Predicting Company Bankruptcy

Springboard DSC - Capstone 3 Proposal

Grace Tang

March 2021

Business Problem:

Predicting company bankruptcy is critical in any financial institution, especially those involved in lending. For such companies, being able to predict whether a business will succeed or fail can result in successful business loans that grow companies, create jobs, and bolster the economy; or result in millions of dollars in losses.

I am interested in developing a binary classification model to predict whether a business is at risk for bankruptcy, what factors contribute to its risk for bankruptcy, and by how much do those factors play a role. I will use a dataset from the Taiwan Economic Journal.

Taiwan's economy is a developed capitalist economy with most government firms being privatized. In that respect, there are many similarities between the US and Taiwanese economies. Furthermore, despite Taiwan's population ranking 57th largest (equivalent to 0.31% of the total world population), it is the 7th-largest in Asia and 20th-largest in the world by purchasing power parity. Taiwan is also the most technologically advanced computer microchip maker in the world. It is definitely an economy worth studying, and may have many insights that carry over to our US economy.

More broadly, the results of this project can be applied to not only finance, but any kind of classification problem with imbalanced data.

Data:

"Company Bankruptcy Prediction: Bankruptcy data from the Taiwan Economic Journal for the years 1999–2009"

- https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction
- The data is historical data from 1999 to 2009, collected from the Taiwan Economic Journal.
- Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.
- The data is imbalanced, with ~3.2% bankruptcies and ~96.8% non-bankruptcies.

Anticipated Data Science Approach:

Predicting company bankruptcy is a binary classification problem: either a company will go bankrupt or it will not. We can use various algorithms that are suitable for binary classification:

- Logistic Regression
- k-Nearest Neighbors
- Decision Trees (Random Forest, XGBoost)
- Support Vector Machine
- Naive Bayes

Since the data is highly imbalanced we will use methods to upsample the minority class (bankruptcies), and downsample the majority class (non-bankruptcies). Possible methods we can use are:

- Synthetic Minority Oversampling TEchnique (SMOTE)
- Borderline-SMOTE
- Adaptive Synthetic Sampling (ADASYN)
- Safe-Level-SMOTE
- Majority Weighted Minority Oversampling TEchnique (MWMOTE)

Not all of the listed methods will be used, and some hybrid methods may be implemented. The models will be evaluated on metrics such as accuracy, precision, recall, and ROC-AUC. Precision and Recall are especially important as our data is highly imbalanced.

Deliverables:

- All code I will develop (python code in the form of Jupyter notebooks).
- A written final report.
- A presentation slide deck.
- Tableau dashboard, showing the results of the modeling and any informative trends or patterns found in the data.