Project Midterm Report - ORIE4741
Company Bankruptcy Prediction

Name	NetID
Christopher Hales	cmh342
Mathias de Bang	md794

11.01.2021

Predicting Company Bankruptcy

Introduction

Our project seeks to predict company bankruptcy based on data collected on bankruptcies from the Taiwan Economic Journal between 1999 and 2009.

Our objective is to answer whether it is possible to accurately predict whether a company will go bankrupt given the financial information that we have about all of the companies in the data set. Bankruptcy prediction is crucial in the financial industry for individuals and financial institutions make better assessments when it comes to credit and loans. Generally, the costs of these decisions involve a lot of resources for the institutions, and the large bankruptcies could have an affect on the rest of the market and economy as a whole. Therefore, the value of being able to accurately and efficiently predict which companies will end up bankrupt can not be overstated.

Fundamentally, these decisions have traditionally relied entirely on financial theory, but the increase in the amount of available data makes the application of machine learning enticing. The data set we have selected can make it possible to understand whether patterns in the financial data would enable us to predict which companies are at risk of bankruptcy (which will also be of interest to the company itself or other potential stakeholders).

Some specific questions of interest involve how we balance the ratios of false positive and negatives in our predictions to achieve the best results and most valuable predictions. We are mainly interested in identifying high-risk companies, which means we must consider the value of identifying a company with a high bankruptcy risk compared to falsely categorising a company as being at the risk of bankruptcy (false positives). We also hope to gain insights into the most valuable predictors of bankruptcy. This could add to the wider understanding of financial statement analysis.

Data and Features

The data set that we are working with includes 6819 observations and 95 features about the financial and managerial accounting information for companies in Taiwan, including features such as the debt ratio percentage (liabilities / total assets), accounts receivable turnover, cash flow to sales, the ratio of gross profit to sales, and others, and of course if the company ended in bankruptcy or not. In all, it contains the quantifiable data about all of the financial information for each company recorded. The data was collected by the Taiwan Economic Journal for a decade: from 1999 - 2009. A description of the data shows that about 3.22% of companies went bankrupt during this time period.

Our data set does not contain any Null values, but there are two columns that do not accurately represent our data: 'Liability-Assets Flag' and 'Net Income Flag'. The 'Liability-Assets Flag' column has the Boolean value 0 for all but 8 of the rows, whereas the 'Net Income Flag' column has the Boolean value 1 for all entries.

Our next plan was to create a correlation matrix of all of our data. We did so over all of features in the data, and it was subsequently too small and convoluted to reasonably understand and interpret. As a result, we created a Series of all columns and their respective correlation to the Boolean column 'Bankrupt?' (returning 1 if bankrupt, 0 else), and then returned the ten features with the highest correlation coefficients.

These columns are as follows: Debt ratio % (0.2502), Current Liability to Assets (0.194), Borrowing dependency (0.1765), Current Liability to Current Assets (0.1713), Liability to Equity (0.1668), Current Liabilities/Equity (0.154), Current Liability to Equity (0.1538), Liability-Assets Flag (0.1392), Total expense/Assets (0.139), Equity to Long-term Liability (0.139). Noticing identical correlation values, further exploration of the columns shows that the columns Current Liabilities/Equity and Current Liability to Equity are identical columns, so we can ignore one of them. Thus, the top variables that we will be looking at are 'Debt ratio %', 'Current Liability to Assets', 'Borrowing dependency', 'Current Liability to Current Assets', 'Liability to Equity', 'Current Liabilities/Equity', 'Liability-Assets Flag', 'Total expense/Assets', 'Equity to Long-term Liability'.

Below are three different histograms representing the frequencies of different bins of the data for specific columns: namely, 'Debt ratio %', 'Current Liabilities/Equity', and 'Borrowing dependency'.

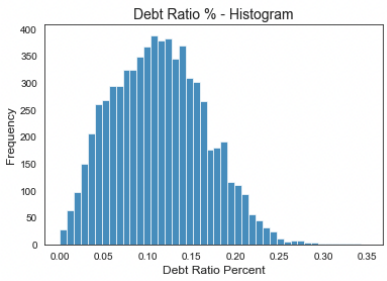


Figure 1: Debt Ratio %

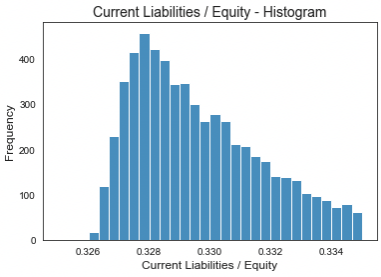


Figure 2: Current Liabilities/Equity

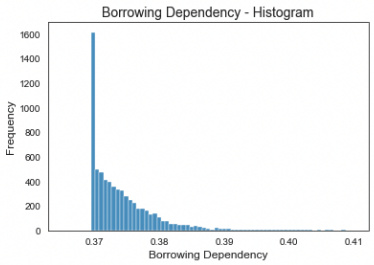


Figure 3: Borrowing Dependency

Model selection and validation

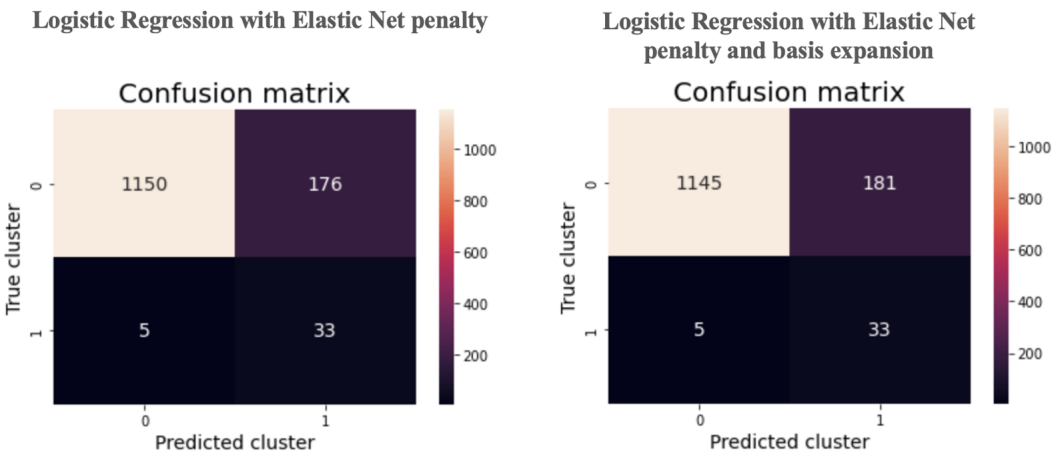
Avoiding over- and underfitting
In order to avoid both over- and underfitting, we need to select the right model, meaning the one that generalizes best to new data. This usually includes testing different models and choosing the optimal values for the model parameters such as a regularisation value in a penalized regression or number of trees in a tree ensemble model. The goal is to build a model that balances bias and variance in an optimal way where we control the complexity of the model while capturing the patterns in the data. One way to do so is by using a cross-validation, which allows us to test different models and use all the data. The data is split in training and test sets. The different models we wish to test are then fitted to the training sets and tested on the last part to get an independent performance estimate. For each split we train and test a model for the grid of parameters we wish to optimize over. The advantage of using a cross-validation compared to a simple hold-out test set, is that we are able to use all the data to both test and train our models, while avoiding biasing the performance estimates. After choosing a model, we can get an estimate of the error by testing the model on a completely separate validation set. Another way to do this is by using a nested cross-validation which performs both model selection and validation. Depending on the computational power needed, we might try this option.

Performance Metrics
An important part of selecting and validation a classification model as the one in our case, is to decide one or more performance metrics. As mentioned in the last section, our data set is imbalanced and only 3.2% of the outcome variable belong to the positive class, meaning only 3.2% of the companies in the data set went bankrupt. This makes Accuracy a very poor performance metric as a very high score could be obtained with a very poor model. Instead, we will look into using an F1-measure, which takes the Precision and Recall into account. A good and unbiased measure is also just to look at a confusion matrix, which is easily made in our 2-class example.

Preliminary Modelling

In order to get more information on important features and a benchmark for further modelling, a couple of supervised classification models were tested on the data. Based on the exploratory data analysis, we saw that several features had a correlation with the response variable close to zero. Several of the features also had a significant intercorrelation. It was therefore decided to fit a Logistic Regression model with an elastic net penalty, which includes both an L1 and L2 penalty. Ideally, this gives some sparsity in the features while shrinking highly correlated features. Besides the penalty, we used balanced class weights to account for the imbalance in the data. The “balanced” mode uses values of the output variable to adjust the weights so they are inversely proportional to size of the class prevalence (gives higher weight on the underrepresented class).

As another test, we tried the model with a polynomial basis expansion of our features so they included all interactions and second order effects. The regularization of the model is expected to avoid over-fitting with the additional features. A 5-layer cross-validation was used to produce he following results:



As shown in the figure above, the model does do an alright job of identifying the companies that went bankrupt. Out of 38 companies in the test data, 33 are identified correctly. However it also comes at the cost of having many false positives, so the overall performance is not great. We also note that it actually does not make a big difference to include polynomial features or interaction effects.

Next Steps

The next steps in the project is to continue with the modelling. Results show that there is a potential for creating a valuable model, but also that there is room for improvement. Specifically, we want to reduce the amount of false positives we see here in the first model. New different kinds of models should be tried, which could be classification trees (incl. boosting or random forest models) or Support Vector Machines. More work should also be put into doing a more thorough search for optimal model parameters to select the best model.