Chunilall Kissoon, Angela Liang, Li Lin, Daniel Luriye, Carmen Ruan

Andrew Treadway

CIS 9660 - Datamining for Business Analytics

May 13, 2021

**Bankruptcy Prediction**

Bankruptcy is a legal proceeding carried out to allow individuals or businesses freedom from their debts, while simultaneously providing creditors an opportunity for repayment. The COVID-19 pandemic has caused many companies to shut down or file for bankruptcy and as such, we would like to see which factors besides the COVID contribute to bankruptcy. Short term ratios are not good indicators of distress, therefore in this project we will use data mining to identify relationships to assess the financial condition of a company in the long term. This would be useful in the real world in determining which businesses will succeed, and what variables they would need to change to avoid bankruptcy.

The dataset was obtained from Kaggle and uploaded by the InClass Prediction Competition. Sebastian Tomczak created the dataset from the Emerging Markets Information Service (EMIS) based on Polish companies’ bankruptcy data. The data was used to showcase companies that went bankrupt and train the dataset to then predict companies that could go bankrupt based on different financial measures. We used the original "train data", as it actually had an attribute to indicate bankruptcy ("class"). Each observation in the dataset are financial ratios derived from account balances and other financial statement accounts. As you walk through the code and the remainder of the report, please refer to Exhibit 1 found at the end of the report, which is replicated in our attached data dictionary.

Our group chose this bankruptcy dataset because we wanted to predict if a company will go bankrupt, what are the causes of company bankruptcy, and what are some of the factors that will increase the chance of bankruptcy. While the dataset is about Polish companies, we believe the general course of business is relatively similar to that of the United States. The attributes, or financial ratios, are also calculated in a similar way. There may however be slight differences in the calculation due to differences between United States generally accepted accounting principles (US GAAP) and international financial reporting standards (IFRS) in recognition of accounts and transactions.

To begin our analysis and modeling, we read the dataset using pandas read\_csv function. After calling our data, we saw there were 65 columns with 64 being input features and 1 output feature, and 10,000 rows. From there we used a value count method to determine how many bankruptcies there were in the dataset. We discovered there were 203 bankruptcies. By creating a pie chart, we were able to visualize this was only 2% of the companies in our entire dataset. Next, we began cleaning our data to ensure we didn’t run into any issues later on. The first thing we checked was the data types of the columns to ensure there were no strings or objects. All our columns turned out to be float types which work great with the models we decided to use later on. The dataset we chose was preprocessed for a competition by removing missing values and standardizing columns, therefore we did not see any nan values. After cleaning up our data we moved on to the next step of data exploration.

Our first step in data exploration was to create a correlation matrix and heatmap to see how our variables affected each other and to summarize the dataset to locate any potential patterns. This was our starting point for our regression analysis. However, due to the number of variables it was hard to determine which values had a significant impact on each other on the dataset. As a result, we moved another method which was to create histograms for all the columns in the dataset to plot their frequency. What we found is a lot of the variables had the same distribution with a few outliers. Due to the issues, we ran into while trying to explore the data using the correlation matrix, heatmap, and histograms we determined to obtain useful information we would need to reduce the number of variables.

In the next phase, we began engineering some new features based on models we found that could assist in predicting bankruptcy. These included the current ratio, the operating cash flow to sales ratio, the debt-to-equity ratio, cash flow to debt ratio, Edward Altman's Z score, and a few other key metrics. To visualize some of these metrics we created a bar chart to highlight the mean of the sales to total assets ratio. The ratio measures the ability of a business to generate sales on as small a base of assets as possible. The higher the asset turnover ratio, the better the company is performing. In the visual, the average for a bankrupt company is a little over 0.2 but we didn’t put too much emphasis on this financial ratio since we could be looking at a variety of industries where the ratio doesn’t apply. For example, retail and consumer staples have relatively small asset bases but have high sales volume. Additionally, we created a scatterplot utilizing seaborn to outline the differences between the bankrupt and non-bankrupt businesses using the current ratio. The current ratio assesses a company’s short-term liquidity with respect to its current assets and short-term liabilities. A current ratio lower than the industry average may indicate a higher risk of distress or default. In this visual, bankrupt companies did relatively consistently among other bankrupt companies, but also didn’t put too much emphasis on this ratio since weaknesses of the current ratio include the difficulty of comparing the measure across industry groups and overgeneralization of the specific asset and liability balances.

After the implementation of our new features, we began feature selection to determine the most important variables for our model. We examined our correlation matrix again but this time selected values that had a correlation greater than zero. This is due to the fact that any correlation less than zero would actually be preventing a company from going bankrupt. Using a scatterplot, we were able to see that this data would be considered neutral. However, due to the outliers in the dataset the scatterplot was considered positive. We kept these values as they could be the ones that are the causes of bankruptcy. Using these values, we ran a logistic regression model which contained high accuracy and AUC scores but abysmal precision, recall and F1 Score.

As such, we moved on to our next model Random Forest with Hypertuning to see if there was a better combination of parameters, we could use to improve our metric scores. We split our smaller dataset into training and test sets and standardized. Overall, this was helpful in boosting the AUC score in the training set but not in the validation set. Outright eliminating variables didn’t help so we ran a logistic regression with an L2 penalty. Even though we didn’t eliminate any variables, our results using logistic regression actually gave us lower AUC scores while the accuracy remained the same. The other metrics continued to be a challenge as we saw no improvement in precision, recall or F1 score. Given this challenge, we moved on to our next model, extreme gradient boosting. After several runs, we had cases where both the training and validation evaluation metrics returned perfect ones which should be impossible, given how the accuracy should drop off for the validation after a certain time. However, we saw that conducting this model has drastically improved the precision score for the validation set. Recall and F-1 scores have also risen above their usual zero values.

To evaluate our models, we focused on using AUC, accuracy, precision, recall, confusion matrix, and F1 score. Accuracy determines the overall predictability of the model and AUC values the model’s ability to distinguish between positive and negative classes. While our AUC and accuracy was relatively high all the other metrics were relatively low especially on our validation data. This seems to show there was an issue in our model in accurately predicting bankruptcy.

One issue we encountered while running our logistic regression model was when evaluating our results, the F1 score, recall and precision was zero while the accuracy and AUC was relatively high. Another issue we ran into when performing logistic regression with the L1 penalty was it found the majority of attributes to not be important. Additionally, the features we engineered based on models used to predict bankruptcy were also eliminated and determined to not be important. To see if the number of variables were a significant factor, we also tried to run PCA on our model with a 90% variance to keep features that were more consistent. However, this also lowered the accuracy of our data and was something we decided not to further pursue because we wanted our model to be as accurate as possible since this analysis was created to find what features are the most important in predicting bankruptcy.

One of the patterns we noticed in the correlation matrix is that the first ten attributes had positive correlations. These attributes were mainly about current and noncurrent assets. The next set of attributes that were highly correlated were mainly based on sales and profit. At the start of our analysis with the correlation heatmap we observed this pattern was simply because the formulas used the same account or transaction, but this was justified in our models. Additionally, in comparing the models we ran we found that one of the most important features was the operating expenses / total liabilities ratio, which appeared in all our models. Other key features included sales / total assets, (total liabilities - cash) / sales and EBITDA / total assets. If businesses were to focus on making these values positive metrics, they would be less likely to go bankrupt.

Overall, our model required far more tuning for it to be considered viable. We believe that one of the keys would be to analyze the correlation matrix more intently and eliminate variables that are too closely correlated with each other. Furthermore, utilizing PCA might serve to, at the very least, determine the optimal number of variables needed to explain the variance in the data. Additional changes we might need to make in order to move our model into production is to gain more data over a longer time period to determine which variables place the highest correlation in determining bankruptcy. Our analysis would have also benefited from a more recent dataset because it is possible some companies that went bankrupt during COVID already began to fall apart before the pandemic hit. The dataset we currently have may contain a sample size that is too small, leading to a poor performance. To monitor its performance while deployed we would use real information based on the company's annual performance. If companies reported a decrease in employee size or stock price and the key metrics, we found that determined performance was doing poorly, then we would see our model actually works. This would indicate companies should focus on these metrics to improve sustainability in the market.

If we had more time, we would try to implement our model on U.S. data in order to see if there is a systemic difference between two countries and if different features are important for bankruptcies amongst them. Additionally, we would continue to tune our model to see if we can improve our scores on the evaluation metrics by engineering other features that may be more important that are not included in the dataset.

**Exhibit 1**

* attr1 - net profit / total assets
* attr2 - total liabilities / total assets
* attr3 - working capital / total assets
* attr4 - current assets / short-term liabilities
* attr5 - [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] \* 365
* attr6 - retained earnings / total assets
* attr7 - EBIT / total assets
* attr8 - book value of equity / total liabilities
* attr9 - sales / total assets
* attr10 - equity / total assets
* attr11 - (gross profit + extraordinary items + financial expenses) / total assets
* attr12 - gross profit / short-term liabilities
* attr13 - (gross profit + depreciation) / sales
* attr14 - (gross profit + interest) / total assets
* attr15 - (total liabilities \* 365) / (gross profit + depreciation)
* attr16 - (gross profit + depreciation) / total liabilities
* attr17 - total assets / total liabilities
* attr18 - gross profit / total assets
* attr19 - gross profit / sales
* attr20 - (inventory \* 365) / sales
* attr21 - sales (n) / sales (n-1)
* attr22 - profit on operating activities / total assets
* attr23 - net profit / sales
* attr24 - gross profit (in 3 years) / total assets
* attr25 - (equity - share capital) / total assets
* attr26 - (net profit + depreciation) / total liabilities
* attr27 - profit on operating activities / financial expenses
* attr28 - working capital / fixed assets
* attr29 - logarithm of total assets
* attr30 - (total liabilities - cash) / sales
* attr31 - (gross profit + interest) / sales
* attr32 - (current liabilities \* 365) / cost of products sold
* attr33 - operating expenses / short-term liabilities
* attr34 - operating expenses / total liabilities
* attr35 - profit on sales / total assets
* attr36 - total sales / total assets
* attr37 - (current assets - inventories) / long-term liabilities
* attr38 - constant capital / total assets
* attr39 - profit on sales / sales
* attr40 - (current assets - inventory - receivables) / short-term liabilities
* attr41 - total liabilities / ((profit on operating activities + depreciation) \* (12/365))
* attr42 - profit on operating activities / sales
* attr43 - rotation receivables + inventory turnover in days
* attr44 - (receivables \* 365) / sales
* attr45 - net profit / inventory
* attr46 - (current assets - inventory) / short-term liabilities
* attr47 - (inventory \* 365) / cost of products sold
* attr48 - EBITDA (profit on operating activities - depreciation) / total assets
* attr49 - EBITDA (profit on operating activities - depreciation) / sales
* attr50 - current assets / total liabilities
* attr51 - short-term liabilities / total assets
* attr52 - (short-term liabilities \* 365) / cost of products sold)
* attr53 - equity / fixed assets
* attr54 - constant capital / fixed assets
* attr55 - working capital
* attr56 - (sales - cost of products sold) / sales
* attr57 - (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
* attr58 - total costs /total sales
* attr59 - long-term liabilities / equity
* attr60 - sales / inventory
* attr61 - sales / receivables
* attr62 - (short-term liabilities \*365) / sales
* attr63 - sales / short-term liabilities
* attr64 - sales / fixed assets
* class - the response variable Y: 0 = did not bankrupt; 1 = bankrupt