Sean Olson

MBA706 – Data Analytics

Github link: <https://github.com/oldmanoldson/DataAnalytics/blob/Homework-4/SeanOlson_Homework4.ipynb>

Predicting Corporate Bankruptcy

# Exploratory Data Analysis

Appendix 1 shows the exploratory data analysis performed in the received dataset of firm bankruptcy data. The first thing done was to perform descriptive statistics on the data-frame. This showed the overlap of mean, standard deviation, and the percentiles. Once it was confirmed that there were no string-based variables in the dataset the next exploratory analysis performed was determining the ratio of failed to survived companies in the data. Although the case itself said that there was an even split, this was an easy way to confirm with a bar chart. A correlation matrix was then created to identify variables that were strongly correlated (greater than 25% correlation) with the dependent variable, renamed “Failed Firms”. After multi-collinearity was removed from the variables of interest then histograms were generated to ensure that there were no binary variables amongst the explanatory variable list.

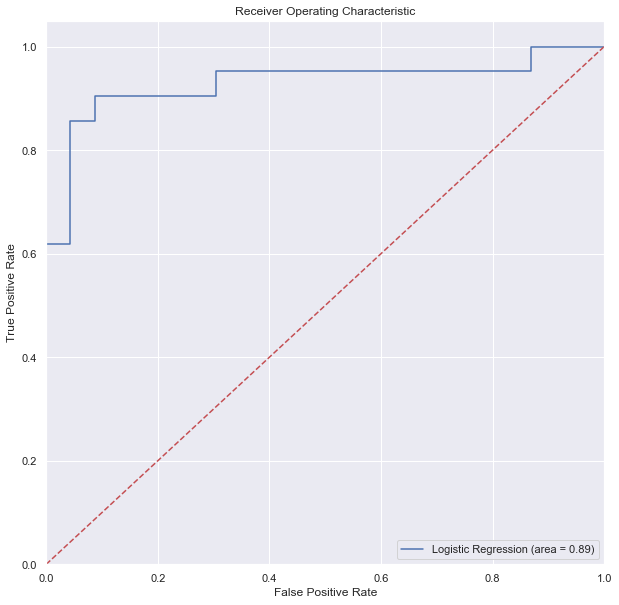
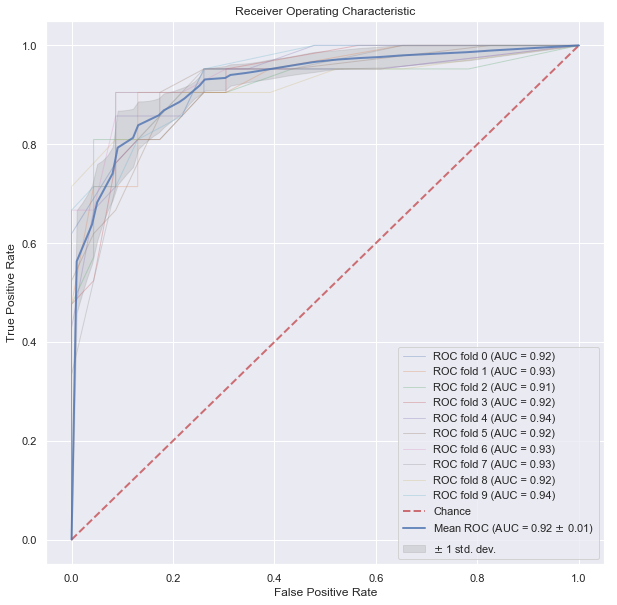
# Logistic Regression – Bankruptcy Prediction

The predictors for the final model were chosen using the correlation matrix method. Predictors were chosen if they were correlated at greater than 25% with “Failed Firms” because this was half the ratio of firms that did not survive according to the dependent variable. This narrowed the initial list down to 13 potential variables, which were then narrowed down to 5 variables after collinearity was established. The final logistic regression model was: . CURASS/CURDEBT ended up being the most significant predictor with a p-value of approximately 0.0002401. The rest of the output of the logistic regression model can be found in Appendix 2. To test model performance a confusion matrix analysis was performed as well as a classification report and ROC (a population independent lift curve). The confusion matrix showed a correct prediction of 39 out of 44 validation or test observations, this results in an 88.7% model accuracy figure, indicating a non-overfit model. The confusion matrix can be found in Appendix 2 along with the classification report and the logistic regression ROC curve.

# Random Forest – Bankruptcy Prediction

Instead of performing a simple tree classification model, a Random Forest classification model was built. Simply put, a Random Forest model is many decision trees based upon the given data that are then averaged to find the predicted value. Random Forests are great at avoiding multi-collinearity and overfitting due to the inherent mean absolute error calculations that go into building each decision tree. The resulting Random Forest model called INC/ASSETS and WCFO/DEBTS as the two most important variables to explain firm bankruptcy. By building the forest from all the explanatory variables, except “Firm ID”, a model is generated that is roughly 86% accurate according to the confusion matrix (38/44 predictions correct). Model validation was also performed in the form of a classification report to confirm the findings from the confusion matrix, as well as a ROC curve was generated showing different slices of the random forest and the mean area under the curve. The ROC curve is a better indicator of model performance because it looks at the false positive rate as the model is extrapolated. The Random Forest ROC curve showed a mean accuracy of 92%. These matrices and plots can be found in Appendix 3.

# ROC Curve (Lift Chart) Analysis

The ROC curve on the left was produced from the logistic regression, while the one on the right was produced from the Random Forest model. The Random Forest model outperformed the logistic regression model when looking that the probability of whether a predicted point was randomly generated or predicted. This is due to the inherent calculation structure of a random forest versus logistic regression model.

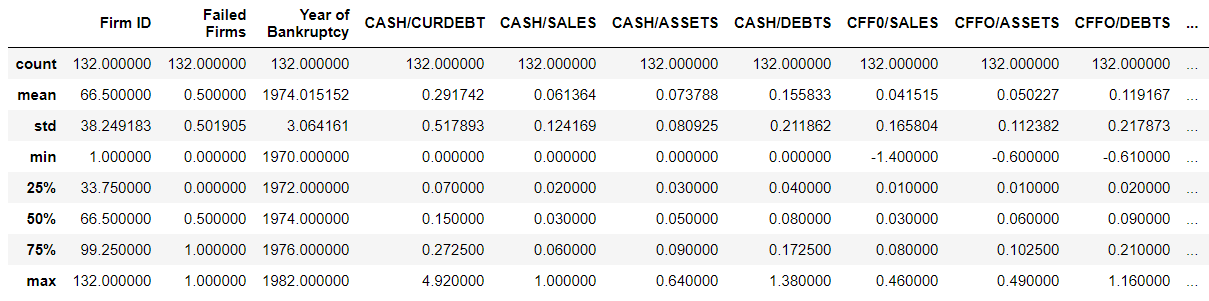
# Logistic Regression vs. Random Forest

Determining which model to use depends on the amount of available data. For a dataset of this current size, logistic regression is much more suited because it doesn’t need a lot of data to get as accurate as possible. Regression analysis was designed to be a standard method for all datasets, regardless of the size. For larger datasets, upwards of 1000 observations or more then Random Forest would be my model of choice. This is because the efficiencies of the algorithm are such that they perform very well on large datasets. Another factor to consider too is the number of predictor variables. This dataset had 25 predictor variables to choose from, which was becoming time consuming using a correlation matrix; sure, Recursive Feature Elimination could have been performed but when a lot of the variables correlated well with the dependent variable, RFE would have only narrowed the field a little bit. This is where the power of Random Forest was crucial. By taking all the variables and performing decision tree analysis on all the possible combinations highlighted variables that were originally nixed because of bias from the correlation matrix and the selector. When comparing the model performance of each they were almost identical in the correct number predicted, but Random Forest beat out Logistic Regression when looking at the ROC curve. Combined with automatic pruning of variables and automatic determination of variable importance, I would choose to go with Random Forest for this analysis.

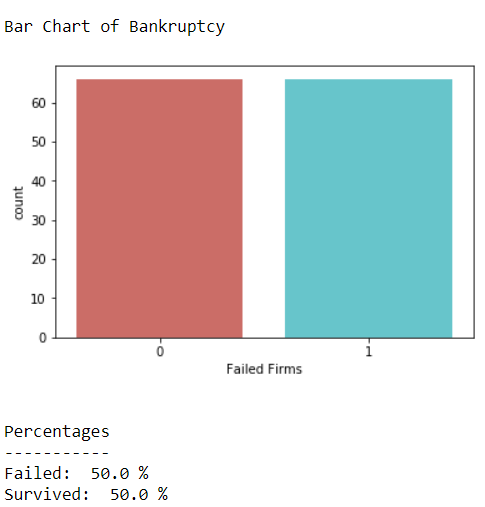
# Insights

According to the logistic regression a lot of the CASH/X variables were important to the prediction of the bankruptcy. This contrasts with the Random Forest model which puts a lot of emphasis on the INCDEP/X variables. And when looking at the exploratory analysis from step 1, we see both sets of variables correlated with the dependent variable but also with each other. Thus, enters the issue of multi-collinearity between these two sets of potential explanatory variables. Given the fact that logistic regression modeling must avoid colinear variables in order to produce an accurate model, a lot of explanatory variables are potentially being lost. Whereas Random Forest does not have the issue of multi-collinearity amongst predictor variables, we see insights from this model that weren’t apparent due to the natural constraints of the logistic regression model.

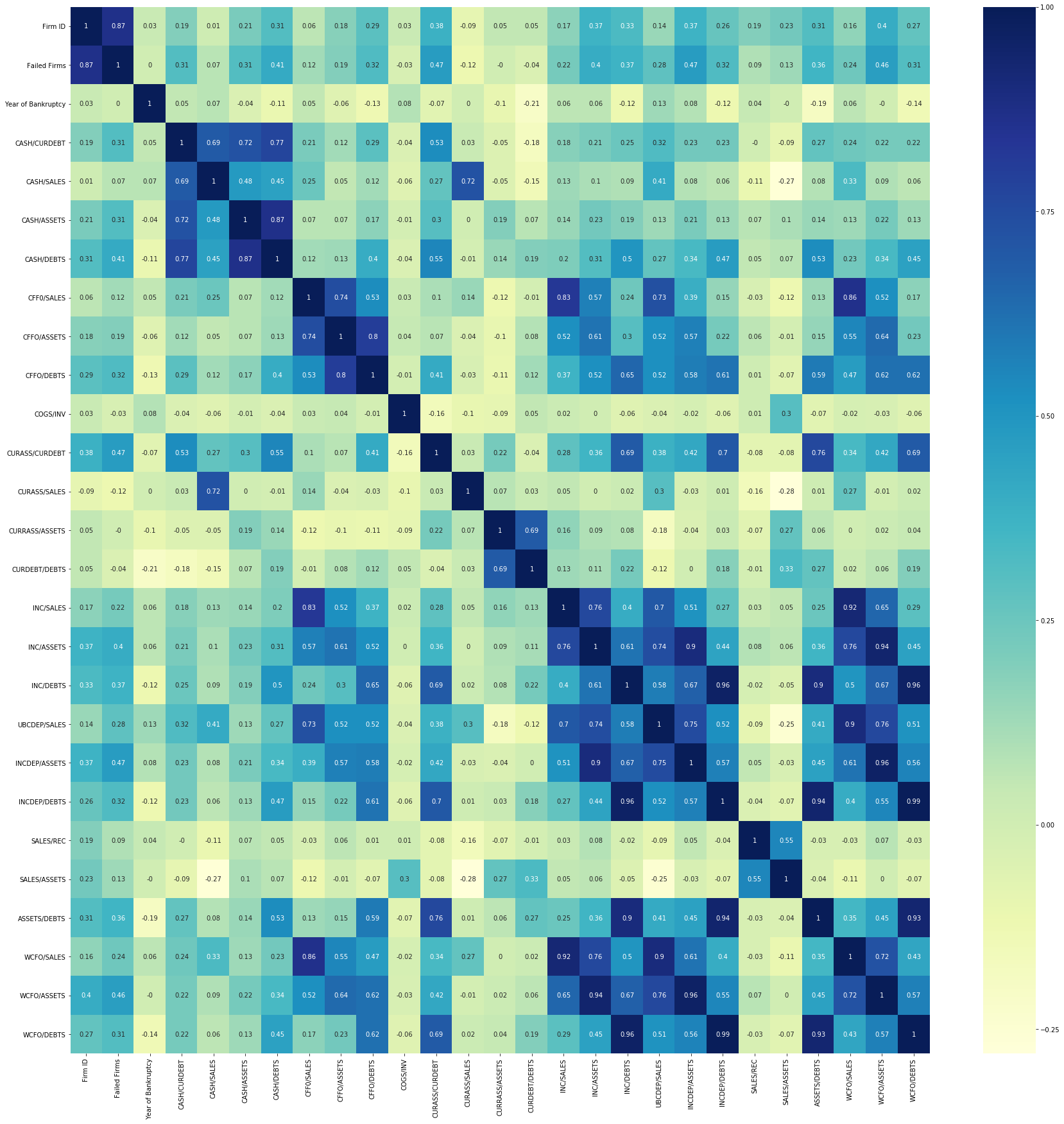
# Appendix 1 – Exploratory Data Analysis



Snippet of dataframe descriptive statistics

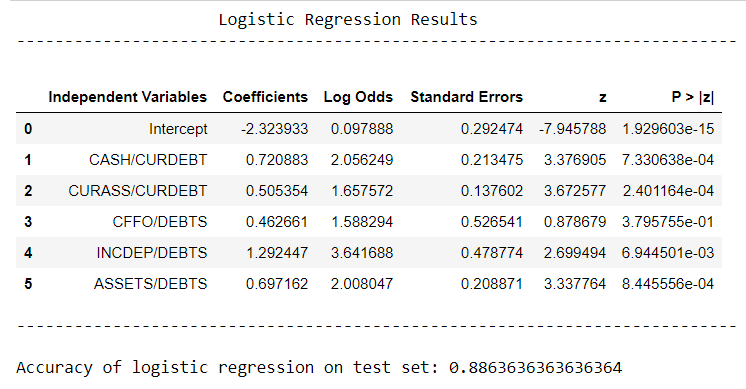


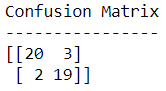
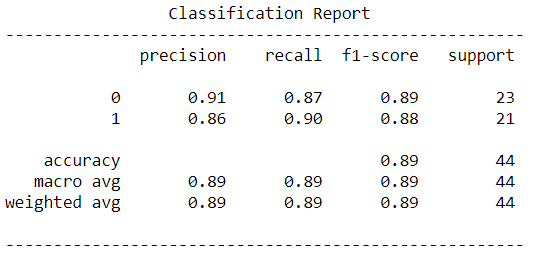
Bar Chart of dependent variable ratio

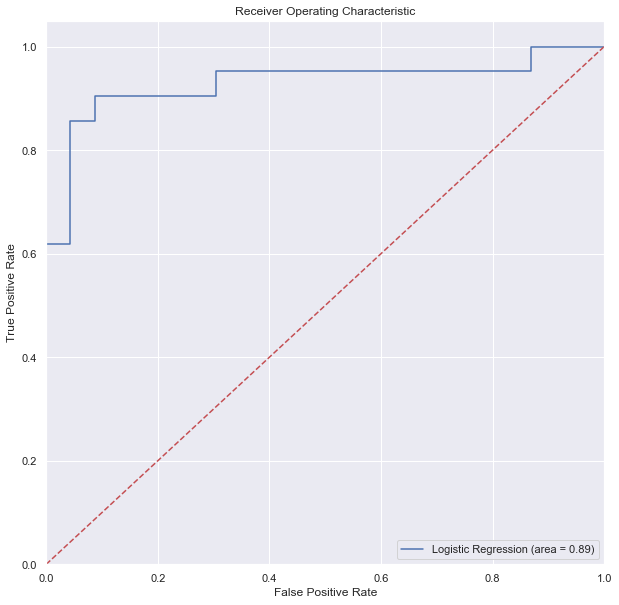


Correlation Matrix of variables found in dataframe

# Appendix 2 – Logistic Regression Model





# Appendix 3 – Random Forest Model

