Machine Learning Project: Corporate Bankruptcy Prediction

Authors: Matthieu Hanna Gerguis, Renaud de l'Epine, Ilian Segoin
Class: IF3, ESILV
Date: November 2025

I. Introduction

Predicting corporate bankruptcy is a major challenge in finance because undetected failures can lead to heavy losses for lenders, investors and institutions. Our objective in this project is to build a machine learning pipeline capable of identifying companies at risk using historical financial indicators while prioritizing the early detection of distress.

The dataset presents a significant difficulty: the *bankrupt* class represents only about 3% of all observations, creating a strong class imbalance and making recall the most critical performance metric.

This pre-project presents the context, data source, methodology and expected outcomes of our work.

II. Business Context

Financial institutions rely on early warning systems to anticipate financial distress. Missing a bankruptcy event can result in unpaid liabilities, cascading failures across the supply chain and inaccurate risk assessments. Machine learning models offer the ability to analyze large volumes of financial data and uncover patterns that often remain invisible using traditional methods.

Our project seeks to develop such a system by training predictive models capable of identifying the warning signs that precede corporate failure.

III. Dataset Description

The dataset originates from the Taiwan Economic Journal (TEJ) and is publicly available on Kaggle under the name *Company Bankruptcy Prediction*, accessible at:
`https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction.` .
It contains financial indicators recorded between 1999 and 2009 and includes:

- 6,819 companies,
- 95 numerical financial features,
- 1 binary target variable "Bankrupt?"

All features are numerical and no missing values are present, which simplifies preprocessing. However, the dataset remains challenging due to its high dimensionality, correlated variables and extreme class imbalance.

These characteristics justify the use of variance filtering, correlation analysis, dimensionality reduction and resampling methods.

IV. Project Objectives

The main goal is to build a reliable and interpretable bankruptcy prediction model. To achieve this, the project includes:

- exploring and analyzing the dataset to understand its structure,
- cleaning and reducing dimensionality using variance and correlation filtering,
- standardizing all financial features and applying PCA to stabilize the learning process,
- training baseline models such as Logistic Regression, Naive Bayes and Random Forest,
- addressing class imbalance with techniques including Random Oversampling and SMOTE,
- evaluating models focusing primarily on recall for the minority class,
- applying hyperparameter tuning to improve the performance of selected models.

The final model should provide interpretable insights into the financial indicators most associated with bankruptcy.

V. Methodology Overview

A. Exploratory Analysis

We analyze variance, distributions, zeros and correlations to identify irrelevant or redundant features and prepare the dataset for modeling.

B. Preprocessing

All variables are standardized and PCA is applied to retain 95% of the variance while significantly reducing dimensionality.

C. Modeling

Three baseline models are evaluated both with and without PCA:

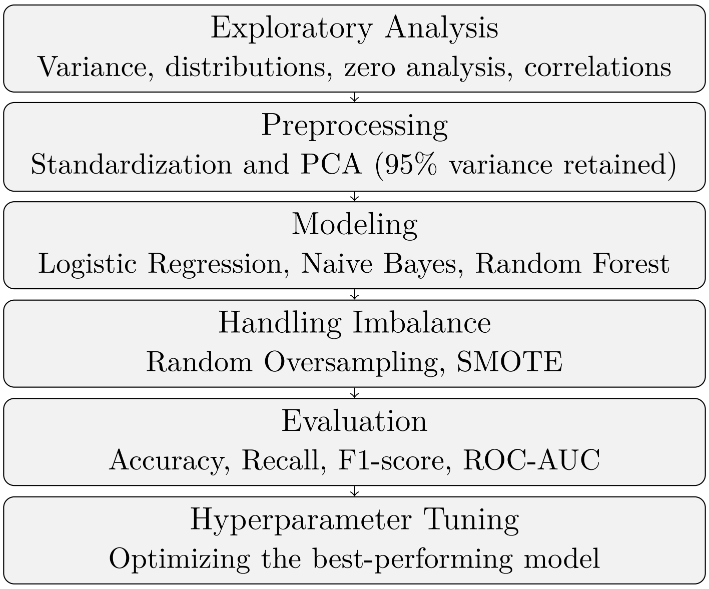
- Logistic Regression,
- Naive Bayes,
- Random Forest.

D. Handling Imbalance

Techniques such as Random Oversampling and SMOTE are applied to enhance minority-class recall and balance the training process.

E. Evaluation

Models are evaluated using accuracy, recall, precision, F1-score and ROC-AUC, with particular attention to the detection of bankrupt companies.


Exploratory Analysis
Variance, distributions, zero analysis, correlations
Preprocessing
Standardization and PCA (95% variance retained)
Modeling
Logistic Regression, Naive Bayes, Random Forest
Handling Imbalance
Random Oversampling, SMOTE
Evaluation
Accuracy, Recall, F1-score, ROC-AUC
Hyperparameter Tuning
Optimizing the best-performing model

VI. Expected Contribution

This project aims to deliver:

- a complete, reproducible end-to-end machine learning pipeline,
- insights into which financial indicators best predict bankruptcy,
- an analysis of how preprocessing and resampling techniques impact performance in imbalanced datasets.

The project will emphasize both the predictive ability and interpretability of classical machine learning models on real-world financial data.

VII. Conclusion

This pre-project summarizes the motivations, data and methodology behind our approach to bankruptcy prediction. The full project will evaluate the limitations of baseline models, measure the impact of resampling techniques on recall and investigate more advanced methods such as XGBoost and calibrated probabilistic models.

Our objective is to build a robust and interpretable predictive system capable of supporting financial decision-making, while clearly outlining its limitations and potential future extensions.