STAT 482 Final Report: The Effect of Trending on the Stock Market

Chase Brown, Michael Flora, Adam Punnoose, & Guillermo Ali Garcia

February 2021

1 Introduction

The birth of technologies such as machine learning and neural networks has with it created new opportunities for massive data sets to predict future events, categorize groups, and just gain insight in general. A commonly used data set for this kind of analysis is Twitter because of its ability to offer up-to-date public opinion across the world. One use case for this kind of analysis is predicting stock market changes before they happen. There have been numerous papers published about the success of this strategy. One such example [1] dates back to 2010 and found that they were able to make predictions at an accuracy of 86.7 percent. 11 years later, the advancements in not only twitter usage, but also algorithm strength and integrity gives us the idea that we can potentially do better.

In the first leg of our project, we focused on getting our datasets complete enough to answer the set of exploratory questions we brought up in our proposal. Through this effort, there were roadblocks that necessitated a slight shift in our exploratory research. These roadblocks affected a few details of our original expected dataset which will be explained in more detail in the next section. These dataset changes, also in turn changed our exploratory analysis plans. Because of this some of our initial exploratory questions were deemed out of scope.

After seeing promising results in a select group of stocks, and working to reducing the limiting factors of our dataset, we expanded our exploratory research to include any $TICKER or #TICKER mentions in our 1% sample of twitter. This resulted to just shy of 400,000 tweets including 2967 different stocks for sentiment analysis. The results of this were much more telling than the previous dataset, but it is not without its limitations.

We tested this data across different models and set up our final dataset to be usable for trading in the future. These changes, as explained below, allow the model to collect more data over time.

2 Our Dataset

2.1 Exploratory Analysis

Our dataset for exploratory analysis consists of 393,928 tweets that we ran through 2 separate sentiment analysis packages. One that only offers positive and negative and another that offers "Happy","Angry", "Sad", "Surprise", "Fear". These tweets were pulled from the dates we were able to download. The tweet was pull if the stock ticker was mentioned with a # or a $. The point of this was to ensure as best as possible that the dataset was free of irrelevant tweets. In situations where the stock ticker was a word, such as $CAKE, there was too much noise to filter out, so those were removed from the dataset.

As far as stock data goes, while we are using the Yahoo Finance API for stock data, we have slightly altered how we picked our stocks. We pulled tweets for every stock listed in the Nasdaq, then eliminated any stock with less than 100 days of data in our twitter set.

2.2 Our Model

For our Model, we further reduced our dataset to stocks that were showing a higher correlation between their stock and tweet data than the others. This consisted of stocks one might expect, such as AAPL, TSLA, AMZN, and about 10 others. We also switched to live hourly data for our model as that is a much more reliable data source, and it is the only practical option for the model in the real world. This consisted of roughly 300,000 tweets over the course of 2 weeks.

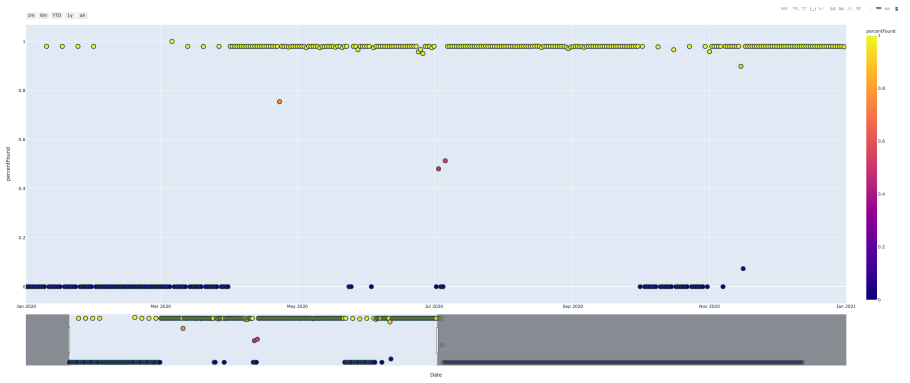


Figure 1: Exploratory Data Collected

3 Exploratory Analysis

3.1 Initial Analysis

A majority of these past few weeks have been spent solving data preprocessing issues and downloading data. So far we have been able to download and analyze most of 2020, as shown in Figure 1.

To start our analysis, we used the data we had collected and created a data frame with the following features:

Features Date
Opening price of stock
Highest price of stock that day
Lowest price of stock that day
Closing price of stock that day
Volume of stock traded that day
Volume of tweets collected for that stock that day
Daily average happy score
Daily average angry score
Daily average surprise score
Daily average sad score
Daily average fear score
Daily average sentiment score (number between 0.5 and 1; 0.5 is neutral and 1

is very confident in either positive or negative direction)
Whether the stock price increased that day
Whether the stock price increased by 5 percent that day
Yesterday's average happy score
Yesterday's average angry score
Yesterday's average sad score
Yesterday's average fear score
Yesterday's average surprise score
Yesterday's volume of tweets collected for that stock
Yesterday's volume of stock traded

Below we have included a correlation matrix showing the magnitude of the correlation between some of the variables from the data set. This will be useful as we move forward through the analysis:

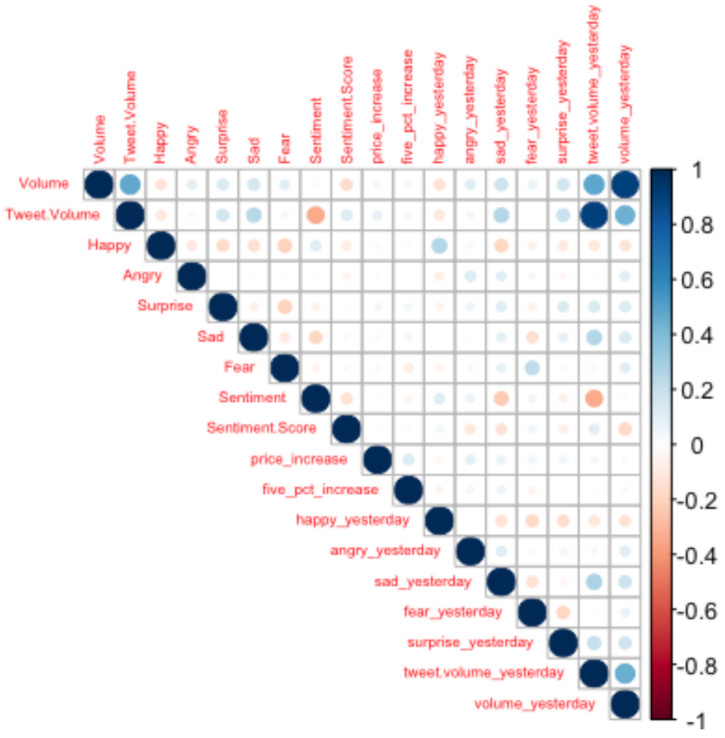


Figure 2: Feature Correlation

One interesting thing that caught our attention in Figure 2 is that the correlations between emotion scores from yesterday and whether or not the stock price increased that day are extremely similar as the correlations between emotion scores from the same day and whether or not the stock price increased that day. Our initial belief would've been that the emotion scores from yesterday would have had a higher correlation since people's emotions change and then the stock price reacts. Determining whether the emotion scores are predictive, reactive, both, or neither will be something that we will be focusing on a lot as our analysis continues.

Next, I want a quick glance at the distributions of the daily averages of the emotion scores. This will also be useful, so we can determine how strongly an emotion is for a tweet compared to other tweets

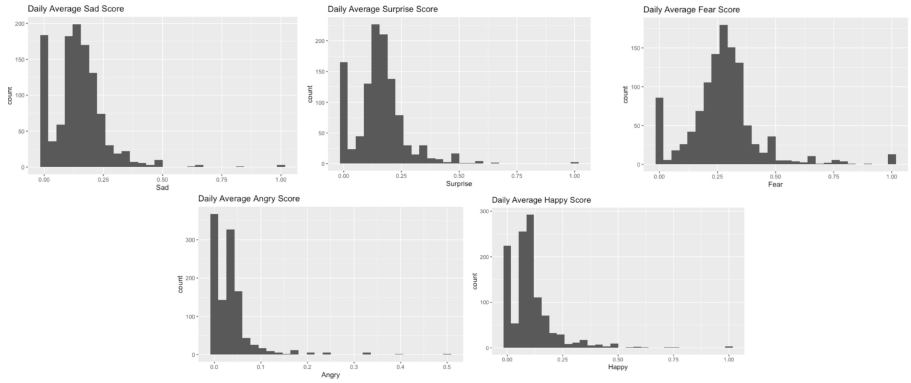


Figure 3: Averages of Emotion Scores

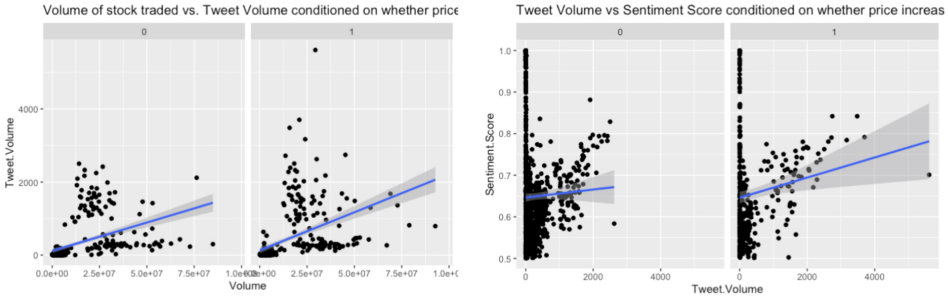


Figure 4: Volume

Next, we have a graph that shows the relationship between the volume of tweets collected for said stock during that day and the daily average sentiment score with it conditioned on whether the stock price increased that day. At first glance, it appears interesting that there is a steeper slope for when the price did increase than when price did not increase. Following that, we have a graph that shows the relationship between the volume of stock traded that day and the tweet volume conditioned on whether the stock price increased that day. At initial glance, it appears that both relationships look similar, but further analysis should be done.

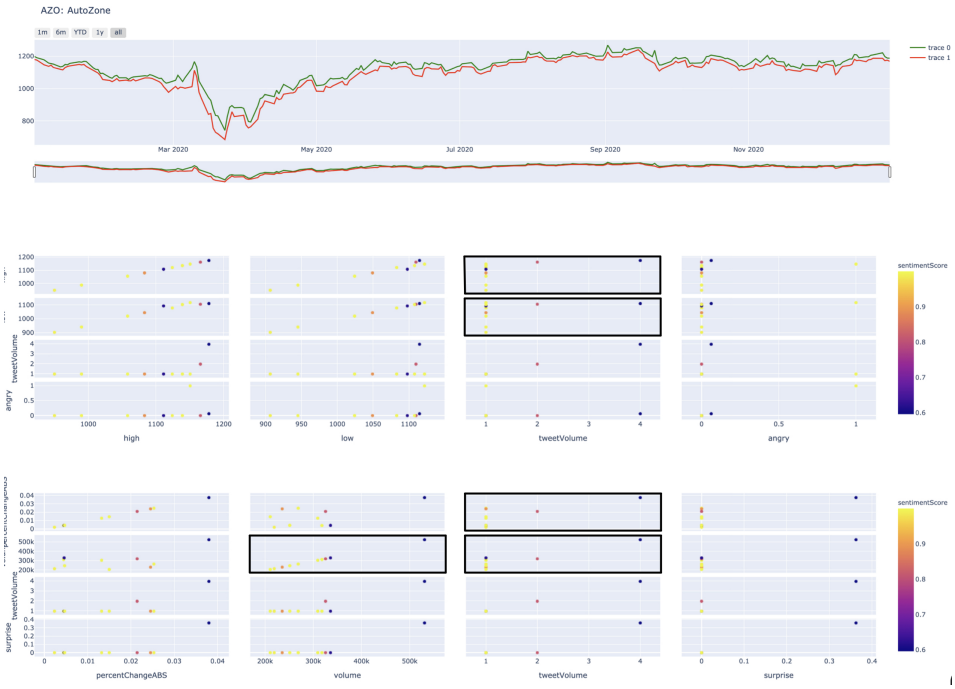


Figure 5: AutoZone Analysis

When looked at across all of our stocks, it is clear that there is no general rule of a relationship between our twitter sentiment variables and stock price. This is not the case when you look at the stocks individually, as shown in Figure 5. We will look into more stocks in our next batch of analysis and find more consistent trends.

3.2 Narrowed Analysis

After gathering our dataset, we looped through each stock and ran correlation tests on 3 versions of the dataset. We made changes to our dataset that let us see how well our twitter data correlated with the previous days stock data and the day after. The goal of this is to test whether or not twitter is tracking the stock market live, or whether there are predictive or reactive elements to it. In short, after we eliminated any correlation test that had less than 100 data points, leaving us with 3627 features to test, we found that 44 features had medium and 2 had strong correlation between stock elements tweet elements. Every single time it was a connection between stock price and daily stock volume being compared to tweet volume. Here are some of those findings more close up:

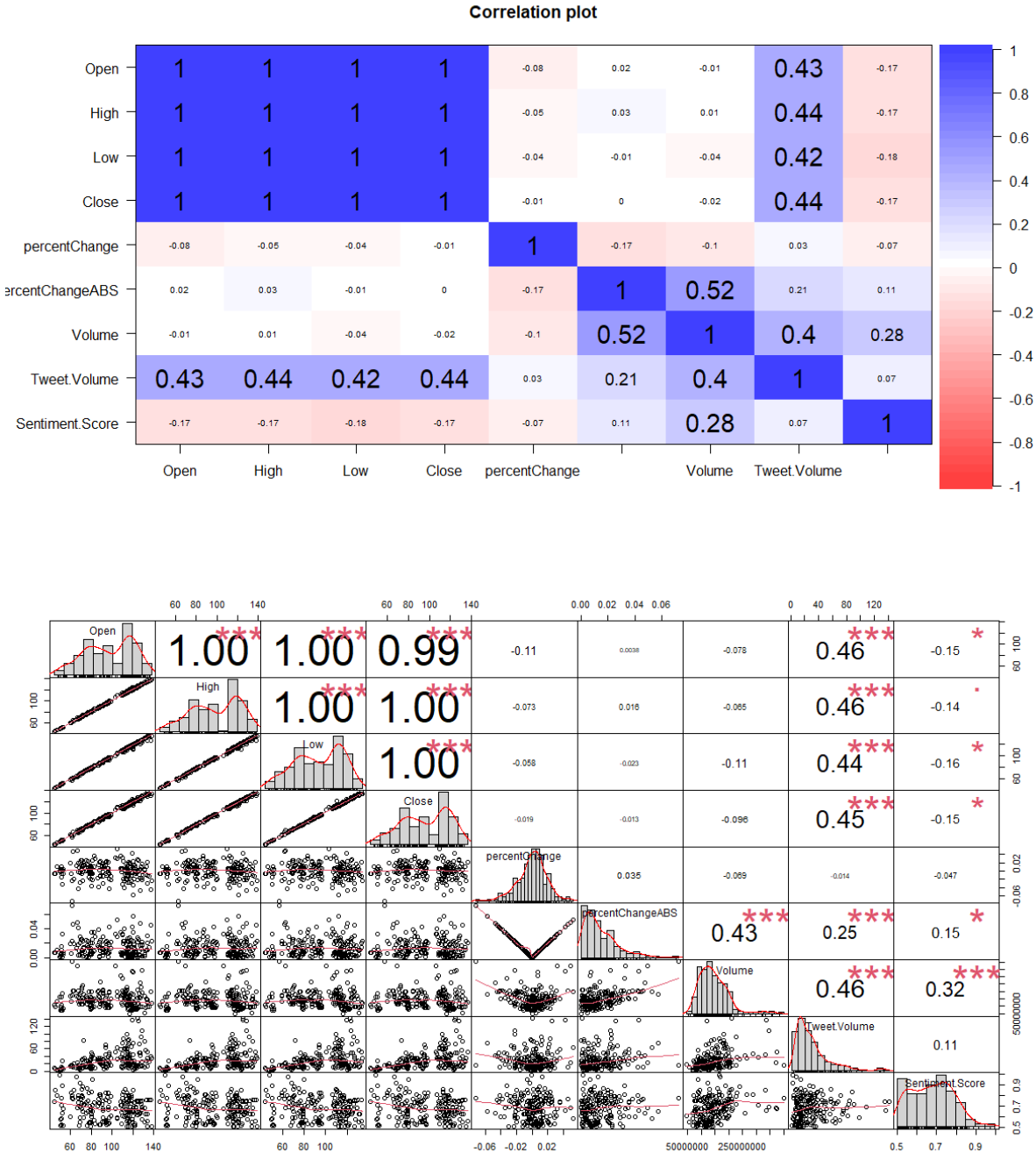
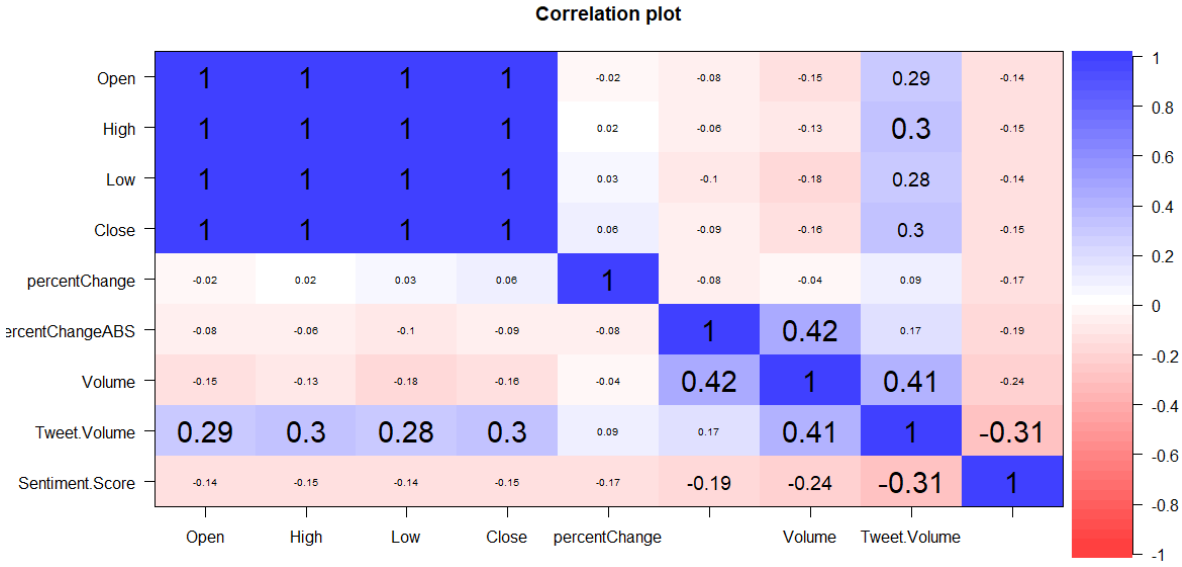


Figure 6: AAPL



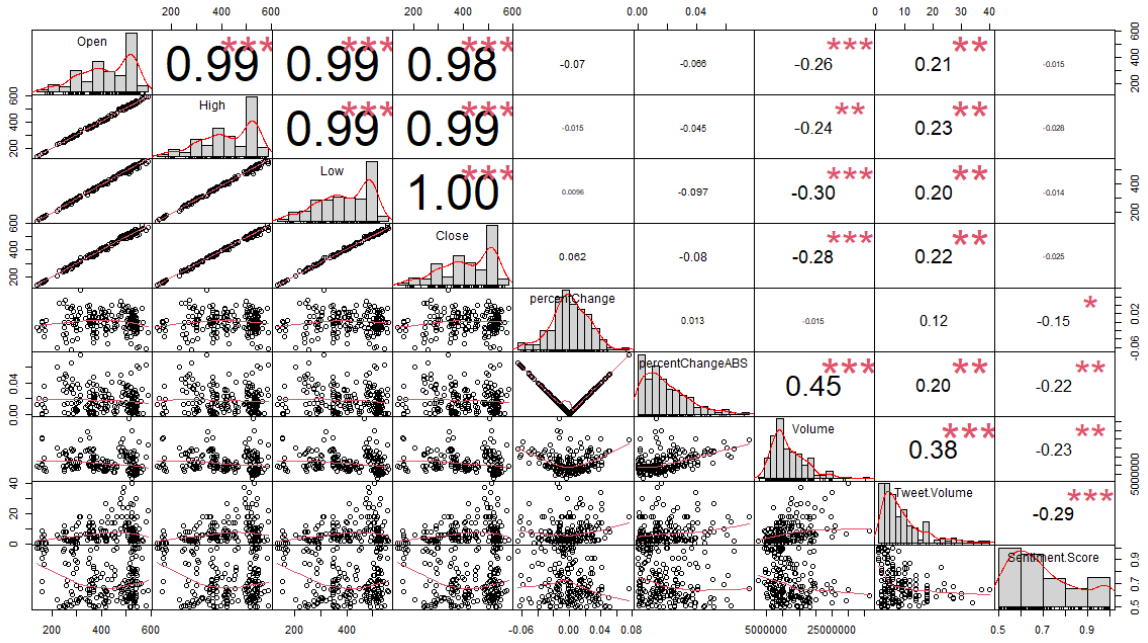


Figure 7: NVDA

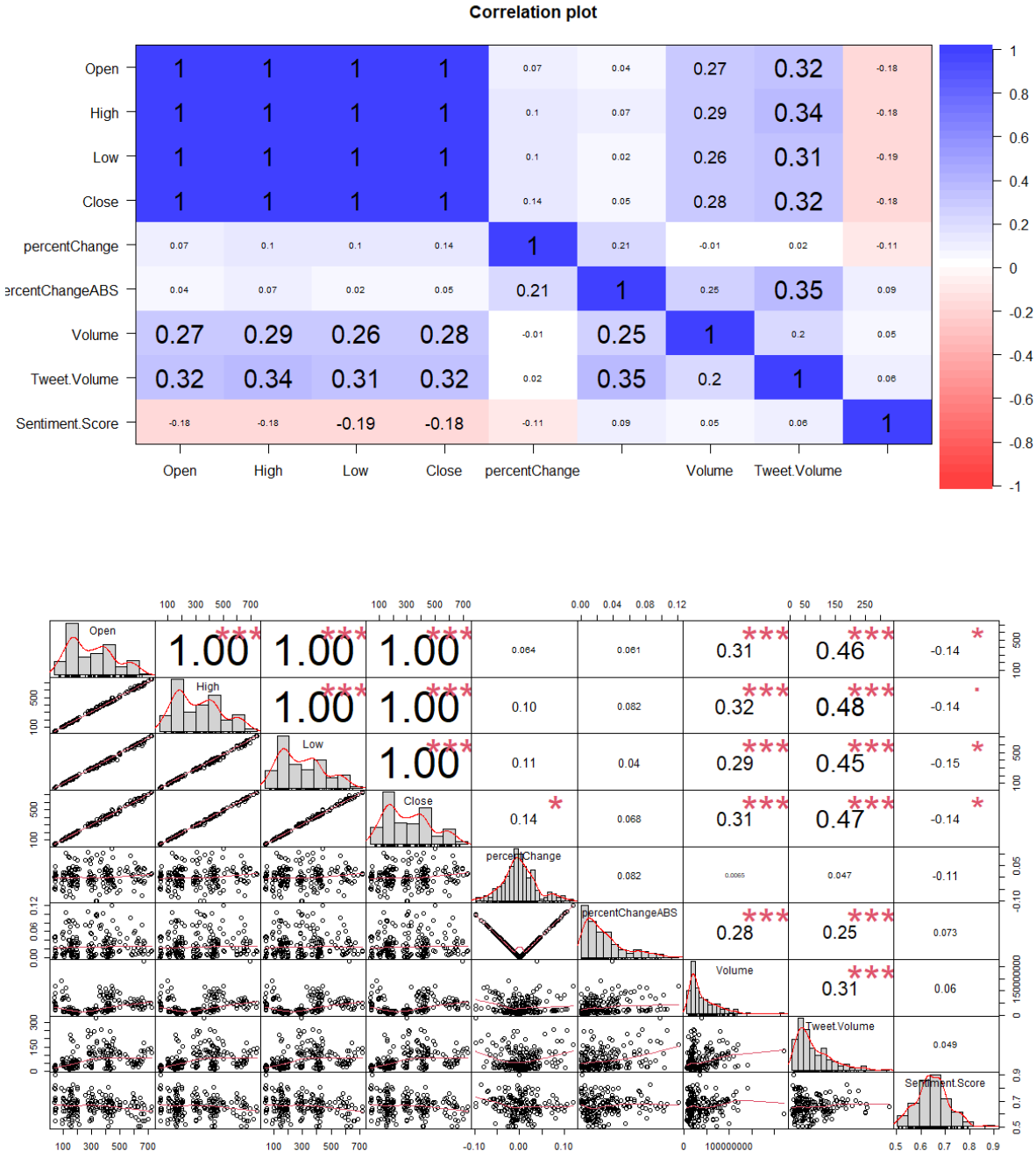


Figure 8: NFLX

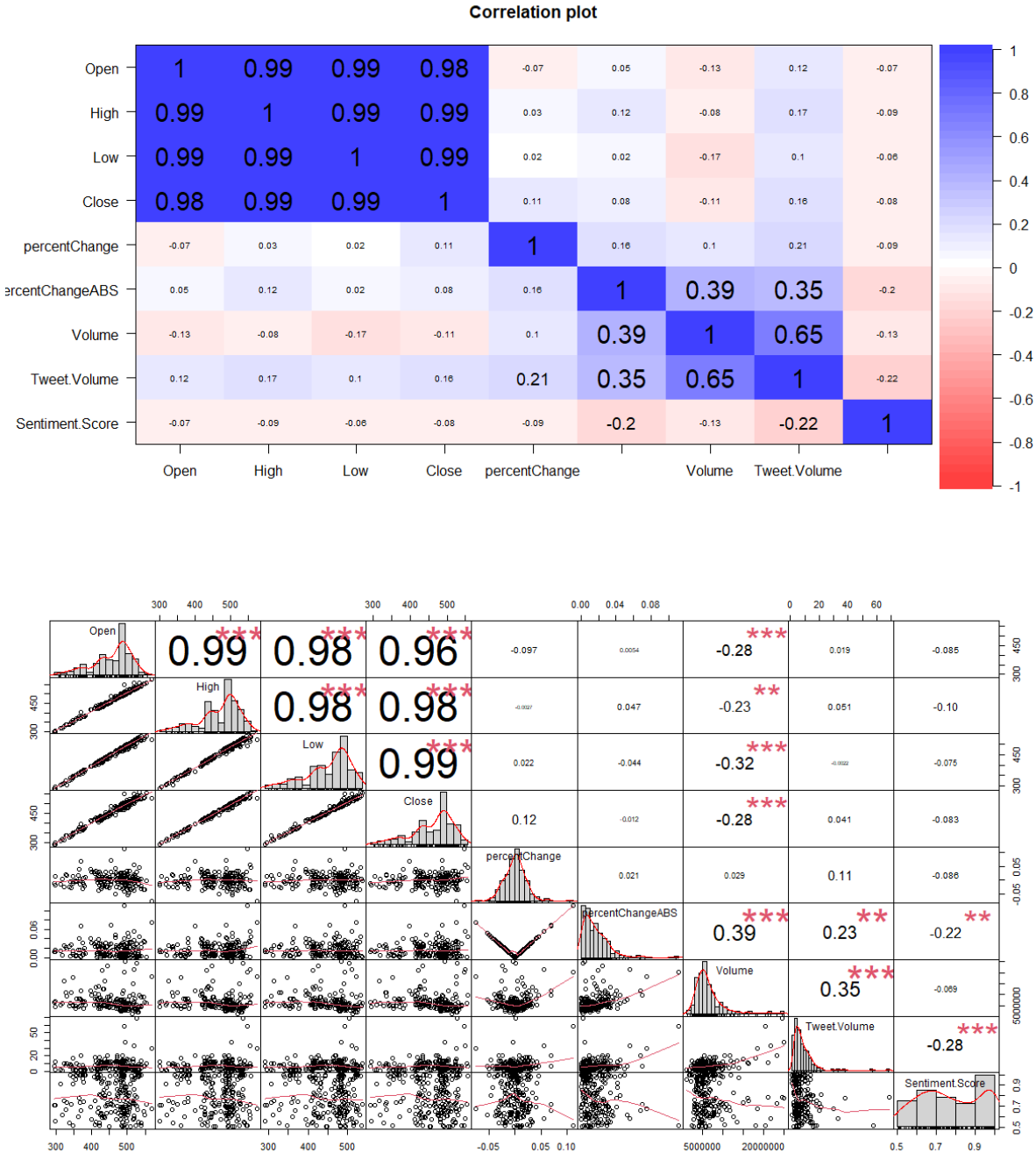
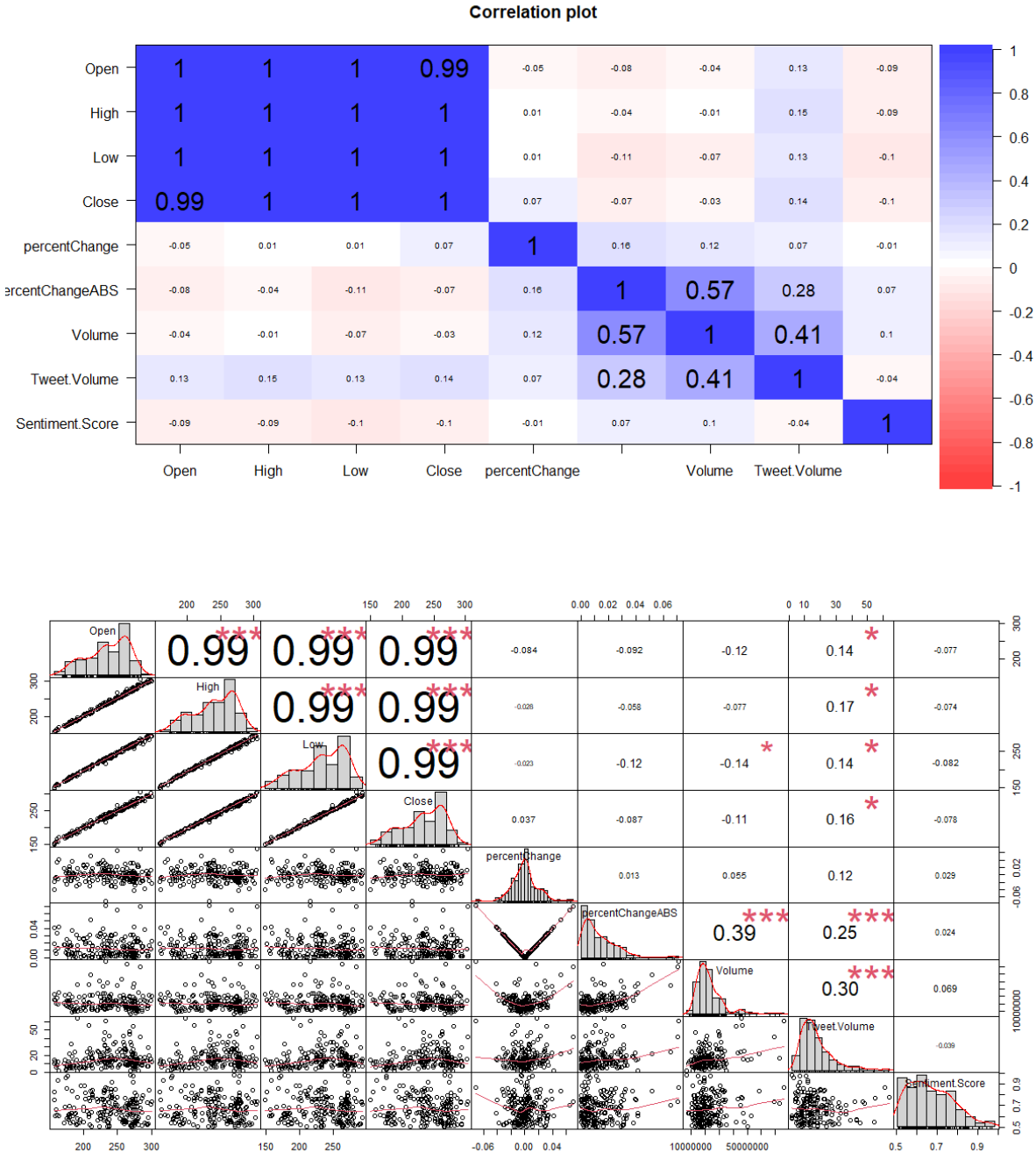


Figure 9: TSLA


Correlation plot
Open
High
Low
Close
percentChange
ercentChangeABS
Volume
Tweet.Volume
Sentiment.Score
1 1 1 0.99 -0.05 -0.08 -0.04 0.13 -0.09
1 1 1 1 0.01 -0.04 -0.01 0.15 -0.09
1 1 1 1 0.01 -0.11 -0.07 0.13 -0.1
0.99 1 1 1 0.07 -0.07 -0.03 0.14 -0.1
-0.05 0.01 0.01 0.07 1 0.16 0.12 0.07 -0.01
-0.08 -0.04 -0.11 -0.07 0.16 1 0.57 0.28 0.07
-0.04 -0.01 -0.07 -0.03 0.12 0.57 1 0.41 0.1
0.13 0.15 0.13 0.14 0.07 0.28 0.41 1 -0.04
-0.09 -0.09 -0.1 -0.1 -0.01 0.07 0.1 -0.04 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Open
High
Low
Close
percentChange
percentChangeABS
Volume
Tweet.Volume
Sentiment.Score
0.99*** 0.99*** 0.99*** -0.084 -0.092 -0.12 0.14* -0.077
0.99*** 0.99*** -0.028 -0.058 -0.077 0.17* -0.074
0.99*** -0.023 -0.12 -0.14* 0.14* -0.082
0.037 -0.087 -0.11 0.16* -0.078
0.013 0.055 0.12 0.029
0.39*** 0.25*** 0.024
0.30*** 0.069
-0.039

Figure 10: FB

The number of medium or strong correlated features for the day before and after dropped from 9 each. This pretty conclusively shows that that is little to no correlation between our twitter data and the market moving the day before or after. The main interpretation at this point is there is a partial correlation between stock volume and tweet volume in reference to that stock and this happens live. While we did not find any correlation between our sentiment analysis and the market, there is a lot of potential error to keep in mind. First off, the data set is sampled based on all of twitter, not just stock tweets, so there is not a consistent proportion of stock tweets at any given time. Another issue is the limitation in sentiment analysis technology. Given the size of our dataset, it is possible errors were overlooked.

4 Our Model

In order to build out a model, we first needed to clean and prepare our data significantly. A huge thing that we learned from the project was that cleaning and preparing the data might be the most important aspect of building a successful machine learning model. We will explain the steps in chronological order below.

4.1 Steps

Step 1: We chose 21 stocks that we found to have the strongest correlations with twitter sentiment than the other stocks did. These are more popular stocks that are talked about on the news and on social media. The stocks are American Airlines, Apple, Amarin, Amazon, Dynamic Materials Corporation, ENGlobal Corporation, Facebook, Gilead Sciences, Heat Biologics, Inovio Pharmaceuticals Inc, Intel, JOST Werke, Marathon Digital Holdings, Netflix, Novavax, Nvidia, Penn National Gaming, Plug Power, Sorrento Therapeutics, Trillium Therapeutics, and Tesla. Intuitively, it makes sense that bigger and more well known stocks are more correlated with twitter than others. A lot of these stocks such as Apple, Amazon, Facebook, Netflix, and Tesla, we were able to collect a significant amount of tweets associated with them. The more data that we could collect, the better our model would perform in theory.

Step 2: We collected hourly data for each weekday for each stock. We did this using the Twitter API. Since the Twitter API only allows you to retrieve tweets going back 7 days, we collected this data over the course of the month of April. At the beginning of our data collection process, we were collecting a maximum of 500 tweets per stock each day. This was due to data retrieval timing. As we collected data throughout the month, we tried a different number of max tweets to pull in order to find the most efficient and effective number of tweets to pull. At the end of our data collection, we were pulling a maximum of 10,000 tweets per stock each day. While 10,000 was the maximum, none of the stocks ever had 10,000 tweets associated with it. The tweet sentiments were averaged for the

stock_symbol	date	time	Open	Close	Happy	Angry	Surprise	Sad	Fear	Sentiment.Score	Stock.Volume	Tweet.Volume
TSLA	2021-04-07	04:00:00	NA	NA	0.1111110	0.0000000	0.2222220	0.2222220	0.4444440	0.3068980	NA	9
TSLA	2021-04-07	05:00:00	NA	NA	0.0000000	0.2000000	0.0000000	0.3000000	0.5000000	0.2921740	NA	10
TSLA	2021-04-07	06:00:00	NA	NA	0.4000000	0.0000000	0.2000000	0.0000000	0.6000000	0.6332030	NA	5
TSLA	2021-04-07	07:00:00	NA	NA	0.1666670	0.0000000	0.3333330	0.0000000	1.0000000	0.4791040	NA	6
TSLA	2021-04-07	08:00:00	NA	NA	0.1666670	0.0000000	0.5000000	0.0000000	0.1666670	0.6467930	NA	6
TSLA	2021-04-07	09:00:00	NA	NA	0.2500000	0.0000000	0.0000000	0.0000000	0.3750000	0.2599350	NA	8
TSLA	2021-04-07	10:00:00	688.905	689.690	0.1428570	0.0000000	0.1428570	0.0000000	0.2857140	0.5378050	2104095	7
TSLA	2021-04-07	11:00:00	686.950	684.280	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.8121140	1522432	2
TSLA	2021-04-07	12:00:00	684.380	676.975	0.0000000	0.0000000	0.2000000	0.0000000	0.6000000	0.5995010	2342371	5
TSLA	2021-04-07	13:00:00	675.084	678.072	0.3750000	0.0000000	0.1250000	0.1250000	0.5000000	0.6928640	1445008	8
TSLA	2021-04-07	14:00:00	674.475	676.436	0.0555556	0.0000000	0.1111110	0.2222220	0.3888890	0.5692120	2135673	18
TSLA	2021-04-07	15:00:00	673.490	669.430	0.1724140	0.0000000	0.1379310	0.1034480	0.5517240	0.6700020	1824408	29

Figure 11: Figure A

stock_symbol	date	starting_open	ending_close	mean_happy	max_happy	min_happy	sd_happy
TSLA	2021-04-06	687.239	692.300	0.13060186	0.333333	0.0000000	0.10415093
TSLA	2021-04-07	688.905	669.430	0.11672060	0.400000	0.0000000	0.11346220
TSLA	2021-04-08	679.490	682.260	0.10564777	0.235294	0.0000000	0.07674672
TSLA	2021-04-09	676.370	674.350	0.22061432	0.769231	0.0588235	0.20366753

Figure 12: Figure B

hour from all the tweets collected for that hour. Figure A shows a subset of the hourly data. It shows Tesla's hourly data for April 7th, 2021. As you can see from the figure, we have open price for the hour, closing price for the hour, mean happy score, mean angry score, mean surprise score, mean sad score, mean fear score, mean sentiment score, volume of stock traded that hour, and volume of tweets collected that hour.

Step 3: We took the daily mean, maximum, minimum, and standard deviation of the hourly emotion sentiment scores for each stock. A subset of this data is shown as Figure B. It represents the data collected from the week of April 6th to April 9th for Tesla's stock. From the figure, you can see that we have the starting open price of the stock that day, the ending close price of the stock that day, the mean happy score, the maximum happy score, minimum happy score, and standard deviation happy score. While the figure only shows happy, we also have those features for angry, surprise, sad, and fear as well.

Step 4: We created a new column called "increase." This column is TRUE if the stock price closes higher on the next workday than it did today. It is FALSE otherwise. This is our target variable. It is worthy to note that the breakdown of the increase column was 55% FALSE and 45

Step 5: We omitted rows with NAs.

Step 6: Using correlations (albeit very weak), we tried multiple machine learning models to try to predict the increase column. When trying these machine

learning models, we tried a large number of different combinations of the features to find the best performing model that we could. We ended up trying 4 different types of models.
1. The XGBoost was by far the worst performing model that we tried. This was somewhat surprising, but this model was no better than a random guess and had an ROC of around 0.52
2. The ridge regression model was the best performing model of the bunch. We will get into the metrics and how it performed below.
3. The lasso regression model performed decently well, but not as well as ridge.
4. The neural network model performed pretty similarly to the lasso model. We did consider using this model as our final model, but opted for ridge since the ridge regression model is significantly more interpretable than a neural network. The ridge regression model also ran instantaneously, while the neural network model took anywhere from a few seconds to a few minutes to run depending on the parameters we inputted.

4.2 Ridge Regression Model Coefficients

Lambda = 0.1172175
min_angry = -215.906087
sd_surprise = 1.576678
mean_sad = -3.575914
sd_sad = -1.059764
mean_fear = 1.914781

4.3 ROC Curve

The ROC curve for our model is labeled below as Figure C. The ROC curve plots the false positive rate against the sensitivity, which is the true positive rate for different values as the cutoff point. Our area under the curve is 0.58. While this may seem like not a great value, it seems decent given the context of the problem. We are trying to predict the stock market. This is something that experts have been trying to predict for 100s of years. If we were able to predict the stock market with a very high AUC, we'd be millionaires. Given that we are trying to predict a money making environment, an AUC of 0.58 looks promising in our opinion.

4.4 Confusion Matrix

Confusion Matrix: Using the optimal cutoff point that maximizes the accuracy of the model, we created a confusion matrix that is denoted as Figure D. It is worthy to note that since the majority class is FALSE, the confusion matrix labels FALSE as the positive class. To clarify on some terminology, the no information rate is the percentage of the target variable that is the majority class.

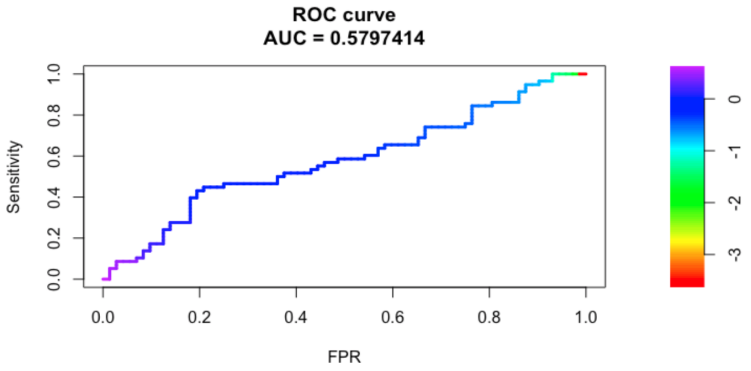


Figure 13: Figure C

A model that predicted the FALSE class every single time would be correct 55.38% of the time. At a minimum, a model that we build should be better than that.
Accuracy: 63.85%
95% Confidence Interval for accuracy: (54.96%, 72.09%)
P-Value = p(Accuracy ¿ No Information Rate) = 0.03
58 True Positives, 25 True Negatives, 14 False Positives, and 33 False Negatives

4.5 Machine Learning Metrics

Figure F shows the observed value of increase given the predicted value. When our model predicts FALSE, it is correct 64.10% of the time. When our model predicts TRUE, it is correct 63.74% of the time. Keep in mind that this model is just another tool to help you in the stock market. It is not an end all be all to investment strategies. Additionally, our model does not detect the magnitude of rises and falls. It just detects whether or not a stock price goes up or down. It is possible that 36% of wrong predictions could equal the same magnitude as 64% of correct predictions. While further exploration could be done using linear regression to detect magnitude of increases or decreases, we focused on classification models.

5 Conclusion

The big takeaway from this project is the data limitations when dealing with historical tweets. It is either expensive or needs to be planned years ahead of time for data collection. While that is the case, we found traces of a connection in our exploratory analysis and out model looked even better. It is safe to say that there is a connection between certain stocks and what is being talked about on twitter.

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE    58   33
     TRUE     14   25

               Accuracy : 0.6385
                 95% CI : (0.5496, 0.7209)
    No Information Rate : 0.5538
    P-Value [Acc > NIR] : 0.03120

                  Kappa : 0.2444

 Mcnemar's Test P-Value : 0.00865

            Sensitivity : 0.8056
            Specificity : 0.4310
         Pos Pred Value : 0.6374
         Neg Pred Value : 0.6410
             Prevalence : 0.5538
         Detection Rate : 0.4462
   Detection Prevalence : 0.7000
      Balanced Accuracy : 0.6183

       'Positive' Class : FALSE
```

Figure 14: Figure D

	Precision	Recall	F1
FALSE	0.8055556	0.6373626	0.7116564
TRUE	0.4310345	0.6410256	0.5154639

Figure 15: Figure E

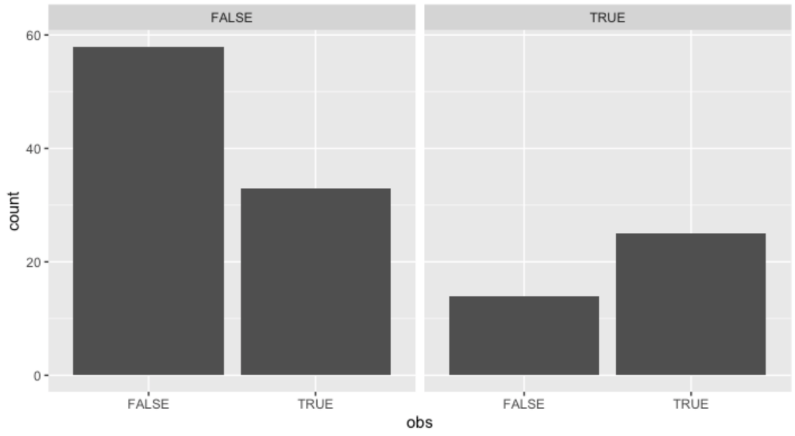


Figure 16: Figure F

References

[1] Johan Bollen, Huina Mao, Xiao-Jun Zeng: Twitter mood predicts the stock market. https://arxiv.org/pdf/1010.3003&embedded=true