Bankruptcy Prediction with Statistical Learning

MA678 Midterm Project

Zixuan Mary Liu

**Personal Statement**

Prediction of firm bankruptcies have been extensively studied in the field of accounting to monitor the financial performance by all shareholders. Especially I want to start my own business after graduate. I want to find out the change that I will success to build my own company. The aim of this project is to examine the relationships between these parameters and develop an effective prediction model which allows forecasting the bankruptcy condition of a firm in the near future.

**Question**

In the project, I apply and compare some widely known statistical imputation techniques, such as Decision Tree, K-nearest Neighbor, Logistic Regression and Linear Regression and evaluate the performance of these techniques by their accuracy rates.

**Data Source and Description**

The dataset I use is called `Polish Companies Bankruptcy Data Set` which is hosted by UCI Machine Learning Repository and collected from EMIS, a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013. In this project, I will use partial data called `3year` for bankruptcy prediction. It contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years.

The data `3year` contain contains 64 variables and 10503 observations in total. The dependent variable is the class variables with levels 0 or 1, indicating the company bankruptcy or not. Some variables as financial ratio could affect the company be classified as bankruptcy or not. For example, the first variable is "net profit/total assets" which is return on assets (ROA), a financial ratio that shows the percentage of profit a company earns in relation to its overall resources. It is possible that the higher the ROA, the less likely the company will be bankrupt.

**Proposed Timeline of Work**

1. EDA: Nov 12th
2. Data processing: Nov 19th
3. Modeling and Validation: Nov 26th
4. Write up: Dec 3rd