DEFAULT OF CREDIT CARD CLIENTS

- **ASDS 6303 – Data Mining**
- Group 9 Members:
  - Phuong Trinh
  - Kyra Stolarski

WHY THIS PROBLEM MATTERS?

- Credit card default prediction is crucial task in risk management.
- Financial institutions need accurate models to identify clients likely to default.

INTRODUCTION

- **Credit Card Default Prediction**
    - Dataset from **UCI Machine Learning Repository**
    - Includes demographic, credit limit, repayment history, bill amounts, and payment data for customers in Taiwan
    - **Target (Y):**
        - 1 = Default next month
        - 0 = Non-default
- **Goal**
    - Build predictive models—logistic regression, CART decision tree, and random forest—to predict a client's default status.
    - Help financial institutions identify high-risk clients, reduce losses, and improve credit strategies.

OBJECTIVES

Perform data cleaning, descriptive statistics, and EDA.

Build three classification models to predict default status.

Evaluate model performance using:
- Confusion Matrix
- Train and Test Accuracy, Sensitivity, Specificity
- ROC Curve
- Area Under Curve (AUC)

Interpret model results:
- Feature importance
- Tree structure

Provide actionable insights based on findings.

DATASET OVERVIEW

- **Observations:** 30,000 credit cards clients
- **Features:** 23 predictors + ID
- **Target:** Default payment (Y) for next month
- **Variable Groups:**
    - Demographics: sex, education, marriage, age
    - Credit limit: LIMIT_BAL
    - Past payment behavior: PAY_1 to PAY_6
    - Bill amounts: BILL_AMT1–6

DATA PREPROCESSING

- Fixed incorrect header row and standardized column names.
- Converted all character columns to numeric.
- Recoded categorical variables:
  - **SEX:** Male / Female
  - **EDUCATION:** GradSchool, University, HighSchool, Others
  - **MARRIAGE:** Married, Single, Others
  - Converted repayment history (PAY_1 to PAY_6) into ordered factors.
  - Renamed target variable to **y** (NoDefault / Default).
- **No missing values** found in any variable.

DESCRIPTIVE STATISTICS OVERVIEW

Average credit limit ~ **167k**, but highly right-skewed.

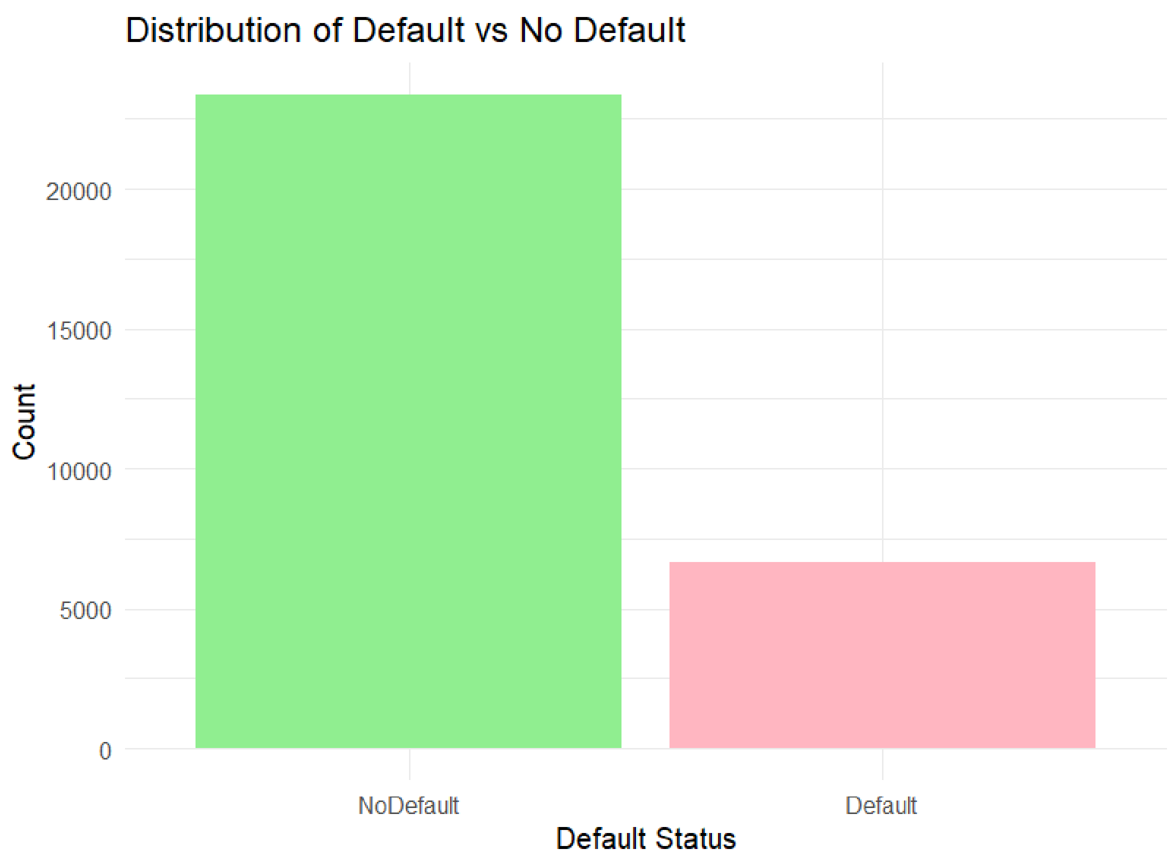
Majority age range: **30–41 years old**, median age = 34.

Majority are **Female**, **University-educated**, **Single** or **Married**.

Payment status mostly on-time or slightly delayed.

Bill and payment amounts show large outliers.

CLASS IMBALANCE



Default: **6,630 clients (22%)**

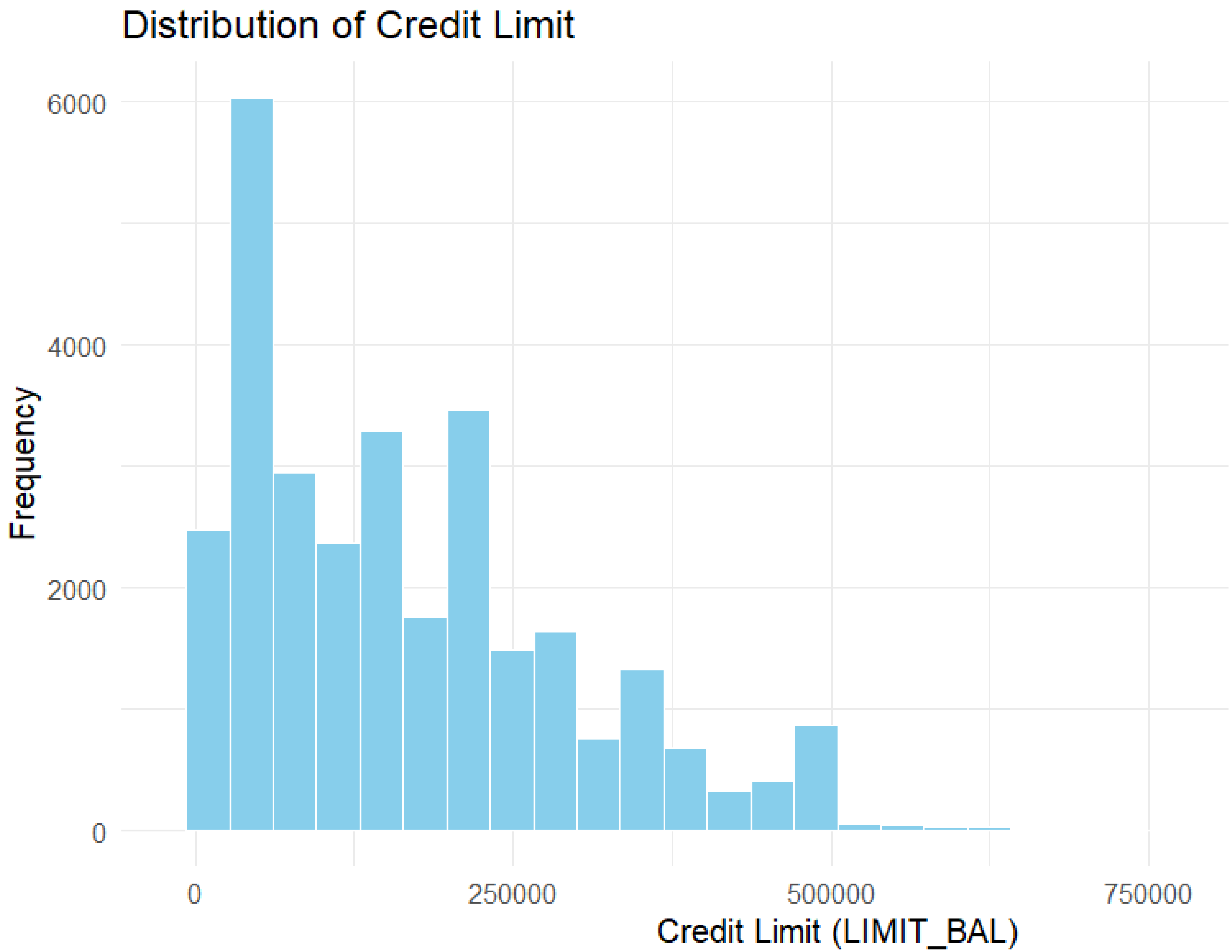
Non-Default: **23,335 clients (78%)**

Moderate class imbalance → impacts model performance.

Addressed using **upSampling** for balanced training

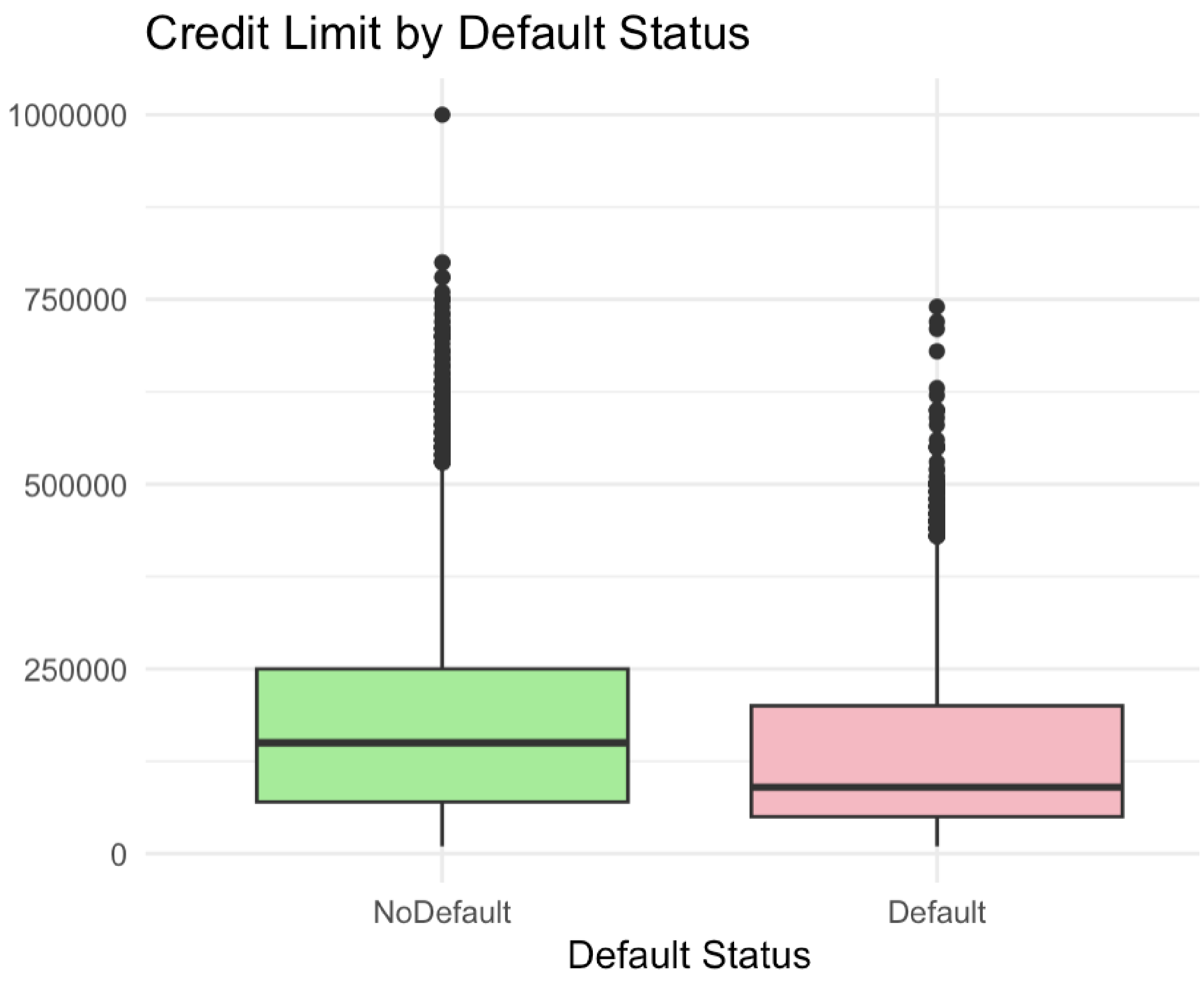
EXPLORATORY DATA ANALYSIS

- Credit limit distribution is right-skewed; many clients have small limits; few have very high limits.



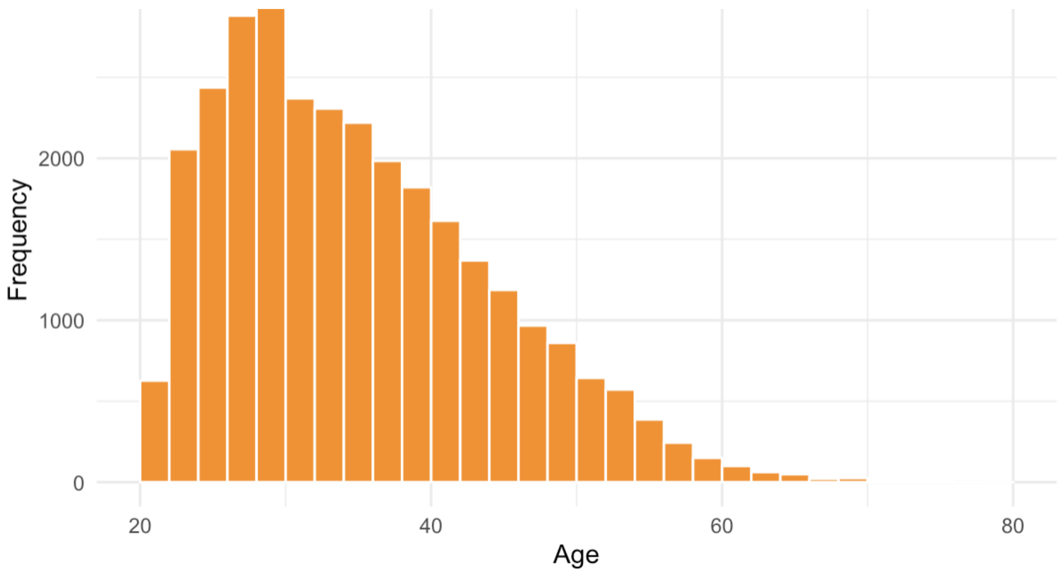
EDA

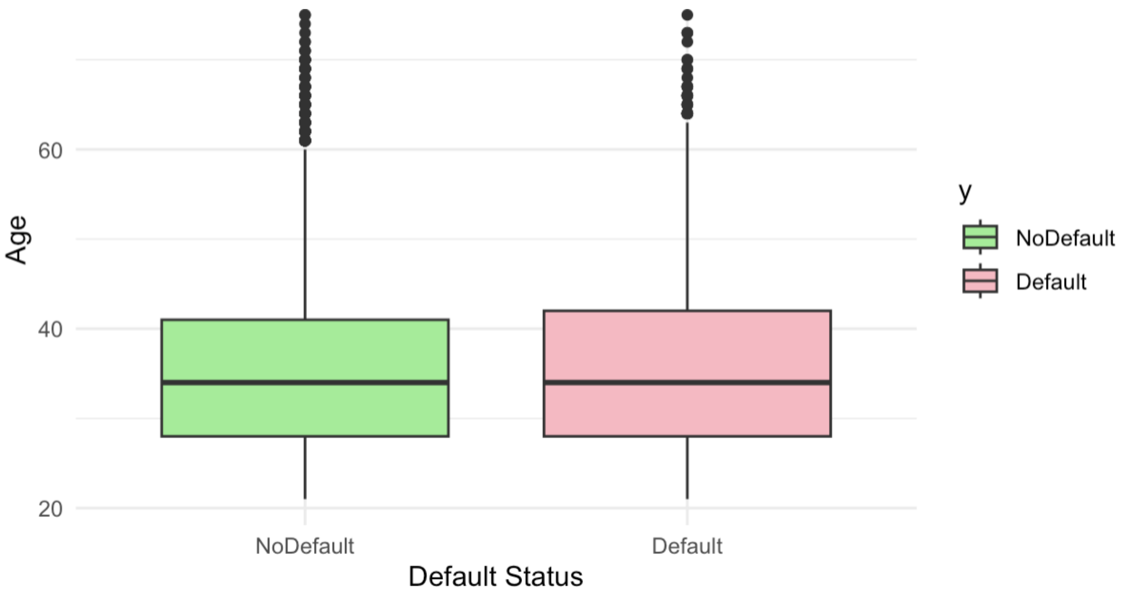
- Defaulting clients tend to have **lower credit limits**.



AGE DISTRIBUTION

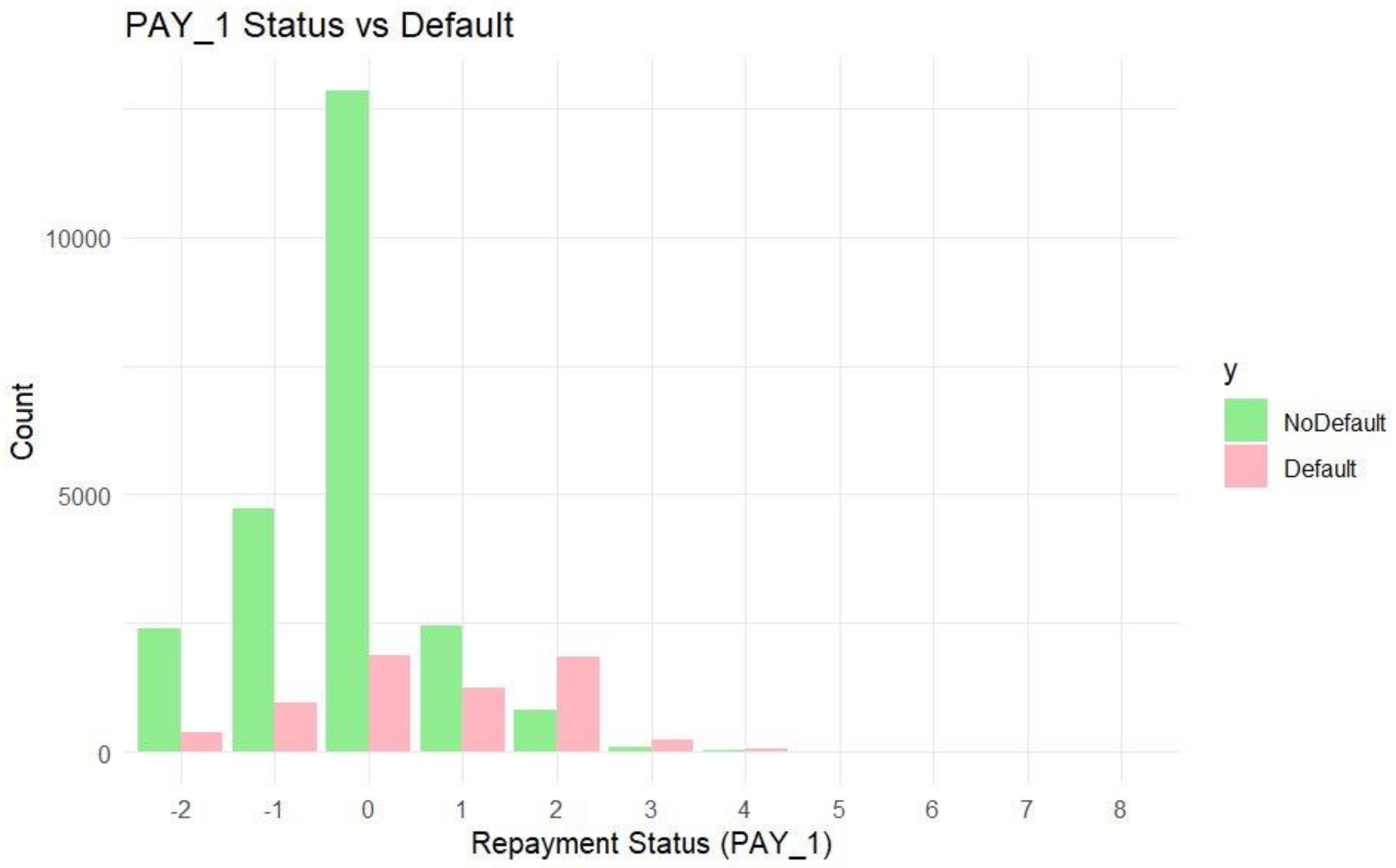
- Age is mostly concentrated between 25–40.
- Age differences between default/no-default are small.



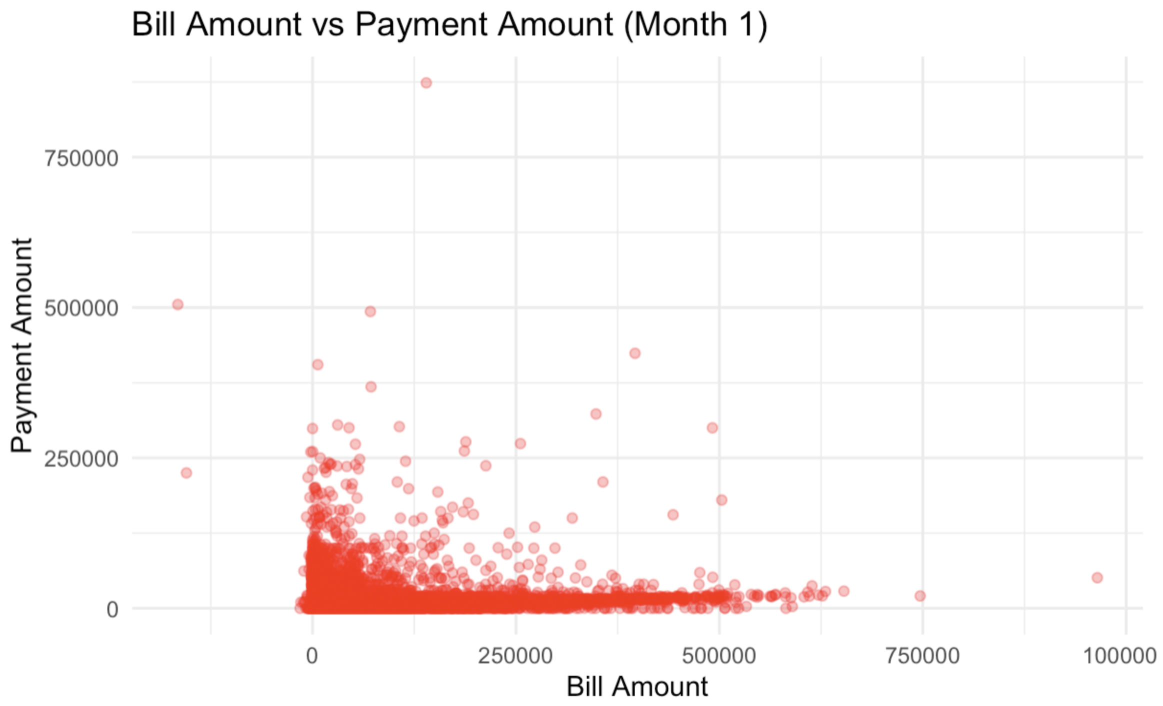


EDA

- Most recent repayment behavior strongly correlates with default risk
- PAY_1 = 1 or more → high default probability.



KEY EDA FINDINGS



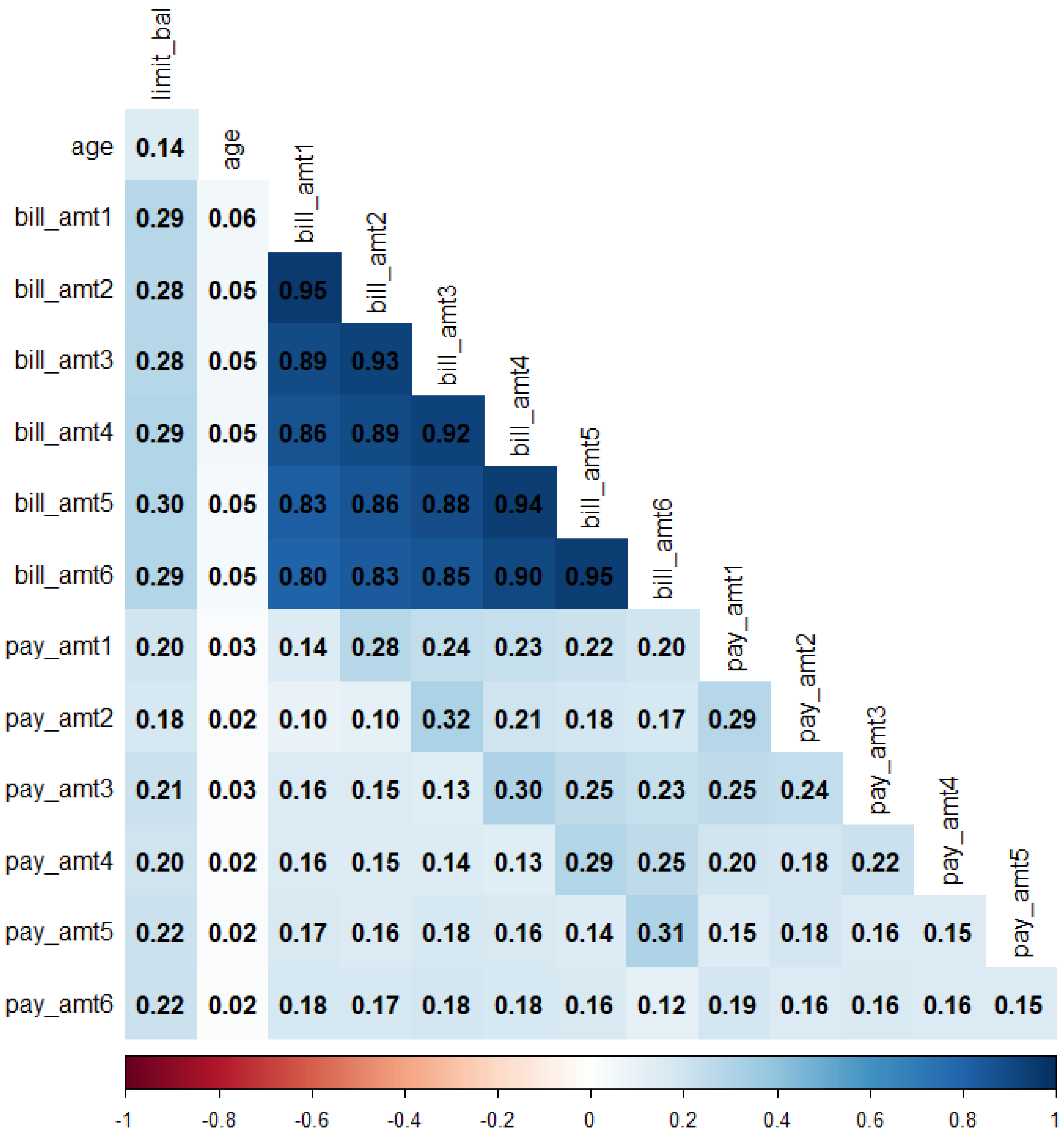
- Bill amounts and payment amounts show a positive but highly variable relationship.
- Many clients with high bill amounts make large payments, but the spread indicates inconsistent repayment behavior.
- Several extreme outliers appear in both bill and payment amounts, reflecting diverse financial patterns.

CORRELATION HEATMAP

•Strong correlations among bill amount variables (BILL_AMT1–6).

•Removed highly correlated features using **caret::findCorrelation()**.

•Remaining variables prevent multicollinearity in models.



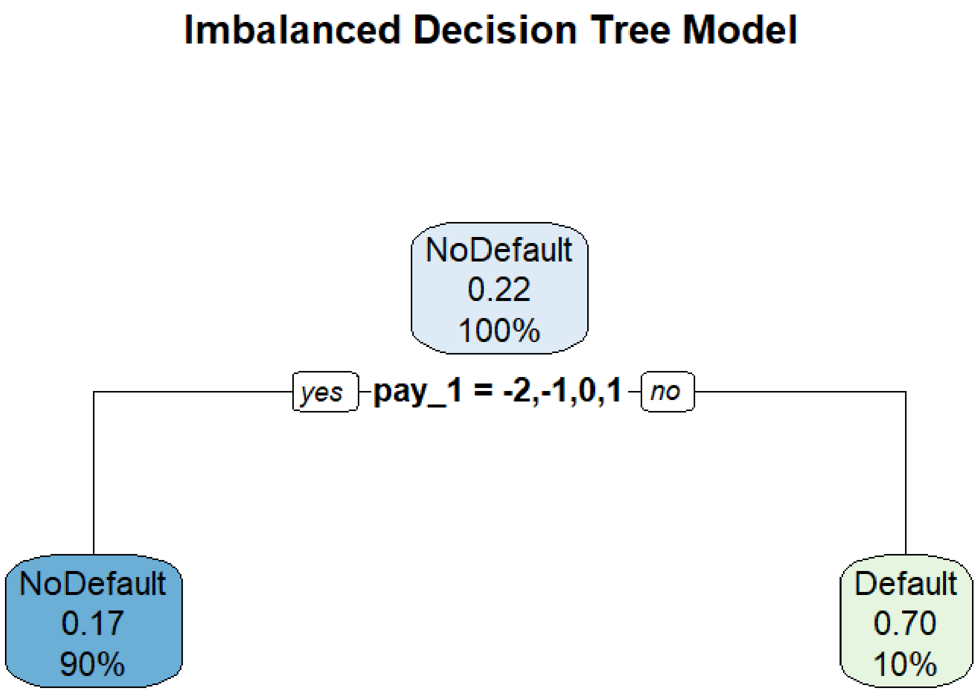
TRAIN/TEST SPLIT & BALANCING

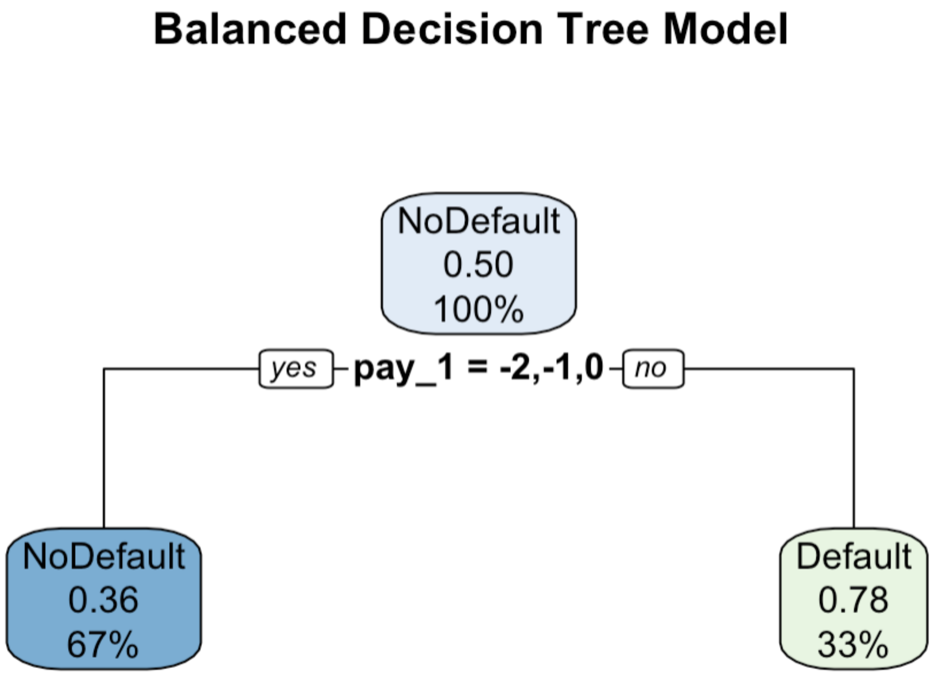
- Dataset split into **70% training** and **30% testing** using stratified sampling.
- Original class distribution in training set:
  - NoDefault: **16,335**
  - Default: **4,641**
- To handle imbalance, applied **upSampling**:
- Balanced class distribution:
  - NoDefault:**16,335**
  - Default: **16,335**
- Ensures models don't overwhelmingly predict "NoDefault."

LOGISTIC REGRESSION

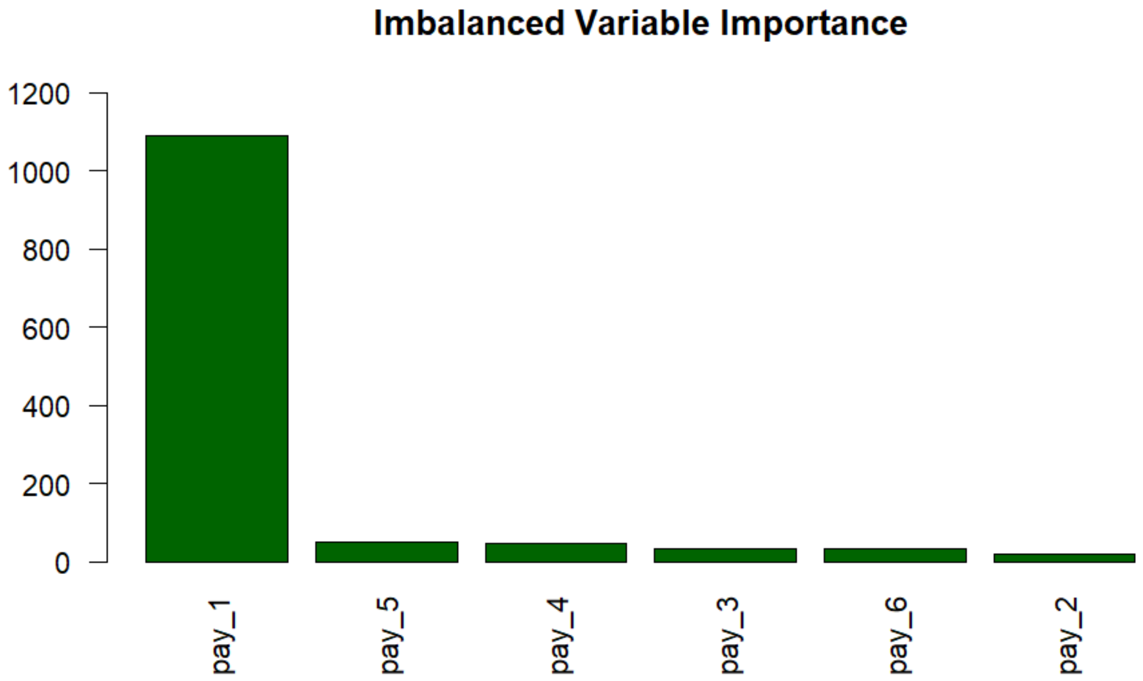
- Imbalanced – Statistically Significant Attributes
  - limit_bal
  - sexFemale
  - educationOthers
  - marriageSingle
  - age
  - bill_amt1
  - pay_amt1
  - pay_amt2
  - pay_amt3
- Balanced – Statistically Significant Attributes
  - limit_bal
  - sexFemale
  - educationOthers
  - marriageSingle
  - age
  - bill_amt1
  - pay_amt1
  - pay_amt2
  - pay_amt3
  - pay_amt4
  - pay_amt5
  - pay_amt6

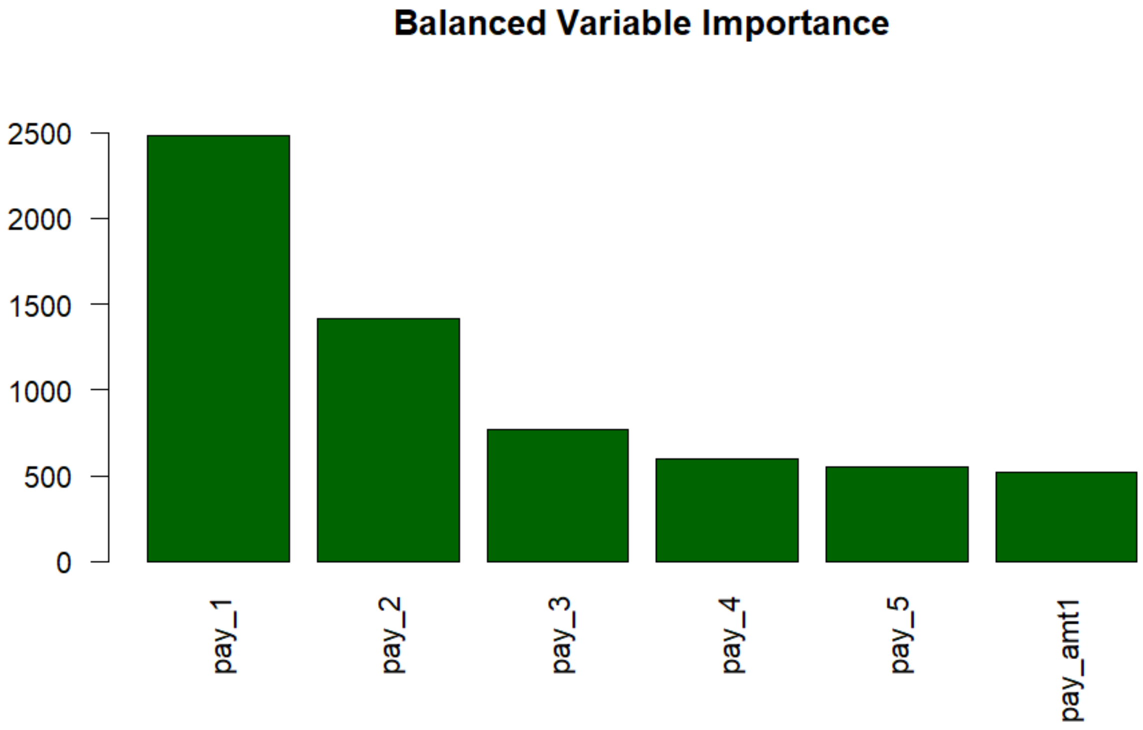
DECISION TREE



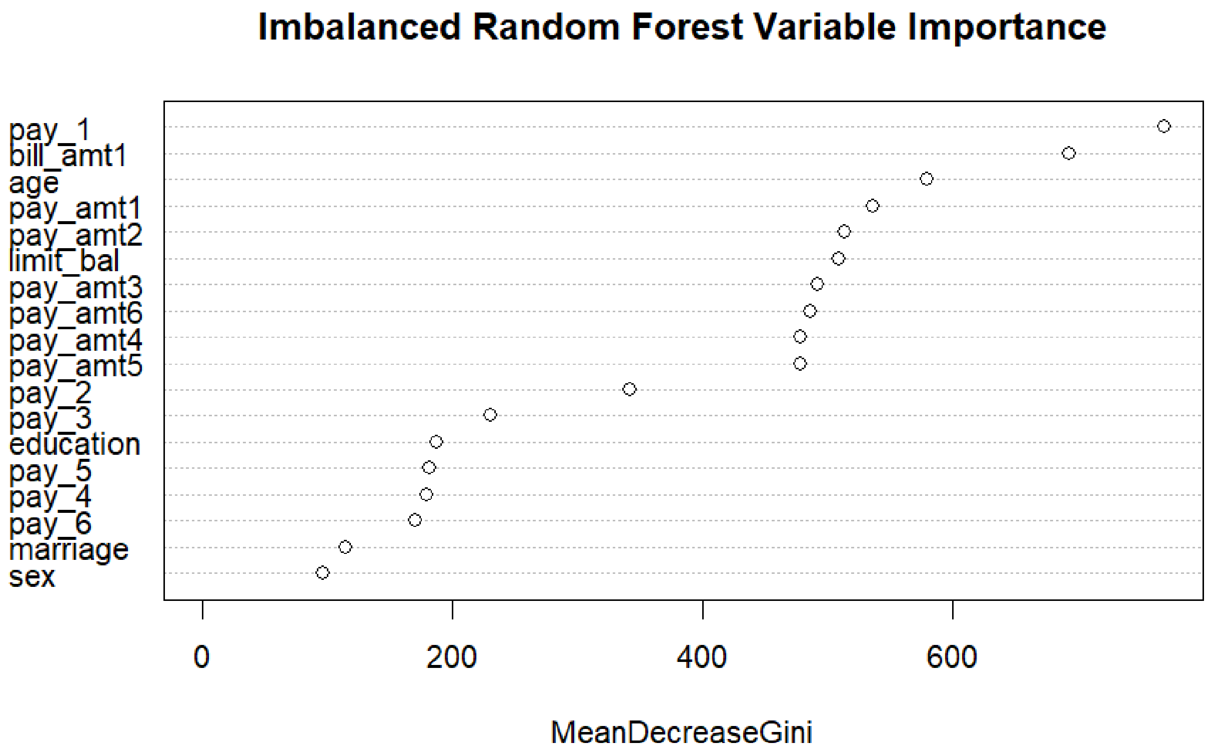


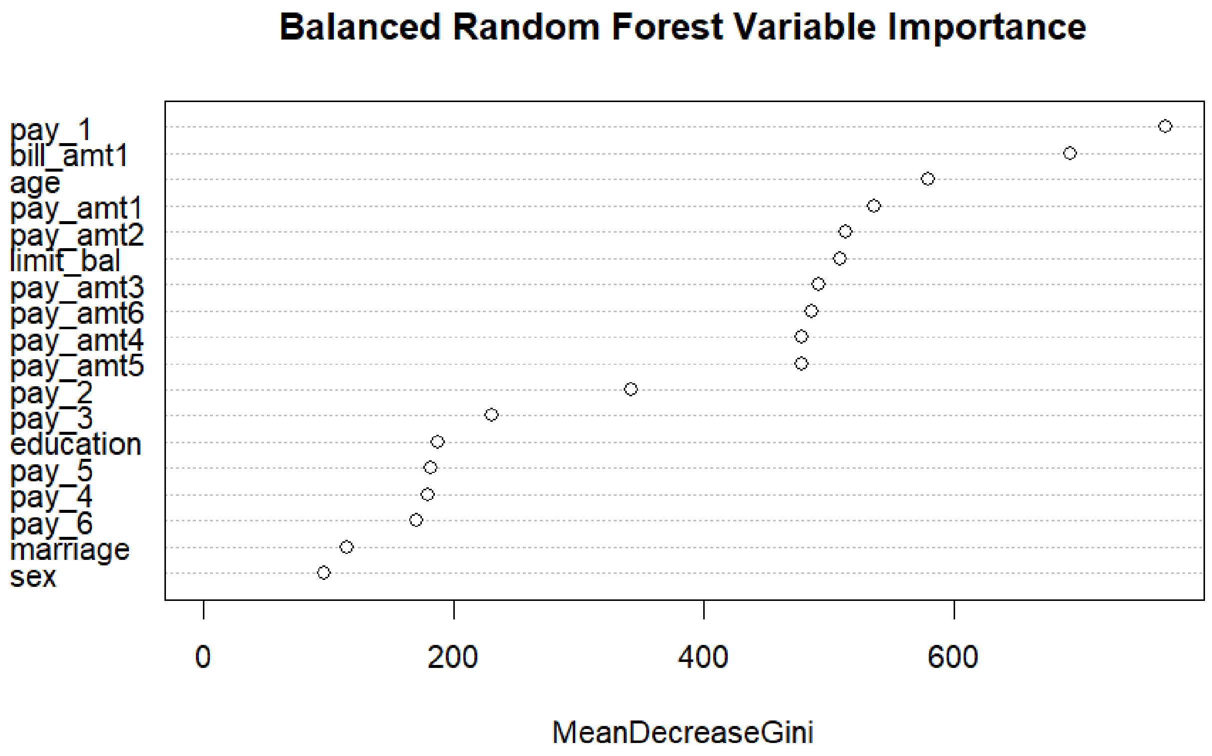
DECISION TREE





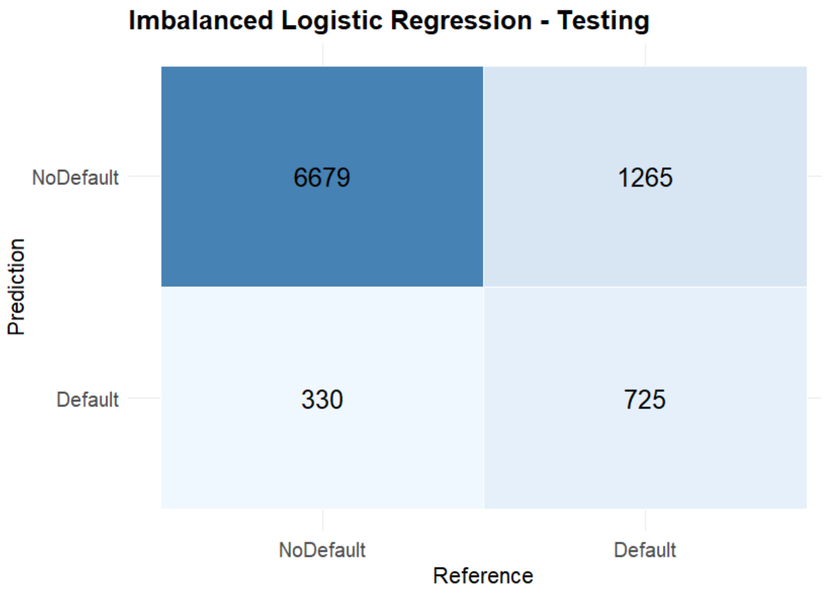
RANDOM FOREST

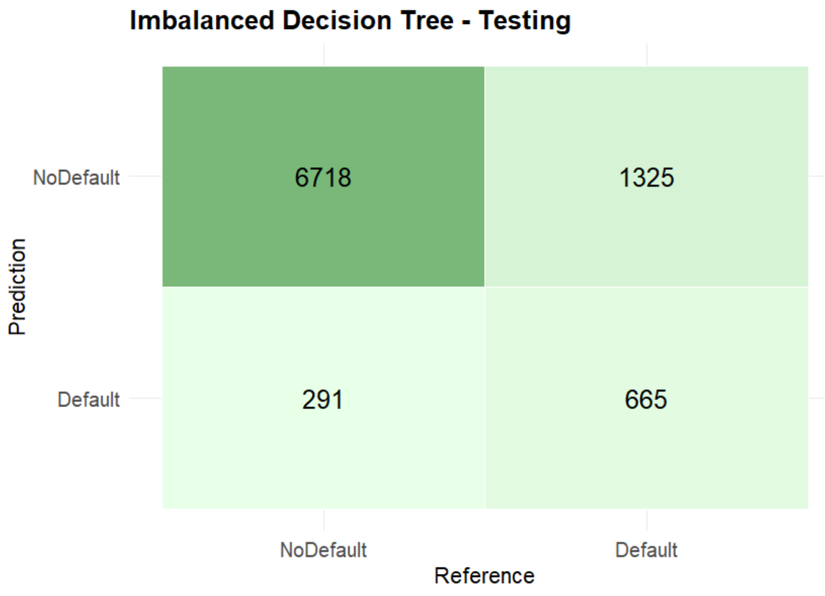


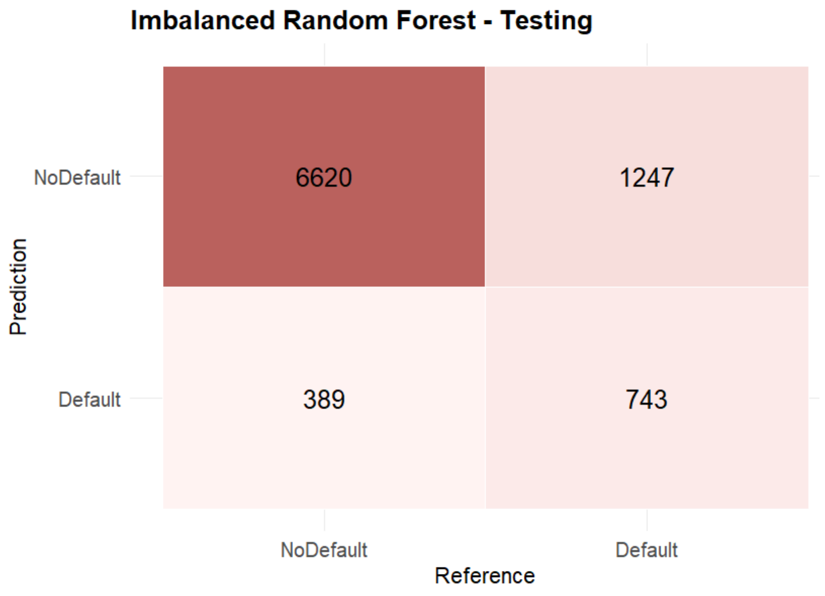


MODEL PERFORMANCE – IMBALANCED DATA

Model	Train Accuracy	Test Accuracy	Sensitivity	Specificity
Logistic Regression	0.821	0.823	0.364	0.953
Decision Tree	0.819	0.820	0.334	0.959
Random Forest	0.820	0.818	0.373	0.945

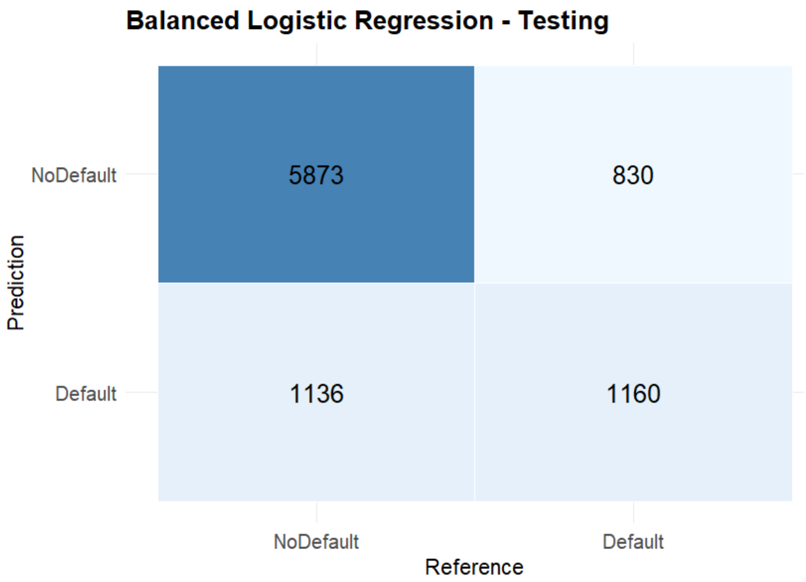


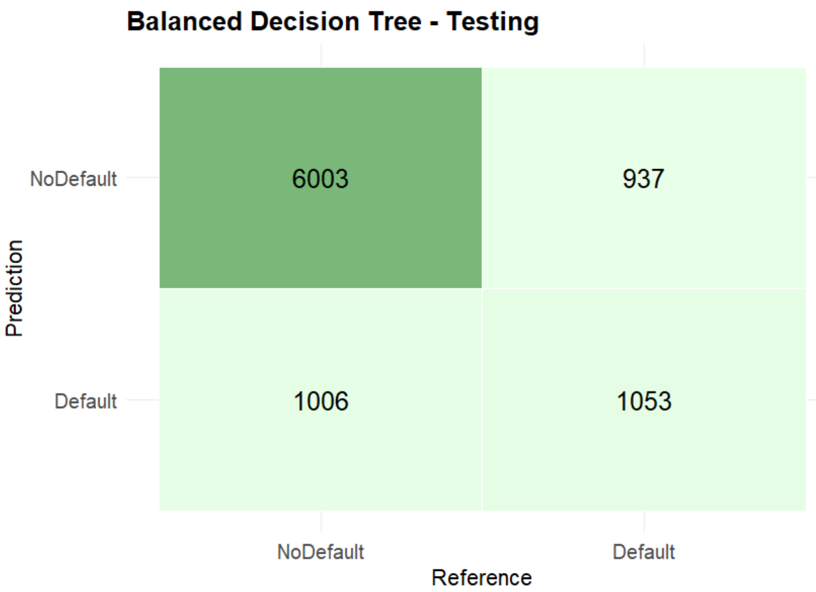


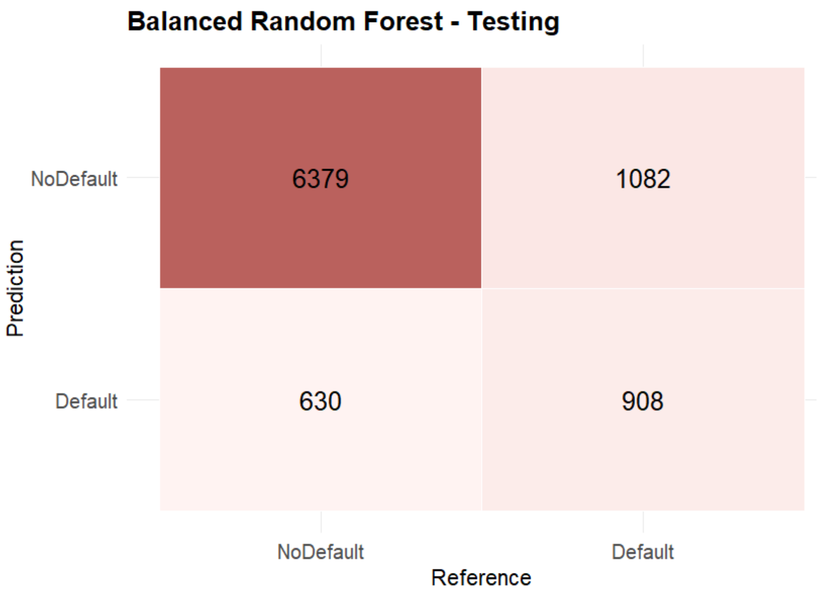


MODEL PERFORMANCE – BALANCED DATA

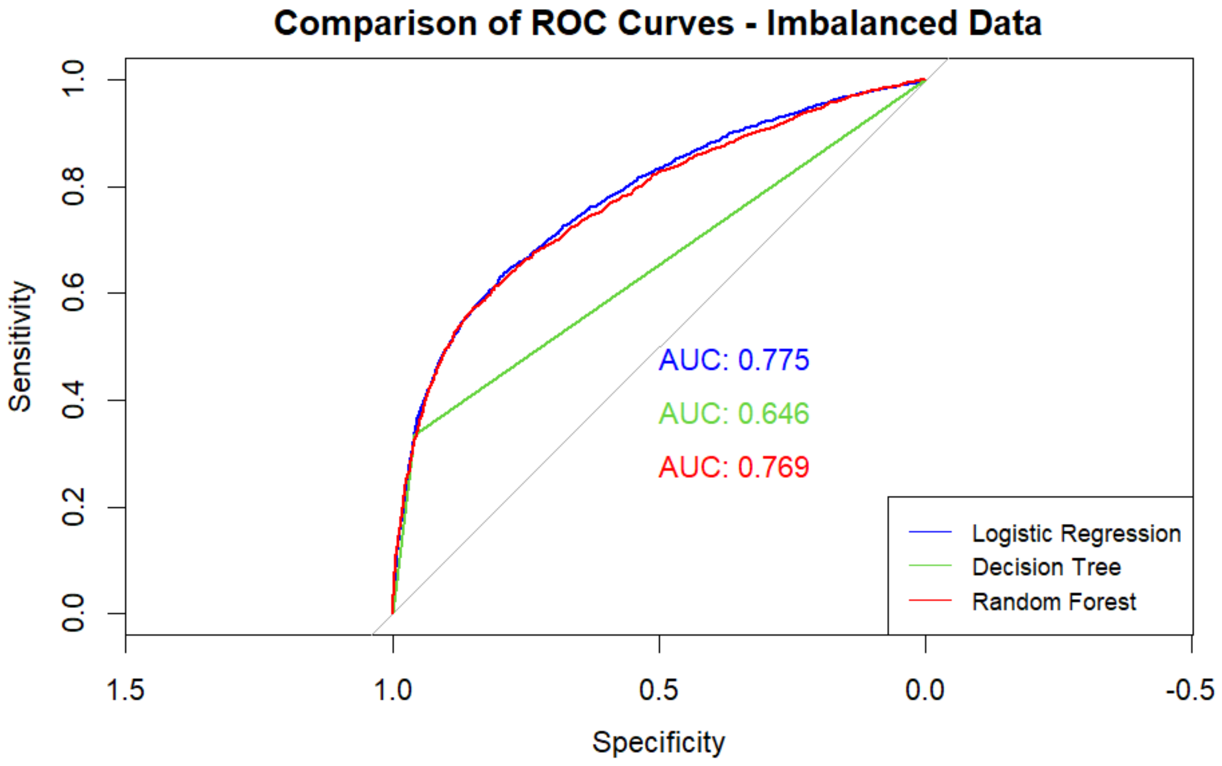
Model	Train Accuracy	Test Accuracy	Sensitivity	Specificity
Logistic Regression	0.779	0.782	0.583	0.838
Decision Tree	0.778	0.784	0.529	0.857
Random Forest	0.992	0.810	0.456	0.910

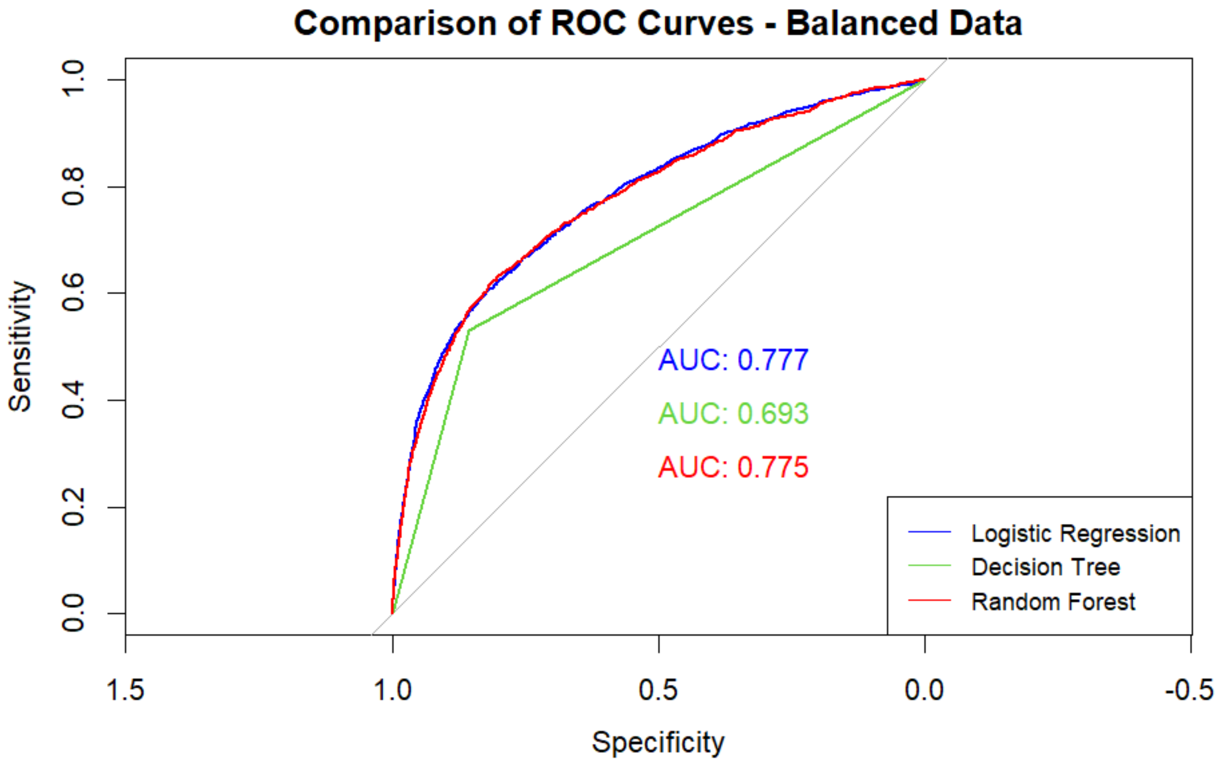






MODEL PERFORMANCE – ROC CURVES





MODEL COMPARISON SUMMARY

- **Best test accuracy:** Random Forest (Imbalanced) — 81.8%
- **Best fairness (sensitivity):** Decision Tree (Balanced) and Logistic (Balanced)
- **Best AUC:** Random Forest (Balanced & Imbalanced) — ~0.77–0.78
- **Most interpretable:** CART Decision Tree

CONCLUSIONS

- Data shows repayment history (PAY variables) is the strongest predictor of default.
- Logistic Regression provides a useful baseline and has the best sensitivity score.
- Decision Trees give clear interpretability but limited depth of patterns.
- Random Forest models achieve the best overall accuracy and AUC, but lower sensitivity.
- Balanced versions of models significantly improve detection of Default cases.
- Final takeaway: **Logistic Regression (Balanced)** offers the most reliable and fair default prediction.

FUTURE WORK

- Incorporate additional financial features such as income, spending patterns, and transaction history.
- Explore advanced models (XGBoost, Gradient Boosting, Neural Networks) for improved predictive power.
- Try SMOTE or hybrid resampling to better handle class imbalance.
- Perform feature selection or dimensionality reduction to simplify models.
- Add explainability tools (SHAP values) to interpret complex models.
- Validate performance with cross-validation or time-based splits.


THANK YOU