Aditya Shah, C.J. D’Amico, Rahil Virani, Jesmyn Zhang

Gen Bus 656

Team Assignment 1

**Altman Z-Score**: The Altman Z-score is used to calculate the likelihood of bankruptcy in two years for any given company using their financial data with only specific data points. The Kaggle dataset is made up of over 8,000 US companies spanning from 1999 to 2018. The formula to calculate it uses 5 ratios: working capital to total assets, ratio of retained earnings to total assets, ratio of earnings before interest, taxes, depreciation, and amortization (EBITDA) to total assets, ratio of market value of equity to book value of total liabilities, and ratio of sales to total assets. Each of these has a corresponding coefficient and are summed up to get the final Z-score. We transformed this Z-score into a probability using a CDF function and found that the average expected probability of the companies that did go bankrupt within 2 years was 0.54 while the average expected probability of the companies that did not go bankrupt within 2 years was 0.80. This model on average was right 81% of the time, however there was a strong likelihood of type II errors since the model was not good at predicting companies that went bankrupt. Although the model was correct 82% of the time for companies that didn’t go bankrupt within 2 years, it was only correct 44% of the time for the companies that did go bankrupt. This model is unfortunately not accurate enough since in a real-world application, it tends to be more important to find the companies that will go bankrupt and either need assistance or avoidance.

**Altman During Recession:** During the recession, many companies went under even though they may have had stronger financial records. To calculate the accuracy of the Altman Z-score during a recession, we looked at all the companies that were observed in 2008 and 2009 to see if they went bankrupt in that time. For companies that did go bankrupt, Altman was accurate 78% of the time during the recession and 70% of the time outside of the recession. For the companies who didn’t go bankrupt, Altman was correct 42% and 48% during and outside the recession respectively. The data suggests that Altman was more accurate at finding the bankrupting companies during the recession and finding the companies that were safe outside the recession. However, the probabilities are within 8% of each other and don’t provide sufficient evidence that the model is affected by the recession, meaning it does hold over periods of economic instability.

**Constructing Our Model**: To build a model, we used a crude version of backwards stepwise selection, taking all variables and removing the ones we wanted to. The first thing we did to look for these select features was look for collinearity between all features using a correlation heatmap. This helped visualize which variables have high collinearity with every other variable and paired with the Variance Inflation Factor (VIF), we were able to spot the features with high probability of collinearity. After running the data through this filter, we saw 6 potentially problematic features: cost of goods sold, EBITDA, net sales, total revenue, total liabilities, and total operating expenses. We removed these variables from our model as reduce unnecessary noise and kept all other variables given in the dataset. For all our models, we used a rolling time-series cross-validation method where it tests the data on a given year using training data from all years prior. We started our models at 2005 as to give the training at least 5 years of data. This also accounts for the issue of multiple observations in the dataset for each company. Since the test set is based on one year, there is only one observation for each company that is used to predict and any other observations from previous years are only used to train the model. For all our models, we also tried 3 methods of predicting bankruptcy: logistic regression, linear discrimination analysis, and quadratic discrimination analysis. LDA was by far the most accurate out of the three, most likely since much of the collinearity was removed earlier when using the heatmap and VIF.

**Other Models**: We also saw that there are other interactions (ratios commonly used to assess a company’s financial strength) that could mean more in financial analysis. We split these up into two sets: the first – which I will call the short-term model – focusing on ratios that assess short-term liquidity (the ability to pay off short-term financial obligations) and the capital structure of the company, and the second – which I will call the long-term model – focusing on long-term health of the company and a comprehensive analysis of all cash flows for the company. One problem we faced was seeing which one was more important in determining bankruptcy for a given company. All the models we created were very accurate with LDA, however there was much more variation in the logistic regression and QDA. Assessing the accuracy of those two in both models (the two using ratios rather than individual variables with no interactions), logistic regression was far more accurate for the short-term model whereas QDA was far more accurate for the long-term model. This showed us that both models have something different to offer and the ratios for each are somewhat important. We built off of the short and long-term models by meshing them together and putting all the ratios into one model. This model acted as expected, more of a balance for logistic regression and QDA than the short and long-term models.

**Our Models During Recession**: Going purely off of the accuracy of the model placing the recession years in the test set, it performs very similarly to how the combined model worked before. This shows the model does still work during the recession as logistic regression and QDA are similar (excluding 1999-2001) and LDA is still averaging 99% accuracy.

**Limitations and Improvements**: As we discussed in class, there are millions of other models that can be created using interactions between the variables and we wouldn’t know for sure that the model we created is the best possible one with the data we have until we look at the accuracy of all other possibilities. There are also other variables we could use that can help predict bankruptcy in companies. Macroeconomic data such as GDP and inflation can be good predictors of the general economy of the US. When these factors are off of ideal numbers, they can mean the US is in a bad economy and can greatly improve the long-term model especially during the major 2008 recession as well as other smaller recessions the US has faced. Market data such as stock price deltas can also be solid features to assess financial stability in the short-run which would improve the short-term model.