Capstone Project 1: Data Wrangling

Loading File for Analysis

The data used for this project is the 5 year financial data of 5910 Polish companies, some of which have filed for bankruptcy. The data file is called "5year.arff". The file extension, .arff, stands for Attribute Relation File Format in which each column of data is an attribute that may be a numeric value, a nominal specification, a string, or a date. In this file, the 64 features are assigned as numeric values, and the target is classified as a nominal value of either 0 or 1 specifying that a Polish company did not file (0) or filed (1) for bankruptcy.

To successfully convert the .arff file to a dataframe, a collection of Python libraries are utilized. Specifically, the requests, zipfile, io, and scipy.io.arff.loadarff libraries allow the "5year.arff" file to be converted into a dataframe. However, when the dataframe was inspected, the target column of nominal values was blank with no data. For an unknown reason, the nominal column was lost in the conversion. To fix this problem, the file was opened and read using a dummy variable. Then, the 'replace' method was used to replace the nominal value, {0,1}, with 'numeric' before converting the file to a dataframe. Following the conversion, the target column was correctly filled with 0s and 1s.

Handling of Outliers and Missing Values

Before handling the missing values, the data was examined for the presence of outliers, in which, several were present. The outliers that fell outside the 95% confidence interval (outside $\mu \pm 2\sigma$) were removed. These values were transformed into Nan values that are addressed below.

There are several missing values in the data. Since the features are numeric financial ratios, a simple way to handle the missing data is to take the mean value of the observations per feature. In addition, considering that there are 5910 observations in the dataset, removing any rows that have a missing value may be explored as long as there are enough observations to create an accurate model. After removing the rows with missing values, the number of observations was reduced by 52% to 2814. Thus, I will model both sets of data (where the missing data was filled with the mean per column and where it was removed) and evaluate the significance of the two datasets. Furthermore, feature “Attr37” ([current assets – inventories] / long-term liabilities) only contains 3362 out of 5910 observations. This represents over 43% of missing observations. With such a large number of missing observations, this column may not be useful in predicting bankruptcy, but future analysis will be conducted to resolve this occurrence.