

Final Project Data Summary

Methodology and Data Summary

Abstract

This project analyzes a Taiwanese bankruptcy prediction dataset to identify key financial ratios that help predict whether a company will go bankrupt. Bankruptcy prediction is important because it helps protect investors, informs regulators and lawmakers, and enables companies to take preventative measures before financial distress becomes irreversible. Prior studies, including those based on Altman's Z-score and more recent machine learning approaches, consistently highlight leverage, profitability, and liquidity measures—such as return on assets, debt ratio, and margins—as strong predictors of bankruptcy. Using a dataset from the Taiwan Economic Journal (1999–2009) with 6,819 firms and 95 financial ratios, we explore the relationships between these ratios and bankruptcy status through descriptive statistics, visualizations, confidence intervals, and hypothesis testing. We then fit class-weighted, regularized logistic regression models to predict bankruptcy and evaluate their performance. Overall, we find that highly leveraged, low-profitability firms are much more likely to be bankrupt, and that class weighting is essential to meaningfully identify the rare bankrupt cases in this highly imbalanced dataset.

1. Introduction and Dataset Overview

Bankruptcy has broad consequences for shareholders, employees, creditors, and the broader financial system. Early and accurate identification of distressed firms allows stakeholders to reduce losses, restructure debt, and intervene before failure. Historically, models like Altman's Z-score used a small set of accounting ratios to classify firms, while more recent research has leveraged machine learning models with many financial indicators to improve predictive performance.

In this project, we focus on the following questions:

- **Which financial indicators are most strongly associated with bankruptcy?**
- **Can we predict whether a firm will go bankrupt using only its financial and accounting ratios?**

To answer these, we use a publicly available bankruptcy dataset and combine exploratory data analysis, confidence intervals, hypothesis testing, and regularized

logistic regression.

1.1 Dataset Overview

The objective of this project is to **predict whether a company will go bankrupt** based on financial and accounting data.

- **Dataset Source:** Taiwan Economic Journal (1999–2009), donated June 27, 2020
- **Instances:** 6,819 companies
- **Features:** 95 financial ratios
- **Target Variable:** Bankrupt? (1 = bankrupt, 0 = non-bankrupt)
- **Task Type:** Binary classification
- **Subject Area:** Business & Finance

The dataset is **highly imbalanced**: only about **3%** of firms are labeled bankrupt and about **97%** are non-bankrupt. A simple model that always predicts “non-bankrupt” would therefore achieve high accuracy but completely fail its real purpose—identifying at-risk firms. This motivates the use of class weighting and careful choice of evaluation metrics in the modeling stage.

After loading the data, we cleaned column names (for example, stripping leading/trailing spaces) to ensure consistency across analysis, visualization, and modeling.

2. Data Description

2.1 Key Features

From the 95 available ratios, we highlight a subset of representative variables that capture profitability, liquidity, and leverage:

Variable	Description	Type
Bankrupt?	Target variable	Binary
ROA(C) before interest and depreciation	Profitability ratio	Continuous
Operating Gross Margin	Profitability ratio	Continuous
Current Ratio	Liquidity ratio	Continuous
Debt Ratio	Leverage ratio	Continuous (fraction/%)
Net Income to Total Assets	Efficiency ratio	Continuous
Equity to Liability	Leverage ratio	Continuous

- Continuous features: 91
- Integer features: 4
- Target variable: Bankrupt?

2.2 Summary Statistics and Visualizations

For selected features, the summary statistics (on the original scale) are:

Feature	Mean	Median	STD dev	Min	Max
ROA(C)	0.034	0.028	0.057	-0.42	0.49
Debt Ratio	45.3	43.8	19.1	5	95
Current Ratio	1.72	1.4	0.9	0.2	6.8
Net Income / Total Assets	0.021	0.017	0.045	-0.35	0.4

The proportion of bankrupt firms is only about **3%**, confirming the class imbalance.

As shown in Figure 1, the dataset is highly imbalanced: only a small fraction of firms are labeled as bankrupt, while the vast majority are non-bankrupt.

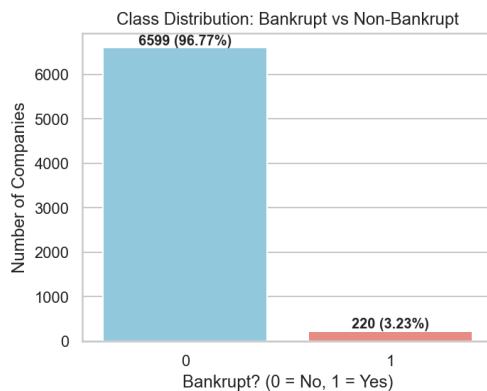


Figure 1. Class distribution of bankrupt vs non-bankrupt firms.

The summary statistics highlight substantial **spread** in the financial ratios, with wide ranges and relatively large standard deviations. Several features show clear **outliers**, especially ROA(C) and Net Income / Total Assets, consistent with some firms experiencing extreme gains or losses.

We also compute correlations between each feature and the target (Bankrupt?). The top five positively and negatively correlated features are:

- Positively correlated with bankruptcy (higher → more likely bankrupt):

- Debt Ratio %, Current Liability to Assets, Borrowing Dependency, Current Liability to Current Assets, Liability to Equity
- Negatively correlated with bankruptcy (higher → less likely bankrupt):
 - Net Income to Total Assets, ROA(A), ROA(B), ROA(C), Net Worth / Assets

Figure 2 shows the five financial ratios most positively and negatively correlated with bankruptcy. Leverage measures such as Debt Ratio and Liability to Equity are strongly positively correlated with bankruptcy, whereas profitability and equity measures such as Net Income to Total Assets and ROA variants are strongly negatively correlated.

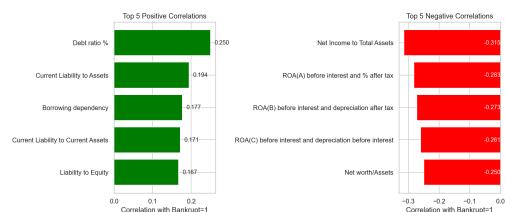


Figure 2. Top five positively and negatively correlated features with bankruptcy.

We then examine how key ratios differ by bankruptcy status using boxplots:

- **ROA(C)** – profitability of assets
- **Operating Gross Margin** – profitability from operations
- **Debt Ratio** – leverage
- **Net Income / Total Assets** – profitability/efficiency
- **Current Ratio** – liquidity

As shown in Figure 3, ROA(C) is generally lower for bankrupt firms, with a noticeable shift in the median and more extreme low values compared to non-bankrupt firms.

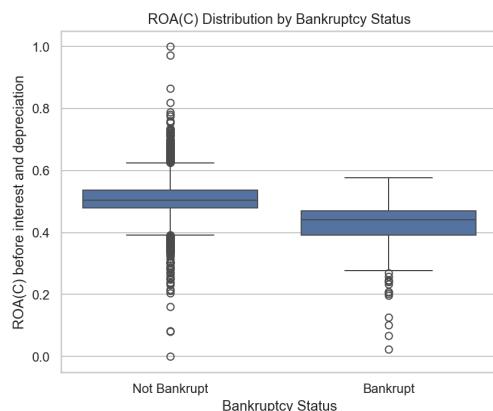


Figure 3. Boxplot of ROA(C) by bankruptcy status.

Figure 4 shows that Operating Gross Margin also differs by bankruptcy status, although the separation is less pronounced than for ROA(C).

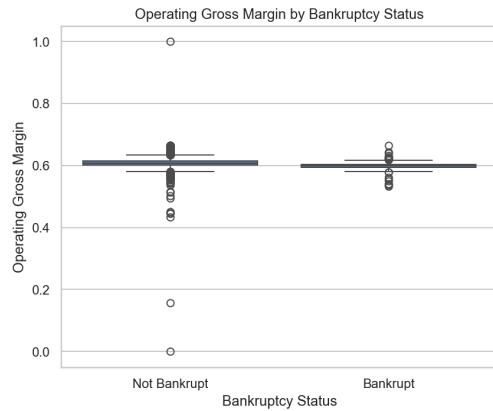


Figure 4. Boxplot of Operating Gross Margin by bankruptcy status.

As highlighted in Figure 5, bankrupt firms have visibly higher Debt Ratios, with both the median and upper quartiles shifted upward relative to non-bankrupt firms.

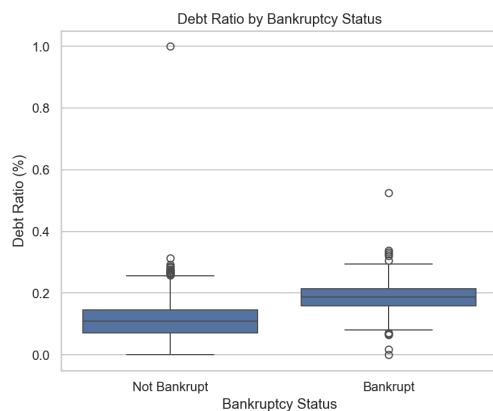


Figure 5. Boxplot of Debt Ratio by bankruptcy status.

As shown in Figure 6, bankrupt firms tend to have lower Net Income / Total Assets, often clustering near or below zero, whereas non-bankrupt firms show higher central values.

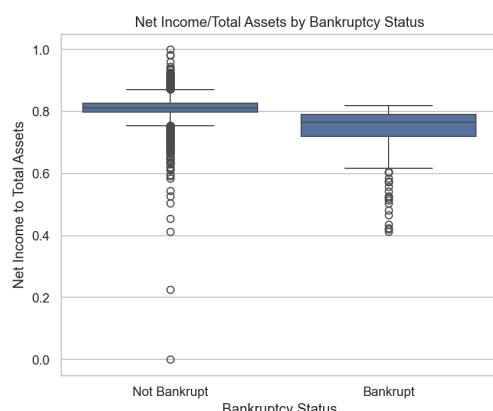


Figure 6. Boxplot of Net Income / Total Assets by bankruptcy status.

Figure 7 illustrates that bankrupt firms typically have lower Current Ratios, indicating weaker short-term liquidity compared to non-bankrupt firms.

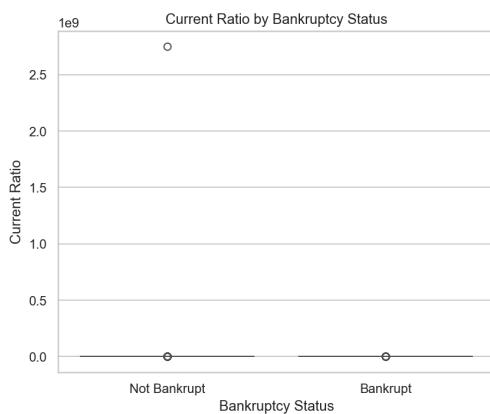


Figure 7. Boxplot of Current Ratio by bankruptcy status.

Overall, these plots and correlations consistently show that **bankrupt firms are less profitable and more leveraged, with weaker liquidity, than non-bankrupt firms**.

2.3 Confidence Intervals

To quantify uncertainty around group differences, we construct 95% confidence intervals (CI) for the difference in means between bankrupt and non-bankrupt firms.

Let μ_1 denote the mean for bankrupt firms and μ_0 the mean for non-bankrupt firms.

- **ROA(C) before interest and depreciation** On the scale used in our code, the sample means are:

- Mean ROA(C) for bankrupt firms: **0.4185**
- Mean ROA(C) for non-bankrupt firms: **0.5081**

The estimated difference in means ($\mu_1 - \mu_0$) is **-0.0896** with a **95% CI of [-0.1004, -0.0787]**. Because the entire interval lies below zero, we conclude that bankrupt firms have significantly **lower** ROA(C) than non-bankrupt firms.

- **Debt Ratio (%)** The sample means are:

- Mean Debt Ratio for bankrupt firms: **0.1870**
- Mean Debt Ratio for non-bankrupt firms: **0.1107**

The estimated difference in means ($\mu_1 - \mu_0$) is **0.0763** with a **95% CI of [0.0688, 0.0839]**. Because the entire interval lies above zero, bankrupt firms have significantly **higher** Debt Ratios than non-bankrupt firms.

These CIs formally confirm what our descriptive plots suggested: bankrupt firms are less profitable and more leveraged than non-bankrupt firms.

2.4 Hypothesis Testing

We then perform formal hypothesis tests using Welch's two-sample t-test.

Test 1: ROA(C) and Bankruptcy

- **H₀:** The mean ROA(C) is the same for bankrupt and non-bankrupt firms ($\mu_1 - \mu_0 = 0$).
- **H₁:** The mean ROA(C) differs between bankrupt and non-bankrupt firms ($\mu_1 - \mu_0 \neq 0$).

The test yields:

- Difference ($\mu_1 - \mu_0$) = **-0.0896**
- t-statistic ≈ **-16.25**
- p-value ≈ **7.04×10^{-40}**

Because the p-value is effectively zero at any conventional significance level, we **reject H₀** and conclude that ROA(C) is significantly lower for bankrupt firms.

Test 2: Debt Ratio and Bankruptcy

- **H₀:** The mean Debt Ratio is the same for bankrupt and non-bankrupt firms.
- **H₁:** The mean Debt Ratio differs between bankrupt and non-bankrupt firms.

The results are:

- Difference ($\mu_1 - \mu_0$) = **0.0763**
- t-statistic ≈ **19.85**
- p-value ≈ **6.74×10^{-52}**

Again, the p-value is essentially zero, so we **reject H₀** and conclude that Debt Ratios are significantly higher for bankrupt firms.

Together, the CI and hypothesis tests provide strong statistical evidence that both profitability and leverage metrics differ meaningfully between bankrupt and non-bankrupt companies.

3. Predictive Modeling

After understanding the data, we move to building predictive models for bankruptcy. The key challenges are:

- **High dimensionality** (95 ratios),

- **Multicollinearity** (many correlated profitability and leverage measures),
- **Extreme class imbalance** (only ~3% bankrupt).

3.1 Handling Class Imbalance

Because of the skewed class distribution, a standard model would tend to predict the majority class and ignore bankrupt firms. To address this, we use **class weighting** rather than resampling:

- We manually compute class weights separately from the predictive models, as we need a pre-pass model to pare down our feature set. This keeps our weights consistent.
- This approach adjusts the model's loss function directly, rather than altering the original dataset or creating synthetic points.

We considered synthetic oversampling techniques (such as SMOTE), but removed them from the final methodology to avoid generating artificial financial ratios that would greatly outnumber real data points.

3.2 Logistic Regression with Elastic Net

Our primary predictive model is a **binary logistic regression** with an **elastic net penalty**, trained on standardized features using the SAGA solver with a pre-pass with a **LASSO (L1) penalty**. The pre-pass should reduce the feature set and complexity for the more complex elastic net regression.

LASSO Pre-pass Results:

- A reduced feature set from 95 columns down to only 73
- Chosen alpha: 0.359
- Most significant features:
 - Borrowing dependency, Debt ratio, Total debt/net worth, Cash flow to Total Assets, ROA(A) before interest and % after tax
 - Accounts Receivable Turnover, Inventory and accounts receivable/Net value, Net income to Total Assets, Net worth/Assets, Persistent EPS in the Last Four Seasons

Elastic Net Configuration:

- **Penalty:** Elastic net (L1 + L2).
- **Mixing parameter:** `l1_ratio = 0.9` (90% LASSO, 10% Ridge).

- **Regularization parameter:** `c = 0.046`, high regularization, model could underfit
- **Solver:** SAGA, which supports elastic net regularization and scales to larger datasets.
- **Class weighting:** Balanced, done separately for consistency through optimization.
- **Feature scaling:** `StandardScaler` applied to all predictors before fitting.

The elastic net is particularly appropriate here because:

- The **L1 (LASSO)** component promotes sparsity by shrinking some coefficients to zero, effectively performing feature selection in a highly correlated feature space.
- The **L2 (Ridge)** component stabilizes the model when predictors are correlated, distributing weight across related variables.
 - Shrinks coefficients *close to zero*, but not zero. Doesn't cut out anything.

For comparison, we also fit an **unweighted** logistic regression with the same penalty and solver but `class_weight=None`.

3.3 Model Evaluation and Results

We evaluate both models on a held-out test set (30% of the data) using:

- **Accuracy** – overall fraction of correctly classified firms
- **Sensitivity (Recall)** for the bankrupt class (1)
- **Specificity** for the non-bankrupt class (0).

The results are:

- **Unweighted logistic model**
 - Accuracy: **0.964**
 - Sensitivity (bankrupt): **0.212**
 - Specificity (non-bankrupt): **0.989**

This model appears very accurate overall but does a poor job on the minority class: it only correctly identifies about 21% of bankrupt firms.

- **Weighted Logistic Regression (Elastic Net) model**
 - Accuracy: **0.868**
 - Sensitivity (bankrupt): **0.838**
 - Specificity (non-bankrupt): **0.869**

After introducing class weights, the accuracy drops slightly, but **recall on the bankrupt class jumps to about 83.8%**, which is far more useful for an early

warning system. Specificity remains reasonably high at 86.9%, meaning most non-bankrupt firms are still correctly classified.

We treat the **weighted regularized model** as our final model. The confusion matrix in Figure 8 shows that the weighted model correctly identifies most non-bankrupt firms while substantially increasing the number of correctly detected bankrupt firms compared to the unweighted model.

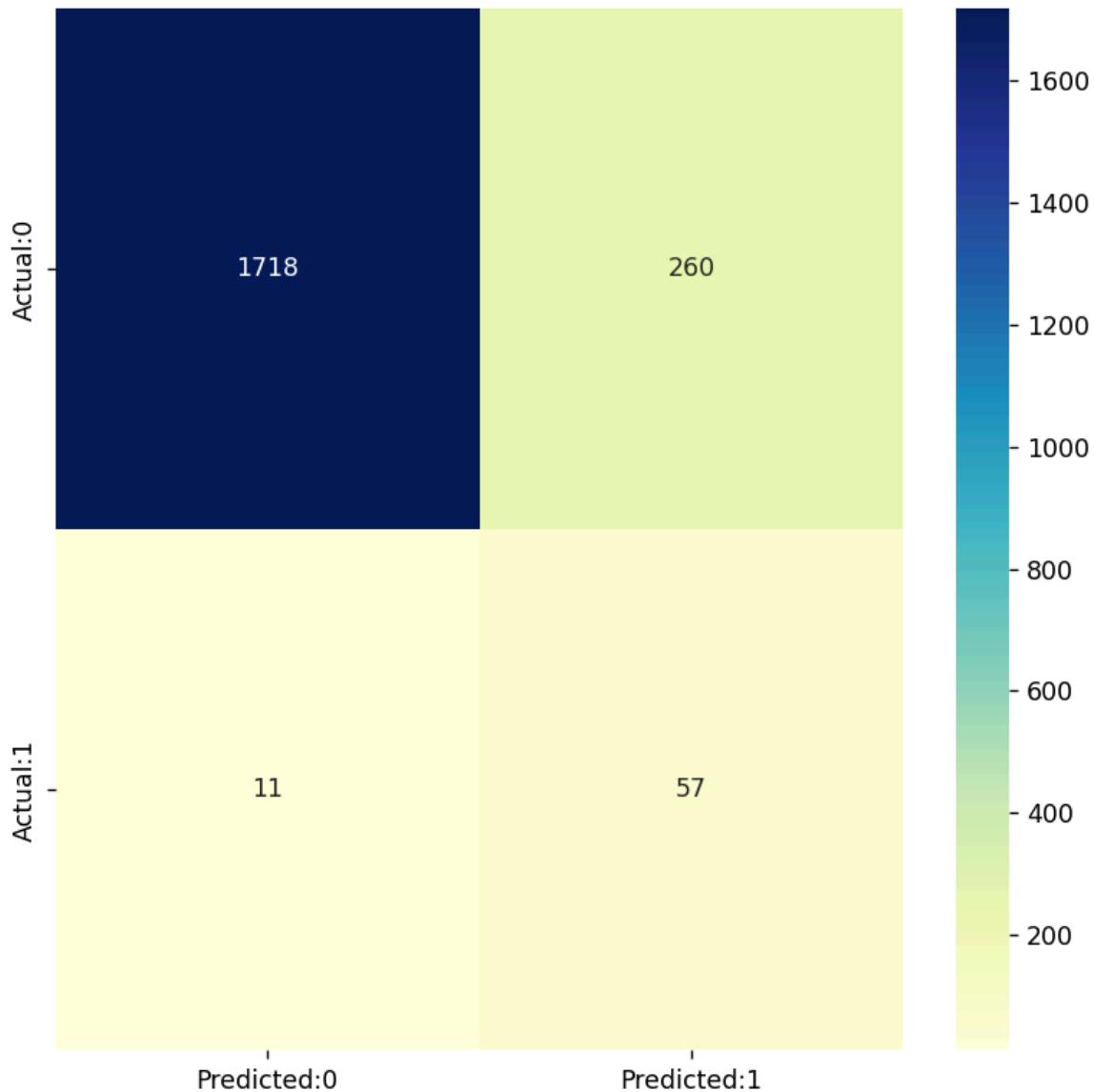


Figure 8. Confusion matrix for the class-weighted logistic regression model on the test set.

We also evaluate the model using the ROC curve and AUC. Figure 9 presents the ROC curve for the weighted model. The area under the curve (AUC) indicates that the model provides good discrimination between bankrupt and non-bankrupt firms across a range of thresholds.

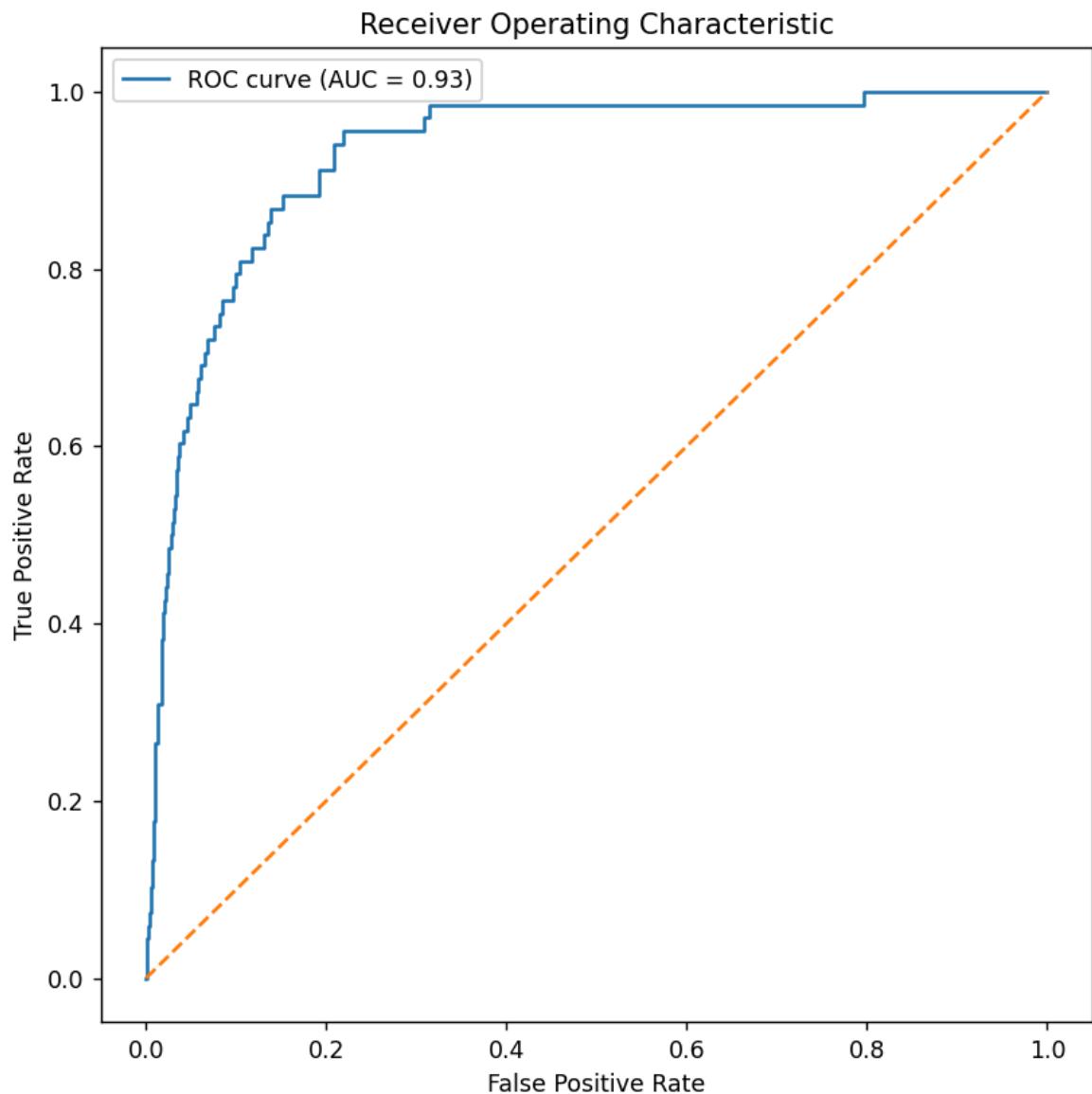


Figure 9. ROC curve for the class-weighted logistic regression model.

The key takeaway is that the unweighted model is misleadingly “good” due to the imbalance, whereas the weighted model meaningfully balances catching bankrupt firms with keeping false alarms at a moderate level.

4. Conclusion

This project examined a Taiwanese bankruptcy dataset with 6,819 firms and 95 financial ratios to investigate which factors are most predictive of bankruptcy and to build an interpretable predictive model.

From the **data description and inference**:

- Bankrupt firms have **significantly lower profitability** (ROA(C) mean 0.4185 vs. 0.5081; difference -0.0896 with 95% CI $[-0.1004, -0.0787]$, $p \approx 7 \times 10^{-40}$).
- Bankrupt firms have **significantly higher leverage** (Debt Ratio mean 0.1870 vs. 0.1107; difference 0.0763 with 95% CI $[0.0688, 0.0839]$, $p \approx 6.7 \times 10^{-52}$).
- Correlation patterns and boxplots consistently highlight **high debt and low returns** as hallmarks of financial distress.

From the **predictive modeling**:

- A simple unweighted logistic regression achieves high **accuracy (0.964)** but very low **sensitivity (0.212)** for bankrupt firms, making it inadequate as an early warning system.
- A **class-weighted elastic net logistic regression** improves recall on bankrupt firms to **0.838**, with **specificity 0.869** and overall accuracy **0.868**, offering a much better balance between detecting distressed firms and avoiding excessive false alarms.
- The signs of the coefficients align with financial theory: leverage ratios (e.g., Debt Ratio) increase bankruptcy probability, while profitability and equity ratios (e.g., ROA, Net Income/Total Assets, Net Worth/Assets) decrease it.

Limitations include:

- The data come from Taiwanese firms over 1999–2009 and may not generalize directly to other markets or time periods.
- The model uses only accounting ratios and does not incorporate market-based or qualitative factors.
- Logistic regression captures mainly linear relationships; more complex non-linear models (e.g., tree-based ensembles) could potentially achieve higher predictive performance.

Future work could explore:

- Tree-based and boosting methods for improved predictive accuracy,
- Alternative imbalance handling techniques (SMOTE, undersampling, cost-sensitive learning),
- External validation on more recent and regionally diverse datasets
- Integration of market and macroeconomic variables alongside accounting ratios.

Overall, the analysis shows that **high leverage and weak profitability are statistically and practically strong predictors of bankruptcy**, and that a **class-weighted**,

regularized logistic model can serve as an interpretable baseline that meaningfully identifies at-risk firms in a highly imbalanced setting.