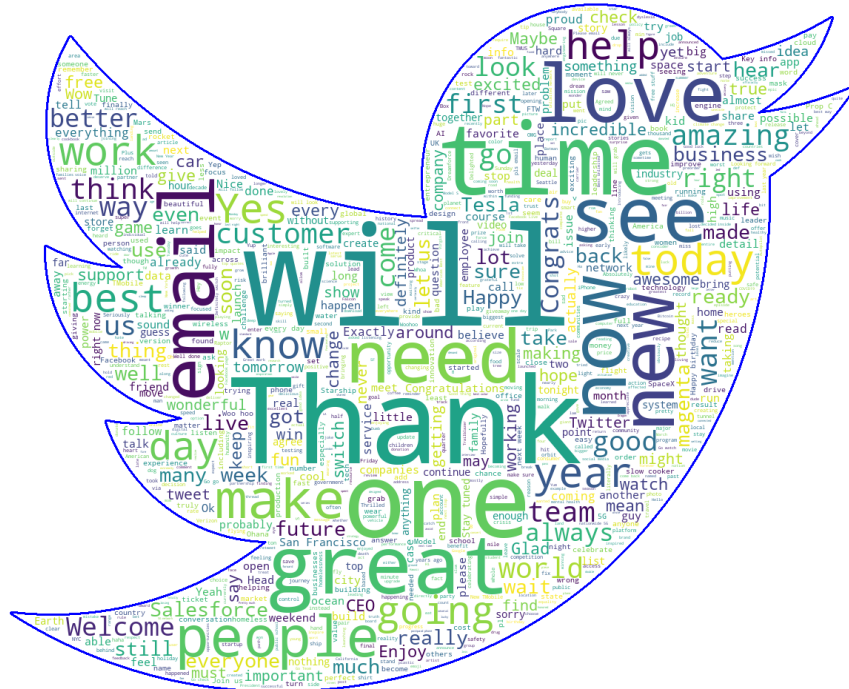


# Predicting Company Stock Movements with CEO Tweets



Tyler Poore  
Oct, 2021

<b>1. Introduction</b>	<b>3</b>
<b>2. Business Scenario</b>	<b>3</b>
<b>3. Data Information</b>	<b>4</b>
Data Collection	4
<b>4. Data Wrangling</b>	<b>5</b>
Stocks	5
Tweets	5
Final DataFrame	5
Additional Notes	6
<b>5. Exploratory Data Analysis and Inferential Statistics</b>	<b>6</b>
Distributions	6
Bivariate Distributions	8
Correlation Matrices	9
Feature Importance	10
<b>6. Preprocessing and Machine Learning</b>	<b>11</b>
Feature Engineering	11
Text Preprocessing	11
Feature Preprocessing	12
<b>7. Modeling</b>	<b>12</b>
Selecting the Right Metrics	12
Training/Testing Splits	12
Model Selection	13
Hyperparameter Tuning	13
Model Testing	14
<b>8. Summary</b>	<b>16</b>
Interpretation	16
Explanation 1: Predictive Word (Virgin)	16
Results	17
Explanation 2: Misclassifying Move as Stay	18
Results	18
Threshold Scenario	19
Explanation 3: Misclassifying Stay as Move	20
Results	20
Threshold Scenario	21
<b>9. Conclusion</b>	<b>22</b>
Major Findings	22

Model Implementation	22
Next Steps/Assumptions	23
Improving the Current Model	23
Future Work	24
<b>10. Sources</b>	<b>24</b>

---

## 1. Introduction

In August of 2018, Elon Musk, CEO of the electric car company, Tesla, sent out the tweet "Am considering taking Tesla private at \$420. Funding secured." The result of this tweet was a \$40M fine, Musk losing his chair at Tesla, and an 11% increase of Tesla's stock price at the day's close [1]. In spite of his censorship, Musk continued to send out tweets that directly impacted his own company's stocks. And in 2020, Tesla's stock fell sharply by 10.8% at the close when Musk famously tweeted "Tesla stock price is too high imo" [2]. Investors that have, in the past, carefully done their due diligence in deciding how to allocate their funds are now faced with this new element in how stocks perform: that of social media. Highly influential individuals can, and have, voiced their opinions on the market through various social media platforms, causing prices to rise or fall within hours of publication, and many have claimed this is outright market manipulation, calling for regulators to get involved. Nonetheless, today's investing strategies now have to rely on closely monitoring social media to alert them for these instances.

Tesla CEO Elon Musk got into hot water with regulators back in 2018 for infamously tweeting about taking the company private at a nice round stock price of \$420. ... In May 2020, Musk tweeted that "Tesla's stock price is too high imo." That day Tesla's stock price closed 10 percent lower than the day before.

Elon Musk tweets have been in the news because of how they are affecting company stock prices. The sentiment now is really a big deal. Social media has changed how stock trends are analyzed. <https://www.washingtonpost.com/technology/2020/05/01/musk-tesla-stock/>

---

## 2. Business Scenario

The accessibility and ubiquity of social media has created a new problem for investors, in which influential individuals can cause dramatic changes in the market. For example, on January 28th, shares for the CD Projekt developers of Cyberpunk 2077 surged over 12% after Elon Musk tweeted that he liked the game the night before. In another example, Tesla shares fell over 10% when Musk tweeted that the company's valuation was too high. The rapid market changes from these tweets has caused investors to be fixated on monitoring social media for similar occurrences.

Andy Swan writes in Forbes that there are 4 types of tweets to look for that influence stocks.

- Influential people or groups tweets about a company's stock
- Company tweets about its own stock or by influential investors
- Policy events that are tweeted by government insiders or observers
- Influencer tweets an opinion about a particular brand or product

For this application, it will be determined if an executive's tweets can be an accurate predictor of stock changes for their company. A CEO that has a well-known public image can potentially influence their stock prices from social media (Elon Musk as an example), and would fit within the types of tweets we want to be looking for that Andy Swan mentions.

---

## 3. Data Information

### Data Collection

A maximum of 5-years of stock data were collected from Yahoo! finance using **yfinance**. All stocks were collected between *August 23rd, 2016* to *July 20th, 2021* with the exception of SPCE (Virgin Galactic) where the earliest public stock price is *September 29th, 2017*.

Twitter data was scraped via **twint**. Twint has no time restrictions when scraping data, so the corresponding stock start dates were used. Company executives were selected with the criteria that they must have at least 1000 tweets over the scraped time, and they are or were high-level executives in their company with a strong, publicly-known association. The following CEOs were picked using their twitter handles, along with the number of tweets that were scraped and they corresponding stocks:

	Twitter Handle	Company Affiliation	Number of Tweets	Company Ticker	Number of Stock Days
Elon Musk	@elonmusk	CEO and product architect of Tesla, Inc.	11622	TSLA	1259
Aaron Levie	@levie	Co-founder of Enterprise Cloud Software Box	1736	BOX	1259
Jack Dorsey	@jack	CEO of Twitter, Inc.	5124	TWTR	1259
Marc Benioff	@Benioff	Founder, chairman and CEO, Salesforce	4871	CRM	1259
Richard Branson	@richardbranson	Founder of Virgin Galactic	7711	SPCE	981
John Legere	@JohnLegere	Former CEO T-Mobile US	37302	TMUS	1259

*Figure 1. Summary of Collected Data*

## 4. Data Wrangling

### Stocks

The open and close stock prices with their dates and times were combined into a single dataframe for each collected stock. The index was reset to a date index spanning the first and last dates of the collected stocks. Dates in which the market was closed were set to NaN. The percent change between each stock price was calculated, which we will use as our dependent feature.

### Tweets

The tweets were organized into Pandas DataFrames containing 37 features for each pulled tweet. Tweets in languages other than english were removed and the features pertaining to the text in the tweet (e.g. hashtags) were selected. The features were converted from lists to integers counting feature occurrences per tweet. Empty lists were set to 0.

The tweet and stock dataframes were merged together by date. As we are hypothesizing that tweets will have an impact on the opening or closing stock price, posts were grouped together based off this hypothesis, i.e. tweets that occur between 9:30 am and 4:00 pm on a weekday were grouped into one tweet assuming they would impact that day's closing price. Posts between 4:00 pm the next stock open were grouped together as another tweet. This could mean multiple days could be grouped together as one tweet since the number of days between the last closing price and the next opening one varies due to weekends and holidays. The total number of tweets collected per group was added.

### Final DataFrame

The final dataframe with the collected tweets, tweet features, and associated stock percent changes was saved for further processing. The dataframe contains 8445 rows with 11 features. A sample row is shown below.

	date	tweet	username	mentions	hashtags	video	photos	urls	number of tweets	price	percent change
1650	2019-12-11 16:00:00	@ItalianPali1 @SievertMike Make one for us! Can you say 3-PEAT? 🏆🏆🏆 @TMobile has been awarded 'Best Place To Work' by @Glassdoor for the 3rd year in a row!! Did I mention that we're also the #1 place to work in wireless?? #AreYouWithUs <a href="https://t.co/iDGLSWQ5vQ">https://t.co/iDGLSWQ5vQ</a> We're at the HALFWAY point of #JollyJohn!! Blast your favorite holiday song with 🎧 @Samsung Galaxy Ear Buds!! 🎵 To win, tell me what present @BraxtonCarter got in the gift exchange? Follow him and use #JollyJohn in your answer below! <a href="https://t.co/onCEmihc2Z">https://t.co/onCEmihc2Z</a>	JohnLegere	4	3	1	1	1	3	74.620003	-0.006127

Figure 2. Sample Row of DataFrame

## Additional Notes

Recently the twint api has been unable to scrape data correctly, and so the twitter data used is from a previous save state, as is the stock data to ensure both have matching date ranges.

# 5. Exploratory Data Analysis and Inferential Statistics

The goal of EDA is to explore the relationships between variables in a dataset. There are some questions that we will want to answer as we explore the data.

1. How is the data distributed?
2. Are there any features that correlate with the percent change in stock prices?
3. What does binning the target feature for classification do to the distributions?
4. Do the CEO tweets differ in any significant way?

## Distributions

The distributions of the 6 different features were examined. As can be seen in figure X. The features are heavily skewed. A log transformation was used to better symmetrize the data (fig. X

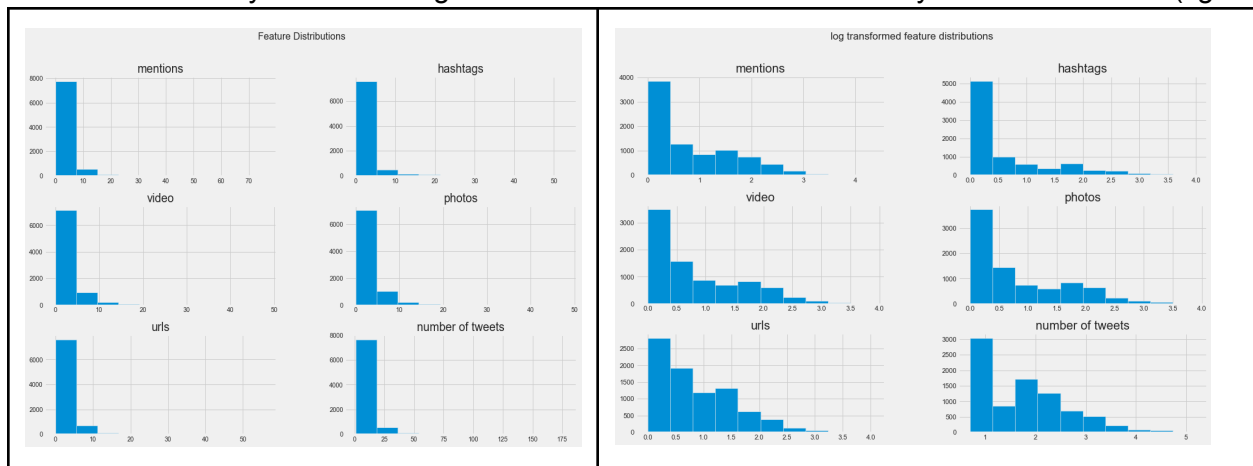


Figure 3. Distribution of Features (Left) with Histogram. Log Transformed Features (Right)

We next looked at the mean feature distributions for each CEO. What stuck out is that Richard Branson's usage of additional tags and media in his tweets, (e.g. pictures, hashtags, urls) seems disproportionate with how frequently he tweets.

The average of the sums of the url, video, hashtag, mention, and photo features was compared against the average number of tweets for each CEO (fig X).

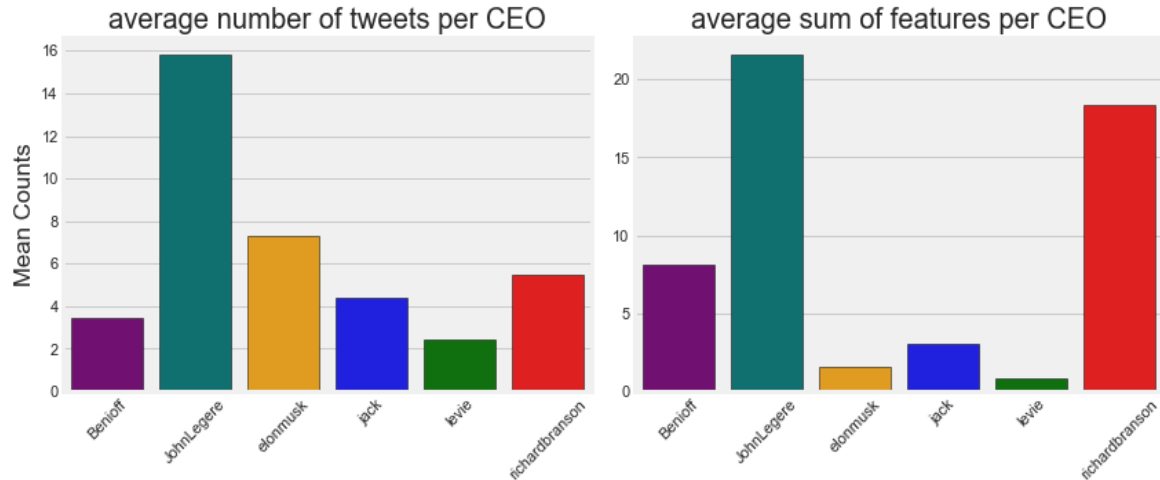


Figure 4. Average Number of Tweets per CEO (left), Average Sum of Hashtags, Mentions, Urls, Videos, and Photos per CEO (right)

The distribution of the percent change values was next examined. In the histograms shown below, stocks that moved above or below a percent change threshold were labelled as move and colored in green. Stocks that did not move beyond this threshold were labelled stay and colored in red.

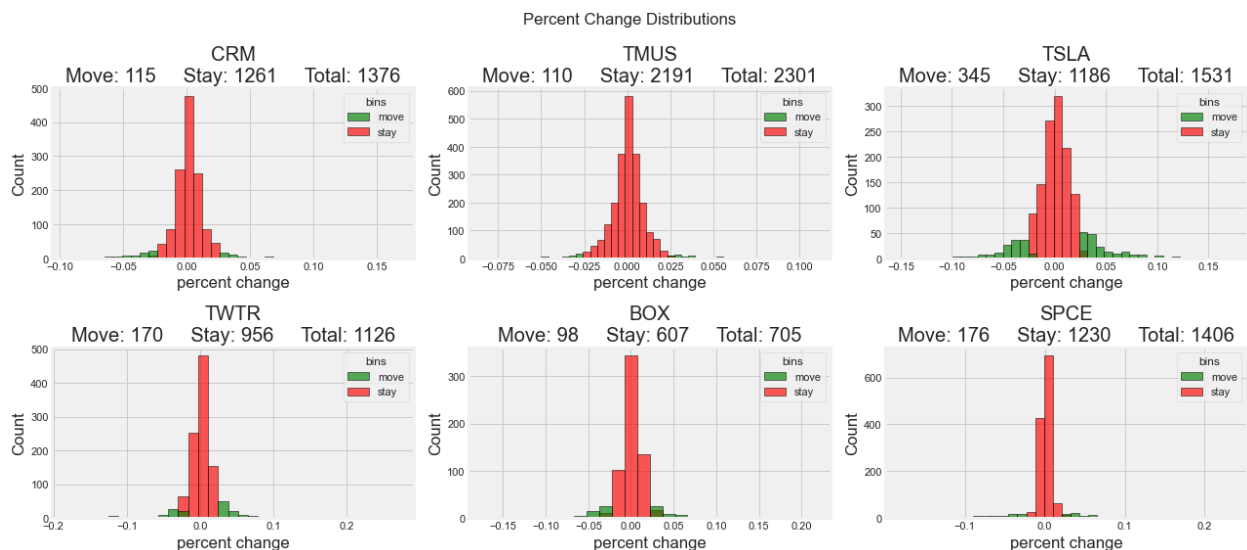


Figure 5. Distributions of Stock Returns (Percent Change) in Green for Move, Red for Stay. Cutoff of bins is +/- 2.5%

It was decided to bin the percent change values as to whether or not they indicated a significant stock movement. In this case, a move of  $\pm 2.5\%$  was chosen for the following reasons.

- **Differentiation**

We hypothesize that stock movement from tweets will depend on the features in the tweet. (e.g. a higher number of hashtags in a tweet predicts stock price movement). To allow this, we want to capture the largest percent change cutoff between stocks that move beyond that cutoff (labelled as move) vs those that are under that cutoff (labelled as stay). If our hypothesis is true, then the feature distributions should be much more different between stocks that move only a little, vs stocks that move a lot.

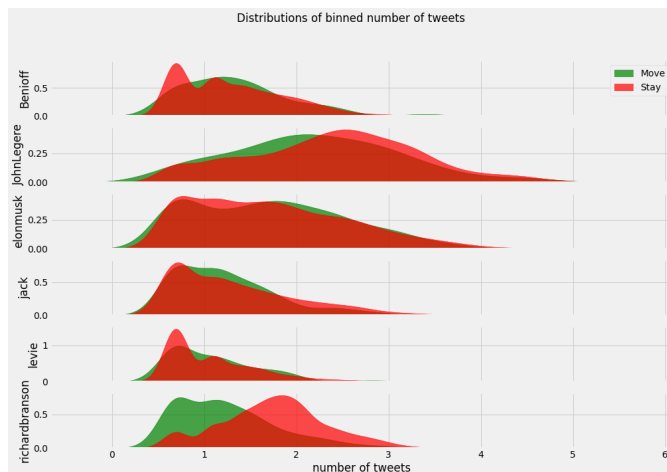
- **Sample Size**

We are trying to predict stock movement through classification, and both the disparity between the two classes (move and stay) and the number of positive samples (move) we can use for prediction will impact our model. Though we would ideally like balanced data, this restricts the variation between the features that we want to capture.

We can change the bin cutoff for testing purposes, but for the purpose of maximizing feature differences and the number of samples in the move class, 2.5% is used moving forward.

## Bivariate Distributions

The kernel density estimations of the different features were plotted together with features where stocks move colored green and those that don't in red. The KDE for the number of tweets for each CEO is shown below.



*Figure 6. KDE Plots of Number of Tweets per CEO*



We can see that the distribution in the number of tweets varies significantly with stock movement for Richard Branson. Even though there is a large overlap in the binned values, the visible density difference in a lower number of tweets before an open/close price seems to indicate that the stock will move beyond  $\pm 2.5\%$ .

## Correlation Matrices

Heatmaps of the feature correlations for each CEO are below. The bin categories, stay and move, were added as binary features for visualization. The heat color corresponds to Pearson's  $r$  correlation value between each feature.

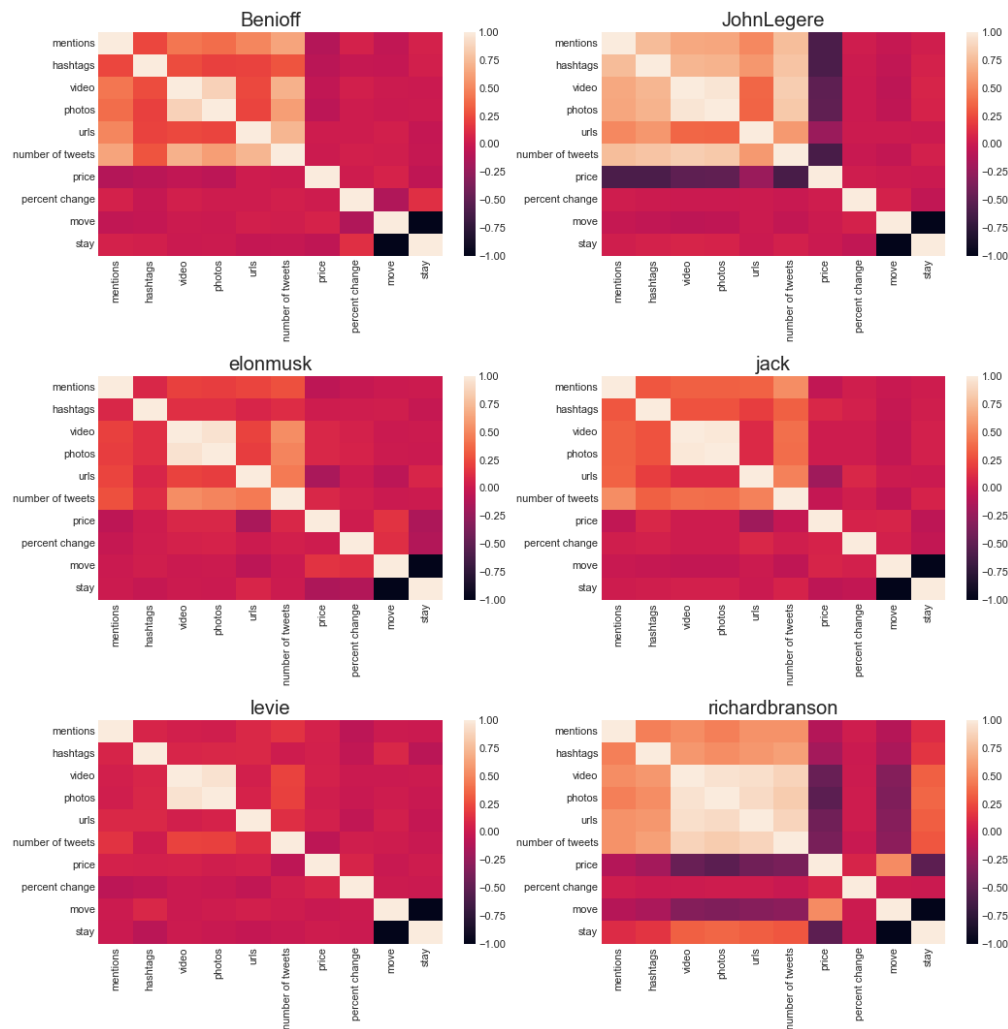


Figure 7. Heatmaps of Features per CEO. Metric used is Pearson Correlation Coefficient.

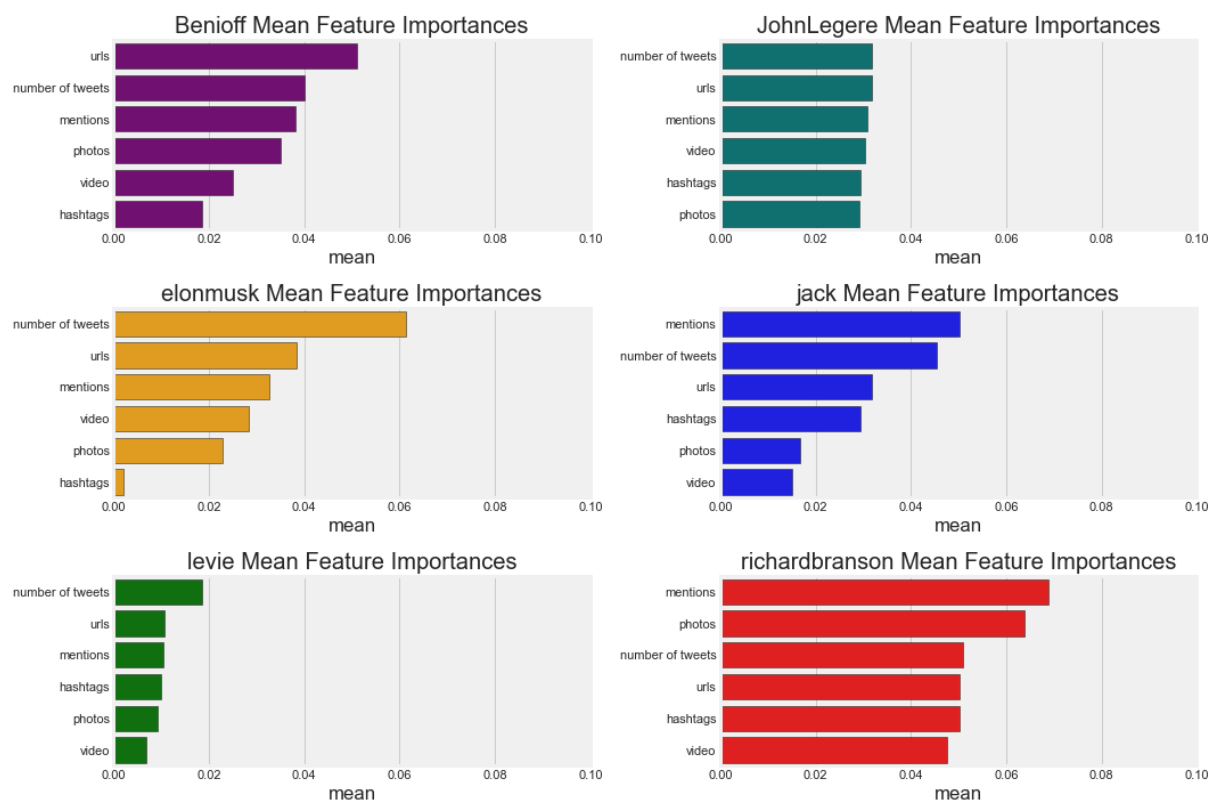
Strong correlations between the mentions, hashtags, video, photos, urls and number of tweet features can be seen for each executive in varying degrees. There are very strong correlations between Richard Branson's media features and the number of tweets, and they look to be positively correlated with stocks that don't move and negatively correlated with stocks that do. This strong

correlation between the features and bins isn't present in any of the other CEO heatmaps. It is also interesting that the price and move bins are both negatively correlated with the other features in the dataframe.

The percent change values themselves don't have any strong correlations across any of the scraped tweet features, save for a small positive correlation in Elon Musk's tweets, but the tweets collected from Richard Branson seem to have some unique correlations that may impart some predictability in our machine learning model.

## Feature Importance

The feature importance was determined through random forest classification using scikit learn for each of the CEOs. The higher the value, the more important the feature. The importance of a feature is computed as the Gini importance.



*Figure 8. Mean Feature Importances per CEO via Random Forest Classification*

The degree of feature importance for each CEO, note the degree of feature importance across different CEOs (e.g. Richard Branson vs Aaron Levee). Number of tweets, urls and mentions seem to be important across all different tweets.

---

## 6. Preprocessing and Machine Learning

### Feature Engineering

In an attempt to tease more data from the raw tweets, the counts of various features derived from the text were created:

- emoji counts
- number of sentences
- number of words
- avg number of words per sentence

Additional features using Spacy were extracted, using the en\_core\_web\_lg model for natural language processing:

1. number of nouns
2. number of verbs
3. number of stopwords
4. number of organizations mentioned in tweets (e.g. Tesla, SpaceX, Twitter, etc.)

### Text Preprocessing

The twitter data needs to be cleaned prior to machine learning. In this effort, special characters, punctuation, numbers, and emojis were removed. The text was converted to lowercase and stop words, words shorter than 1 character, and words appearing only once in a document were removed. The final text is lemmatized before returning the cleaned text data.

Tweet Increases Stock by 10.391%

=====

date	tweet
2018-08-07 16:00:00	Investor support is confirmed. Only reason why this is not certain is that it's contingent on a shareholder vote. @MindFieldMusic Def no forced sales. Hope all shareholders remain. Will be way smoother & less disruptive as a private company. Ends negative propaganda from shorts. @thatspacegeek Yes, but liquidity events would be limited to every 6 months or so (like SpaceX) Shareholders could either to sell at 420 or hold shares & go private @heydave7 Absolutely. Am super appreciative of Tesla shareholders. Will ensure their prosperity in any scenario. @Gfilche My hope is all current investors remain with Tesla even if we're private. Would create special purpose fund enabling anyone to stay with Tesla. Already do this with Fidelity's SpaceX investment. @EvotoRentals Yes @FredericLambert I don't have a controlling vote now & wouldn't expect any shareholder to have one if we go private. I won't be selling in either scenario. Good morning 🌞 @LizClaman @Tesla 420 Am considering taking Tesla private at \$420. Funding secured.

*Figure 9. Raw Sample Tweet*

date	cleaned
2018-08-07 16:00:00	investor support confirm reason not certain contingent shareholder vote def force sale hope shareholder remain way smoother disruptive private company end negative propaganda short yes liquidity event limited month like spacex shareholder sell hold share private absolutely super appreciative tesla shareholder ensure prosperity scenario hope current investor remain tesla private create special purpose fund enable stay tesla fidelitys spacex investment yes not control vote not expect shareholder private not sell scenario good morning consider take tesla private funding secure

*Figure 10. Sample Tweet after Cleaning and Pre-Processing*

## Feature Preprocessing

The cleaned text was converted into a sparse vector matrix to allow for machine learning. Only unigrams were considered, and the min\_df was kept to the default of 1. Both of these features can be modified to see if any improvements in prediction capability can be made by looking at bi and tri-grams and the minimum number of documents that can be kept, but for this project unigrams were kept as support vector machine fit times scale quadratically with sample size. Because data is limited, the min df was kept to 1. Count Vectorizer was also considered, but early explorations didn't see much of a difference between these vectorization methods.

The natural log of the features was used to assess the dataset. Because the scale of the different count features varies significantly, the scales of the features were normalized between 0 and 1 using Min Max Scaler. Other scaling methods were not explored.

---

## 7. Modeling

### Selecting the Right Metrics

We could look at this problem in one of two ways: if an investor were to use this machine learning approach to evaluate investment opportunities from twitter data, assuming predicting stock price changes are the positive label and no stock changes are negative, an investor would prioritize predicting stock changes to examine investment opportunities, that is to minimize opportunity loss. This would be a reason to look at recall scores for model metrics as it provides an indication of missed positive predictions.

Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions.

For imbalanced learning, recall is typically used to measure the coverage of the minority class. Other metrics like F1 and the ROC-AUC score are also appropriate as it incorporates loss mitigation as a factor (both are indicators of how to optimize both precision and recall).

We will use F1 as our base metric so that we can later tune recall and precision with modifying the threshold.

### Training/Testing Splits

Because of the imbalanced dataset, it's important to partition enough of the testing data for the model to make accurate predictions. It was decided that a 50%/20%/30% Training, Validation, and Tests split would allow for enough data to train the model and make accurate predictions.

## Model Selection

For this report, Logistic Regression, Support Vector Machines, and Multinomial Naive Bayes were used to assess stock change predictions on the processed Twitter data. Other classification methods were tried (see conclusions). We selected the f1 score as our metric and tested the models against the training and validation sets for each CEO. The best score obtained was with a LogisticRegression model from Richard Branson's tweets, with an average f1 score of 46.13 +/- 10.53%.

There is a significant decrease in f1 scores as we move from the second to the third best models. The second best was again Richard Branson with support vector classification ( f1 score 44.92 +/- 13.21%). Third was Elon musk with SVC (f1 score 28.10 +/- 6.53%). Because of this discrepancy in the CEO scores, we will only tune the Logistic Regressor and Support Vector Classifier hyperparameters on Richard Branson's data. As the Logistic Regression and the Support Vector Machine Classifiers performed similarly well, both were optimized by hyperparameter selection through GridSearch.

## Hyperparameter Tuning

	clf	params	solver	kernel	username	mean_test_score	std_test_score	mean_fit_time	mean_score_time
0	LogisticRegression(class_weight='balanced', n_jobs=-1, solver='sag')	{'C': 1.0, 'penalty': 'l2', 'solver': 'sag'}	sag	NaN	richardbranson	0.482597	0.143578	1.559464	0.033663
1	SVC(class_weight='balanced', probability=True)	{'C': 100.0, 'gamma': 0.001}	NaN	rbf	richardbranson	0.473582	0.056790	3.349731	0.269145
2	SVC(class_weight='balanced', probability=True)	{'C': 10.0, 'gamma': 0.01}	NaN	rbf	richardbranson	0.467778	0.049842	3.327912	0.276686

*Figure 11. Best Models after GridSearch. Mean Scores are F1 Scores.*

The best model obtained was a Logistic Regression with an average f1 score of 48.30 +/- 14.40%. Plotting the ROC AUC Curve shows that the model can make relatively good predictions on the training and validation test sets. (AUC = 0.82).

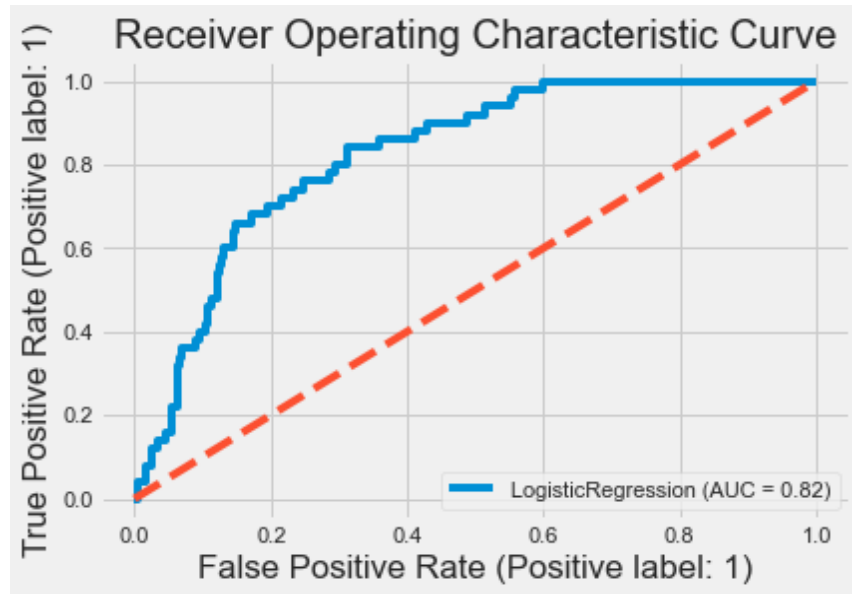


Figure 12. ROC AUC for Best Model

## Model Testing

In our scenario, an investor would want to see stock changes as quickly as possible, so the twitter features number of replies, retweets, and likes should be discarded as a prediction of new stock price changes on trained on historical data wouldn't have nearly as many counts in those features, and thus aren't a good representation of a real-world application.

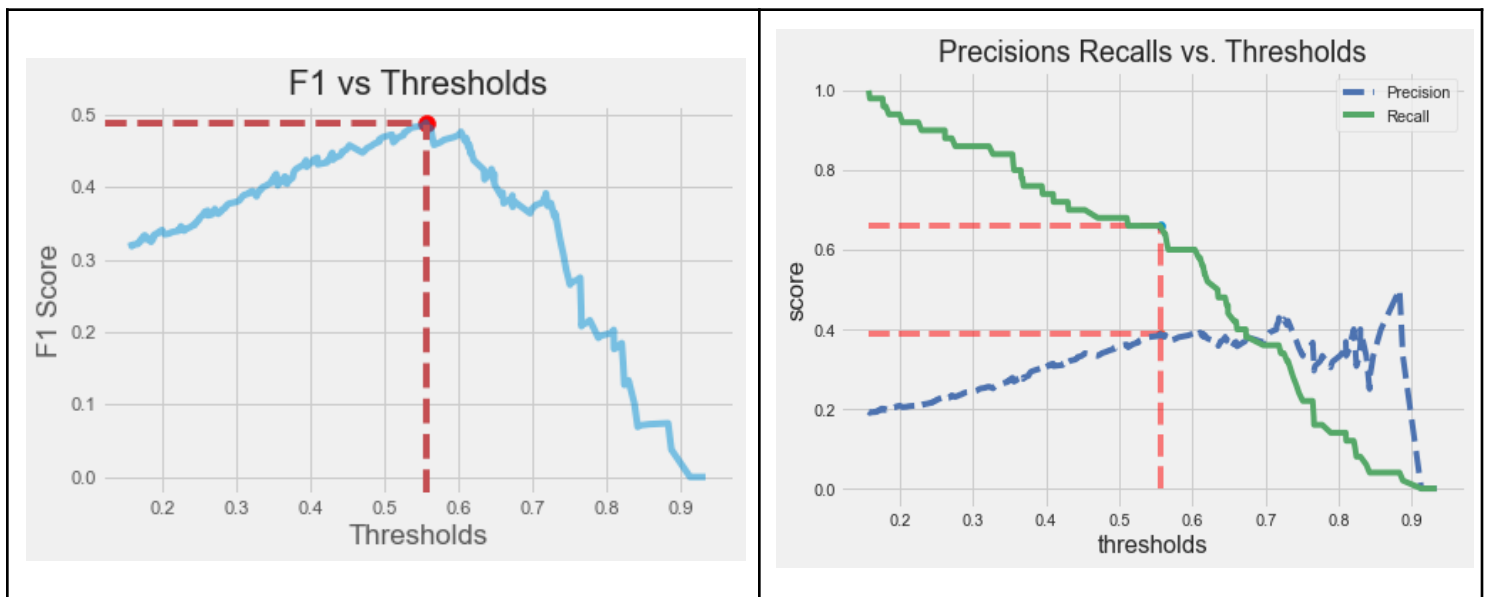


Figure 13. Optimal Threshold: 0.556. Best F1 Score = 0.49

	precision	recall	f1-score	support
0	0.95	0.85	0.90	349
1	0.39	0.66	0.49	50
accuracy			0.83	399
macro avg	0.67	0.76	0.69	399
weighted avg	0.88	0.83	0.84	399

Figure 14. Classification Report after Optimizing F1 Score

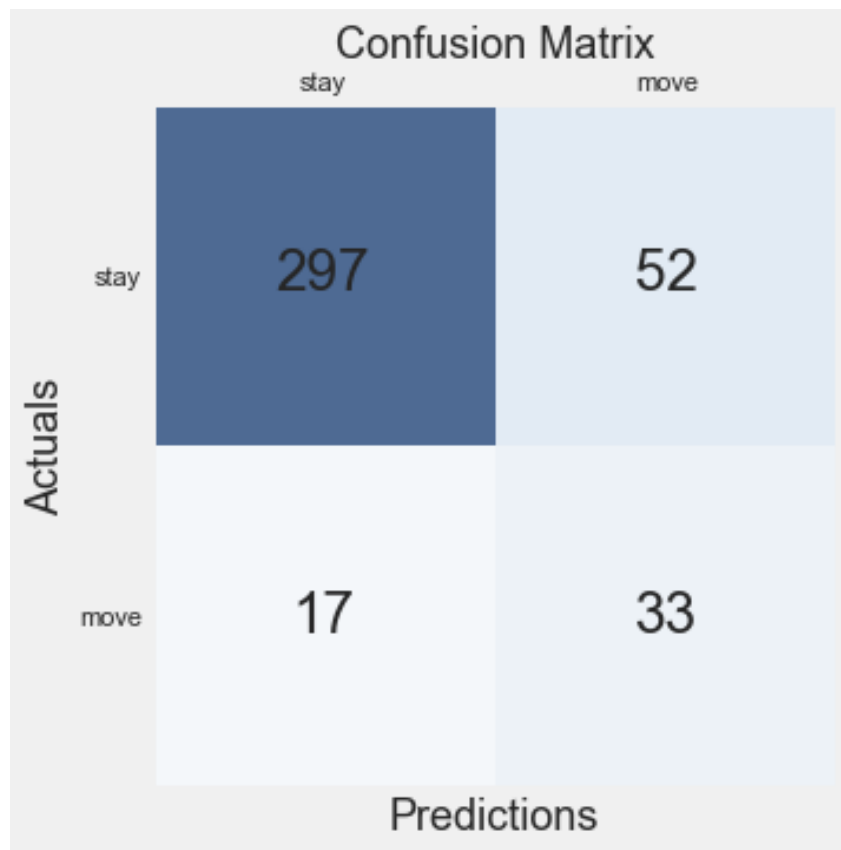


Figure 15. Confusion Matrix for Model

The accuracy isn't representative of the classifier's predictive ability due to the imbalanced data, so the f1 score is used as a better metric. The optimized f1 score takes into account both precision and recall for our positive class. In this case a precision of 39% indicates that when the model predicts move, it is correct 39% of the time, which means more stocks that are predicted to move actually won't exceed the 2.5% cutoff that was set for this model. However, of the actual stocks that move, it

predicts which ones do 66% of the time. Later on, threshold changes will be demonstrated for a given scenario.

---

## 8. Summary

### Interpretation

The most predictive words for the test set were calculated using the best classifier. Words that are highly predictive of stocks to move have an overall positive sentiment, are associated with the Virgin Brand, or are about conversations or thoughts. Words that are highly predicted of no stock movement imply wanting something, like get or need. World is highly predictive for no stock movement, possibly because the topics are about climate change or something unrelated to the stock in question. This is likely true for quote and way as well, in that they appear in tweets that aren't relevant to the stocks.

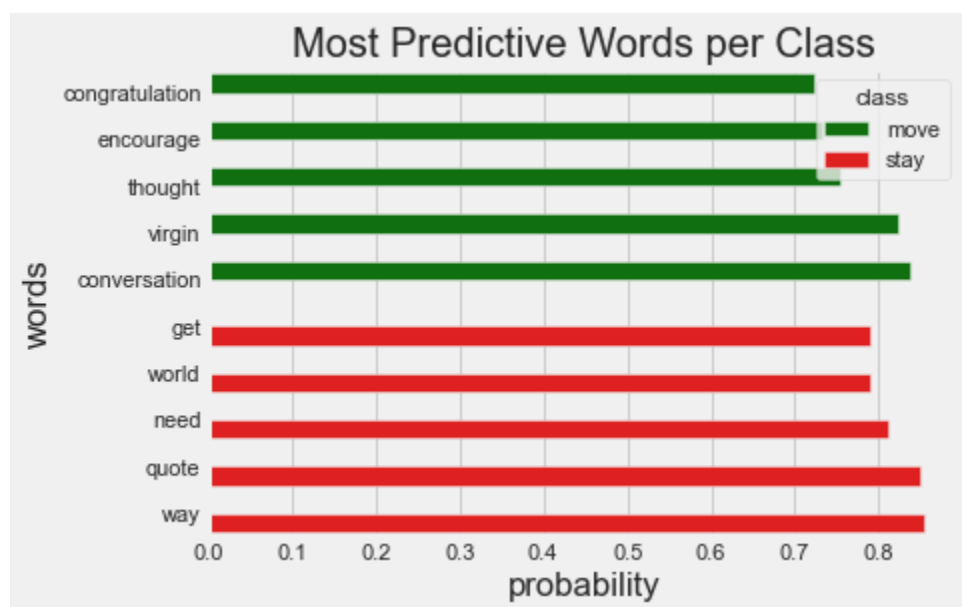


Figure 16. Most Predictive Words for Richard Branson per Class

### Explanation 1: Predictive Word (Virgin)

The raw tweet containing the word Virgin is examined below. It discusses an idea (thought) for a biodegradable bag from the Virgin Megastore. There is a duplicated tweet, but checking the scraped tweet data shows one of the links in the tweet was changed. The rest of the tweet discusses how things were when Virgin started, and the UK getting broadband.



'Love this idea from Virgin Megastore - introducing biodegradable bio-cassava bags <https://t.co/NoKCoLzgqT> @VirginME NA <https://t.co/ONF3N09f8h> When we started @Virgin life was all about hand-written letters, calling advertisers from phone booths, and running across town to speak with bands and banks... <https://t.co/vloy7AKZAq> <https://t.co/1kDduGydR> x When we started @Virgin, life was all about hand-written letters, calling advertisers from phone booths, and running across town to speak with bands and banks... <https://t.co/vloy7AKZAq> <https://t.co/IykSvKBnTx> @stevebartlettsc Thank s for the kind words Steven and hope it all went well! @VirginAtlantic @katherinemills So impressive - still can't wo rk out how she did it. It's incredible to think that by the end of 2021 nearly 15 million homes in the UK will have g igitabit broadband speeds - I remember listening to the dial-up tone and fighting over who could use the phone or the c omputer <https://t.co/vloy7AKZAq> @VirginMedia <https://t.co/kNnvd3a5ib> '

Figure 17. Collected raw tweets before the October 30th, 2019 market close.

## Results

- Return: -4.33%
- Predicted Stay ( $Prob(Move) = 20.00\%$ )
- True class: Move

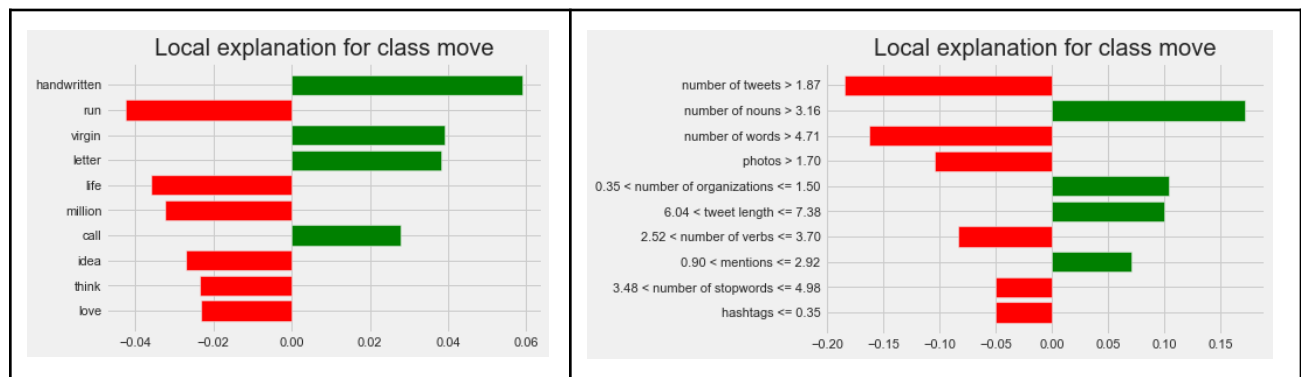


Figure 18. Local Explanation: Text (Left) proba 49% move. Features (Right) proba 29% move.

The local explanations for text features predict stocks move with a 49% probability, and the numeric features predict move with 29% probability. The stock went down by 4.33%, well above the 2.5% threshold, so the prediction is incorrect.

This does point to ways in which the model can be optimized further, though. The text predicts the right class with a higher probability than the features class. So additional implementation of weighting the text features more than the numeric features is a possibility. If stock movement is attributable to the tweets, then it can be assumed that not all the tweets are going to impact the stock price closing as equally, given what the tweet is about and what features it contains. Filtering the tweets by the topic might help to cut down on tweets that introduce noise into the model predictions. The downside is that an already small dataset might be truncated too heavily by the filtering process.

Bivariate data analysis demonstrated that tweets containing fewer features are more indicative of stock movements beyond the 2.5% cutoff, which explains the poorer performance of the numeric features over the text ones. There are a total of 6 tweets Richard Branson posted between October 30th's stock open and its close that are represented above, which introduces more features. This is further reason to implement some type of filtering system in which the

features will be more representative of stocks that are hypothesized to more directly impact tweets.

## Explanation 2: Misclassifying Move as Stay

The classifier incorrectly predicted actual MOVE as STAY 17 times. This can be a particularly bad result if the stock drops and money is lost as a result, or the stock rises and an opportunity to invest in the stock was missed. An example of this missed opportunity is below, in which a total of 4 tweets between the previous night's close and current day's open. He discusses lessons his father taught him and Virgin Startup helping entrepreneurs.

'Remembering my dad and some of the things he taught me: <https://t.co/OkRbHDX8qH> <https://t.co/9UMuIYBd12> One of the most resounding lessons I learned from my father was the importance of dreaming <https://t.co/OkRbHDX8qH> <https://t.co/radjZoP13f> The lessons I learned from my father: <https://t.co/OkRbHDX8qH> <https://t.co/eyCCRT0RMO> Always great to see how @VirginStartUp help entrepreneurs to scale their businesses: <https://t.co/9bJlU6pAou> #Crowdfunding #CrowdBoost <https://t.co/tzAS67Ed5y> '

Figure 19. Collected raw tweets before the March 10th, 2020 market open.

## Results

- Return: 7.05%
- Predicted Stay ( $Prob(Move) = 15.77\%$ )
- True class: Move

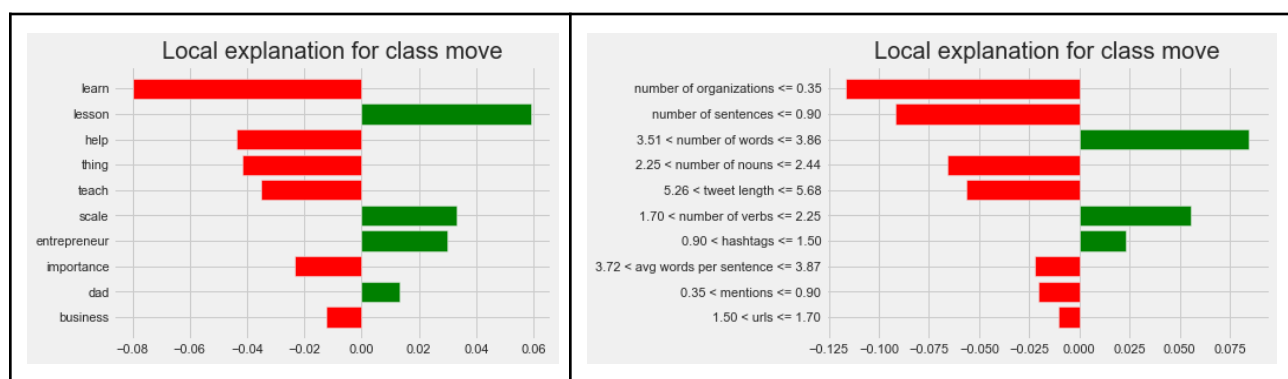


Figure 20. Local Explanation: Text (Left) proba 30% move. Features (Right) proba 15% move.

In examining the content of the raw tweet, it conveys positive messages and discusses how Virgin is helping entrepreneurs scale their businesses. The 7% return may or may not be attributable to Richard Branson's posts, but assuming it is, then further refinement of the model is needed. If the price increase is due to the overall sentiment of the tweet, then it would be good to include a sentiment analysis, using VADER or some other text analyzer, and see if it helps better the predictive power of the model. Mentions and hashtags were excluded in cleaning, but keeping them and processing them in may help to further improve scores as well.

If the tweet is attributable to the Virgin Startup mention, then again it points to being able to filter unnecessary tweets as a means to improve predictions, or at least to prioritize certain words or phrases by weighting them differently.

### Threshold Scenario

By decreasing the threshold to 0.225, we can minimize the number of missed opportunities by the algorithm.

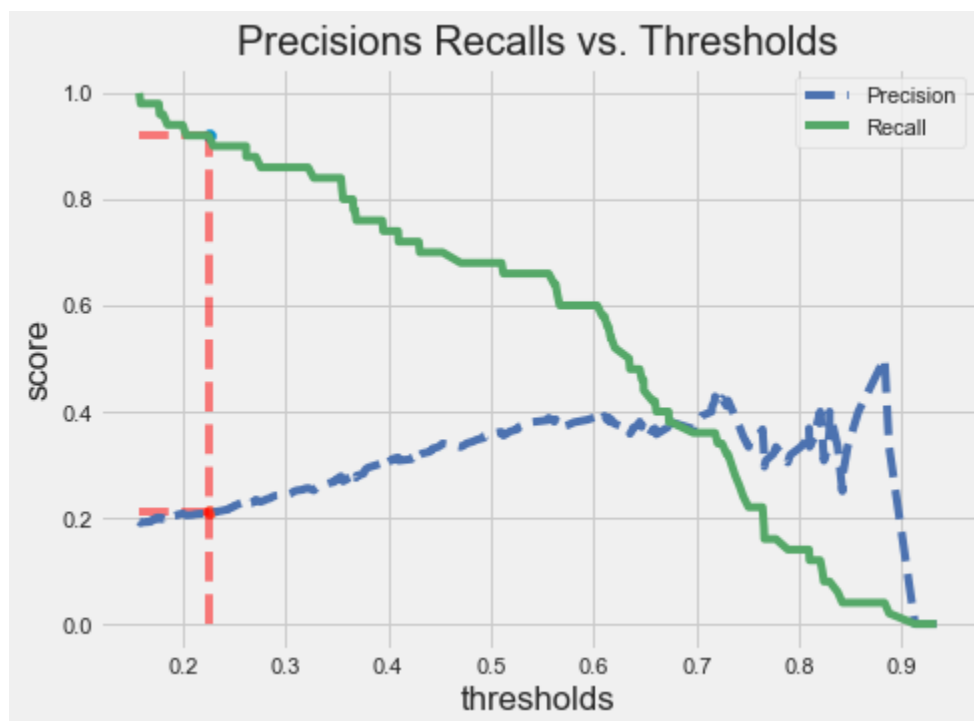


Figure 21. Threshold of 0.225

	precision	recall	f1-score	support
0	0.98	0.51	0.67	349
1	0.21	0.92	0.35	50
accuracy			0.56	399
macro avg	0.60	0.72	0.51	399
weighted avg	0.88	0.56	0.63	399

Figure 22. Classification Report using 0.225 Threshold

With the lower threshold, our recall for the Move class increases by 26%, but at the expense of both precision and the f1-score, which decrease by 18% and 14% respectively. Of the 50 stocks

that actually move, this model correctly predicts 46 of them, with 4 being misclassified as not moving. This greatly mitigates the opportunity loss, but of the 218 stocks that were predicted to move, 170 don't at all. This represents the issue regarding risk. If an investor wanted to be absolutely sure it was a smart decision to sell or invest in a stock, then this model would be a poor choice.

### Explanation 3: Misclassifying Stay as Move

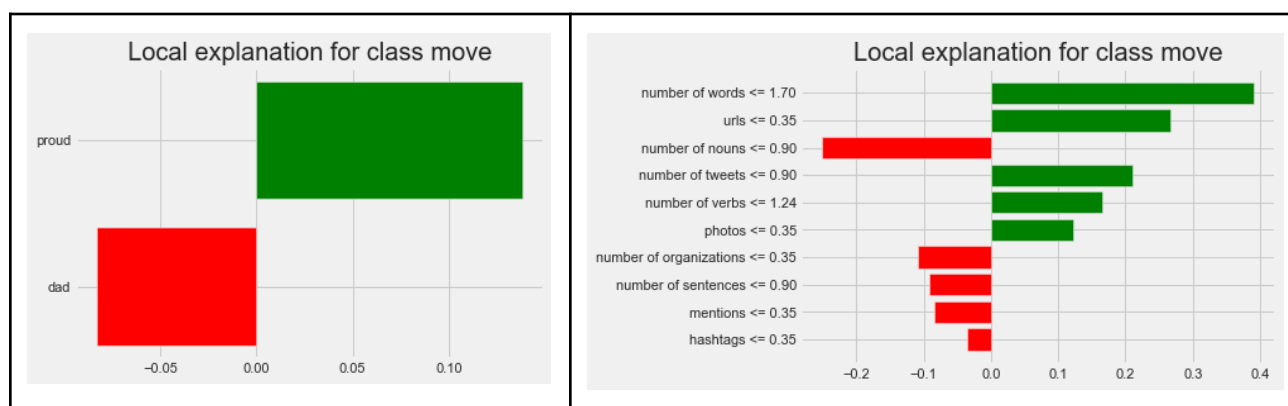
The classifier Incorrectly predicted actual STAY as MOVE 52 times. Were an investor wanting to be absolutely sure that the predicted stock will actually move, this would be an issue. If there are more stocks that are predicted to move than actually do, then betting on stocks moving would result in a greater number of losses than wins. Below is a tweet from a misclassification where a stock was predicted to move when it didn't beyond the 2.5% cutoff

'Proud dad! '

*Figure 23. Collected raw tweets before the December 4th, 2020 market open.*

### Results

- Return: 0.43%
- Predicted Move ( $Prob(Move) = 93.37\%$ )
- True class: Stay



*Figure 24. Local Explanation: Text (Left) proba 59% move. Features (Right) proba 95% move.*

This misclassification really demonstrates some of the limitations of the model. In examining this tweet, it is very unlikely that Richard Branson's tweet about being a proud dad will have any impact on Virgin Galactic's stock price, and indeed the return is only 0.4%, very much below the 2.5% cutoff. And yet here, a single tweet is predicted with 93% probability that the stock price will move.

Short tweets are going to weigh individual words much more heavily, and with proud and dad, we see that proud has greater influence even though the tweet itself has little, if anything, to do with the stock.

The reduced number of features also has a significant impact, having a local prediction of move at 95% probability. Short tweets and fewer tweets collected are likely going to predict move as demonstrated by the EDA analysis. Removing this tweet altogether by filtering the tweet topic, as has been shown previously, would again benefit the model as this introduces noise that impacts the model's overall predictive power.

#### Threshold Scenario

By increasing the threshold to 0.718, we can minimize the number of missed opportunities by the algorithm.

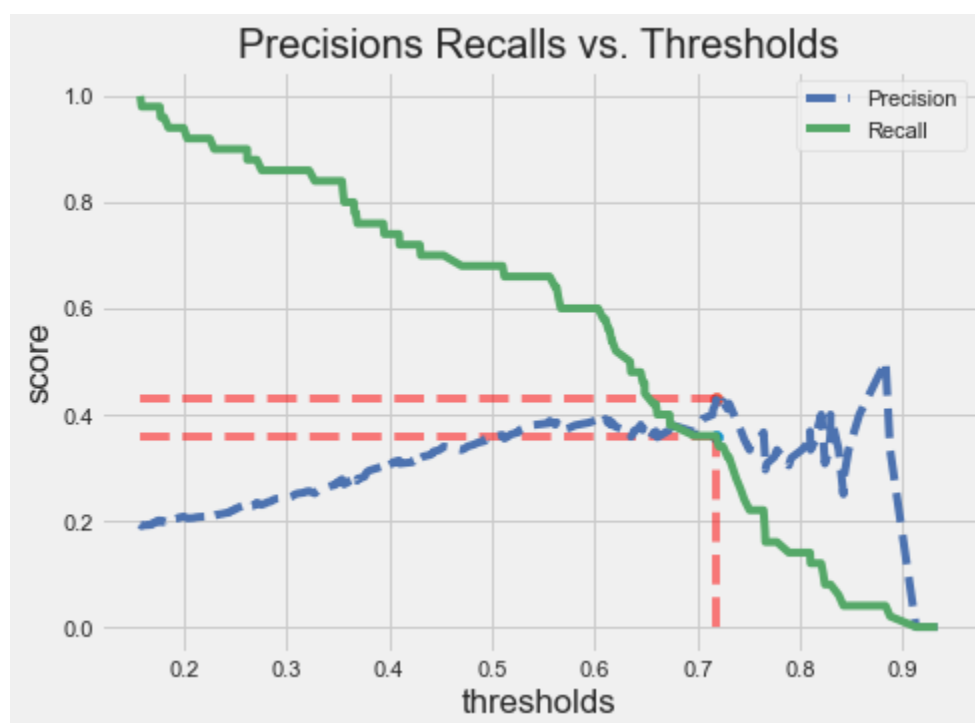


Figure 25. Precisions and Recalls with 0.718 Threshold

	precision	recall	f1-score	support
0	0.91	0.93	0.92	349
1	0.43	0.36	0.39	50
accuracy			0.86	399
macro avg	0.67	0.65	0.66	399
weighted avg	0.85	0.86	0.85	399

Figure 26. Classification Report with 0.718 Threshold

The best precision score that can be achieved without seriously sacrificing recall is 43%, meaning 43% of the total number of predicted stock movements actually move. Of the 42 stocks predicted to move, only 18 actually will. The maximum precision score obtainable with modifying the threshold is 50%, but with a recall of 4%. This means at the very best, there is only a 50% chance that the stocks that are predicted to move actually will. Future implementation will look to improve the precision score as this will best determine the model's success when used for investing purposes.

---

## 9. Conclusion

### Major Findings

Of the tweets collected, Richard Branson's demonstrated the greatest predictive potential with EDA and inferential statistics, for stock movement. Therefore it's not surprising to see that Richard Branson's Twitter data had the highest classification scores (using f1), given the statistically different significance between stocks that move vs. stay. The logistic regression classification performed the best and was significantly faster than the support vector classification.

It was found that fewer features corresponded to a greater probability in stock movement beyond the 2.5% cutoff. This could be interpreted as tweets that are indicative of stock movement have fewer features and therefore less noise from tweets that aren't important for stock movement. It could also be that Richard Branson's focus is on Virgin Galactic when stocks are moving, so he tweets more on his downtime when stock movement is more stagnant.

The logistic regression model had an optimized f1-score of 49%, a recall score of 66%, and a precision score of 39%. For stocks that actually move, this model correctly predicted which ones did 66% of the time. However, out of all the predictions the model made only 39% were correct. Modification of the threshold can greatly improve the recall of the model without drastically reducing the precision score, but optimizing in precision greatly reduces the recall, and at best, the precision score is only 50%.

### Model Implementation

The goal was to determine stock movement using the executive's tweets. Of the tweets gathered, only Richard Branson's demonstrated any predictive capabilities given the distribution of the features. However, this was looking specifically at the differences between stocks that move beyond a threshold, and modification of that threshold or separating stocks that rise vs fall might yield different results and will be examined in future implementations.

While it was shown above that modifying the threshold can change the predictions, keeping the model's threshold at 0.556 maximizes the recall and precision, allowing for a good trade-off between predicting which stocks actually move, and which predictions are correct. As the model is currently better at predicting which stocks move, this can be used in conjunction with other classification methods or analyzers to have an investor make an informed decision as to whether to invest or sell stock.

A downside to this model is that, in order to use it, twitter feeds would need to be analyzed to see if similar distributions in tweet features can be extracted that are similar to Richard Branson's. In this case, a T-Test can be used to see what features are statistically significant for similar analysis.

## Next Steps/Assumptions

### Improving the Current Model

Further optimizations on the model can be performed through text preprocessing. The current model uses Tfidf Vectorization with unigrams, and does not take into account other n-gram ranges, the minimum number of documents a word needs to appear in to be considered, or other vectorization methods. Word embeddings via Word2Vec might provide greater insight into the context within the tweets.

Additionally, other features can be engineered that could provide further insight into the text itself. Using a sentiment analyzer might help to further improve predictions, the tweet reading level, and the duration between the last closing price and the upcoming one are all considerations that can give important information not necessarily provided in the current data.

Random Forests and Stochastic Gradient Descent were also used in early explorations of classification methods, but neither performed as well as the Logistic Regression or the Support Vector Machines. There are other methods that should be examined, namely ensemble and deep-learning methods. Google AI Language researchers developed a context-aware deep learning method called BERT (Bidirectional Encoder Representations from Transformers), which has shown to be incredibly powerful in extracting information from text, and should be the next to explore([link](#)).

Getting more data is important, but this is difficult at present. Twitter only recently became a tool that analysts can use for stock movement, and so data collection is limited. It is also being assumed that the impact a tweet will have on stock prices is time independent, when in fact tweets made early in a company's life might have little impact vs those when the company's stock is being closely watched and traded. Ideally data would have been collected within the confines of a shorter duration, minimizing the change or influence a tweet has over time, and looking at relatively similar stock prices within the lifetime of a company. To this end, collecting intraday stock prices at an hourly frequency, rather than the open and the close as has been

done with the current model, would allow for better separating Twitter posts that have an impact on stock price, vs those that don't, and so would allow for more refined data collection.

At the very least, reducing the noise in the collected tweets is important. There's a lot of information in the text that the model was not able to capture because it was drowned out by needless information. Filtering tweets by using posts that only mentioned the stock by name were tested, but too much data had been cut in the process, which didn't allow for reasonable predictions. The best option is to start with collecting stock prices at greater frequencies, as described in the paragraph above, to better separate the tweets by the prices they impact.

### Future Work

This project looked only at the influence an executive may have on their company's stock price changes, however influential media users can also have an impact on other stocks they mention (e.g. Elon Musk's mention of Cyberpunk 2077), and an analysis of other stocks will be done. The ultimate goal is to predict stocks that rise vs fall, so a multi-labelled classification will be performed after optimizing the current model to better detect price movements.

---

## 10. Sources

<https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>