Bankruptcy Prediction

Felicia Angjaya

Data Understanding & Goals

Dataset from Kaggle

Company Bankruptcy Prediction

Bankruptcy data from the Taiwan Economic Journal for the years 1999–2009

kaggle.com

6819

Companies

95

Financial Indicators

Target Variable : **0 = Not Bankrupt** **1 = Bankrupt**

Data Characteristics

All numerical attributes (financial ratios)

Imbalanced Dataset (Majority of companies are Not Bankrupt, Minority of companies are Bankrupt)

Problem Statement

Many companies fail unexpectedly due to lack of early warning signs. There is a need for a predictive model to detect financial distress based solely on internal financial ratios

Objective

To develop and evaluate a supervised machine learning model that can accurately classify companies as "bankrupt" or "not bankrupt" based on their financial ratios, with a focus on maximizing recall and precision on the minority class (bankrupt)

Goals

To build a binary classification machine learning model to predict whether a company is at risk of bankruptcy based on its financial ratio data

Supervised Machine Learning – Binary Classification

Data Cleaning and Manipulation

Data Manipulation

Technique : StandardScaler from scikit-learn

Purpose

To normalize the scale of all numerical features (financial ratios).

BEFORE › Highly varied feature scales

AFTER › Standardized (Mean ≈ 0, Std ≈ 1)

Data Cleaning

0 Duplicates

0 Missing Values

```
Missing value tiap kolom:
Bankrupt                                                    0
ROA(C) before interest and depreciation before interest     0
ROA(A) before interest and % after tax                      0
ROA(B) before interest and depreciation after tax           0
Operating Gross Margin                                      0
                                                           ..
Liability to Equity                                         0
Degree of Financial Leverage (DFL)                          0
Interest Coverage Ratio (Interest expense to EBIT)          0
Net Income Flag                                             0
Equity to Liability                                         0
Length: 96, dtype: int64
```

Exploratory Data Analysis

Correlation

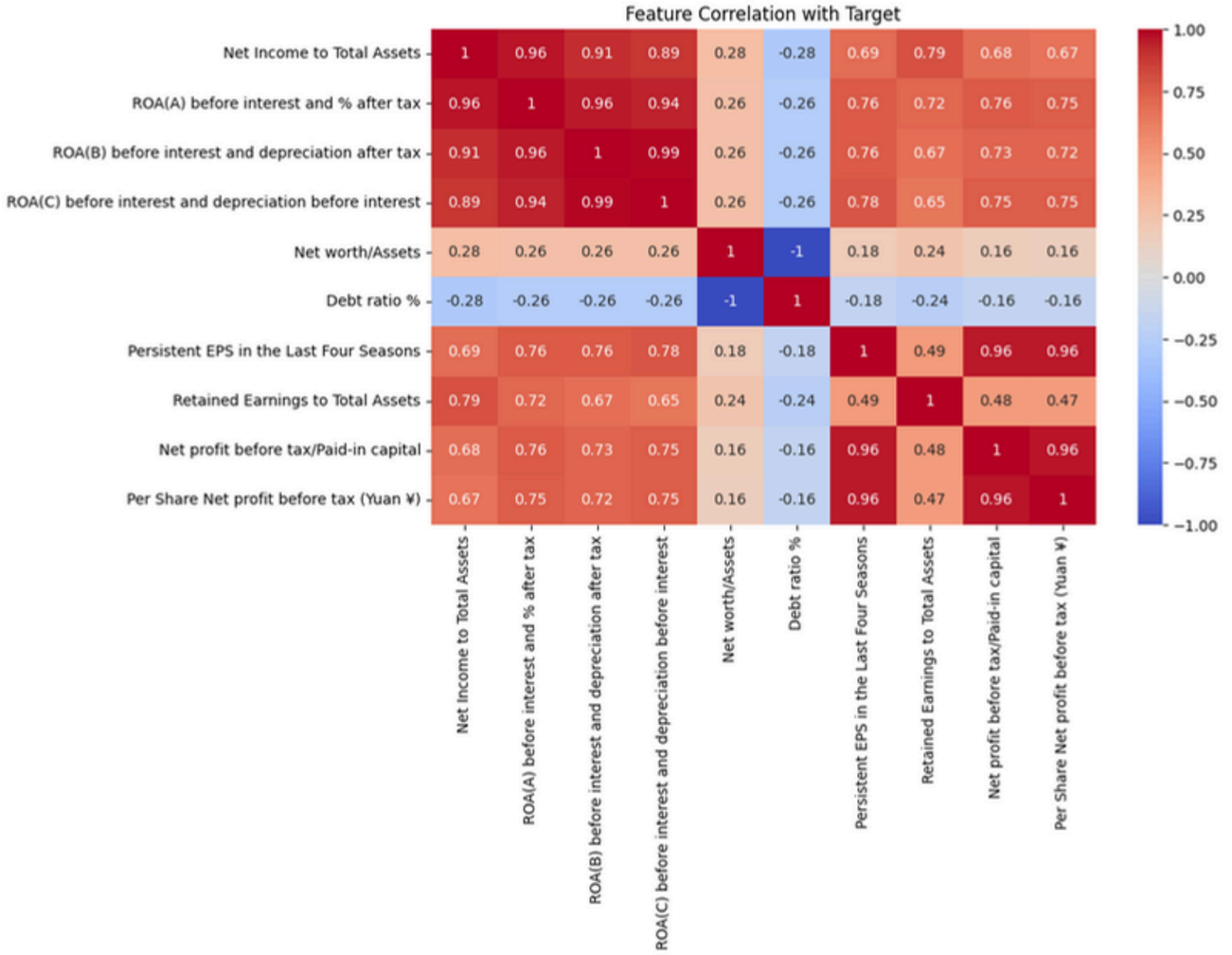
Negative

Net Income to Total Assets — The higher the net income relative to assets, the lower the bankruptcy risk

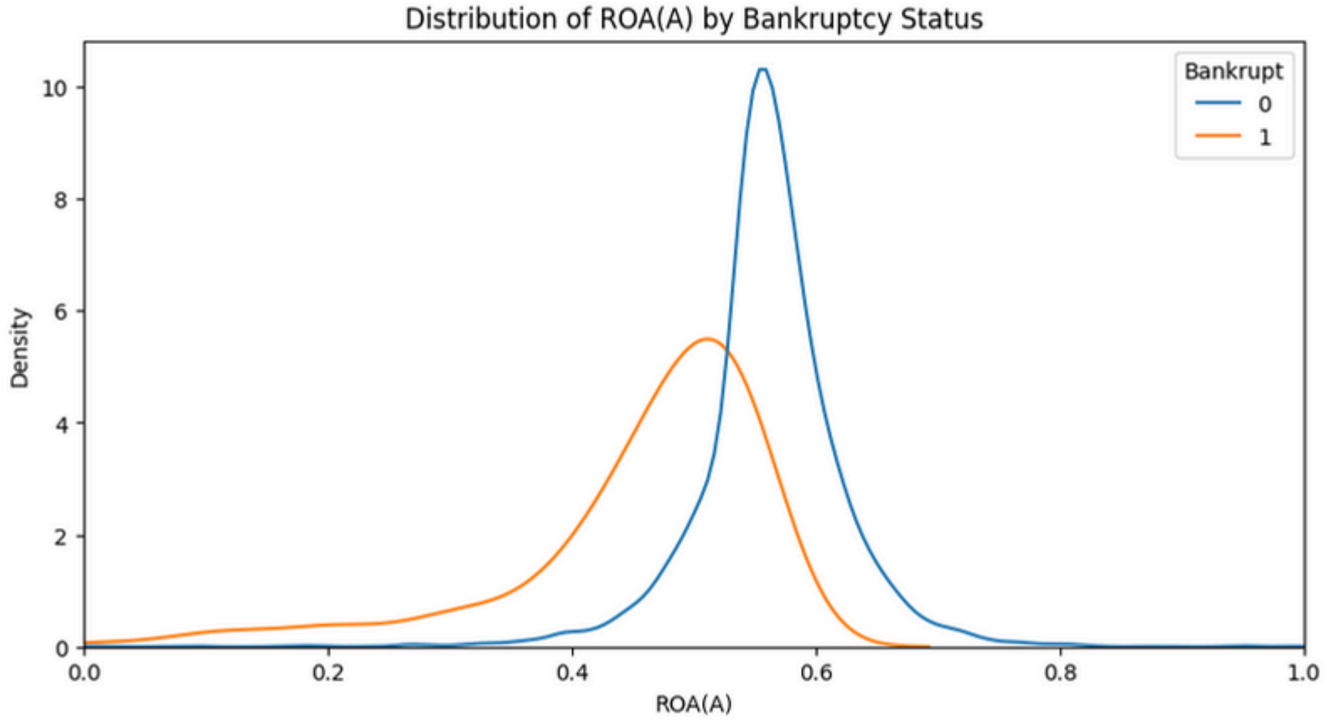
Return on Assets — Higher operational efficiency leads to lower bankruptcy risk

Positive

Debt Ratio — The higher the debt ratio, the higher the risk of bankruptcy



Exploratory Data Analysis



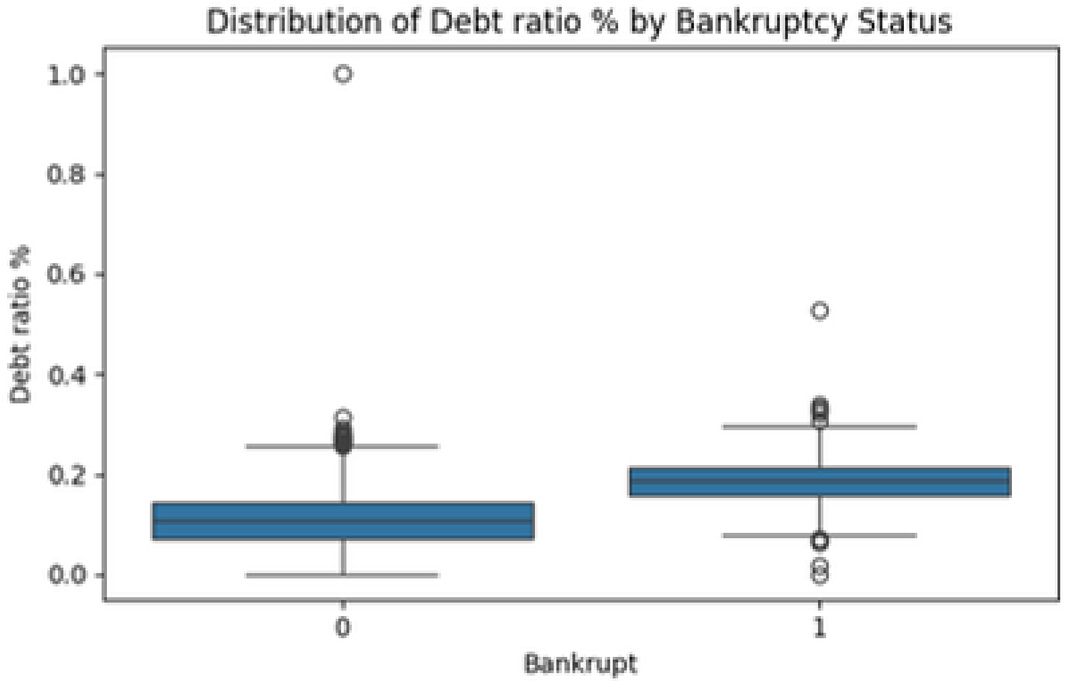
Non-Bankrupt Companies

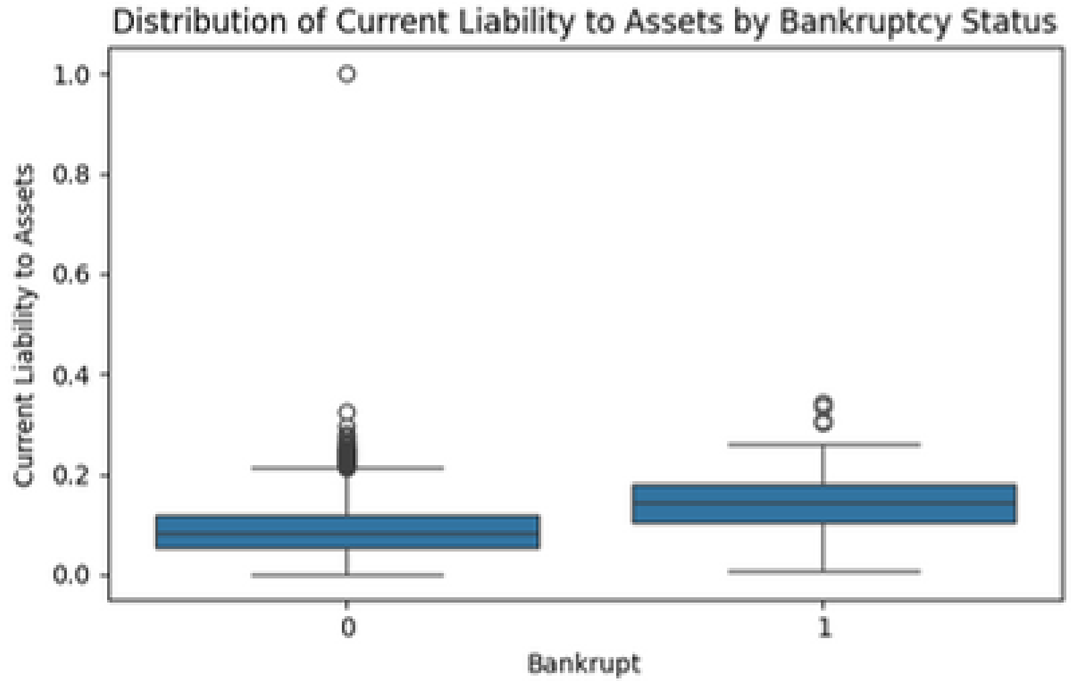
- The peak of the distribution (mode) is around ROA(A) ≈ 0.55 – 0.60
- The curve is narrow and tall, indicating that most companies are stable and efficient
- A long right tail suggests that some companies are extremely efficient

Bankrupt Companies

- The peak of the distribution is around ROA(A) ≈ 0.45 – 0.50
- The curve is wider and shifted to the left, showing that many bankrupt companies have lower ROA
- Almost no bankrupt companies have ROA(A) > 0.60

Exploratory Data Analysis





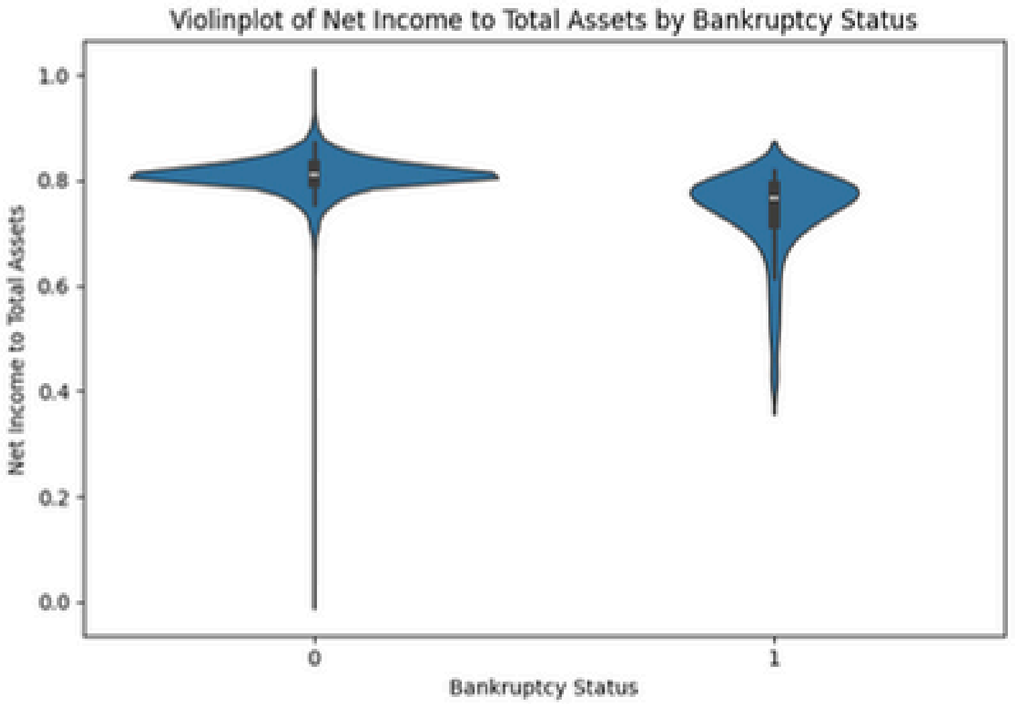
Bankrupt companies (label 1) have a higher median debt ratio compared to non-bankrupt ones (label 0)

The distribution shifts upward, meaning bankrupt companies tend to rely more heavily on debt financing

Bankrupt companies also show a higher ratio of current liabilities to total assets

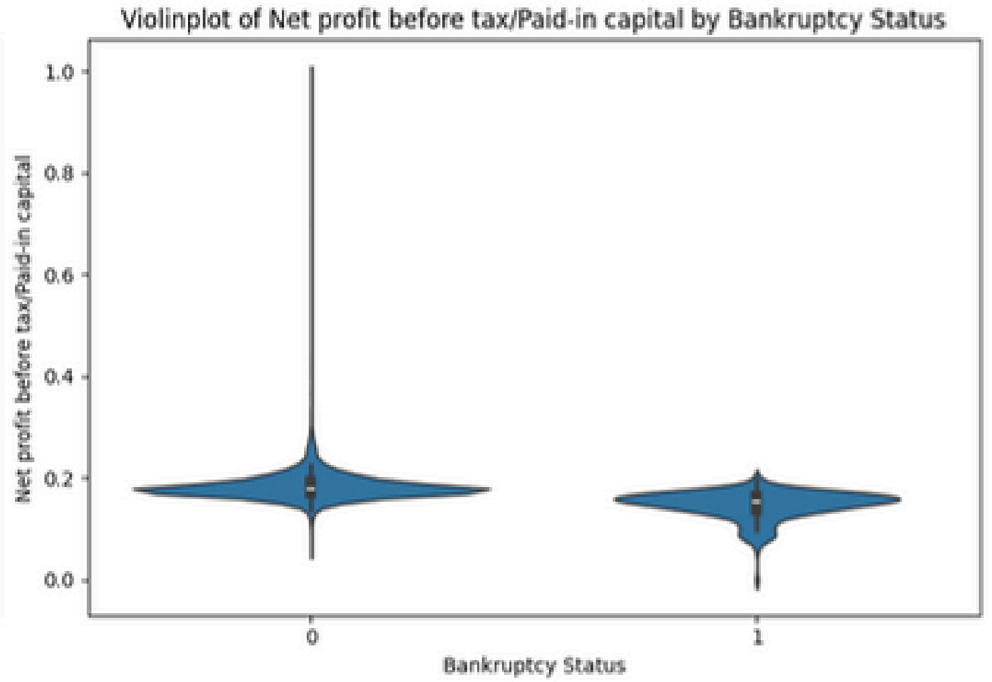
The entire distribution for bankrupt firms is shifted higher, indicating a heavier burden of short-term obligations

Exploratory Data Analysis



Non-bankrupt companies show a higher and tighter distribution (~0.8), indicating strong profitability and efficient asset use

Bankrupt companies have lower values with more spread, reflecting weaker profitability



Non-bankrupt companies are slightly more efficient in generating profit from paid-in capital, with some high outliers

Bankrupt companies mostly cluster at lower ratios, suggesting poor capital efficiency

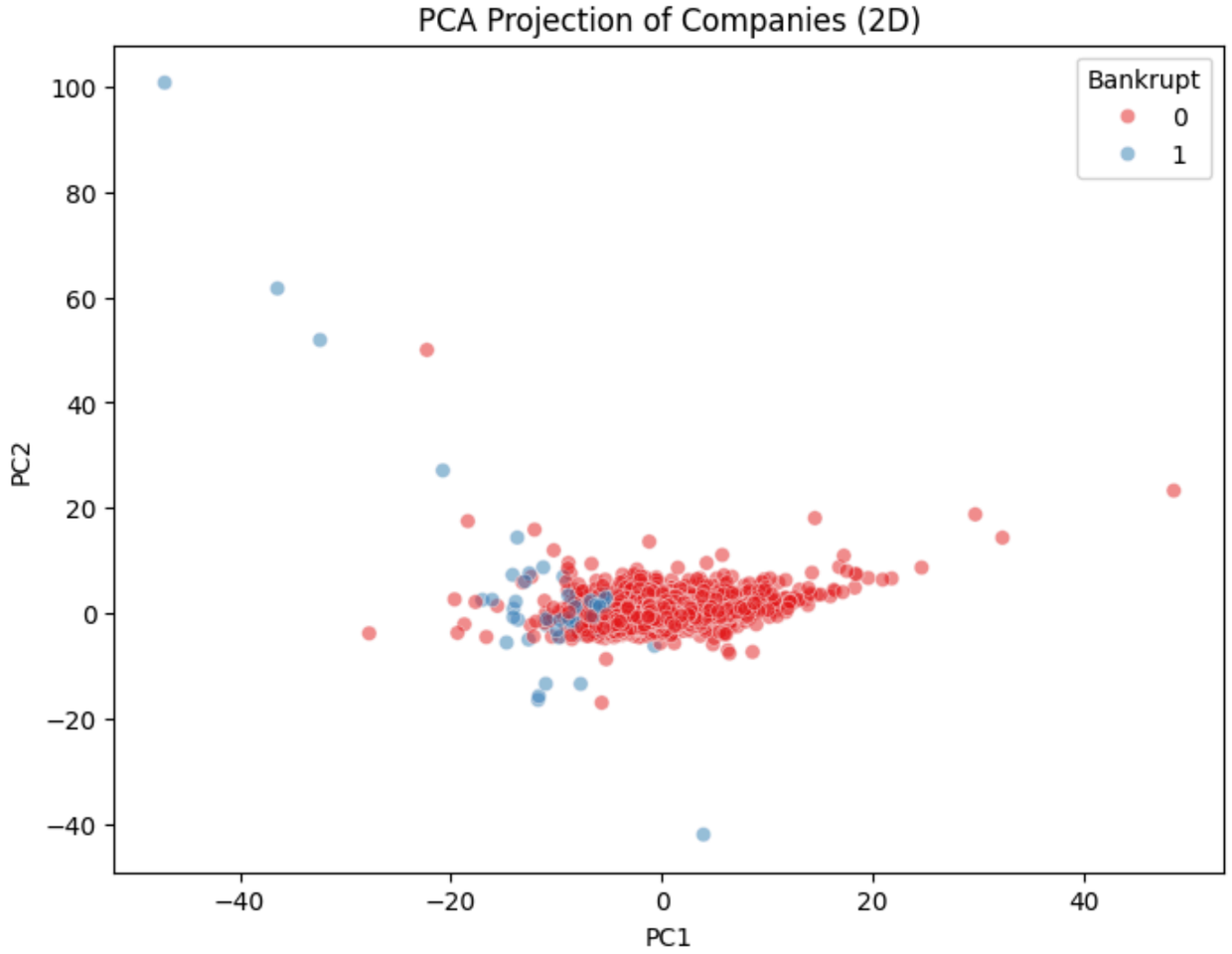
Exploratory Data Analysis

Overlapping Distribution
Blue (bankrupt) and red (not bankrupt) points overlap, PCA's two components don't fully separate the classes

Tight Cluster
Most non-bankrupt companies are densely clustered at the center, indicating financial stability

Bankrupt Spread
Bankrupt firms are more scattered, showing greater financial variability

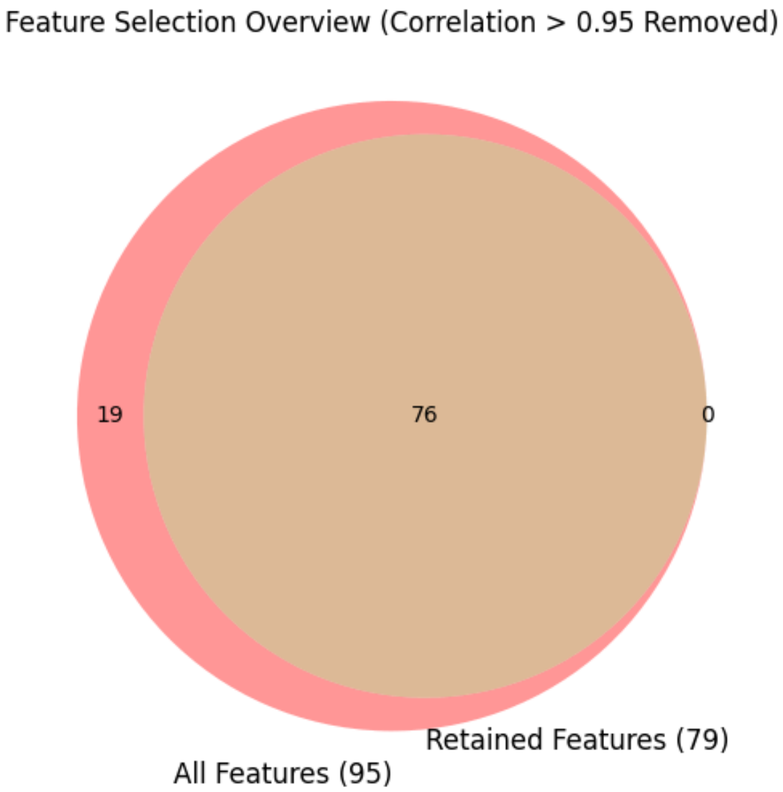
Outliers
Some isolated points (both red and blue) may be anomalies or special cases worth deeper analysis



Feature Engineering

Removed features with correlation > 0.95 to reduce multicollinearity

Objective of Feature Engineering → **Reduce redundant information and prevent multicollinearity to improve model robustness and generalization**



Step	Count
Original Features	**95**
Features Removed (>0.95)	**16**
Final Features Used	**79**

To reduce multicollinearity, only one representative feature from each correlated group was retained

This ensures a simpler, more stable model without redundant signals

Machine Learning

Train-Test Split → **70% train**, **30% test**

Stratified by target (Bankrupt)

Model → **Random Forest Classifier**

⚠ **Problem : Imbalanced Classes** ⚠

Non-bankrupt : 2046 samples

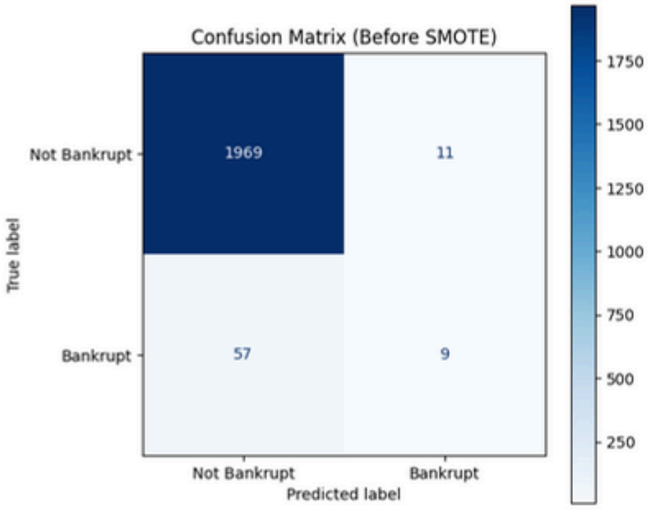
Bankrupt : only 66 samples

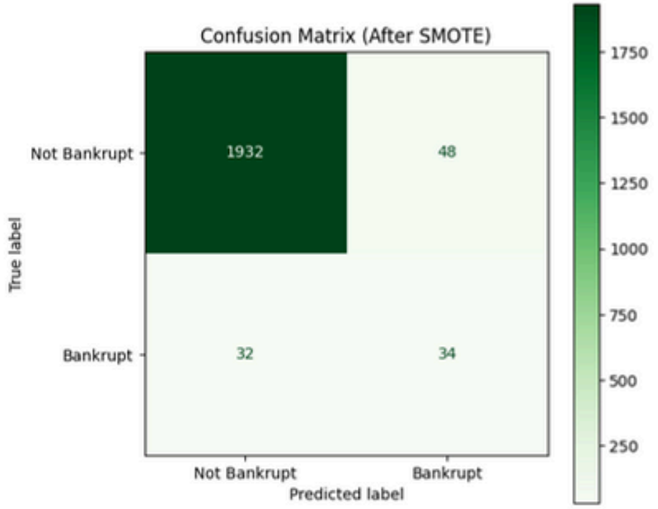
Model tends to ignore minority class (bankrupt companies)

Handling imbalance with SMOTE

(Synthetic Minority Over-sampling Technique)

Generates synthetic data for the minority class and helps model detect bankruptcy patterns





Before SMOTE

- True Bankrupt correctly predicted: **9**
- False Negatives (missed bankrupt): **57**
- **Model underperformed on minority class (bankrupt)**

After SMOTE

- True Bankrupt correctly predicted: **34**
- False Negatives reduced to: **32**
- **Model significantly improved at detecting bankrupt companies**

Machine Learning

Evaluation Metrics

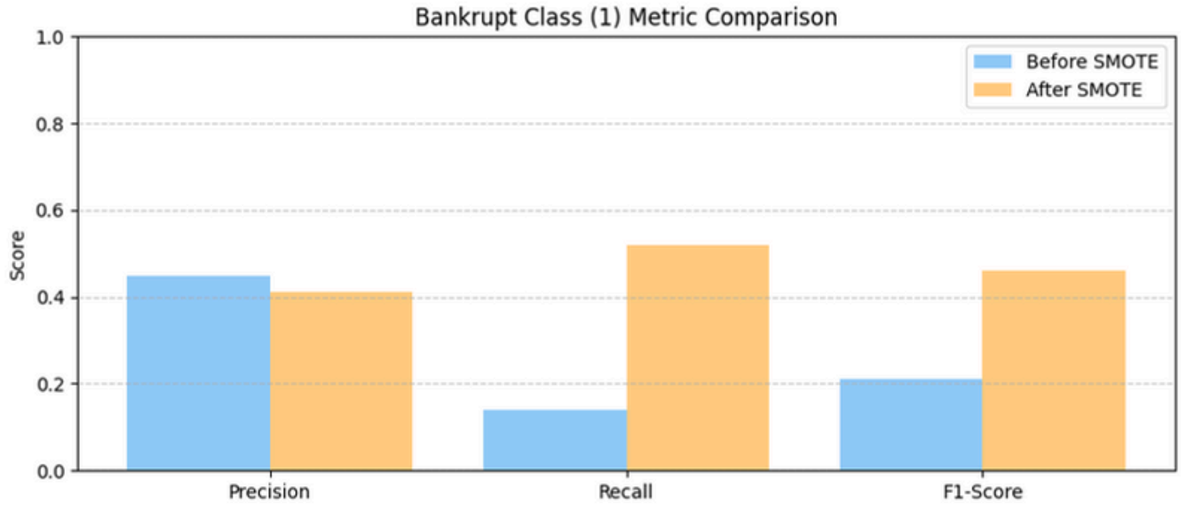
Goal : **Improve detection of bankrupt companies (class 1)**

Before SMOTE:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1980
1	0.45	0.14	0.21	66
accuracy			0.97	2046
macro avg	0.71	0.57	0.60	2046
weighted avg	0.96	0.97	0.96	2046

After SMOTE:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	1980
1	0.41	0.52	0.46	66
accuracy			0.96	2046
macro avg	0.70	0.75	0.72	2046
weighted avg	0.97	0.96	0.96	2046



Model after SMOTE is better at identifying bankrupt companies

Although precision slightly dropped, recall and F1-score improved significantly, better usefulness in bankruptcy detection

SMOTE effectively handled class imbalance and made the model more fair and actionable

Thank You!

Felicia Angjaya