Jason Melnik

12/05/2021

DAT-115

Will Your Company be Profitable?

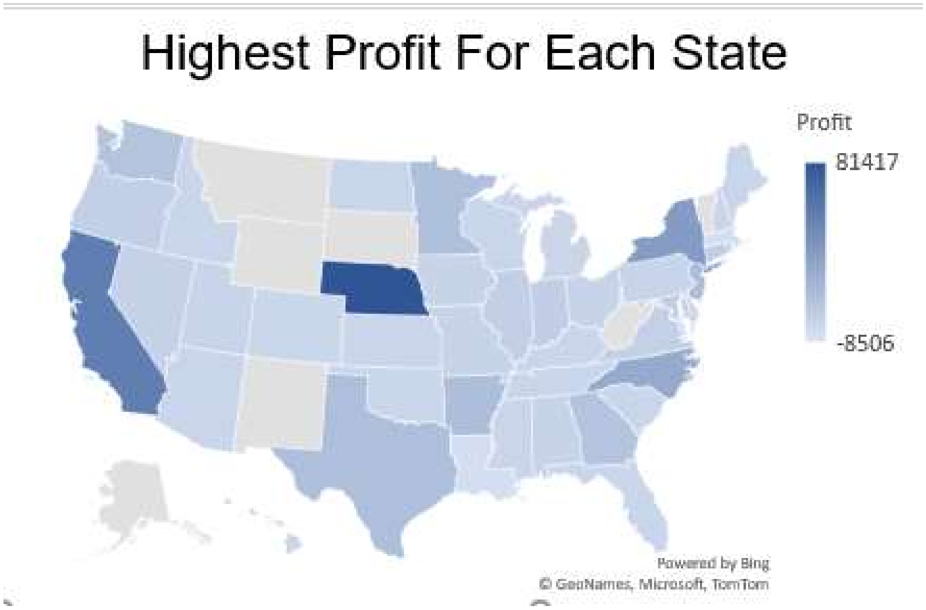
An investor only invests into a company that they know will be profitable. In the report I will be analyzing public variables in companies to determine if that company will be profitable. Using the public information with the fortune 1000 companies and using https://www.kaggle.com/datasets I was able to find data that can be used to determine if a company made profits this year. This is important because many investors would like to know if a company will make them money in the upcoming year.

Exploratory Data Analysis

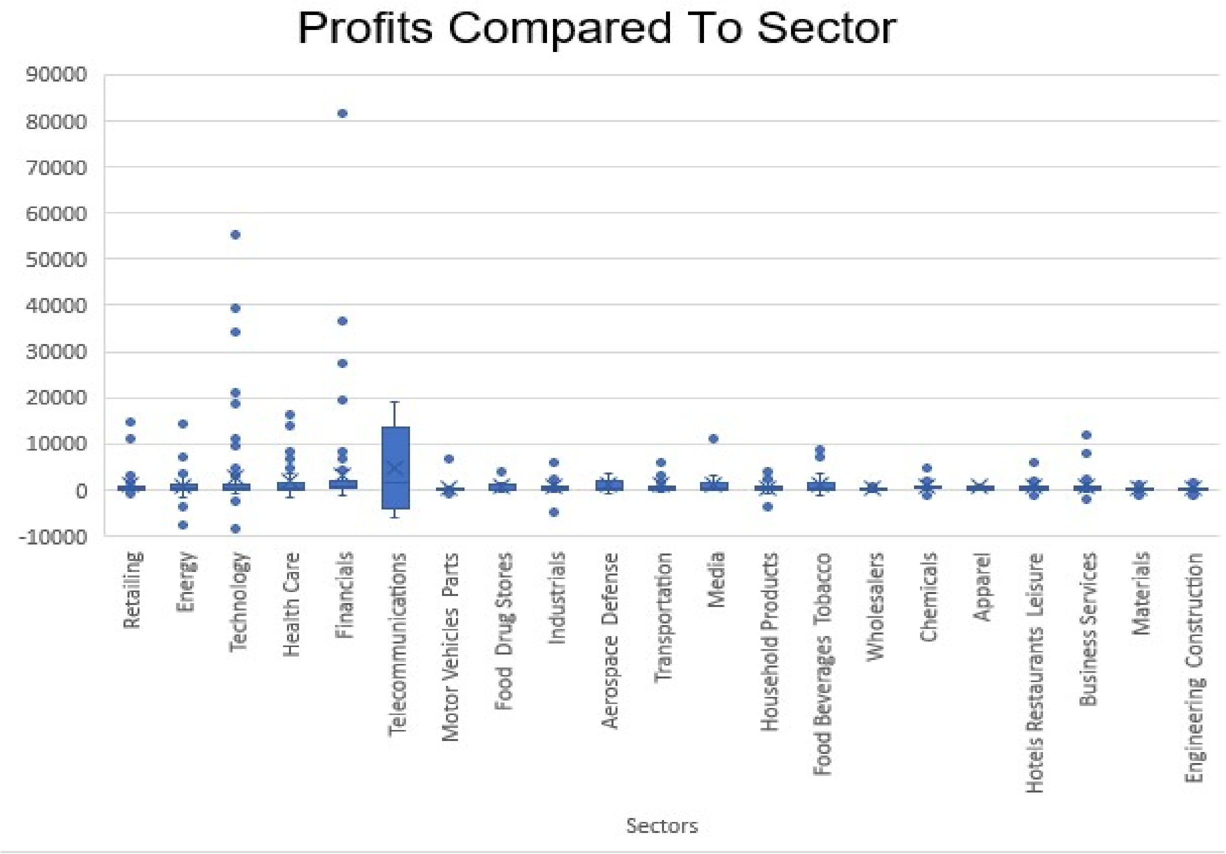
This dataset has thirteen input variables which includes rank, rank change, revenue, state, number of employees, sector, CEO founder, CEO woman, market cap, share price, and average volume. The output variable will be the profitable variable which is a series of "yes" and "no" which is what we will be classifying. There are other input variables such as country, city, company name, and profit which is not included for reason's that will be explained later. The rank is determined by Fortune which is determined by many factors that include the ability to attract and retain talented people, quality of management, innovativeness, and more.

Microsoft Excel

With Excel I was able to find the averages and standard deviations for the integer variables which had very wide ranges. Rank change has an average of -.04 with a standard deviation of 22.56 meaning that it would be normal is a companies rank change fell along -22.52 to 22.60 which is a very wide range. Revenue had a mean of 16778.31 and standard deviation of 37130.93 which makes a wide range of values being between -20,352.62 to 53,909.24.



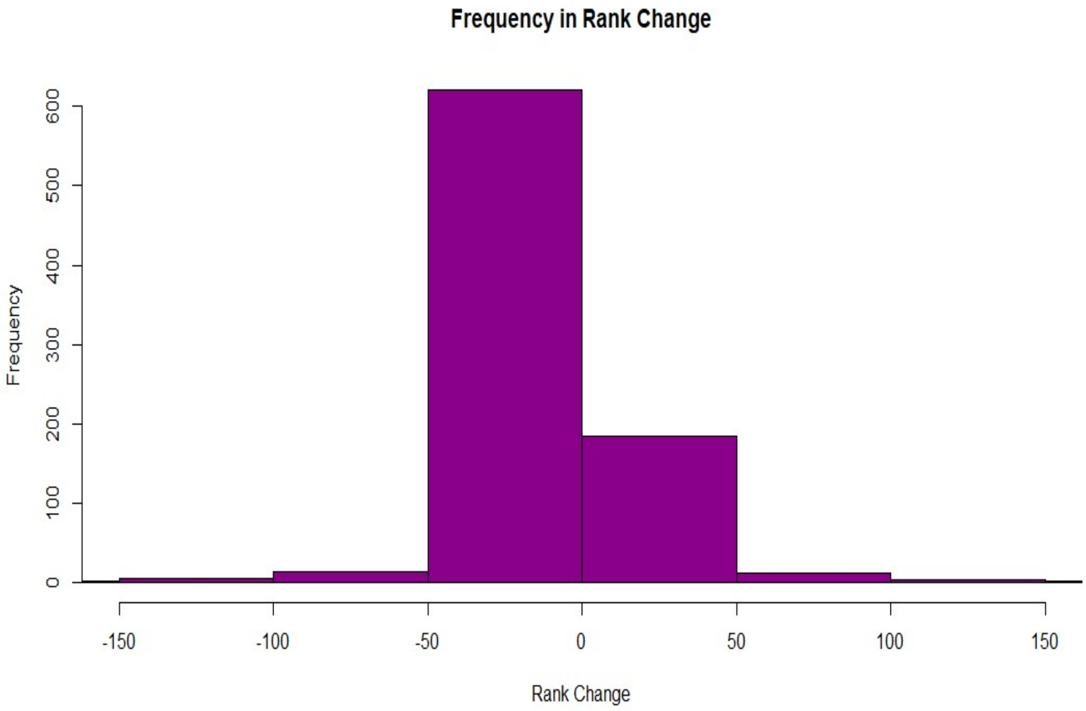
The graph above shows us a map of the states with their corresponding profits. This shows us which states have the highest profits and can help visualize the data much better. When classifying if a company was profitable, knowing the companies state could helpful since every state has a different profit.



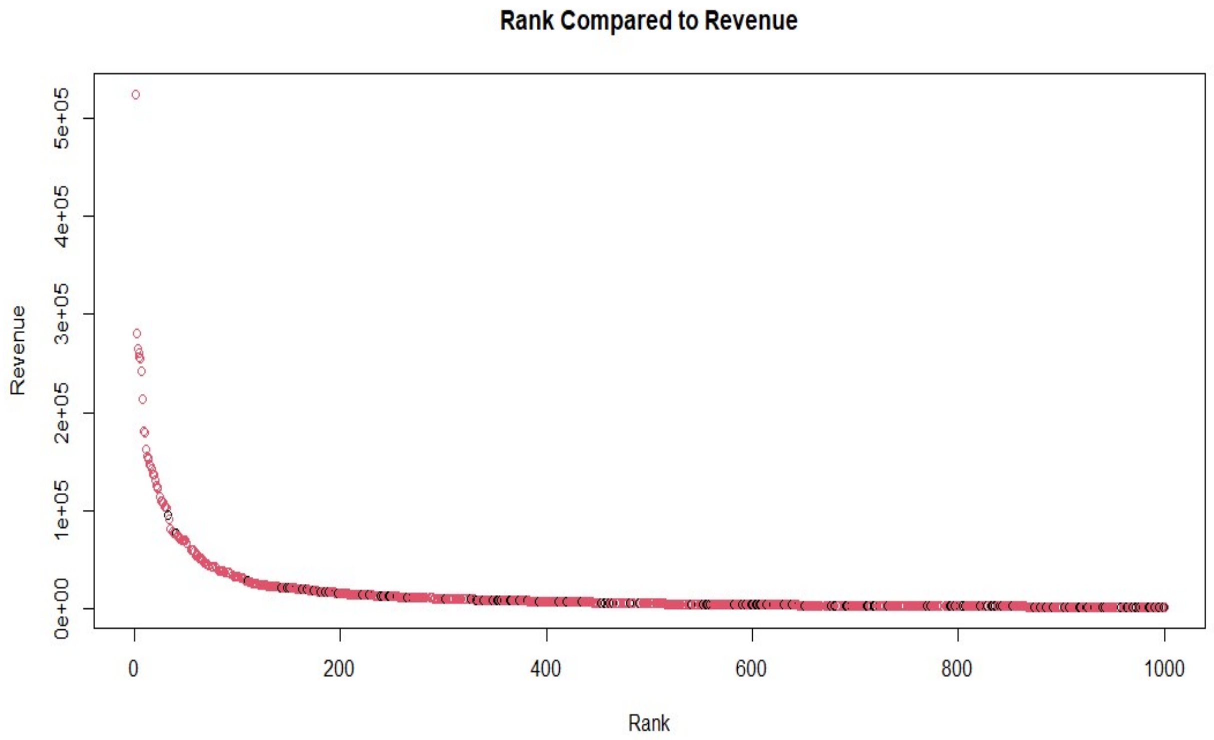
The graph above helps visualize the relationship between profits and sector. When classifying if a company was profitable, knowing the sector can be a huge help since every sector has a different range and average.

R Studio

With R studio I was able to make two graphs that can help analyze more relationships in the data to determine if a company was profitable. These graphs don't analyze profits but instead the relationship of the other variables.



The graph above helps visualize what range most of the rank changes happen. This shows us that most companies have their rank change between -50 to 50. These changes can be important to know if your company is profitable and the range it's in can help determine it.



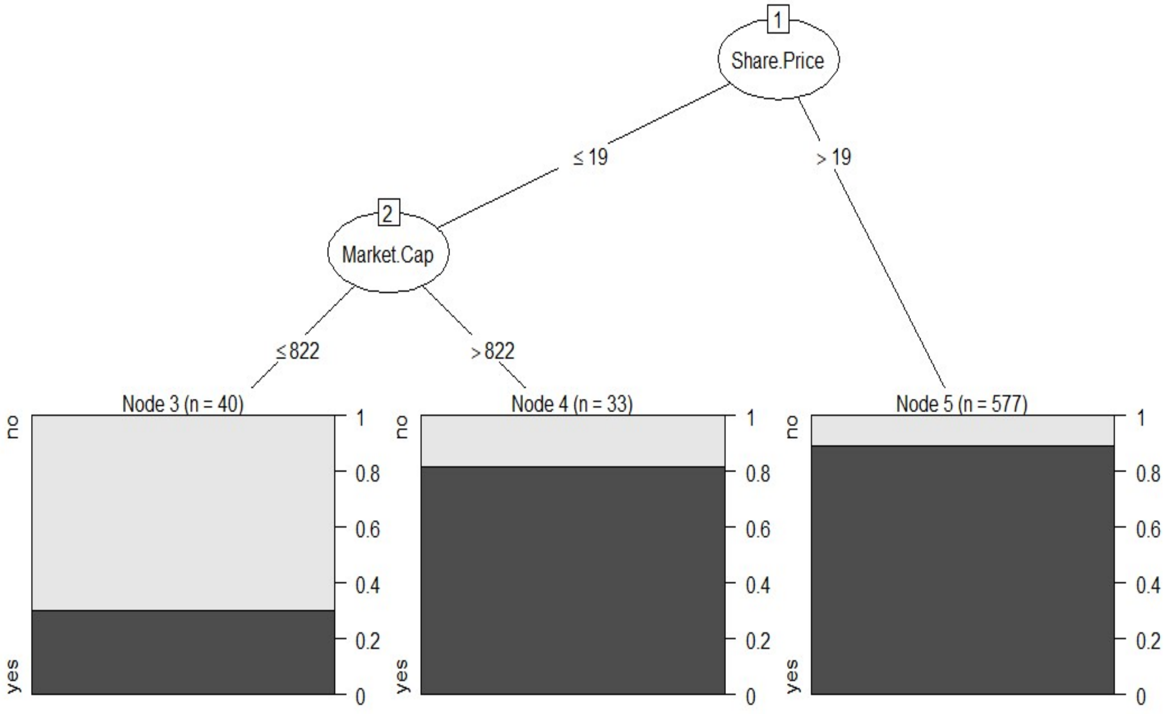
The graph above helps visualize the relationship between rank and revenue. The graph shows us that the companies rank corresponds to their revenue. The lower your rank is the less revenue your company makes. This can help determine if a company is profitable by using their rank and revenue. This graph is a classic example of 1/x which shows us the farther down your rank is the less your revenue will be.

Classification

I used the libraries C5.0 and Random Forest to classify if a company was profitable using the input variables that we have. I didn't include country since all the companies are in the U.S. and removed any companies that had missing data. I also removed profits since that would defeat the purpose of finding out if a company made profits this year. I also removed Company and City since those variables are not important to this classification problem.

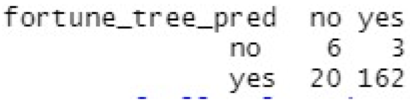
Before creating the classification trees, I removed the column's that would not be used and then used factor on the profitable columns to associate "yes" with 1 and "no" with 2. Then I randomize the data and split them up into training and test sets. The split is important because when we train the data, we want data that the test data has not seen yet so that when we go to test the data there are no variables that the tree has seen already. Doing this removes bias and guarantees that the accuracy you get is the true accuracy on new data.

Decision Tree



Using the library C5.0 I was able to input the training set and it outputted the tree above. This shows us that the share price and market cap are the biggest influencers in finding out if a company made a profit. The tree goes from top to bottom with the ovals being the "leaves" and the lines as the "branches". The "leaves" is the variable, and the branches are what determine which side the company falls under. The first variable is share price and it checks if the company has a price less than or equal to 19 or greater than 19. If a companies share price is greater than 19 then it's most likely to be profitable while if its less than or equal to 19 then it goes to the next variable which is market cap. If the company has a market cap of 822 or less, then it was most likely not profitable while if it was greater than it was most likely profitable.

After making the decision tree we need to test it using the test data which it has never seen. When predicting the test data, we get a nice table that shows us the accuracy of the prediction using the tree created by the C5.0 library.



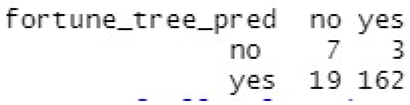
```
fortune_tree_pred  no yes
               no   6   3
               yes 20 162
```

Using this table, we can calculate the accuracy of the prediction. The numbers under the same row and column name need to be added together which is 168, this represents the correct prediction. This is because the prediction and actual line up. Then add all the numbers together which is 191 now using these two numbers we can find the accuracy. The equation below helps visualize the math in finding the accuracy to be 87.96%.

(6+162)/(6+162+20+3)=.8796*100 = 87.96%

Random Forest

Using the training and test data we now try using the random forest library to classify if a company made profits. We want to try using random forest to compare the results to the decision tree method. The random forest had the results:



```
fortune_tree_pred  no yes
               no   7   3
              yes  19 162
```

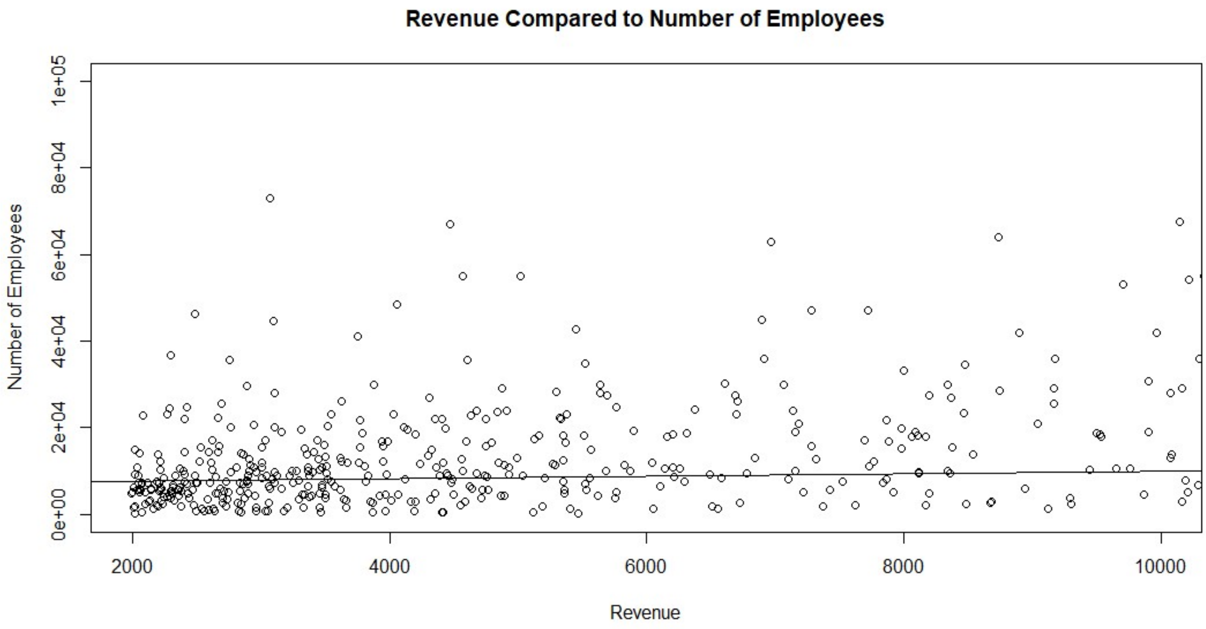
This shows us that the random forest prediction accuracy is 88.48% which is a better accuracy than the decision tree method. So, if you wanted to predict if a company made a profit, then you would want to use the random forest method to get the best accuracy.

Regression

The regression method will be used to determine quantitative data instead of qualitative data. So, we will be trying to predict the revenue variable instead of the profitable variable. The key difference is that the profitable variable is a qualitative variable since it's either "yes" or "no" while the revenue variable is quantitative since it can predict a number within a range of numbers. I used the same split of training and test data to determine the RMSE of the simple regression. For the multiple regression I used training, test, and validation data. The RMSE is the root mean squared error which is important since it tells us how far from the actual the prediction can be. The lower the RMSE is the more accurate the prediction is so using multiple regression we may be able to lower the RMSE.

Simple Regression

I created a simple scatterplot of the relationship between revenue and number of employees of the company. Then with the training data I created a model that can predict the revenue by using the number of employees of the company. This model outputs a linear line that can show us the general trend within the graph.



The above graph visualizes the relationship of the number of employees to revenue with the regression line showing the trend. The linear regression model has a RMSE of $19,319.68 which is a terrible RMSE. The reason is because revenue has a range of $0 to $200,000 so having a RMSE of almost $20,000 doesn't help much.

Multiple Regression

I also used revenue as the predicted variable for multiple regression. For multiple regression you use every variable in creating a regression model. When I used every variable in calculating the revenue with validation data, the RMSE is $9921.20.

To decrease the RMSE I tried to remove some variables in calculating the revenue. This turned out to fail since removing the variables CEO woman and CEO founder then running the validation data the RMSE is $10,002.83.

So, using multiple regression is much better than using simple regression to predict revenue since when predicting revenue using the validation data, RMSE turned out to be better by using multiple regression over simple regression.

Conclusion

This project has shown that investors can use classification to help them pick companies that will make profit. Random forest can give investors an 88.48% chance of investing into a company that will provide them profits in the future. This is an important indicator that investors can use before investing into any company. The use of regression models can also give investors some insight to how a company will perform in revenue. Revenue is important since the more revenue a company makes the higher the profits will be for them. The multiple regression model has an RMSE of $9921.20 which is a great predictor for finding how a company’s revenue will be. If a company is predicted to make anything more than $9921.20 then it’s a safe investment, while anything lower will increase in risk the lower it gets.

https://www.kaggle.com/winston56/fortune-500-data-2021

Multiple Regressions: