DS 5110 - Project Proposal - Spring'22 - Group 8

Taiwan Company Bankruptcy Prediction

Farhan Chughtai | Sai Vineeth Kaza | Nithya Balachandiran | Pratyusha Parashar

Summary:
Bankruptcy can be considered as a curse for the organization and the investors. It is expressed as the inability of a company to pay its debts to its creditors. The bankruptcy of a company and even the possibility of going bankrupt is important for the company's investors and society. Therefore, bankruptcy prediction is a crucial step for each organization before the company goes bankrupt and appropriate models can be built for the development of the organization.

Effective bankruptcy prediction is crucial for the companies to make appropriate business decisions. In general, the input variables (or features), such as financial ratios, and prediction techniques, such as statistical and machine learning techniques, are the two most important factors affecting the prediction performance. While many related works have proposed novel prediction techniques, very few have analyzed the discriminatory power of the features related to bankruptcy prediction.

Goals:

1. Predict bankruptcy effectively and find best performing model.
2. Find key features that lead to bankruptcy.
3. Visualize interesting patterns in the data.

The dataset is about the company's financial data from the Taiwan economic journal for the years 1999 to 2009, which has listed the details of company bankruptcy based on the business regulations of the Taiwan Stock Exchange. It has over 900 listed companies. The dataset has 6819 rows and 93 numerical variables, 2 categorial variables and the target variable is Bankrupt?

Proposed Plan:

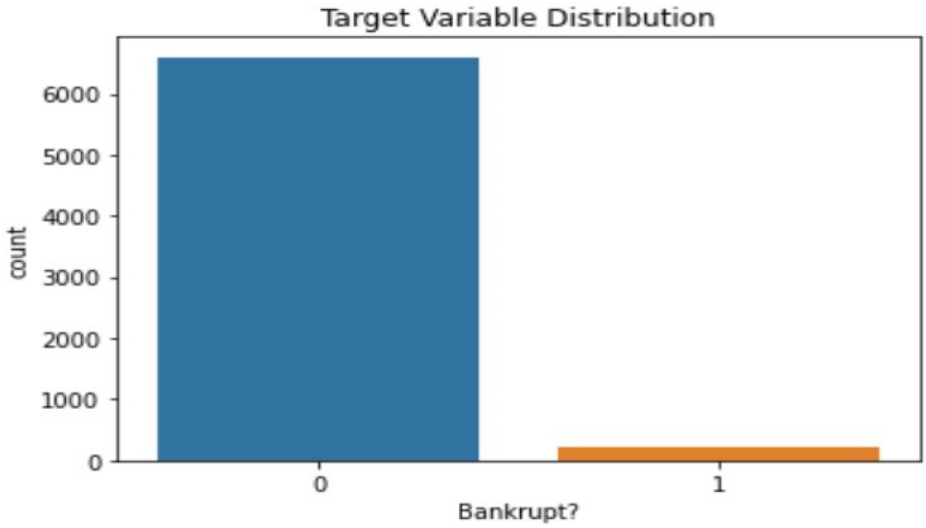
We plan to implement this project in python. After initial analysis, we found that the dataset has no null or empty values. So, we do not need to perform any data cleaning. We plan to do an exploratory data analysis (EDA) on the dataset and try finding some patterns in the dataset, check for outliers, identify any potential clusters etc. As the target label is imbalanced, we plan to employ techniques like SMOTE, oversampling, under sampling, etc.

After EDA, we will use models like Logistic Regression, Decision Trees, Random forests, Support Vector Machines, KNNs and XGBoost to predict whether the company will go bankrupt or not. We will try to look for important features in the dataset, create new features using feature engineering and remove any redundant or useless features. We will also return the most prominent features of the model by performing feature importance.

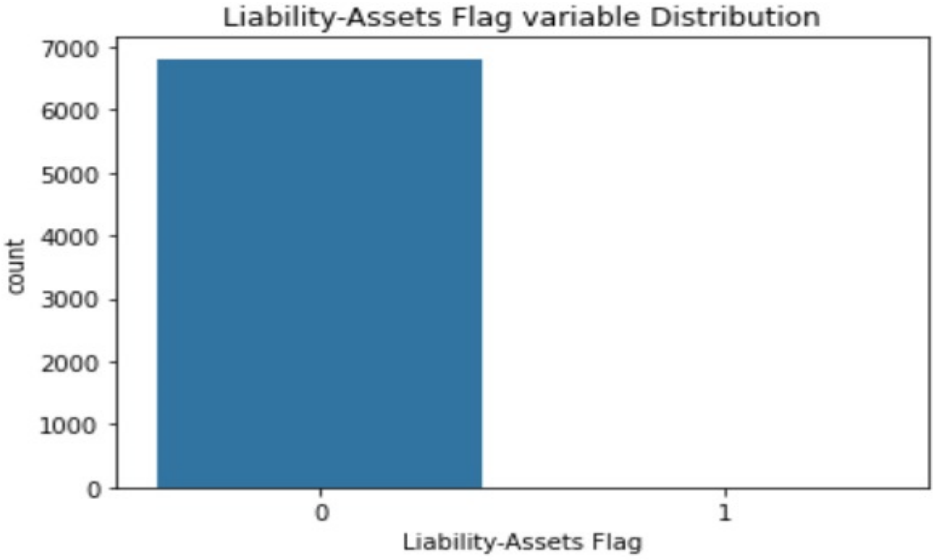
Preliminary Results:

As mentioned above in the proposed plan, there are no null or empty values present in the dataset. We performed preliminary analysis to understand the data better and obtained the following graphs.

DS 5110 - Project Proposal - Spring'22 - Group 8



From the above figure, the target variable is highly imbalanced as there are way more 0s than 1s for this variable. Thus, it is necessary to consider balancing the dataset through "Up-sampling or Down-sampling" techniques.



The "Liability-Assets" flag denotes the status of an organization, where if the total liability exceeds total assets, the flagged value will be 1, else the value is 0. From the above figure, we can observe that majority number of times, organizations/company's assets are more than their liabilities.

References:

1. https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

2. Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572. https://www.sciencedirect.com/science/article/pii/S0377221716000412