Project Proposal

Amar Adusumilli, Kardelen Cicek, Tarush Gupta

2021-04-12

What problem did you select and why did you select it?

For our project, we are using a Kaggle dataset focused on predicting company bankruptcy in Taiwan., spanning from 1999-2009. We will use the methods and models learned during this course to predict bankruptcies, identify the most important determinants of bankruptcy in Taiwan, and to infer the partial effect of these most important determinants towards the probability of bankruptcy. The paper "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study (2016)" published in the European Journal of Operational Research summarizes bankruptcy prediction analysis by using data from Taiwan. Firm bankruptcy hurts both firms and economies. Therefore, understanding and modeling firms' bankruptcy is essential to prevent losses and decline likelihood of default.

The paper (Liang et al. 2016) investigated the power of corporate governance Indicators (CGI) combined with financial ratios (FR) for bankruptcy prediction. A combination of CGI and FR offers the best bankruptcy prediction for better lending opportunities. The dataset allows the exploration of the factors which determine bankruptcy and provides an opportunity to implement the methods discussed in Liang et al. 2016 with real world data.

What database/dataset will you use? Does it need to be cleaned?

We will be using the Kaggle dataset that can be found at the following link (https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction). The dataset contains 6819 observations of 96 variables describing the financial health of numerous corporations in Taiwan. The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. In this data, we have a binary Bankruptcy class variable (1, if the firm went bankrupt; 0, otherwise), and 95 financial predictors. While the data is well-preserved and -organized – i.e., does not require extensive cleaning – we will be preprocessing the data to fix the class imbalance problem of our target variable. We will also need to standardize or normalize the data before we may begin our analyses.

What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?

Roughly speaking, we intend to approach this dataset from two contrasting angles - the data mining/machine learning perspective, and the econometric perspective. With the former, our goal is to accurately predict bankruptcies. We intend to use the algorithms taught to us in class, such as decision tree classification, random forest, and potentially others to best predict bankruptcies. We may employ cross validation as a tool to aid in model selection. Tree based methods will be particularly useful for their in-built measures of feature importance.

With the econometric perspective, our goal is to infer the partial effect of the relevant dependent variables on the probability of bankruptcy. For example, we may ask, 'by how much does a 5% increase in leverage increase the probability of bankruptcy?'. We intend to use a set of limited dependent variable models since our dependent variable is binary. These include the linear probability model (LPM), probit regression, and logistic regression. Using these models, we will estimate the partial effects of the relevant variables within our dataset.

The primary challenge of this dataset is in feature reduction. It is likely that using all variables will result in severe multicollinearity problems in the inferential models, complicating analysis and prediction. We suspect that many of the independent variables are redundant and will add little information at the cost of degrees of freedom. Therefore, exploratory analysis of the features/independent variables is critical. We may also use principal component analysis to reduce the dimensionality of this data.

What packages will you use to implement the network? Why?

We intend to use Numpy, SciPy, Matplotlib, Seaborn Sklearn, Pandas, and Statsmodels. Numpy and Pandas are required for data manipulation. Matplotlib and Seaborn will aid in exploratory analysis. Sklearn and Statsmodels will be used for modeling.

What reference materials will you use to obtain sufficient background on applying the chosen network to the specific problem that you selected?

- Liang, Deron, Lu Chia- Chi Tsai, Chia-Chi Shih, Guan-An (2016) "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study" European Journal of Operational Research, 252(2).
- Kaggle.com/ Company Bankruptcy Prediction

How will you judge the performance of your results? What metrics will you use?

For the predictive models, performance will be judged on accuracy. As only 3% of companies in Taiwan actually went bankrupt in the time interval of the dataset (220/6819), simply predicting '0' for all observations produces an accuracy greater than 96%. Therefore, accuracy in excess of 96% is the goal for prediction. For the econometric models, performance will be judged on whether the assumptions for inference are satisfied, eg, Gauss-Markov assumptions for the LPM. If this is the case for all three econometric models, then the preferred model will be selected on the basis of predictive accuracy.

Provide a rough schedule for completing the project.

For our project, we have distributed the workload into the following components along with our anticipated deadlines to complete them:

- Exploring the data (week of April 12)
- Running regression models (week of April 12)
- Running other statistical models (week of April 19)
- Writing the reports (Week of April 26)