Capstone Project 1: Statistical Data Analysis

Objective

The objective of this statistical data analysis is to shed some statistical light on some financials features of the data. After taking an initial look at the summary statistics, there are some questions about the statistical validity of some of the features, such as their correlations and means. A brief discussion about the correlations of the features will provide some thoughts on the potential of multicollinearity in the data set. Afterwards, a hypothesis is tested with the bootstrap inference method. This test looks to determine if the mean of the Attr39 feature (profit on sales divided by sales) is the same when this feature is split between companies who filed and did not file for bankruptcy. Early indications point to the mean of the feature when the company filed for bankruptcy being much different than the feature's mean when the company did not file. However, the sample size for the feature when the company filed for bankruptcy is much less than the other subgroup.

Correlation

The correlation matrix of features (independent variables) was generated, and a data frame of strong correlations (correlation magnitude ≥ 0.9) was developed. Of the 2016 correlation pairs of the 64 features (64 choose 2), there were only 26 pairs of these strong correlations. This could present a problem in future models due to the existence of collinearity. The presence of collinearity can cause overfitting of models in regression analysis and type II errors. Some of pairs had a correlation magnitude of 1, which was later determined to be pairs of the same data. One of the features of each corresponding pair will be removed in the prediction model as feature duplicity is highly discouraged. However, some of the pairs had correlation magnitudes above 0.95. These will probably be removed in the future as well for the prediction model, but further analysis of the feature and its significance as a financial metric will be conducted before this decision is finalized.

Hypothesis Scenario and Test

In the EDA part of the project, the top ten features with the largest correlation magnitudes with respect to the target were identified. Attr39 (profit on sales divided by sales) had the largest correlation magnitude of 0.29. Although these correlation magnitudes are relatively small, it is important to understand if the means of this feature are the same or different regardless of whether a company filed (Attr39yes) or did not file (Attr39no) for bankruptcy. The difference of the subgroup empirical means (0.053 for Attr39yes and -0.111) was 0.164, suggesting that the means are different. Thus, the null hypothesis is that the means of the two subgroups are different, while the alternative hypothesis is that the means of the two subgroups are the same. After the means of the subgroups were shifted to the concatenated mean, they were replicated 10000 times through a bootstrap inference method, and a distribution of the difference of the bootstrapped means was developed. The corresponding p-value was 0.40, which means that the probability of seeing a mean greater than 0.164 is 40%, and therefore, the null hypothesis is accepted, and the means of these two groups are, in fact, different.