Capstone 1 Milestone Report: Bankruptcy in Poland

1: Background

Poland has historically reported one of the largest number of companies that have filed for bankruptcy on an annual basis in Europe. Between 2016 and 2017, Polish bankruptcy cases have averaged well over 4000, placing them in line for second place with the United Kingdom but well behind Switzerland with over 10,000 bankruptcy filings. Bankruptcies not only hurt the company itself – employees needing to find new jobs, often unexpectedly with short amounts of advanced notice, but also the community at large, such as companies who rely on the goods and services the bankrupt company provide and residents who consume these products and services. Governments and companies must develop a way to use a company's data to predict whether that company will file for bankruptcy to allow the imposition of preventative strategies and intervention before filing occurs.

2: General Information About Collected Data

The dataset that will be used is a dataset representing 64 financial metrics from the 2012 financial statements of 5910 Polish companies. It is a clean dataset that was accessed through the UCI Machine Learning Repository in the form of a text file with no categorical variables. Based on those statements, 410 companies filed for bankruptcy in 2013 representing almost 7% of the companies in the study. Some of the 64 financial metrics that will be used as attributes in this study include the net profit/total assets, earnings before interest and taxes (EBIT)/total assets, working capital, net profit/sales, total assets/total liabilities, equity/fixed assets, gross margin, and profit margin. It is not determined at this time if all the attributes are needed to make a prediction but optimizing the tradeoff between a low bias and a low variance will be considered.

3: Data Preprocessing

The data file for this project is “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. 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, any rows that have a missing value (or NaN values) may be explored removed 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.

4: Exploratory Data Analysis

The next step of exploring the data was useful to help answer some questions about the observations. Some of the questions that were considered were 1.) has the bankruptcy rate in Poland statistically increased over the five-year period from 2008-2012, 2.) if so, what factors may be causing the increase, and 3.) could accurate reporting of financial records provide significant information to help businesses continue operations?

Taking the ratios of bankrupt companies to companies still in business for each of these five years point to an increase in bankruptcy of Polish companies for that time frame (Fig. 1). Looking at the correlation distribution by year of how a company's bankruptcy filing decision is correlated with the reported financial attributes, it is evident that there is relatively low correlation for the first four years (Fig. 2). The tails of these distributions (showing the highest magnitude of correlation) are below 0.15. However, the 2012 correlation distribution shows increased correlation magnitudes (around 0.30), albeit, still relatively low. Thus, what features are directly related to this increase in correlation magnitude?

Figure 3 shows the attributes of the ten largest correlation magnitudes. Interestingly, 80% of those attributes are inversely proportional to total assets, while the other 20% are inversely proportional to sales. This result highlights the importance of those two metrics in predicting if a company will file for bankruptcy. The attribute names of these financial ratios are shown in Figure 4.

5: Further Statistical Analysis

After taking an initial look at the summary statistics, some questions still existed about the statistical validity of some of the features, such as their correlations and means. Further analysis 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.

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.

In the last section, 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.

Figures

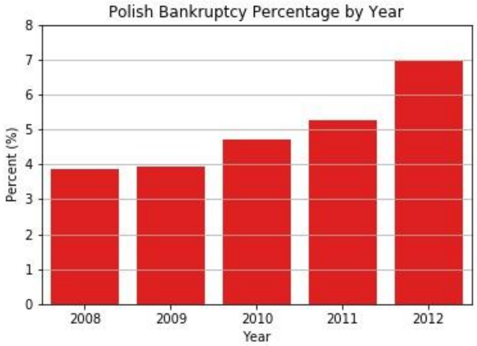


Figure 1. Polish bankruptcy percentage by year from 2008-2012.

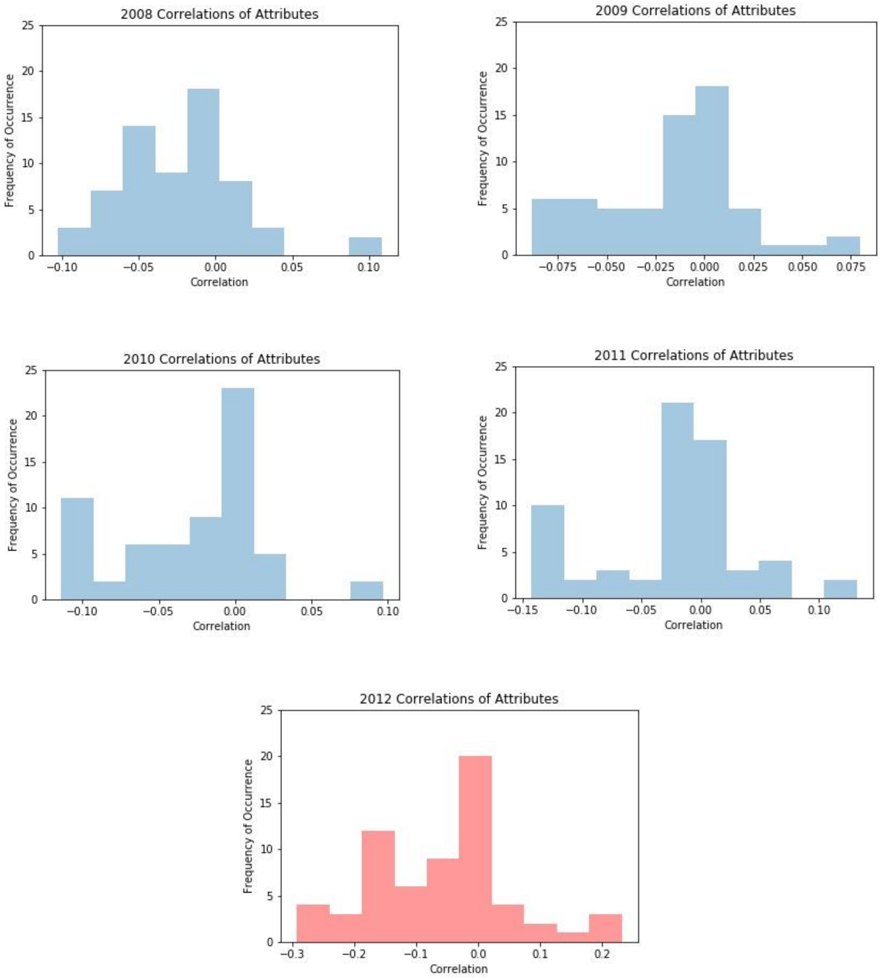


Figure 2. Correlation distributions for each year between 2008-2012.

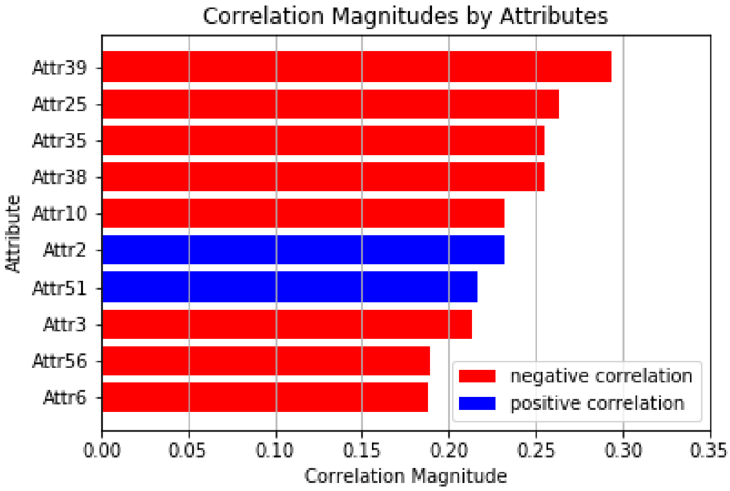
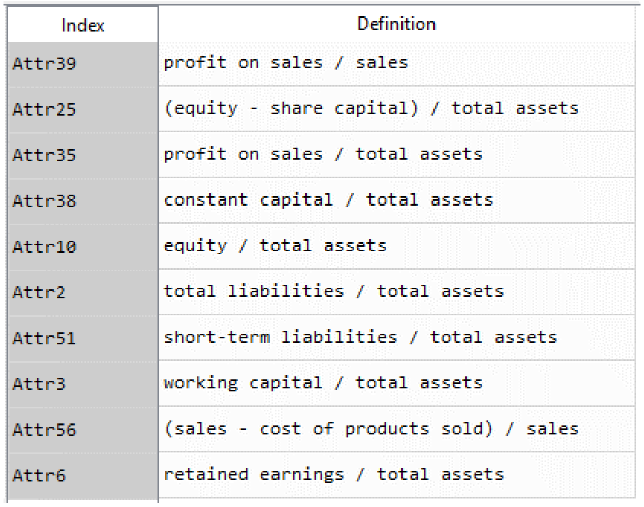


Figure 3. Ten financial attributes with highest correlation magnitude.



Index	Definition
Attr39	profit on sales / sales
Attr25	(equity - share capital) / total assets
Attr35	profit on sales / total assets
Attr38	constant capital / total assets
Attr10	equity / total assets
Attr2	total liabilities / total assets
Attr51	short-term liabilities / total assets
Attr3	working capital / total assets
Attr56	(sales - cost of products sold) / sales
Attr6	retained earnings / total assets

Figure 4. Attribute name of ten financial attributes with highest correlation magnitude.