**SUBMITTED BY** : **Arjun Aji**

**: 100944831**

A screenshot of a computer screen

AI-generated content may be incorrect.

The image shows a **highly imbalanced class distribution** in a dataset labeled "Bankrupt?". Class 0 is dominant (96.8%) with around 6500 counts, while Class 1 is rare (3.2%) with approximately 250 counts.

A screenshot of a computer screen

AI-generated content may be incorrect.

This image is a lower-triangle correlation heatmap showing the relationships between various financial features. Red squares indicate positive correlation, blue squares show negative correlation, and white squares represent weak or no correlation. The darker the color, the stronger the relationship between the features.

A screenshot of a computer screen

AI-generated content may be incorrect.

This image shows a grid of 12 histograms, each depicting the distribution of a different financial feature. Most features are highly concentrated or skewed, with a large number of observations falling within a narrow range. A few features, such as those in the top row, show a more normal-like distribution.

A screenshot of a computer

AI-generated content may be incorrect.

This image displays a series of box plots for the first 10 financial features to detect outliers. The boxes represent the main data range, while the circles outside the whiskers are individual data points identified as outliers. The plots show that many features have a significant number of outliers, indicating values far from the typical range.

A screenshot of a graph

AI-generated content may be incorrect.

This image shows violin plots for the first 10 financial features. The plots' shapes illustrate the data density and distribution, with wider sections indicating a higher concentration of values. The narrow points and long tails represent areas with fewer data points, which can indicate the presence of outlier.

A screenshot of a graph

AI-generated content may be incorrect.

This image is a horizontal bar chart showing the "XGBoost Feature Importance (Top 20)." It ranks the features by their "Importance score," which indicates their influence on the model's predictions. "income per person" is the most important feature with a score of 56.0.

A screen shot of a graph

AI-generated content may be incorrect.

This image is an **ROC Curve** plot comparing the performance of three models: **XGBoost, Logistic Regression, and Random Forest**. The XGBoost model is the best performer with the highest **Area Under the Curve (AUC)** score of 0.95, indicating its curve is closest to the ideal top-left corner. Logistic Regression and Random Forest both have an AUC of 0.92, showing similar performance.

A screen shot of a graph

AI-generated content may be incorrect.

This image is a **Precision-Recall Curves** plot comparing three models. The **XGBoost** and **Random Forest** models (green and orange lines) both perform significantly better than the **Logistic Regression** model (blue line), maintaining a higher precision for a given recall.

**1. Challenges & Solutions**

* **Class imbalance** – More “non-bankrupt” cases than “bankrupt” ones, causing bias.  
  I Solved using class\_weight="balanced" and scale\_pos\_weight for XGBoost.
* **Correlated features** – Some features overlapped in meaning.  
   Checked with VIF and removed unnecessary ones.
* **Model explainability** – Needed to justify predictions.  
  Used SHAP values to highlight key influencing features.

**2. How Lab 4 Helped Lab 5**

* Learned **Logistic Regression** is simple and interpretable → used as baseline.
* **Random Forest** handles complex patterns with minimal tuning.
* **XGBoost** often gives the best performance for tabular data → focused on tuning it in Lab 5.

**3. Recommended Model**

* **XGBoost** performed best overall:
  + Highest ROC AUC score.
  + Managed class imbalance effectively.
  + SHAP plots confirmed logical feature importance, making results trustworthy.