score of the precision and recall. Here are the computed scores: a precision of 0.63, a recall of 0.87, and a F1 score of 0.73. Therefore, the produced model has a classification strength of 73%.

Additionally, an area under the receiver operating curve (ROC) was computed and plotted to further determine the models performance and quality:

ROC Plot - Stepwise Model:

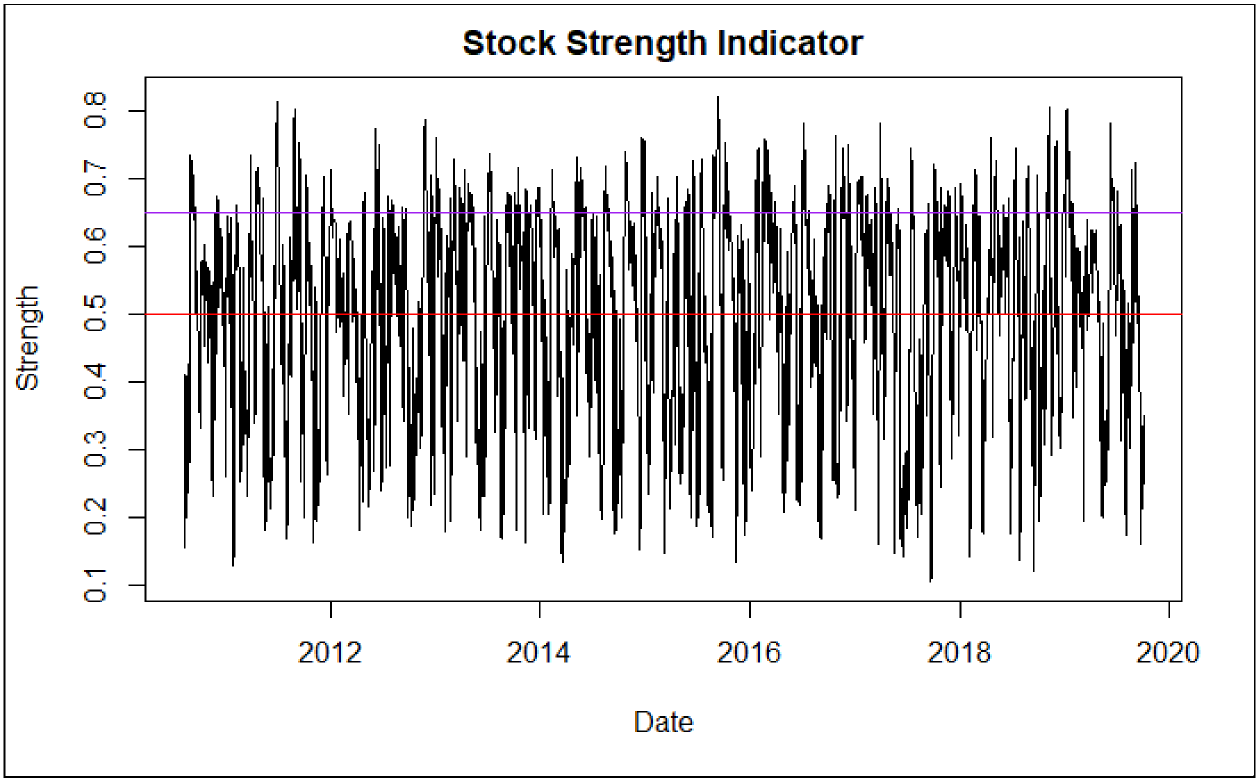


From the graph above, it can be seen that an area under the curve of 0.6709 was achieved by the model, therefore indicating that the model is fairly good at differentiating between daily returns of up or down (up meaning positive percent daily change, and down meaning negative percent daily change).

Stock Strength Indicator:

Now that the model has been shown to have predictive accuracy and evaluated for quality, we can now produce the stock strength indicator. The whole dataset is fed into the stepwise logistic regression model to produce the fitted values. These fitted values are then transformed into the stock strength indicator:

Stock Strength Indicator Over Time:



The graph above illustrates the time-series oscillations of the stock strength indicator (in the circumstance of this report, this indicator only applies to AMZN, for which the model was trained for). The horizontal red line denotes the mean of the stock strength indicator, and horizontal purple line denotes the optimal cutoff. Therefore, the target audience (a hedge fund quant division) may find it prudent to increase their leverage on Amazon when the stock strength indicator is above the red line - and perhaps even more when it is above the purple line - and decrease leverage or exit the Amazon position when it is below the red line; as there is a higher likelihood of negative return days when the stock strength is below the red line and vice versa for stock strength above the red line. To further illustrate the stock strength indicator's ability, a Pearson correlation test was performed between the stock strength indicator and AMZN daily percent returns: p-value < 0.001, and r = 0.38. Thus, the test suggests that there is a significant slightly positive relationship between AMZN daily returns and the stock strength indicator.

Conclusion

For this report, multiple models were tested: logistic regression, stepwise logistic regression, and naive bayes. The naive bayes model was not mentioned within the analysis of this report as the accuracy and performance of the model was worse than the stepwise regression. Additionally, multiple technical features (such as calculating a residual strength indicator, or using % bollinger bands) were researched to be included in the model, but the engineered features that ended up being used in the final model provided by far the strongest relationships and power. But, this report would not be complete without a discussion on data dredging and its implications for this analysis. In defense of the analysis, multiple hypotheses were not tested, we had a clear outcome of creating an insight that had predictive power (the stock strength indicator). A large number of stocks, ETFs, or indicators were not just thrown into the model. Each component was logically selected to be included in the model, and a stepwise variable selection further reduced the model's prediction error. Additionally, all models tested in this analysis did provide significant results, and each model included parameters that were not significant - even after stepwise variable selection. On the other hand, the analysis may include bias in the form of the chosen engineered features; as this is a step that must be performed by the researcher, there could be a bias in the criteria that I used to judge which features were created and thus used to train the model.

While this report was done on the individual stock of Amazon (AMZN), technically the model could take in any stock that is a component of the selected ETFs as a response variable. Thus, this broadly widens the scope of the model and use cases for various hedge fund's quant divisions as a wide variety of stocks could be modeled for the stock strength indicator. Ultimately, the analysis in this report provided evidence of predictive power by the model (as illustrated by accuracy, precise, recall, and F1 scores), and by extension the stock strength indicator can now be used as an insight to inform buy, sell, and leverage decisions.

Critique of Project by Jiajung Li:

What was the initial motivation for tackling the project?

They wanted to explore creating a way of identifying a “honest host” within Airbnb’s ecosystem. This would allow potential customers to identify which Airbnb hosts were being honest about their descriptions of the property and amenities. In their analysis they employed a text data mining and clustering algorithm/strategy.

What datasets were used?

They used datasets from insiderairbnb.com, which included various parameters relating to Airbnb, such as: bedroom/bathroom number, property type, price, etc.

What aspect of the project is considered a data-mining and what is discovered?

They are scraping text and generating an insight that is not easy to see from the data at first glance. This insight, which is identifying “honest hosts” can then be used by potential Airbnb customers to do further research into the potential host/hosting property.

Is there anything you would have done differently?

Due to computational limitations, they were unable to compute clusterings and TF-IDF for their whole dataset; thus, I would have tried to run multiple iterations over random samples from the dataset to at least achieve a sum of parts result. Additionally, I would have found it interesting to pull in data from area specific crime rates to increase likelihood of the “honest host” insight. Furthermore, I would have found it interesting to generate a feature that is somehow related to the quality/number of reviews that were found for each listing. Ultimately, their project was interesting and produced an insight that I would like to use myself the next time I look to stay in an Airbnb.

Github code link: Project Github