Team Zulu
MGSC-310
Dr. Doosti

**Final Report**

## Background Information

Our final project investigated a company based dataset. Our goal was to create a scoring system for propensity to acquire other companies. Intuitively, we know that the ability for a company to acquire other companies would be based on the size, amount of funding they received, or even the last date they received funding on. The data was provided from an employer at Basis State, containing information about various companies including their location, funding amounts, size and other variables. By implementing exploratory data analysis, regressions, classification algorithms, and data mining/cleaning, hidden trends and relationships were discovered within the data.

## Summary of Data(Downsampled)

**Instances:** 852
**Numerical Attributes:** Employee Range, Total Funding Amount, Total Acquired Column
**Categorical Attributes:** Name, Company Type, Industry, Last Funding Date

## Exploratory Data Analysis

*Figure 1*

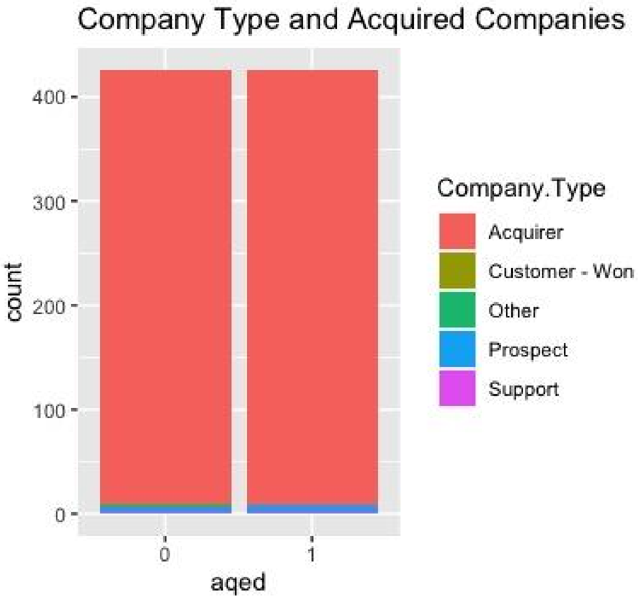
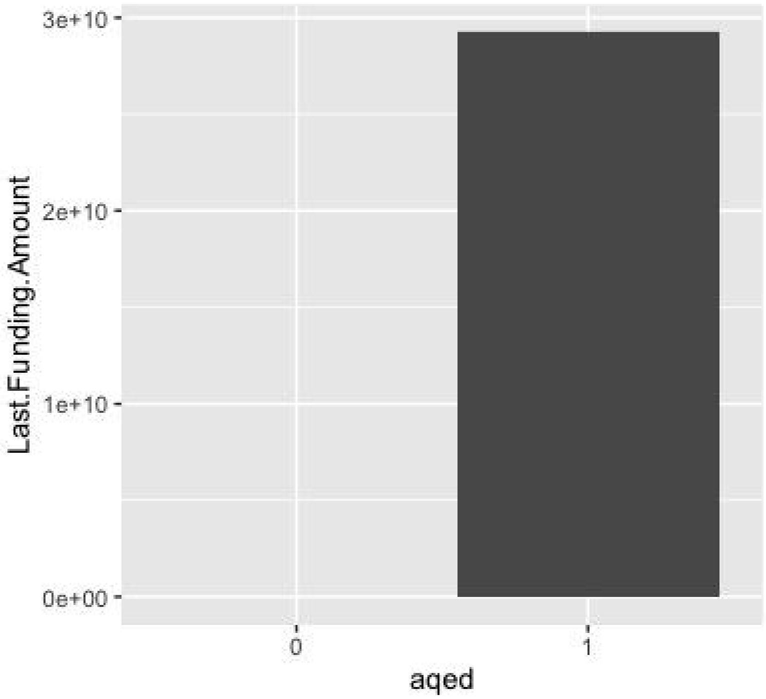


*Figure 1* displays a bar graph that compares the acquired column with the types of companies in the dataset. From the graph we can deduce that the majority of companies are type “Acquirer”. The other types of companies are virtually non- existent, so these companies probably won't really change the accuracy of our model much.

Figure 2

*Figure 2* is an intuitive graph simply proving that the companies that had acquired other companies previously have on average receive more funding than the companies that have ever acquired companies. The amount of funding that the 0 column received was virtually non-existent, or too small to see.



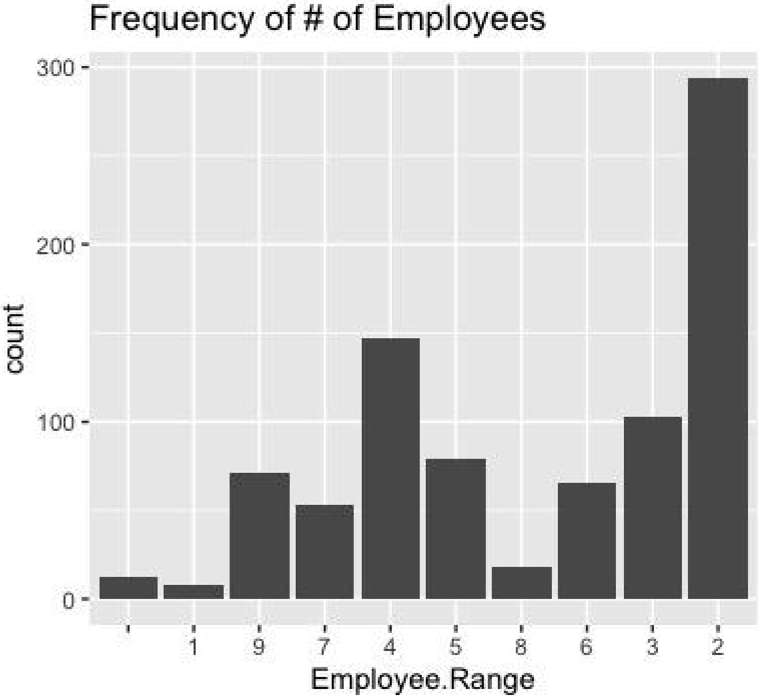


*Figure 3* shows the number of employees that exist(the size) of the companies in our downsampled dataset. The Employee range was converted into levels (1-9) from the string that it was originally in to easily manipulate the data and so that it was easy to use in our models.

*Figure 3*

*Figure 4*

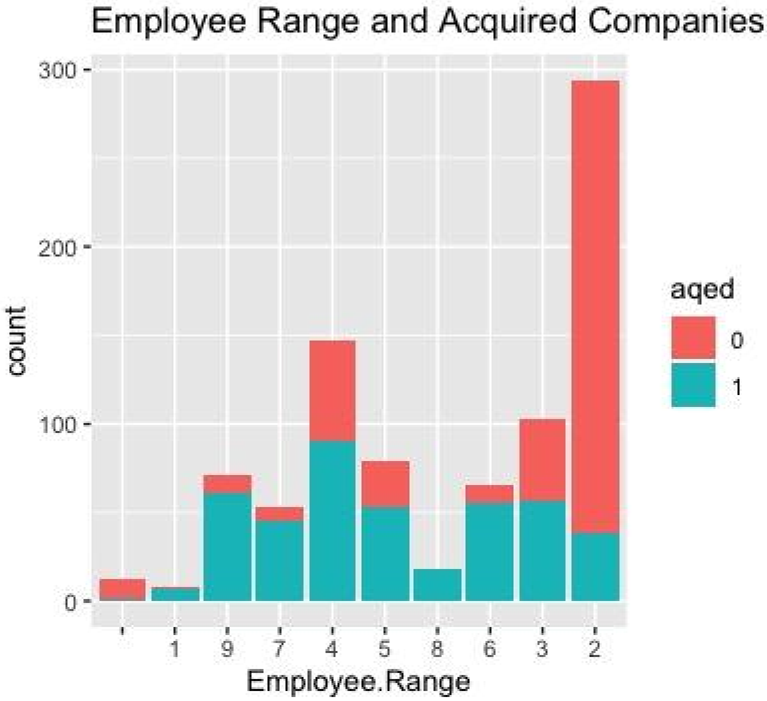
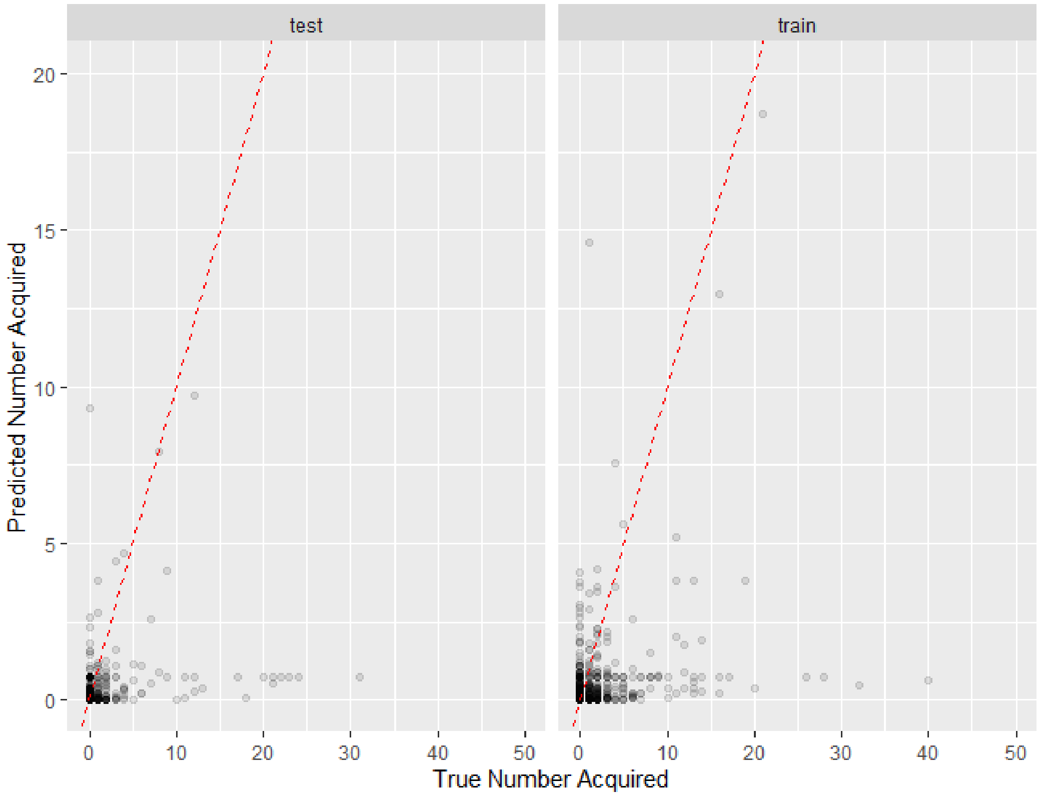


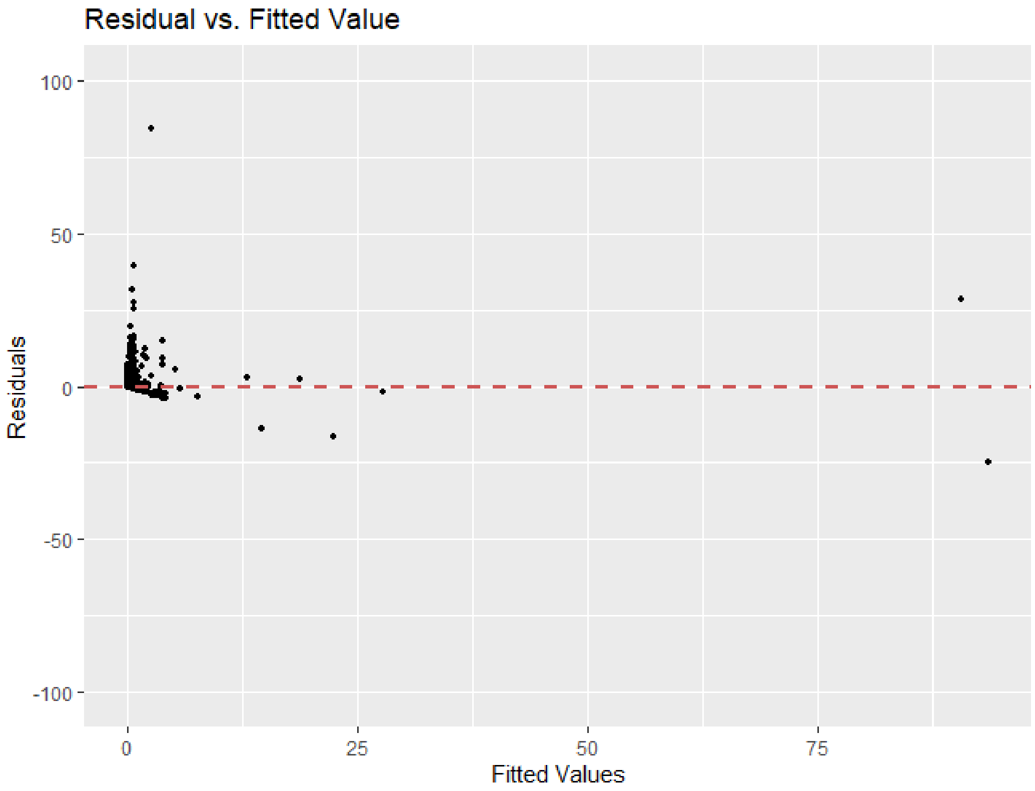
*Figure 4* shows the employee range and versus how many companies have been acquired. The companies under 0 have no companies acquired. This data makes sense because the smaller companies fall under 0, which means they haven't acquired any companies. The companies that were larger were seen to have acquired more companies.

**<u>Model 1 - Linear Regression</u>**

The first model focused on linear regression to determine the significance of the variables in the dataset in regards to a company’s propensity to acquire another company. Initially, all of the variables in the dataset were tested for general comparison. Variable selection was narrowed down with further testing to produce only the significant variables in the model, using a backward elimination method to remove the least significant variables one at a time. With an R^2 of .7844, this model proved to be a fairly strong means for analysis. The interpretation of the coefficient of determination, R^2, means that this model captures 78.44% of the variation in the response variable Number.of.Acquisitions.

This model lead us to the conclusion that the primary significant variables are Employee.Range, Total.Funding.Amount, Company.TypeProspect, and Last.Funding.Amount.


test
train
Predicted Number Acquired
20
15
10
5
0
0
10
20
30
40
50
0
10
20
30
40
50
True Number Acquired


Residual vs. Fitted Value
Residuals
100
50
0
-50
-100
0
25
50
75
Fitted Values

## Model 2- Logistic Regression/ Lasso Regression

The second model focused on developing a logistic regression that determined the top industries for acquirer companies. The model was originally trained using just the variable labeled "Industry." This model did not perform as well as when two variables were used, "Industry" and "Employee.Range." Therefore, the model used the size and industry of the companies to determine how likely the company was to have acquired another. To create this model, the original dataset was downsampled. This did not significantly affect the performance of the model. By downsampling and training the model of fewer observations, the time it took to train the model reduced from several minutes to instantaneous. Another altercation was the removal of every observation that was missing entries for the variables "Industry" and "Employee.Range."

A metric for measuring the performance of a logistic regression model is the AUC-ROC curve, which measures the ability of the model to differentiate between the two classes, i.e. had acquired another company or had never acquired another company. The ROC curve tells the model's ability to separate between the two classes, specifically how much overlap is between the true positive and true negative classifications based on a threshold value. The higher the area-under-the-curve (AUC) value is, the better the model performs. The baseline AUC value is 0.5, which indicates the model classifies the dataset at random and true positives overlap with true negatives.

The model training dataset generated an AUC of 0.9363. These values show that the model performs well. To look at the number of true positives and true negatives correctly classified, a confusion matrix is generated for both data sets. Three metrics can be calculated from the confusion matrix: accuracy, sensitivity, and specificity. The accuracy ratio is calculated as the total number of correct classifications over the total number of observations. The sensitivity ratio is calculated as the total number of true positives over all positive classifications. The specificity ratio is calculated as the total number of true negatives over all negative classifications. For the training dataset, the accuracy was 0.8420, the sensitivity was 0.8472, and the specificity was 0.8367. The close values of the sensitivity and specificity is good, because it shows that the model is not trading one metric for another. Depending on the threshold used in the model, one of these metrics may perform better. For example, a model may be very good at correctly predicting whether a company has acquired another, but because the threshold used is very low, the model overpredicts that many companies are likely to acquire another company. Our model used a threshold of 0.5.

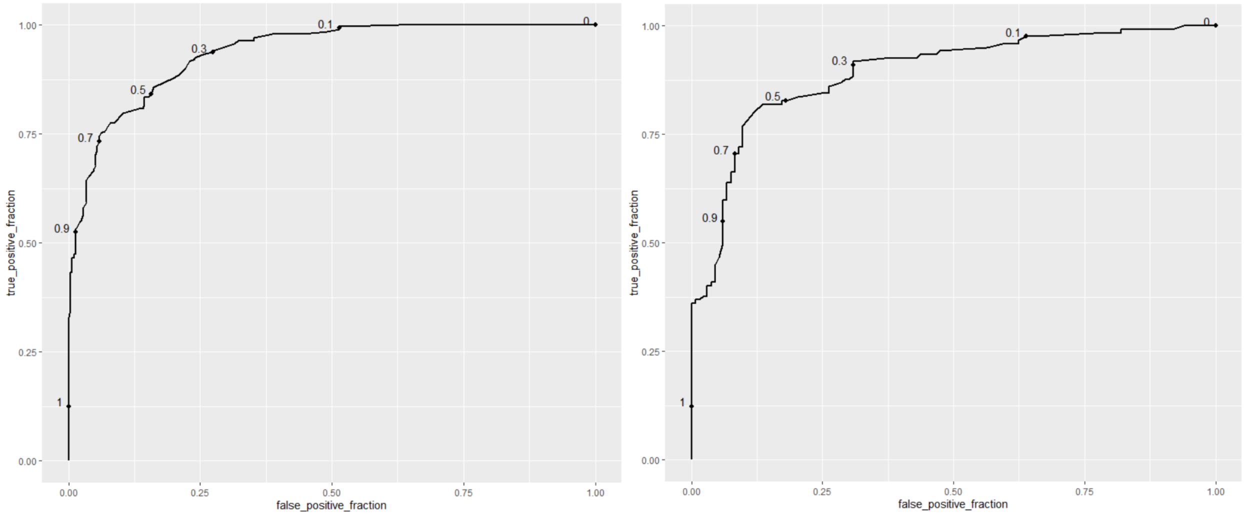


*Figure X: Left graph depicts the ROC curve for the training dataset. Right graph shows the ROC curve for the testing dataset. The ideal ROC curve is pulled to the upper left corner, resulting in an AUC closer to 1 or 100%. The threshold of 0.5 is chosen for the high true positive ratio without trading off too many false positives.*

As well as the training dataset works in the model, it is important to evaluate the performance of the model on new data, the testing data set. The model testing dataset generated an AUC of 0.8924. This is close to the AUC of the training dataset, so there is little indication of overfitting. Overfitting can be seen when the AUC of the training dataset is significantly higher than the AUC of the testing dataset. The accuracy was 0.8471, the sensitivity was 0.8472, and the specificity was 0.8043. The metrics for the testing dataset are not as high as for the training dataset, but the performance of the model was strong.

The purpose of the model was to identify the top industries of acquiring companies. Originally, there were 48 different industries used in the model. A form of regularization was used to drop variables that were less significant to the model. Lasso regularization works by penalizing the coefficients of variables by decreasing them closer to 0. With lasso regularization, the model that produced the lowest cross validation error had 36 variables. The variables can be filtered based on their coefficients: the higher the coefficient, the bigger the effect of the variable on determining the likelihood of an acquirer company. When filtered, the industries with the largest effect was as follows: IndustryInsurance; IndustryEducational Publishing; IndustryTech - Internet - Gig Economy; IndustryTech - Internet - C2C Marketplace; IndustrySAAS - Services Management; IndustryRetail - eCommerce; IndustrySAAS - Contractor Engagement; IndustryFitness Tech.

## **Model 3- Decision Tree/ Random Forest**

Upon inspecting the data, it was decided that a decision tree model would be ideal in an effort to visualize whether the companies studied have/ have not acquired any companies. In order for the decision tree to be generated, the entire data set was cleaned by removing unnecessary company data such as their address and zip code. Additionally, all “NA" values were dropped from the dataset and replaced with zeros in order to prevent any errors. Finally, similar to the previous models, the downsampled data was used. Given that there were over 2000 companies with no acquisitions, and less than 1000 companies had acquisitions, the data needed to be balanced.

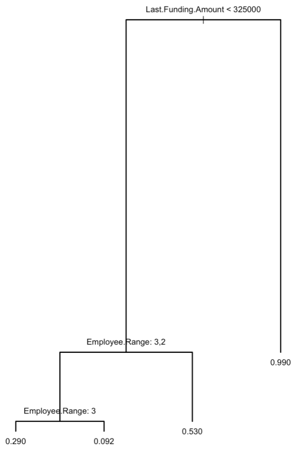


*Figure A: Decision tree generated from the code tree_business_test <- tree(formula = Number.of.Acquisitions ~ Last.Funding.Amount + Total.Funding.Amount+Employee.Range , data = organization_test). In which, the terminal nodes represent the number of companies acquired, and decision nodes represent the employee range, and last funding amount.*

As presented in figure A, the code yielded the given decision tree. The root node was the last funding amount, which split depending whether $ 32500 were funded. The following decision nodes were the employee ranges, splitting depending on the number of employees. At the terminal nodes, the number of acquisitions were found. The number of acquisitions ranged from .29-0.99 probability. Thus, indicating that if the closer to the value of 1, the more likely the company is to acquire other companies. Additionally, the RMSE of both the test & train data were found to be 0.3240147 (Train data) & 0.305317 (Test data). The values were not too far off, indicating that the model is somewhat accurate.

The decision tree utilizes a ranking system in which the higher to the top the explanatory variable, the more important it is. Given the decision tree, the last funding amount is most important. Whereas the employee range was the second most important. This was to be expected, since the data analyst from the Basis State company informed us that the funding amount and employee range are correlated to acquisition numbers.

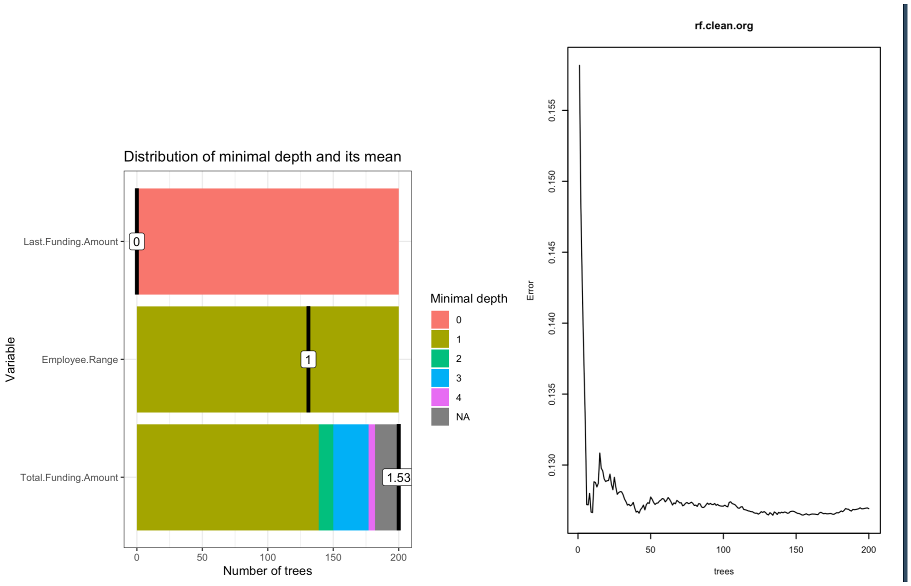


*Figure B: Random forest generated with the code rf.clean.org<- randomForest(formula = Number.of.Acquisitions ~ Last.Funding.Amount +Total.Funding.Amount+Employee.Range + Total.Funding.Amount ,data = organization_train, ntree = 200, mtry = 8, importance= TRUE). Along with its variable depth importance plot.*

The next step after the decision tree was to confirm the importance of the variables as well as to improve the accuracy of the model. A random forest model was generated as presented in figure B. The random forest plot represented the error (y-axis) vs. number of trees (x-axis). Which indicated that when a lower number of trees were utilized, the error was minimized. However, the value has to be higher than 5 trees, or else the error would be way too high. Additionally, the depth importance plot displayed the variables that appeared on the decision tree, in exception with Total.Funding.Amount. As expected, the last funding amount was the most important in predicting whether the company acquired another company. The employee range was the second most important. However, the total funding amount was at the lowest. Which was surprising, since it was thought that the last.funding.amount would be correlated to the total.funding.amount. Lastly, in an effort to determine the improvement of the model, the RMSE of both the train and test data were calculated again, 0.3443756 (train data) and 0.3533169 (test data). Since the RMSE are very close to each other, the model accuracy was improved.

## Conclusion

______In conclusion, the three models constructed performed relatively well in trying to predict whether or not a company acquired another company. The linear regression performed the worst, while the logistic regression and the decision tree were able to provide us with important insight, trends, and relationships within the data. The logistic regression provided us with an accurate ROC and AUC curve, from which we could conclude that the model was able to predict over 80% of values. The most important attributes were the last funding amount and the employee range. The models were limited to their performance just because of the way the data was structured. The data was so unbalanced to the point that even downsampling probably wouldn't provide the best results, but once more data that includes companies that have acquired other companies are added, the same models constructed in this project will perform significantly better.