COVID-19's impact on the US Stock Market

Nikola Danevski

Muhammad Usama Ijaz

May 9, 2020

Contents

1 Introduction

COVID-19 is a disease that originated in late 2019 in China and at the time of writing of this paper there are more than 3.8 million cases globally. Through drastic measures, China managed to reduce the number of new daily cases from its peak at 14,000 cases on February 12, 2020, to only 11 cases per day one month later. However, the virus rapidly spread through the rest of the world, heavily hitting western Europe and the United States of America. With the fear of the virus spreading throughout the USA, the US stock exchange has been heavily impacted [10]. The Dow Jones, one of the most popular stock market indices, hit its lowest point in more than a year on March 12, 2020.

No previous infectious disease outbreak, including the Spanish Flu, has impacted the stock market as powerfully as the COVID-19 pandemic. The major stock indexes declined after major tech companies reported how they have been affected by the COVID-19 pandemic. The Dow Jones Industrial Average declined by more than 622 points, the S&P 500 by more than 81 points, and the Nasdaq composite index by more than 284 points. Apple failed to provide financial guidance to investors for the first time since 2003. Amazon also revealed how the pandemic has been affecting the company. While the tech giant's revenue grew 26% during the first three months of the year, the company will spend the entirety of its $4 billion profit on COVID-19 related expenses between April and June. [1]

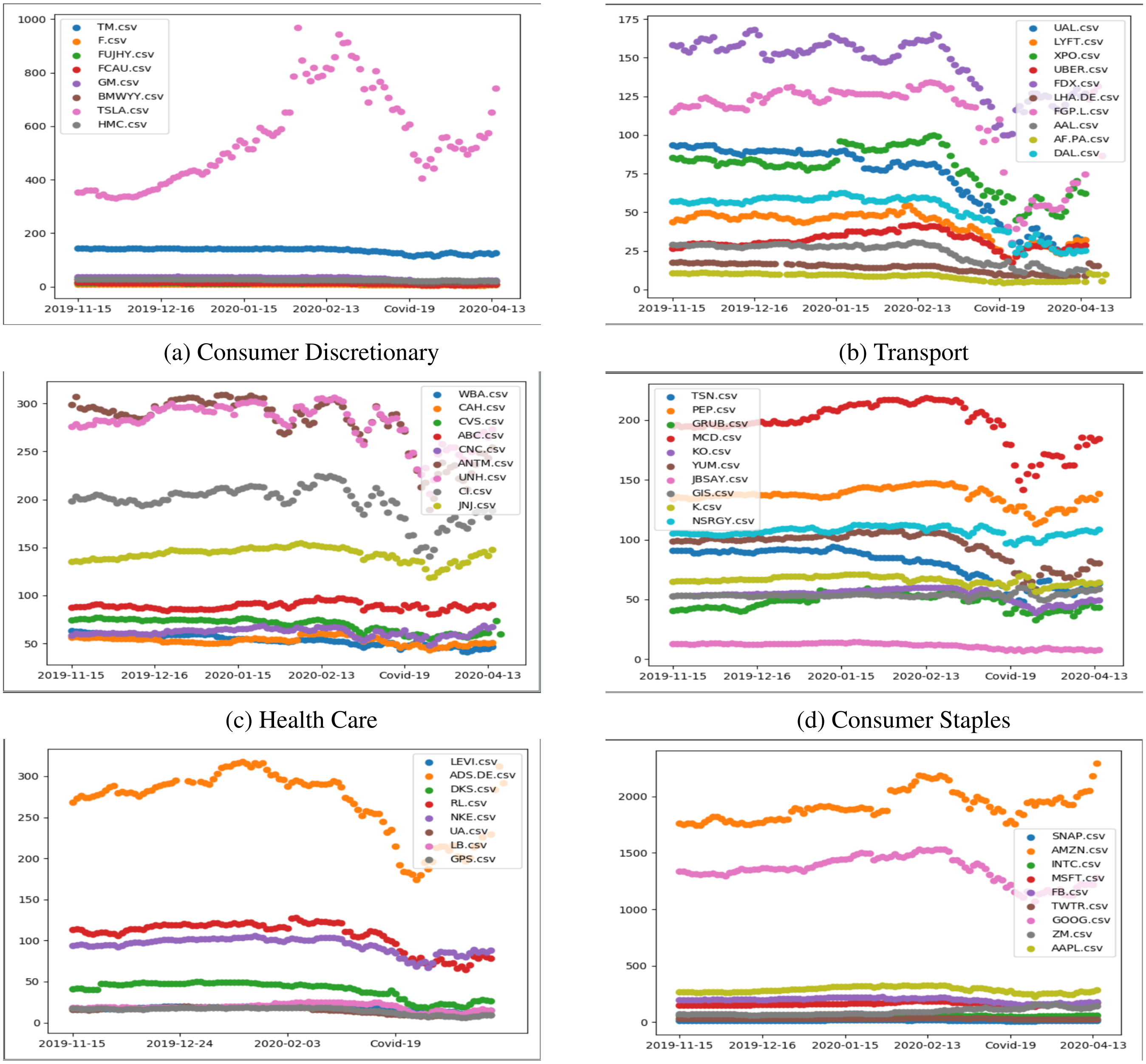
1.1 Background and Motivation

The Federal Reserve announced on March 12, 2020, that it will launch a series of massive cash injections into funding markets and begin buying long-term government bonds as panic over the coronavirus pandemic increasingly stresses the nation's financial system. The Federal Reserve Bank of New York took steps to inject more than $1.5 trillion into the markets in a bid to calm investors who are fearful of the economic impact of the novel virus. [2].

With $1.5 trillion being pumped into the stock market, stocks briefly pared their steep losses. But the extremely volatile market activity continued: The Dow Jones industrial average and S&P 500 index were down 7.8% and 7.3%, respectively. We argue that there might be a more efficient way to inject money; if we can find a common attribute to classify or cluster on. This

1.2 Problem Statement

We use stock-market high prices for each day in a 5-month period and apply data mining methods to better understand the impact of COVID-19 on the S&P 500 companies.
Understanding which attributes are important is a well-known problem that has been studied, however, we argue that it is harder to use any of those statistical methods with continuous data. That is why we apply classification methods to understand which attributes of a certain company affect its stock market price increase or decrease. The attributes we get in the end are the ones that the Federal Reserve should be looking at when deciding how and where to inject the money.



(a) Consumer Discretionary

(b) Transport

(c) Health Care

(d) Consumer Staples

(e) Clothing

(f) Information Technology

Figure 1: Change in Stock Prices during the last six months

1.3 Problem Importance

It is immensely important to find the behavior of the stock market under (natural) disasters. Understanding which sectors in the industry are affected by how much can support making better decisions to get the stock market back to its normal as soon as possible. We can use this analysis in the future to make rational decisions in a case of another possible pandemic.

1.4 Related Work

Many researchers have started looking into the novel coronavirus and its impact on the economy. Baker et.al (2020) [5] studied the newspaper articles for three months to examine each daily move in the U.S. stock market greater than 2.5 percent, up or down. This was done by classifying the journalists' explanations into one of the 16 predefined categories. They find that the current pandemic has grave implications for public health and the economy due to the ease with which COVID-19 spreads, and the non-negligible mortality rate among those who contract the virus.

Yan et.al (2020) [14] analyze the past outbreaks and conclude that often markets will react adversely to these such incidents in the short run but that in the long run, markets eventually correct themselves and grow. They find that if people can predict that COVID-19 will incapacitate them and prevent them from working then stock prices will fall. A similar trend was seen when SARS broke out in 2003, and the stock prices experienced a sudden decline, however, it remained for a short amount of time as after 6 months stock prices were higher than they were at the start of the outbreak.

Giglio et.al (2020) [7] surveyed wealthy retail investors who are clients of Vanguard in mid-February 2020, around the all-time stock market high, and then again on March 11 and 12, after the stock market had collapsed by over 20%. The average investor turned more pessimistic about the short-run performance of both stock markets and the economy. In contrast, investors' expectations about the long run remained largely unchanged, and if anything improved. They find among those investors who were very pessimistic before the crash, i.e. those who ex-ante were expecting negative 1-year stock market returns, about half become more optimistic following the crash, showing that as the market crashed they revised their expectations for the future upwards.

Stephany et.al (2020) [13] identify the industries that are most in need of support to achieve higher effectiveness with cyclical and fiscal policy measures. They find that the travel and restaurant industry is likely to face a complete shutdown, whereas the pharmaceutical, logistics, and health sectors are likely to continue to operate at full capacity.

Procacci et.al (2020) [11] discuss the impact of the COVID-19 outbreak from the perspective of the market-structure. They observe that the US market-structure has dramatically changed during March 2020 and that the level of change has followed the number of infected cases reported in the USA. They find that currently the market-structure resembles most closely the structure during the middle of the 2008 crisis but there are signs that it may be starting to evolve into a new structure altogether.

2 Methodology

2.1 Data Collection

We collected information for the S&P 500 companies from Yahoo Finance. [4] Since we wanted to see the effect of COVID-19 on the stock market, we considered stock prices before the COVID-19 outbreak, and during. This period ranges from 11-15-2019 to 04-20-2020.

2.1.1 Database Attributes

Companies have attributes, such as Sector, ZIP code, regular market price, 50-day moving average, etc. As main attributes representing a company for our research, we decided to use the following:

- *Sector*: a categorical variable that across 11 different sectors;
  Industrials, Health Care, Information Technology, Consumer Discretionary, Utilities, Financials, Materials, Real Estate, Consumer Staples, Energy and Telecommunication Services.
- *Full-Time Employees*: a continuous variable
- *Enterprise Value*: a continuous variable
- *Profit Margins*: a continuous variable
- *Market Cap*: a continuous variable
- *Ask Size*: a continuous variable
- *Average Increase/Decrease* of shares' price: an ordinal variable

2.1.2 Calculating Average Price

We collected the daily high price for each day in the 6-month period. Then, we calculated the average price before and after 03-16-2020, which is when the outbreak started in the United States. From the calculations, we find whether the average high price increased or decreased after the outbreak and by how much. At this point, the average price is a continuous variable that we represent in percentage. However, we decided to apply binning method and to modify it to an ordinal one. We grouped the values into a 25% range per value, resulting in five values, as shown in Table 1:

Average Increase/Decrease (in %)	Class
(-95,-70]	Extreme Decrease
(-70,-45]	Major Decrease
(-45,-20]	Moderate Decrease
(-20,5]	Minor Decrease
(5,30]	Minor Increase

Table 1: Class names for average increase/decrease

2.1.3 Data Cleaning

Financial data is usually very clean, however, we had to do some preprocessing. We removed companies for which we could not access the prices and for which the majority of the attributes we were looking for were not present in the Yahoo Finance Database.

2.2 Attribute Selection Measures

There exists a good amount of work done on attribute selection measures. A fair amount of it is done in decision trees where the data is usually asked to be nominal. However, in our case, most of the data is continuous which makes these methods not applicable. If we tried to use the binning method to separate the algorithms, we would run into some trouble. Namely, if we decided to use binning by frequency, we would see that entropy for every class of the attribute will be equal according to equation (1) [9]. This is in no situation good because every attribute selection measure depends on entropy.

$$H(X) = -\sum_{i} P(x_i)log(x_i). \tag{1}$$

If we decided to use another method for binning, for a precise answer we would have an exponential complexity of possible outcomes. It is possible to use approximations, but we decide on estimating the importance of attributes via classification.

2.3 Classification

As we have previously defined classes for each range of average increase and decrease, we find that we have an imbalanced data set, i.e. the classes are not represented equally. This is shown in Figure 2.

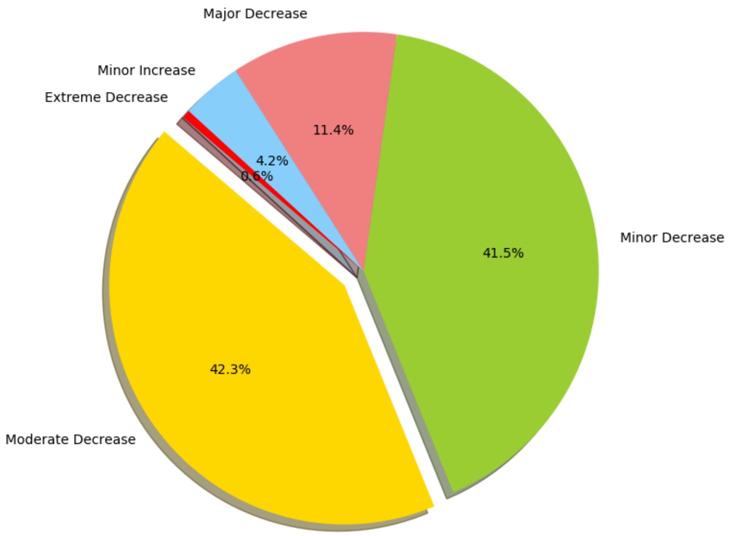


Figure 2: Class Imbalance

We decided to run multiple classification algorithms because only one is not always enough to get novel information. In order to decide which ones to use, we had to do careful consideration regarding the advantages and disadvantages of each algorithm. Our final choice was as follows:

1. **Bayesian network Classifier**
2. **LogitBoost Classifier**
3. **Random Forest Classifier**
4. **Multi-Layer Neural Network**

We use 10-fold cross-validation to train and test the data for each of the classifiers using Weka [8]. The next subsections describe the results we got from our algorithms.

2.3.1 Bayesian Network Classifier

Based on Bayes' Theorem, this classifier uses probability to classify a data point. The model can predict the class given the information about the other attributes using the Bayes Theorem. [12]
The Bayesian network classifier returns an accuracy of 64.06%. The general network structure is shown below in Figure 3. After analyzing the results from this algorithm we managed to get some interesting knowledge about the attributes. Since *Sector* is a categorical variable, the algorithm creates probability tables for each of the 11 sectors and 5 classes. However, all of the other attributes, i.e. *Full Time Employees*, *Enterprise Value*, *Profit Margins*, *Market Cap*, and *Ask Size* are continuous variables. For them, the algorithm creates the most optimal splits (if any) and learns the conditional probabilities according to those splits. The compelling thing to notice here is the number of splits for each of the continuous data. They are summarized in the following table 2:

Full Time Employees	**Enterprise Value**	**Profit Margins**	**Market Cap**	**Ask Size**
0	1	2	2	0

Table 2: Number of splits per attributes for Bayes-Net Model

Since the accuracy of the model is fairly good, a reasonable conclusion that can be made is that Profit Margins and Market Cap are much more useful attributes than Enterprise Value, and Full Time Employees and Ask Size do not contribute any valuable knowledge to the model. That can be seen from their probability tables, which include nothing but 1's.

The classifier's performance is displayed in Figure 4 as a confusion matrix. As we can see, over the diagonal our classifier performs reasonably well, especially for the classes that we have enough training data on. Moreover, our classifier usually misclassified a data point to a one which is in a class either above or below it (from an ordinal point of view). It can be inferred that even for the misclassified data points, we usually have a good estimate. The confusion matrix has one big anomaly, which is the prediction of a major decrease class as an extreme decrease one. This is because we only have two data points in the entire data set classified as an extreme decrease.

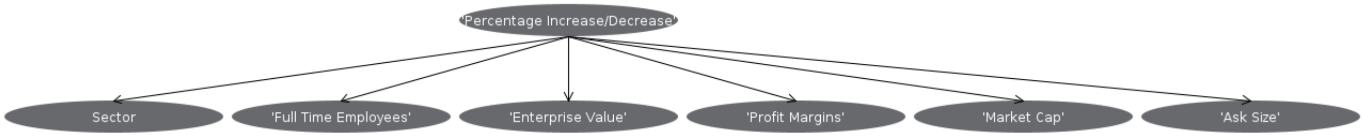


Figure 3: Visual Representation of Bayes-Net Model

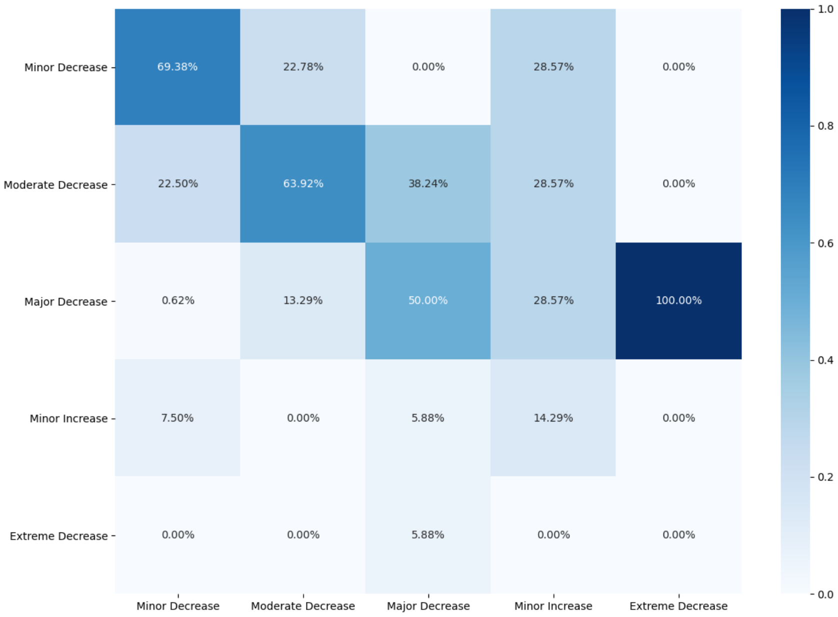


Figure 4: x-axis=Predicted label, y-axis=True label

2.3.2 Random Forest classifier

As part of the wide group of decision tree classifiers, random forest is designed to be faster than boosting and more robust to errors and outliers. [9]

Random Forest algorithm gave an accurate classification in 65.46% of the time. It was run with bagging with 100 iterations, i.e. 100 trees were created. Even though the way the algorithm works is by selecting a node from a list of random candidate attributes, it does indeed choose an optimal one from those. After running the algorithm we noticed an interesting pattern in choosing the root node of the tree, demonstrated in Table 3.

Sector	Market Cap	Enterprise Value	Profit Margins	Ask Size	Full Time Employees
44	29	17	10	0	0

Table 3: Root Node in Random Forest Model

The results of the most important attribute here indeed are convincing. The Sector has a much bigger chance of being chosen as one when compared to all the others, with Ask Size and Full Time Employees being the root node not even once. This strongly demonstrates that sector as a variable is a much more significant one than the other variables.

The confusion matrix regarding Random Forest is shown in Figure 5. As expected, random forest coped with outlier points better than our previous algorithms and we indeed have a stronger diagonal. The problem appears when we need to classify minor increase and extreme decrease because these classes appear the least in the training set.

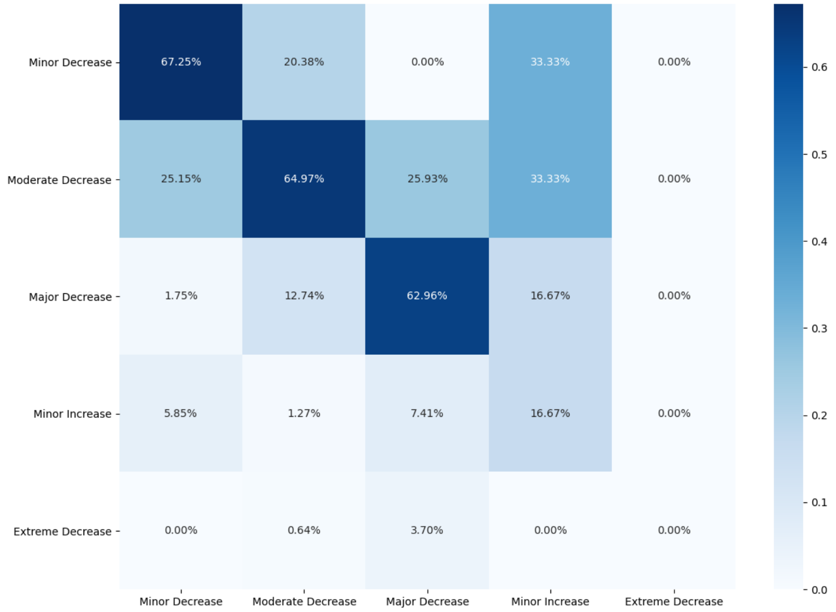


Figure 5: x-axis=Predicted label, y-axis=True label

2.3.3 LogitBoost Classifier

To complete our classification analysis we added a boosting algorithm. We chose LogitBoost, which is based on AdaBoost but applies the cost function of logistic regression. LogitBoost optimizes the log-likelihood of AdaBoost [6] and is the reason why we consider it in our analysis.

The LogitBoost Classifier returns an accuracy of 66.02%. This is in fact, the largest accuracy we got from applying our classification algorithms. To classify a data point, LogitBoost uses decision stumps (one-level decision trees) of the attributes. We look at the frequency of each attribute as a decision stump root. We noticed that LogitBoost looks at the attributes with frequency as shown in Table 4:

Full Time Employees	Market Cap	Profit Margins	Enterprise Value	Sector	Ask Size
14	12	11	6	5	2

Table 4: LogitBoost attribute selection

Thus, we can infer that *Full-Time Employees* and *Market Cap* are the attributes of great importance. This result is very useful, albeit contrary to the ones we have acquired so far, as discussed later on.

The confusion matrix representing the performance of LogitBoost is shown in Figure 6. It is fairly similar to the ones we have seen so far; a reasonably good prediction of three of the five classes, and a poor prediction for the very rare cases.

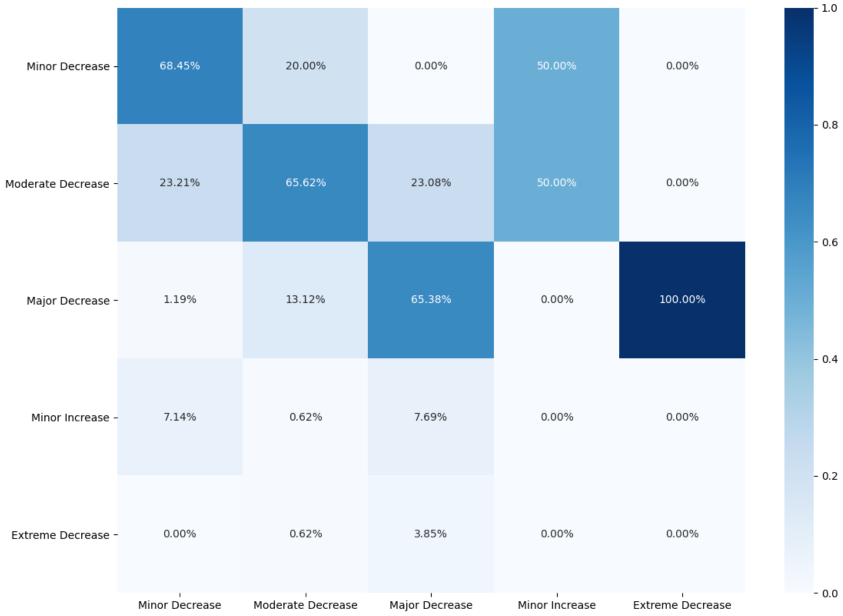


Figure 6: x-axis=Predicted label, y-axis=True label

2.3.4 Multi-Layer Neural Network

We ran the Multi-layer Perceptron (MLP) algorithm on the data. MLP utilizes a supervised learning technique called back-propagation to classify instances.

The two historically common activation functions are both sigmoids and are described by :

$$Logistic(x) = \frac{1}{1 + e^{-x}}$$

and

$$Logistic(x) = tanh(x)$$

[12] [3]

The Multi-layer Perceptron returns an accuracy of 61.28%. Even though the accuracy was decent, we decided that a multi-layer neural network is not very useful for our problem of comparing attributes and their importance. Since neural networks act as black boxes, we can't understand what the layers fully represent, rather give an educated guess. For this reason, we decided not to continue further with this approach.

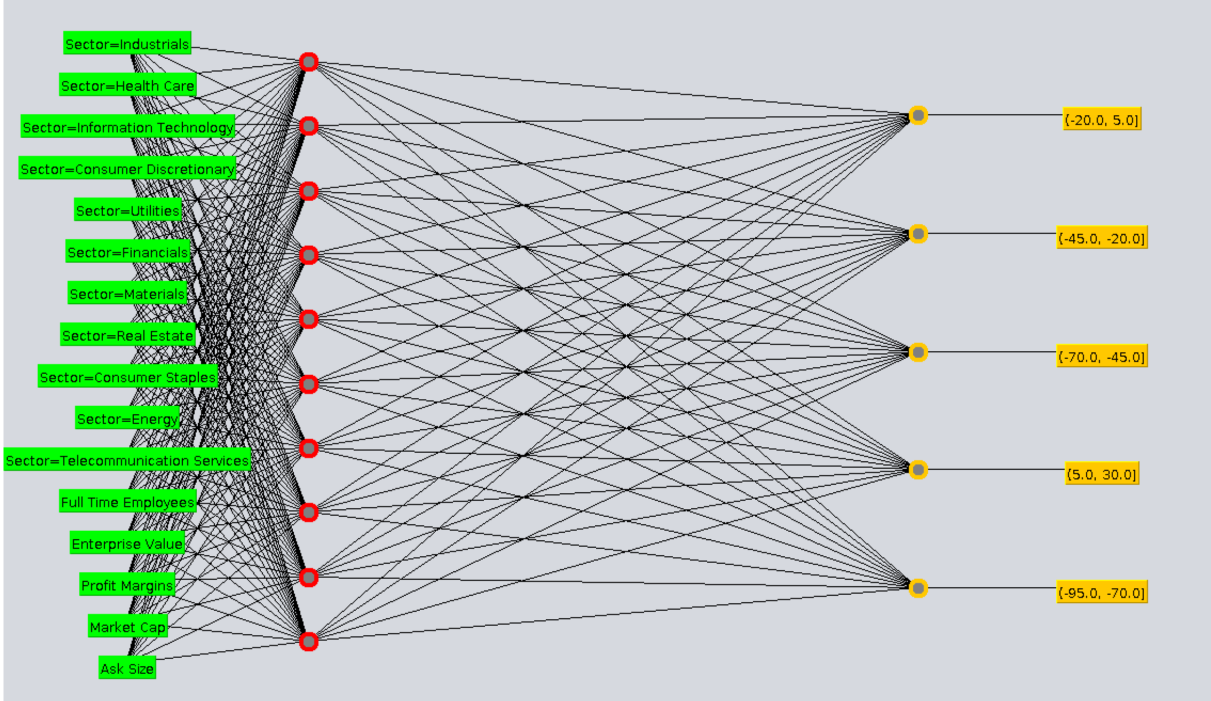


Figure 7: Visual Representation of Multi-Layer Neural Network

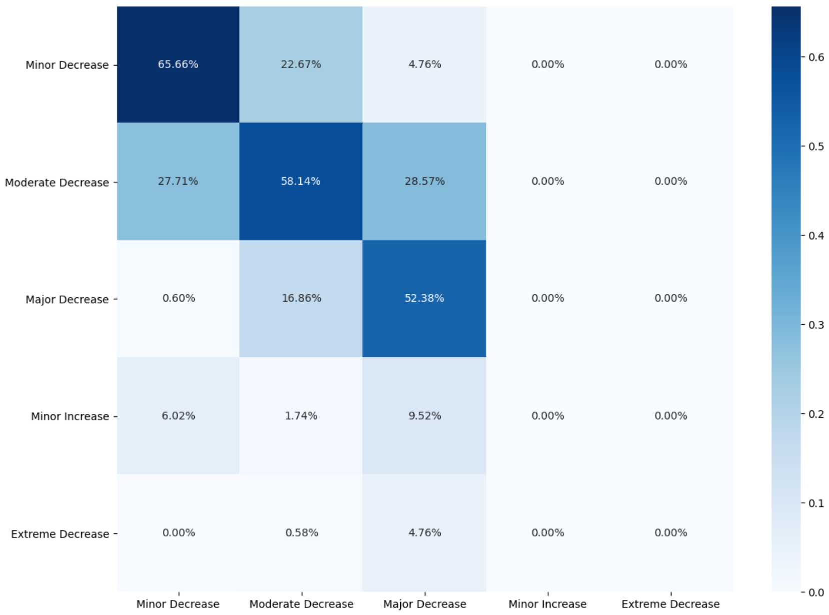


Figure 8: x-axis=Predicted label, y-axis=True label

3 Discussion

Research conducted by Yan et.al (2020) [14], Giglio et.al (2020) [7], Procacci et.al (2020) [11], indicated that there is going to be a substantial decrease in stock prices for the majority of stocks. We used this as background knowledge for the study as we assumed that the majority of stock prices will decrease after the COVID-19 outbreak across several industries. Moreover, we expanded on the work done by Stephany et.al (2020) [13] as we explored not only the impact on industries such as pharmaceutical, logistics, travel and restaurant, but also built a model that predicts the impact on a stock when given certain attributes of the stock. This model is more generic and considers more attributes to classify on the impact of the disease .

Most of our classifier had an accuracy of around 64%. We could see the similarity between the four confusion matrices was great; strong in the first three classes, but very weak in the fourth and fifth. The reason behind this is the class imbalance of the distribution over the *Average Increase/Decrease* class. Across the 500 companies we looked at, there was not enough data points with a Minor Increase or a Major Decrease, leading to some imprecision in the model's prediction accuracy. Nevertheless, the models give a good estimate of a data point's class. That is why we can build on top of them to understand the importance of attributes. From Bayesian-Network, Random Forest and LogitBoost we were able to get a sense of what choices the model makes when it tries to classify an unseen data point, unlike Multi-layer neural network which acts as a black box. We extracted information about the number of splits per attribute in Bayesian-Network classifier, (Figure 2), root nodes in the trees created by Random Forest (Figure 3), and frequency of each attribute as a decision stump root in LogitBoost (Figure 4). We combined this information to assign a rank from 1 to 6 to each attribute according to how important that classifier considered it. In the end we found the cumulative rank of each attribute, summarized in Figure 9.

We can see that *Sector*, *Market Cap* and *Profit Margins* as attributes are more informative than the other ones. Thus, if we need to choose attributes, these are the most important and descriptive ones about our data set. Another interesting point to note is that the cumulative result of all the attributes is identical to the result from Bayes-Network. Therefore, the best classifier for our purpose is Bayes-Network. This study can be improved by adding more attributes of a stock to see whether a different set of attributes would give better classification results.

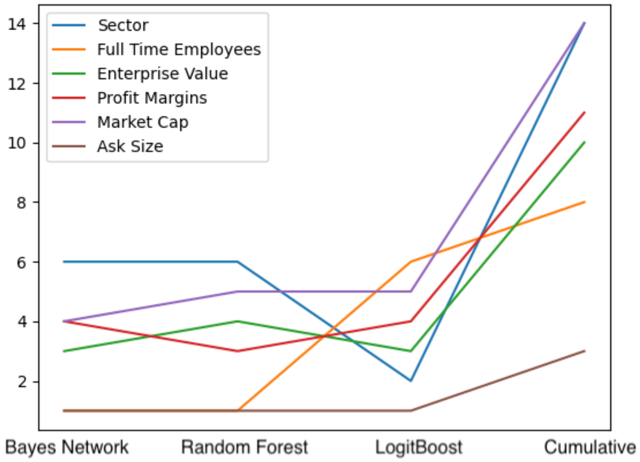


Figure 9: x-axis=Classifier, y-axis=Ranking

4 Conclusion

As stated above, we measure the impact in percentage increase/decrease in stock price after the outbreak of COVID-19. This is an important finding as it provides the government with a metric at the impact level. The classification models provide a prediction on the impact of COVID-19 on the stock price, given its *Sector*, *Market Gap*, and *Profit Margins*. It can be utilized to design policies and allocate funds in a proportion corresponding to the impact level of the stock. This can potentially help stabilize the economy by efficient injection of funds into the stock market.

5 Acknowledgments

We would like to thank Professor Jiebo Luo, and Muhammad Ahmad for the support extended to us in numerous ways while doing this study.

References

[1] Marketplace - market reactions to covid-19: Stocks end the week in decline. 2020.

[2] Politico - fed announces massive cash injection to relieve u.s. debt market. 2020.

[3] Wikipedia - multilayer perceptron. 2020.

[4] Yahoo finance - stock market live, quotes, business amp; finance news. 2020.

[5] Scott R Baker, Nicholas Bloom, Steven J Davis, Kyle J Kost, Marco C Sammon, and Tasaneeya Viratyosin. The unprecedented stock market impact of covid-19. Working Paper 26945, National Bureau of Economic Research, April 2020.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.*, 28(2):337–407, 04 2000.

[7] Stefano Giglio, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus. Inside the mind of a stock market crash. 2020.

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[9] Jiawei Han, Micheline Kamber, and Jian Pei. 8 - classification: Basic concepts. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 327 – 391. Morgan Kaufmann, Boston, third edition edition, 2012.

[10] Bohdan M. Pavlyshenko. Regression approach for modeling covid-19 spread and its impact on stock market. 2020.

[11] Pier Francesco Procacci, Carolyn E. Phelan, and Tomaso Aste. Market structure dynamics during covid-19 outbreak. 2020.

[12] Stuart Russell and Peter Norvig. Artificial intelligence - a modern approach. 20100.

[13] Fabian Stephany, Niklas Stoehr, Philipp Darius, Leonie Neuhäuser, Ole Teutloff, and Fabian Braesemann. The corisk-index: A data-mining approach to identify industry-specific risk assessments related to covid-19 in real-time. 2020.

[14] Heather Yan, Andy Tu, Logan Stuart, and Qingquan Zhang. Analysis of the effect of covid-19 on the stock market and potential investing strategies. 2020.