Predicting Corporate Bankruptcy

Gilles Delpy SOP Barthélémy BENGLOAN

IF ESILV October

Introduction

In today's financial landscape, predicting corporate defaults holds critical importance for financial stakeholders, particularly investment funds and credit insurers. Identifying companies likely to default in advance enables informed decision-making, reduces financial risks, and optimizes portfolios. This project aims to leverage historical financial data to build a predictive model capable of detecting early signals of default, thereby assisting in proactive risk management.

Business Challenge

The business challenge we aim to address with this project is to develop a system capable of identifying companies with a high risk of default to minimize potential losses for financial stakeholders. Such a predictive model would be invaluable for investment funds and credit insurers by alerting them to early financial risk signals from companies before they enter distress. By anticipating potential defaults, stakeholders can adjust their investment and portfolio management strategies while optimizing their exposure to risks.

Dataset Description

The dataset for this project, sourced from Kaggle, contains financial data from American companies listed on the NYSE and NASDAQ. It includes accounting records from 8,262 companies over the period 1999 to 2018, totaling 78,682 firm-year observations.

Companies are labeled as bankrupt (**1**) if they filed for bankruptcy under *Chapter 11* or *Chapter 7* of the U.S. Bankruptcy Code, or as operating normally (**0**) if no bankruptcy occurred. The dataset is divided into three subsets for model training and evaluation:

- **Training Set** : 60% of the total dataset, selected randomly. This set is intended for training the model, allowing it to learn from a diverse range of economic conditions over time, thus enhancing its predictive generalization.)
- **Validation Set** : 20% of the total dataset, randomly chosen from the remaining data. This set will be used to validate and fine-tune the model parameters through cross-validation techniques, reducing the risk of overfitting and improving model robustness.

– **Test Set** : 20% of the total dataset, consisting of randomly selected samples among recent observations. This set is exclusively reserved to assess the model's performance on new, contemporary cases, providing a realistic estimate of its effectiveness under current economic conditions.

Key financial variables available in the dataset include current assets, cost of goods sold, EBITDA, inventory, net income, total receivables, market value, net sales, total assets, long-term debt, gross profit, total revenue, and total liabilities. These accounting indicators will serve as features for training the predictive model.

Project Objectives

The primary objectives of this project are as follows:

1. Develop a machine learning model to predict default risk based on companies' financial variables.
2. Identify the most influential variables that precede bankruptcy cases to understand financial risk signals.
3. Provide investment funds and credit insurers with an early detection tool to minimize financial risks associated with investment portfolios.
4. Test various machine learning algorithms, such as logistic regression, decision trees, random forests, and K-Nearest Neighbors, to determine which performs best in this context.
5. Further the analysis by integrating reinforcement learning techniques to adjust the model and adapt the default detection strategy in real-time.

Project Organization and Workflow

To ensure effective collaboration, we will follow a structured workflow combining independent work and weekly team meetings. Our project repository on GitHub will centralize code, documentation, and updates for seamless synchronization.

Each member will work autonomously on assigned tasks, with weekly meetings to review progress, address challenges, and set goals for the following week.

The next deliverable, due on **November 22, 2024**, will include an analysis of the initial model's results, feature refinement, and testing of additional algorithms to enhance prediction accuracy.