ORIE 4741 Midterm Project Report

Ray Weng (rw525), Dhruv Girgenti (ddg35)

Cleaning data:

Our initial dataset contained information about 3762 company IPOs, with 1665 different columns. Most of these columns contained information about the first 365 days of stock data (open, close, high, low prices), which are all response variables, rather than predictive features. Many potentially useful features about company financials had blank/missing values, which we could tell by summing up all NA values in the columns. These features could not be used because over 80% of the companies in our dataset had blank values for these columns. As a result, we were left with only a couple features that had dense enough data to be used.

In cleaning the data, we first removed all companies with less than 100 trades on the first day, as most of these companies had missing price data on the first day. Next, we created a new column that calculated the percentage change between closing price on day 1 and opening price on day 1. We then created another indicator variable for whether the price increased or decreased on day 1.

For the usable variables we identified (i.e. revenue, net income, sector, etc), we dropped all companies where any of these variables were missing. For other columns, like yearDifference (difference in year between company IPO year and company found year), we transformed all negative-valued yearDifference to 0, as these negative-values were often incorrect.

We actually noticed the dataset had some blatant errors in some columns. The opening price column and the lowest price column were actually swapped for all days, which we noticed by comparing them to values found on Yahoo Finance. As a result, we had to swap these values for these columns. There were also some typos in the dataset, which we manually corrected after looking at histograms for certain features. In the end, we were left with 2000 examples, and 10 potentially useful predictive features.

Exploratory data analysis:

In order to get a better understanding of our dataset, we chose to make several visualizations, illustrating various properties. Since our main interest is in determining what factors influence day 1 success, we first plotted a histogram of percent change in price on day 1 for every stock (Figure 1). The results show what we expected, which is that most stocks do not see much change in price on the first day. This makes sense because, over such a small range of time, there is very little opportunity for change, and many people may be hesitant to buy or sell a new stock. Although we are most interested in the data that falls to the left and right of this mean, these values will still help inform our analyses by offering insight into what features do not affect opening day success. We also wanted to supplement this graphic by providing a visualization to see how many stocks' prices increased on day 1 (Figure 2). From this, we see that most stocks close at a value that is equal to or less than their opening price.

Figure 1:

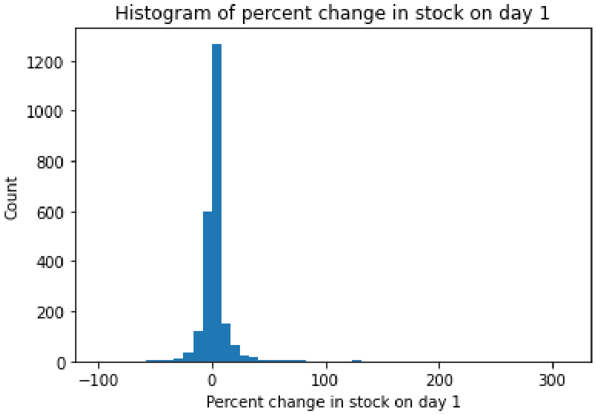
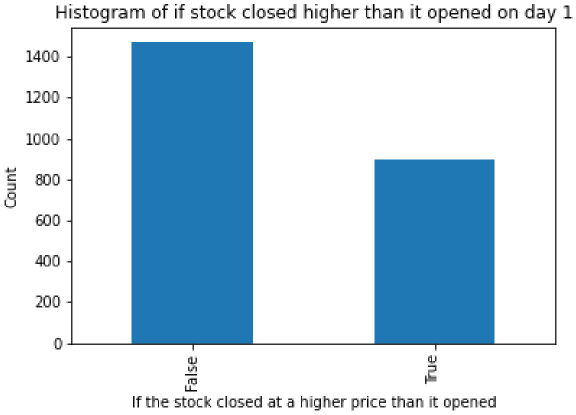


Figure 2:



We also chose to look at the distribution of sectors in our dataset (Figure 3), and where each company is located (Figure 4). In general, most companies are centered in the US, and four sectors dominate our dataset: consumer services, finance, technology, and healthcare. These together provide a better understanding of what kind of companies will play the biggest role in shaping our analyses.

Preliminary data analysis:

We performed some initial analysis on the data, using a 80/20 training/test split on the data. This split will help measure overfitting, find true accuracy for unseen data, and decide which model is the most effective. Overfitting will be carefully measured, and since we have a relatively large amount of data, we will try to keep the number of predictive features small, and use regularization where necessary. We used the following features in our preliminary analysis:

Year, the year the company IPOed; Day, the day of the week the company IPOed; Net income, net income for company; yearDifference, number of years between IPO year and year founded; USA Company, whether the company is US-based or not (one-hot encoding); ector, what sector the company is in (one-hot encoding)

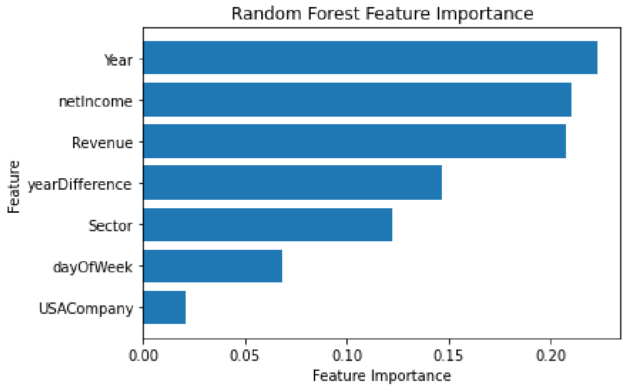
We chose revenue and net income because company fundamentals should determine how good of an investment the company is. USA Company and Sector also seem like important categorical variables - for example, we know that US Tech companies can be very hot investments and may be more likely to increase on the first day. yearDifference could be important, as younger companies' IPOs could have more hype surrounding them than older ones. Year and day of the week are more subtle predictors, but it’s possible that there is some correlation. Our response variable is a boolean value–True or False–on whether the company’s stock closed at a higher price than its opening price on day 1 of trading.

We chose two models to use for our initial analysis on the data: logistic regression and random forests. Both are effective models for predicting categorical variables. The results are summarized below:

Method	Training Accuracy	Testing Accuracy
Logistic Regression	0.53	0.52
Random Forests	0.99	0.59

Based on this data, we see that logistic regression does not overfit, but also does not generalize that well–even on training data, it can only reach 53% accuracy, a bit above average. With random forests, the model is able to have higher testing accuracy of 59%, but it does overfit a lot, with a training accuracy near 100%. These results are not particularly surprising; we know that predicting IPO performance is a hard problem, and we should not expect to have a huge edge over the market using only basic analysis and features. However, having this slightly higher accuracy does seem promising, as we are doing better than random guessing. Figure 5 shows the features in order of importance–surprisingly, year is number one, followed by net income and revenue.

Figure 5:



We also tried a basic linear regression model, using the same set of features, but using day 1 percent change as the response variable instead. This model performed much worse, with an R^2 of 0.01 on the training set, and R^2 of -0.29 on the test set. Given this news, we are more inclined to stick with the categorical response variable instead.

Future plans:

For future analyses, we would like to extend our current model to analyze other predictive features in our dataset, along with possibly searching for new features in other datasets. There are many attributes of a stock that could influence its day 1 success, and we are intent on finding which are the most influential, and the features contained within this dataset alone may not be predictive enough to find meaningful correlations. Although our main focus is on day 1 success, we are also interested in analyzing what factors influence success of a stock over longer time intervals. To do this, we will perform similar analysis methods, except we will measure changes over a week, a month, and a year. We are also interested in how predictive day 1 success is in determining price changes over these periods of time. In order to gain a deeper understanding of the relationships between our features and stock prices, we are also looking into using other analysis methods (beyond logistic regression and random forests) that will help analyze the data in different ways.

Appendix

Figure 3:

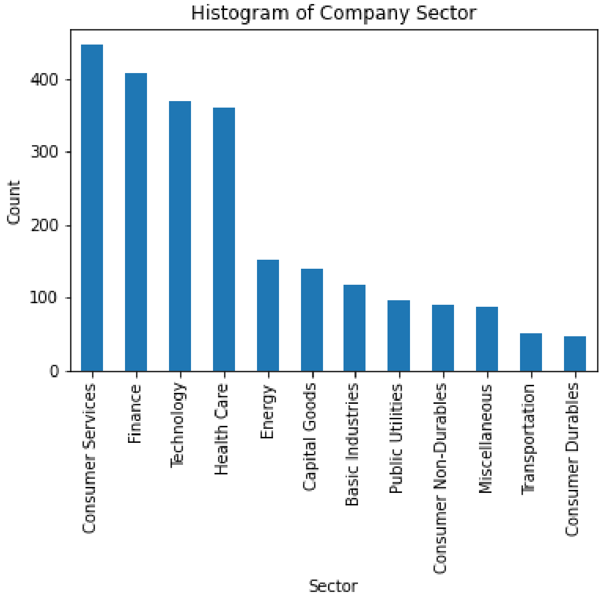


Figure 4:

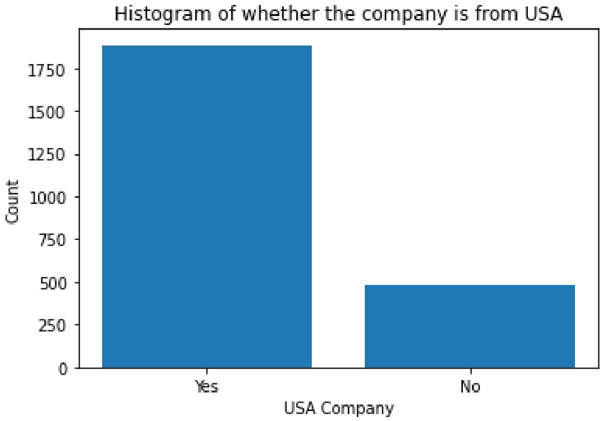


Figure 6:

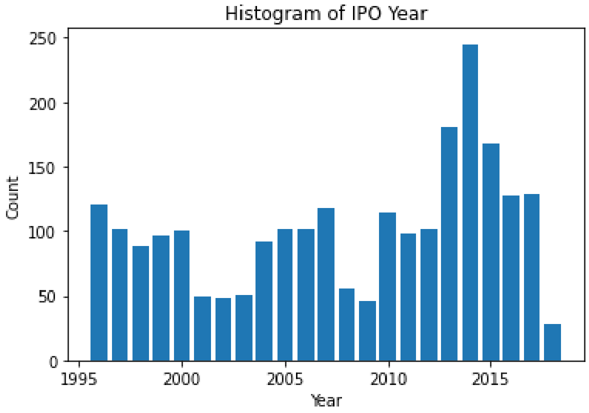
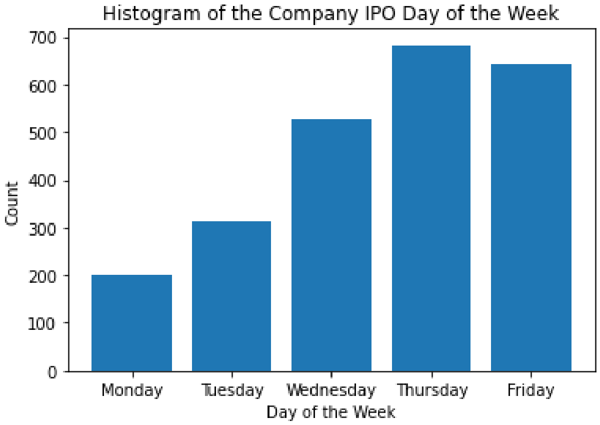


Figure 7:


Histogram of the Company IPO Day of the Week
Count
Monday
Tuesday
Wednesday
Thursday
Friday
Day of the Week