BIG DATA ALPHA MODEL

FIN 6392.001 FinTech & Blockchain

2021.10.06

Philipp Holenstein
Martin Vincent
Zach Burns
Travis McDaniel

Portfolio Securities

The following chapter is intended to provide information on the selected securities. For the stock selection, 4 stocks from different industries were analyzed. The aim was to select shares of well-capitalized companies that are frequently traded on the stock exchange. For each share, the last 5 years were presented in a price chart. The information on the financing structure, including the WACC, was also obtained from Bloomberg and is intended to show how the company is financed and what costs are associated with this.

NCLH (NYSE: Norwegian Cruise Line Holdings Ltd)

Norwegian Cruise Line Holdings Ltd. Operates a fleet of passenger cruise ships. The company offers an array of cruise itineraries and theme cruises, as well as markets its services through various distribution channels including retail and travel agents, international and incentive sales, and consumer direct. Norwegian Cruise Line Holdings serves customers worldwide.

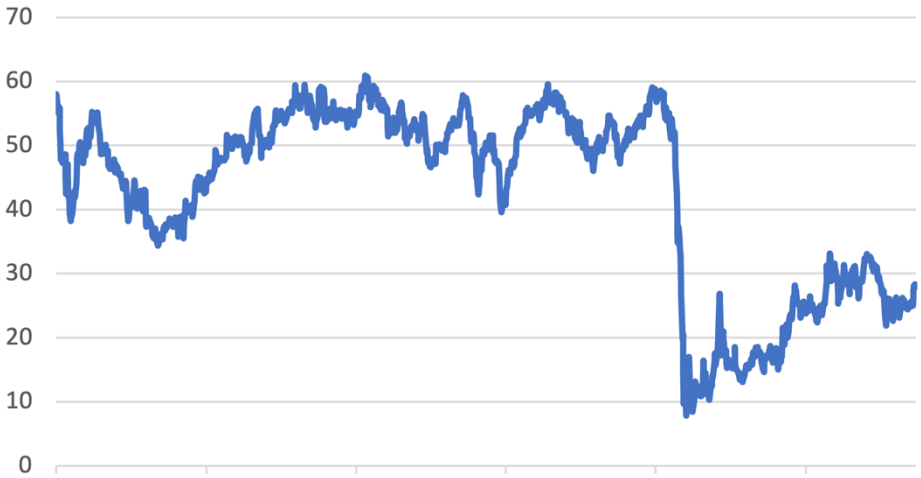


Figure 1: stock price over the last 5 years (Source: Bloomberg)

As can be seen from the price chart in Figure 1, the company has suffered greatly from the Covid-19 pandemic. The share price collapsed in the spring of 2020 and has only recovered slightly since then.

	Weight	Cost	W x C
Equity	46.60%	20.80%	9.69%
Debt Cost	53.40%	2.10%	1.12%
WACC			**10.81%**

Figure 2: WACC composition (Source: Bloomberg)

The uncertainty triggered by Covid-19 regarding the future of the company is also reflected in the composition of the WACCs in Figure 2. The cost of equity capital is very high at 20.80%, reflecting the high expectations of shareholders. One of the largest influencing factors on this stock is the ongoing government regulations regarding cruising globally.

CSCO (NASDAQ: Cisco Systems Inc)

Cisco Systems, Inc. is an American multinational technology conglomerate corporation headquartered in San Jose, California. Integral to the growth of Silicon Valley, Cisco develops, manufactures and sells networking hardware, software, telecommunications equipment and other high-technology services and products.

Figure 3: stock price over the last 5 years (Source: Bloomberg)

The stock reached a lifetime high in 2019 with a fallback due to poor performance of earnings over the course of a year. It did take a short term hit because of Covid-19 yet due to shifts in telecommunication patterns by companies Cisco has been poised to grow at record rates.

	Weight	Cost	W x C
Equity	77.10%	14.60%	11.26%
Debt Cost	22.90%	1.40%	0.32%
WACC			**11.58%**

Figure 4: WACC composition (Source: Bloomberg)

The global pandemic has not severely impacted Cisco's share price relative to other industries. Cisco continues to grow at a stead rate with investor expectations on equity at 14.6%. The challenge going forward will be new technology and staying ahead of competitors in innovation. This rate of equity return investors are expecting is well in line with other established entities in the NASDAQ & NYSE.

GME (NYSE: GAMESTOP CORP.)

GameStop Corporation operates specialty electronic game and PC entertainment software stores. The Company stores sell new and used video game hardware and software, as well as accessories. GameStop markets its products worldwide.

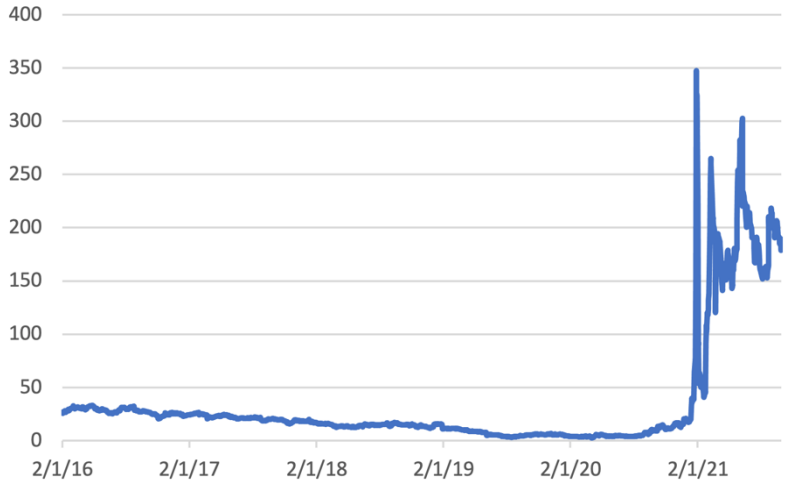


Figure 5: stock price over the last 5 years (Source: Bloomberg)

For the last 5 years prior to 2020 GameStop was a stock in decline with low prospects for returns by investors. The recent "Wall Street Bets" forum on Reddit resulted in an extreme swing in the stock by undertaking trading strategy known as a short squeeze.

	Weight	Cost	W x C
Equity	96.60%	0.00%	0.00%
Debt Cost	5.40%	2.00%	0.11%
WACC			**0.11%**

Figure 6: WACC composition (Source: Bloomberg)

Due to the stock historically being relegated to a dying relic of retail. There had been low expectations placed on it by the market for many years. Along with a very low leveraged capital structure primarily relying upon debt to survive. The question is will management be able to capitalize on this opportunity presented by the fourfold increase in equity value.

KTOS (NASDAQ: KRATOS DEFENSE & SECURITY SOLUTIONS, INC)

Kratos Defense & Security Solutions, Inc. operates as a defense contractor and security systems integrator for the federal government and for state and local agencies. The Company offers services in weapon systems lifecycle support, military weapon range, security and surveillance systems, and IT engineering.

Figure 7: stock price over the last 5 years (Source: Bloomberg)

As with most companies it had a sudden drop in share value due to Covid-19 but experienced a rapid recovery with its share value. Over the last 5 years it has been on a steady state growth trend but has recently accelerated that trend.

	Weight	Cost	W x C
Equity	89.90%	14.90%	13.40%
Debt Cost	10.10%	2.50%	0.25%
WACC			**13.65%**

Figure 8: WACC composition (Source: Bloomberg)

Though it has a high WACC compared to the other 4 stocks in consideration its cost of equity is still inline with market expectations. Its product offering presents a very competitive and modern approach to warfighting systems currently in demand by global defense entities.

Fundamental Data & Sources

FUNDAMENTAL DATA

Consumer Price Index (CPI)

As our primary data source, we utilized the national CPI database provided by the Bureau of Labour Statistics on a monthly basis[1]. This was a critical data point as it represents strong signals in overall market strength, consumer behaviours, fluctuations in market pricing of goods, and permanent vs transitory economic indicators.

TRADING DATA

For each stock we limited the trading data to following parameters for selection of stock and their output data

- Traded on the NYSE or NASDAQ
- Minimum 5 year publicly traded security
- Not involved in the marijuana business
- US Headquartered
- 3 years of financial transactions

It's from these data parameters we identified the four securities to trade and build a factoring model around.

SENTIMENT DATA

RedditExtractoR package

For the sake of sentiment analysis, we utilized the R package RedditextractoR. This package scrapes data from Reddit related to whichever keyword you set, and it shows posts or comments in a specified subreddit or on any subreddit. For the purpose of this project, we used the package to find posts with the keywords NCLH, CSCO, GME, and KTOS in an attempt to assign them sentiment values.

sentimentr package

After scraping relevent posts from reddit, we used the R package sentimentr to use positive and negative words in the post to assign a sentiment score to each post. Using this we averaged the scores for each day for each stock to get a sentiment score for each day for each stock for analysis in the final algorithm step.

[1] US Bureau of Labour Statistics CPI Database

Factors & Technical Indicators

TECHINCAL ANALYSIS

As part of our technical analysis, we chose to include two simple moving averages to catch changes in momentum. This idea was presented to us through Dr. Zhiqiang Zheng's lecture slides on technical analysis. We plotted the 3-day Simple Moving Average and the 13-day Simple Moving Average for each security in our portfolio. Through back testing, we observed that when these two moving averages intersect, there would be a change in direction of the trend. This observation is what led us to create the variables sma_diff, sma3_signal, and sma_signal_duration. Our intention by creating these variables is to identify when sma3>sma13 and then count how many days it did so. We hope that there is some significance to sma_diff being positive due to is indicating an upward trend in momentum.

SENTIMENT ANALYSIS

The goal of sentiment analysis in this project was to find a daily sentiment score for each stock that would be correlated with price changes of the stocks. The reasoning behind this is that finding out how Redditors discuss certain stocks could give insight into changes in stock prices.

The first step in our sentiment analysis was to store the relevant posts related to the stocks we chose. In getting the data from Reddit, we decided it would make sense to filter for "top" posts, since they were "upvoted" to become a top post so other people probably share their ideas, and we chose to search from all subreddits in an attempt to be more unbiased in determining sentiments of the stock. One problem we encountered during this step was that when searching for "KTOS" on all subreddits, many irrelevant posts appeared, since "kto" is a Polish word. To remedy this, we searched only on the subreddit "WallStreetBets" for the "KTOS" ticker to filter for posts about the stock.

The next step was to apply sentiment analysis to the posts. The sentiment() function assigned the posts a value between –1 and 1, positive values corresponding to good sentiment and negative values corresponding to bad sentiment. This method has some shortcomings since the function simply detects positive and negative words and may not be able to understand the actual meaning behind the post, and that a certain sentiment does not necessarily mean that any buy or sell action will actually occur.

The final step was to clean and organize the data. The posts with a sentiment value of 0 were removed, since they seemed to be of no use to the model. The sentiment scores of the posts corresponding to the same stock which were made on the same day were averaged out to find a sentiment score for a particular day for each stock, rather than a score for each post. Following that, the scores were combined into a table with their corresponding date and exported as csv files.

After running the sentiment data in the overall model, we concluded that it was not impactful in explaining or predicting stock prices. The most obvious flaw was the lack of sentiment data that we were able to generate. RedditextractoR only returned about 250 posts for each stock, and after combining them by day and removing 0 sentiment posts, there were only about 50 data points for each stock.

Construction of Alpha Model

Four datasets were created for each stock chosen to be part of the portfolio. Using the rbind function in R, we were able to vertically stack the four datasets into one large panel dataset. Before creating the model, we checked the classes of the columns and noticed that the DATE variable was only being considered as a character datatype. We altered this column to make sure R was reading this as a date datatype. Using panel makes it technically cumbersome to run a regular linear regression. We thought to construct a fixed effect model using SYMBOL as the only reference. Our group felt that accounting for the variation only among the four stocks was not enough to consider the events that occurred in the past three years. We opted for a Random Effect Model and made SYMBOL and DATE a reference. Attempting to use a Fixed Effect model including CPI as a regressor resulted in an error. The CPI data is monthly and does not vary across the stocks, only time. Since we were only fixing the SYMBOL and there was no variation in CPI across the stocks, the model was ignoring CPI variable. The Random Effect model allows us to include CPI and also include any variation that could have been caused by Covid or other events over the past three years.

Figure 9: Data Dictionary

Variable Name	Datatype	Description
DATE	Date	The day of the security's historical data
SYMBOL	Character	Symbol of the stock
OPEN	Numeric	Opening Price
HIGH	Numeric	Highest Price in a day
LOW	Numeric	Lowest Price in a day
CLOSE	Numeric	Price when the market closes
VOLUME	Numeric	Number of shares transacted that day
ADJUSTED_PRICE	Numeric	Adjusted price of the stock
CPI	Numeric	Consumer price index (Inflation Metric)
P_1	Numeric	The Adj Price of the next day
sma3	Numeric	3-Day Simple Moving Average
sma13	Numeric	13-Day Simple Moving Average
sma_diff	Numeric	3-Day SMA - 13-Day SMA
sma3_signal	Numeric	If sma_diff is >0, then 1
sma_signal_duration	Numeric	Counts the number of consecutive days the sma3_signal is 1
log_return	Numeric	the logistic return of a security
SENTIMENT	Numeric	The view/ perception rating of a stock from Reddit

Projections & Performance Results

SUMMARY

Four datasets were created for each stock chosen to be part of the portfolio. Using the rbind function in R, we were able to vertically stack the four datasets into one large panel dataset. Before creating the model, we checked the classes of the columns and noticed that the DATE variable was only being considered as a character datatype. We altered this column to make sure R was reading this as a date datatype. Using panel makes it technically cumbersome to run a regular linear regression. We thought to construct a fixed effect model using SYMBOL as the only reference. Our group felt that accounting for the variation only among the four stocks was not enough to consider the events that occurred in the past three years. We opted for a Random Effect Model and made SYMBOL and DATE a reference. Attempting to use a Fixed Effect model including CPI as a regressor resulted in an error. The CPI data is monthly and does not vary across the stocks, only time. Since we were only fixing the SYMBOL and there was no variation in CPI across the stocks, the model was ignoring CPI variable. The Random Effect model allows us to include CPI and also include any variation that could have been caused by Covid or other events over the past three years.

We foresaw an issue in our sentiment and as result, are presenting to alpha models. Model 1 has SENTIMENT as one of its predictors in the final equation and Model 2 does not include SENTIMENT in its final equation.

MODEL 1 W/ SENTIMENT

In figure 10 the output running the random effect model on all indicators excluding sma_diff. sma_diff was excluded due to it being a computational variable of sma_sig.

We then excluded the variables one-by-one if their p-value was not significant at a 95% CI and monitor changes in Adjusted R-Squared for overfitting. We would have preferred to use a selection procedure such as Forward and/or Backward selection, but we were unable to find a package in R to run this on a plm() model. Figure 11 is the final output.

Figure 11: Output Exclude (Single Variables)

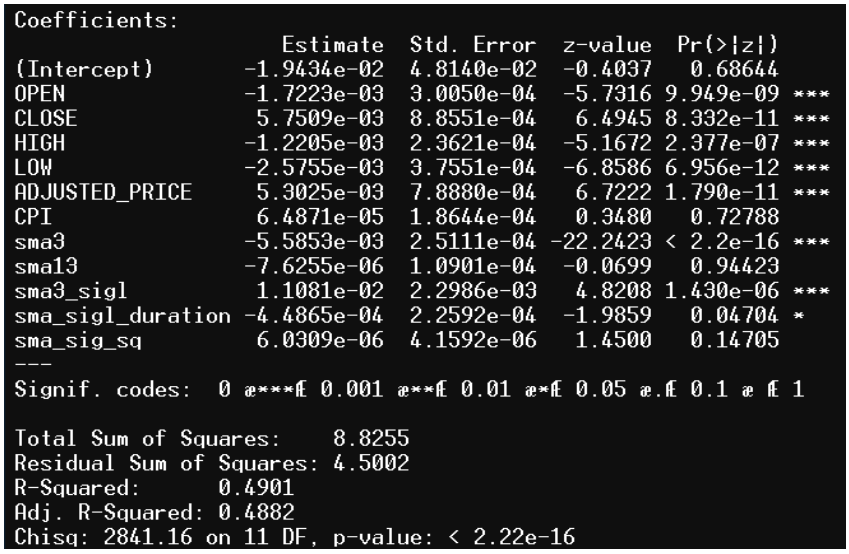
```
Coefficients:
                   Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept)      -0.00329775  0.01139352  -0.2894 0.7722436
OPEN             -0.00253046  0.00069540  -3.6389 0.0002738 ***
HIGH             -0.00118023  0.00056819  -2.0772 0.0377848 *
LOW              -0.00280731  0.00090444  -3.1039 0.0019097 **
ADJUSTED_PRICE    0.01149359  0.00070698  16.2573 < 2.2e-16 ***
sma3             -0.00514718  0.00046036 -11.1809 < 2.2e-16 ***
sma3_sigl         0.05023549  0.01610979   3.1183 0.0018189 **
sma_sigl_duration -0.00468192  0.00217762  -2.1500 0.0315541 *
sma_sig_sq        0.00010935  0.00005880   1.8598 0.0629155 .
SENTIMENT         0.08052158  0.04237348   1.9003 0.0573961 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1.8944
Residual Sum of Squares: 0.38733
R-Squared:      0.79555
Adj. R-Squared: 0.78611
Chisq: 759.229 on 9 DF, p-value: < 2.22e-16
```

The variables sma_sig_sq and SENTIMENT were kept in the final equation due to their p-values being relatively close to the 95% CI and the exclusion of them resulting in decreases of the Adj R-Squared and more insignificant variables.

MODEL 2 W/O SENTIMENT

Figure 12 Model 2 which begins with all variables but SENTIMENT and sma_diff.

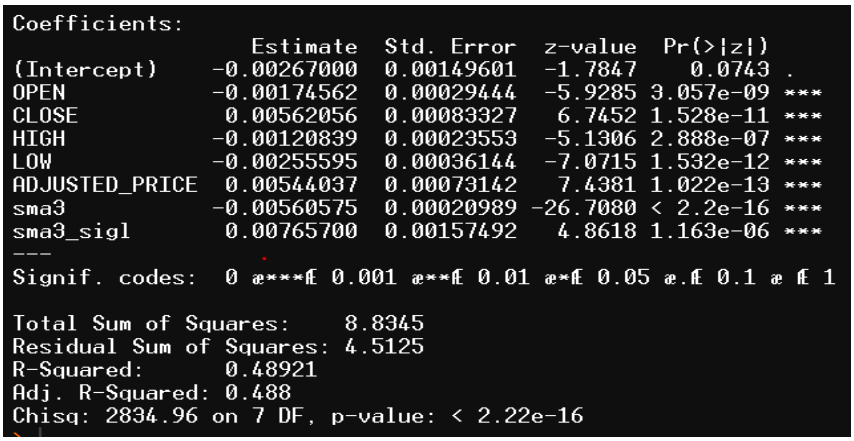
Figure 12: Output Exclude (sma_diff)



```
Coefficients:
                   Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept)      -1.9434e-02  4.8140e-02  -0.4037   0.68644
OPEN             -1.7223e-03  3.0050e-04  -5.7316 9.949e-09 ***
CLOSE             5.7509e-03  8.8551e-04   6.4945 8.332e-11 ***
HIGH             -1.2205e-03  2.3621e-04  -5.1672 2.377e-07 ***
LOW              -2.5755e-03  3.7551e-04  -6.8586 6.956e-12 ***
ADJUSTED_PRICE    5.3025e-03  7.8880e-04   6.7222 1.790e-11 ***
CPI               6.4871e-05  1.8644e-04   0.3480   0.72788
sma3             -5.5853e-03  2.5111e-04 -22.2423 < 2.2e-16 ***
sma13            -7.6255e-06  1.0901e-04  -0.0699   0.94423
sma3_sigl         1.1081e-02  2.2986e-03   4.8208 1.430e-06 ***
sma_sigl_duration -4.4865e-04  2.2592e-04  -1.9859   0.04704 *
sma_sig_sq        6.0309e-06  4.1592e-06   1.4500   0.14705
---
Signif. codes:  0 æ***Æ 0.001 æ**Æ 0.01 æ*Æ 0.05 æ.Æ 0.1 æ Æ 1

Total Sum of Squares:    8.8255
Residual Sum of Squares: 4.5002
R-Squared:      0.4901
Adj. R-Squared: 0.4882
Chisq: 2841.16 on 11 DF, p-value: < 2.22e-16
```

The output, compared to Model 1, begins with an Adjusted R-Squared 0.3 lower which highlights the impact of SENTIMENT. We then conducted another manual selection of the regression based on the same criteria in Model 1. The output of the final regression of Model 2 is shown in figure 13.

Figure 13: Final Regression Model 2



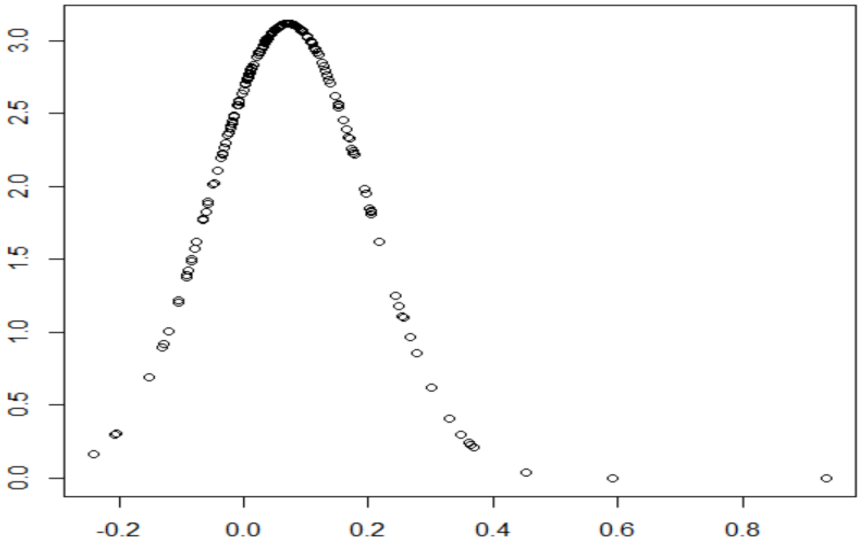
```
Coefficients:
                  Estimate  Std. Error  z-value  Pr(>|z|)
(Intercept)    -0.00267000  0.00149601  -1.7847    0.0743 .
OPEN           -0.00174562  0.00029444  -5.9285 3.057e-09 ***
CLOSE           0.00562056  0.00083327   6.7452 1.528e-11 ***
HIGH           -0.00120839  0.00023553  -5.1306 2.888e-07 ***
LOW            -0.00255595  0.00036144  -7.0715 1.532e-12 ***
ADJUSTED_PRICE  0.00544037  0.00073142   7.4381 1.022e-13 ***
sma3           -0.00560575  0.00020989 -26.7080 < 2.2e-16 ***
sma3_sigl       0.00765700  0.00157492   4.8618 1.163e-06 ***
---
Signif. codes:  0 æ***Æ 0.001 æ**Æ 0.01 æ*Æ 0.05 æ.Æ 0.1 æ Æ 1

Total Sum of Squares:    8.8345
Residual Sum of Squares: 4.5125
R-Squared:      0.48921
Adj. R-Squared: 0.488
Chisq: 2834.96 on 7 DF, p-value: < 2.22e-16
```

In this model all variables are very significant but with a much lower Adjusted R-Squared.

The sentiment data is problematic due to lack of data points. We were only able to gather sentiment for about 50 days for each stock out of over 3,000 records. In figure 14 we demonstrate the SENTIMENT's bell curve.

SENTIMENT has a narrow bell curve and has a low variation. We believe this is the reason for the discrepancy in R-Squared between the two models. Despite SENTIMENT being significant, there is not enough data in our sample for us to be confident we are capturing the true impact of sentiment data. The variables in Model 2 are also significant in Model 1.

Figure 14: SENTIMENT Bell Curve Model 2



Despite Model 1 having a much higher Adjusted R-Squared, we believe that model 2 contains the least amount of bias and therefore, better predicting coefficients.

The interpretation of Model 2 is that the strategy is to monitor prices of the security's prices and the relationship between the 3-day simple moving average and the 13-day simple moving average. The intersection in the two lines signals a shift in momentum and it would be the time to buy. When the 13-day simple moving average surpasses the 3-day simple moving average it would turn the signal off (sma3_sigl=0) and you sell or short the stock. However, if this model is implemented on these four stocks, it is expected to have a negative excess daily return of -0.267%. and the Adjusted R-Squared is low and we would not consider this an accurate model given our dataset.

We can conclude that the model does not provide a profitable trading strategy. The method in which we collected sentiment was limited to packages in R and what we could find on Reddit. This is considered a somewhat a rudimentary method of obtaining sentiment metrics and perhaps there are better methods to use for these stocks. Using a longer time period might have helped our model by making monthly/weekly fundamental data more meaningful and reducing variation with more data points. We believe the strategy in theory is effective, but the model is missing other metrics that would increase its predictive power and accuracy.