Capstone Project Summary

Predicting Potential Business Failure using Machine Learning

Abigail Chen

June 2022

Contents

{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

Predicting Potential Business Failure using Machine Learning Capstone Project Summary

CEU Business Analytics MSc. By : Abigail Chen

1 Problem Definition

Macroeconomic situation will be difficult in the next years and can potentially lead to an increase in bankruptcy and financial distress (Business Failure). This topic is important for shareholders, advisors, investors, and banks . With Covid-19 since 2019, and now with the war between Russia and Ukraine many companies are already in distress and can potentially be more in distress in the following years. The negative effects are expected to continue until there is a clear end to the pandemic. The client is looking to expand its advisory services for companies who can experience these negative financial distress. The advisory services will be for Debt advisory, Lending advisory, and Restructuring advising services for the companies in the manufacturing, automotive, FMCG, and construction industries in Hungary. These new services will allow these companies to plan and negotiate its outstanding loans and debts, which can help them avoid default, bankruptcy, and at the same time take advantage of the better rates that will be given.

Many companies are financed via debt. Due to the macroeconomic changes and new developments, financial restructuring is usually initiated to reorganize a company's assets and liabilities. This will help produce a more beneficial environment for the company. At the same time, this also limits the potential financial harm that a company can undergo. Companies would usually approach the bank to change their payment methods, and this leads to negotiations of loan pay back for the companies, this is where client can step in.

2 Client Introduction

Right now the client is planning to expand their debt advisory services , and restructuring services. The approach using prediction modelling will allow the client to do a proactive approach to the companies at risk. This can open up doors for new clients and lead to new engagements

With the prediction modelling (Business Failure Prediction), it will be a bottom-up approach where the companies who can potential be in distress will be flagged by the model and this is where the business development can step in.

3 Project Workflow

3.1 Extract, Transform, Load (ETL) There were 3 datasets extracted via SQL from the client's database. All the datasets are in Hungarian so it needed to be translated. The first database contained the industry codes and the industry was selected via TEAOR codes which is similar to NACE European economic activity codes. The other database had 100+ variables of financial information. These two datasets were later merged and loaded.

3.1 Adding Financial Ratios

Different financial ratios were added to improve the analysis. These are the financial ratios included: Receivable Turnover, Inventory turnover, Current ratio, Working Capital, Fixed asset turnover, Return on equity, Return on assets, Net profit margin on sales, Debt to total assets, Debt to equity, Financial leverage, and Times Interest Earned (TIE).

3.2 Label Engineering

This is to done to add a classification variable, "0" for non-fail companies and "1" for potential fail companies. To classify the observations, the year on year change of the companies' revenue was computed over the years, and the existence of the companies over the years was also checked.

3.3 Feature Engineering

Removing the unnecessary data was done, by calculating the amount of NA's in the various columns and removing them.

3.4 Data Visualization

Data visualization was done using a data visualization software, to do quick data exploration. The different financial statistics were visualized to get a quick understanding of the data.

3.5 Variable Selection

The clean data was divided per industry. Two industries were used. A t-test was done for variable selection. The p-value was used. If the p-value is less than a certain value then this variable will be selected for the modelling.

3.6 Modelling & Machine Learning

A train-test split with 80-20 ratio is done for the various models. Different models were used: Regression, Decision Tree, Random Forest, K Nearest Neighbors (KNN), Artificial Neural Network (ANN), Linear

Discriminants Analysis (LDA), Quadratic Discriminant Analysis (QDA), SVM with Kernel Sigmoid, SVM with Kernel Radial, SVM with Kernel Polynomial and SVM with Kernel Linear.

4 Project Results

The balanced accuracy was used to rank the models. Sensitivity is also known as the "true positive rate". While specificity is "true negative rate".

$$BalancedAccuracy = (((TP/(TP + FN) + (TN/(TN + FP)))/2$$

$$BalancedAccuracy = (Sensitivity + Specificity)/2$$

The results show that the Artificial Neural Network (ANN) was the better performer for both industries. For the first industry, the results showed with 0.5587 on train and 0.5423 on test. For the second one, the results showed 0.5702 on train and 0.5552 on test.

The accuracy in terms of percentage of comparison with actual values to the predicted was quite high. The balanced accuracy was low for all the models which means that the models were predicting one class with high accuracy but the other one with low accuracy. Here are some possible reasons. The variables with p value is less than 0.1. Usually, this is at p $<$ 0.05. The p value was set at 0.1 because the number of variables selected at 0.05 was very low. The classification was not balanced there were more non failed companies compared to the failed companies hence the low balanced accuracy.