Predicting Payment Default Status of Bank Customers in Taiwan

ASDS 6303 Group 9: Phuong Trinh, Kyra Stolarski

Abstract

Credit card default prediction is a critical task for financial institutions because early identification of high-risk clients helps reduce financial losses and strengthen credit risk strategies. This project uses the UCI Default of Credit Card Clients dataset containing 30,000 customer records, including demographics, credit limits, repayment history, bill statements, and payment amounts. Using extensive preprocessing—including recoding factors, cleaning headers, removing duplicates, addressing class imbalance, and reducing multicollinearity—we developed several predictive models with an emphasis on Decision Trees and comparisons to Logistic Regression and Random Forests. Exploratory Data Analysis revealed strong relationships between repayment status (PAY variables) and default behavior. Results show that the Balanced Logistic Regression model provides the best trade-off between accuracy, AUC, and sensitivity, making it the most reliable option for detecting defaulting clients.

Introduction

Predicting whether a credit card customer will default in the following month is a fundamental challenge in financial risk management. Accurate models allow banks to adjust credit policies, improve approval decisions, minimize losses, and allocate resources efficiently. In this project, we analyze the UCI Default of Credit Card Clients dataset, which is widely used for financial modeling research due to its rich mix of demographic factors, payment behaviors, financial limits, and historical repayment status.

Building on standard data mining techniques, our objective was to construct and evaluate multiple predictive models—specifically Logistic Regression, CART Decision Trees, and Random Forests—and compare performance under both imbalanced and balanced training conditions. The analysis highlights key predictors such as repayment history (PAY_1–PAY_6) and bill/payment amounts. By combining statistical modeling with interpretability and robustness checks, this project aims to identify which model best supports real-world credit risk assessment.

Dataset Overview

The dataset employed in this study is the UCI Default of Credit Card Clients dataset, a widely recognized benchmark for credit risk modeling. It comprises 30,000 observations and 23 predictor variables that capture a broad range of demographic, financial, and behavioral characteristics, including customer age, gender, education level, marital status, credit limit, monthly repayment status, bill statement amounts, and payment amounts. Descriptive statistics reveal that the customer base is predominantly female, university-educated, and either single or married, with ages most commonly ranging from 30 to 41 years and a median age of 34.

Financial variables display considerable variability: the average credit limit is approximately 167,000, although its distribution is highly right-skewed due to a small fraction of clients holding exceptionally high limits. Bill and payment amount similarly show large dispersion and notable outliers, reflecting heterogeneous spending and repayment behaviors across the population. Repayment history is generally stable, with most customers paying on time or experiencing only minor delays, while more severe delinquency levels—though less frequent—play a critical role in differentiating default risk. The target variable exhibits a moderate class imbalance, with approximately 78% of clients categorized as non-defaulters and 22% as defaulters. The dataset contains no missing values; however, several financial variables display significant multicollinearity, necessitating careful preprocessing and variable reduction to ensure model stability and interpretability.

Data Preprocessing

To clean and preprocess the data, we first transformed the column names to have a consistent naming scheme, then removed the unique ID column as it would not provide any insight into the data. Our data did not have any missing or invalid values and did not need to be scaled since all the numerical data was all in the same dollar amount scale. Some of our numeric attributes, such as credit limit and age, were right skewed with outliers. However, we kept the outliers in the dataset without transformation as the outliers were not errors and were representative of the population. The models we chose to test are resistant to outliers, particularly the random forest model.

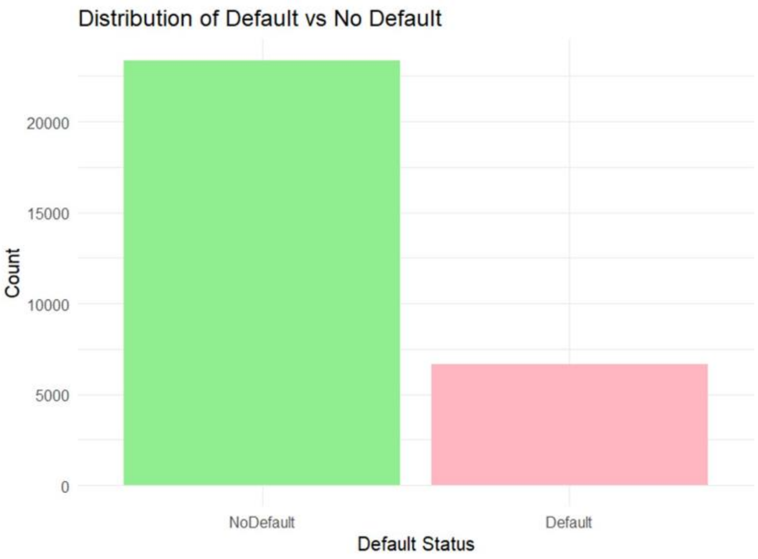
After cleaning the data, we encoded the categorical variables. Attributes such as sex and marriage status were nominally encoded, while the ordered attributes that represented the payment status were encoded ordinally. Then we calculated the correlation between each numeric value to determine if there was any multicollinearity.

Exploratory Data Analysis

A series of exploratory analyses were conducted to understand the distributional properties of key variables, examine relationships between predictors and the default outcome, and identify patterns relevant for model construction. Visualizations focused on credit limits, demographic characteristics, repayment history, and financial activity across billing cycles.

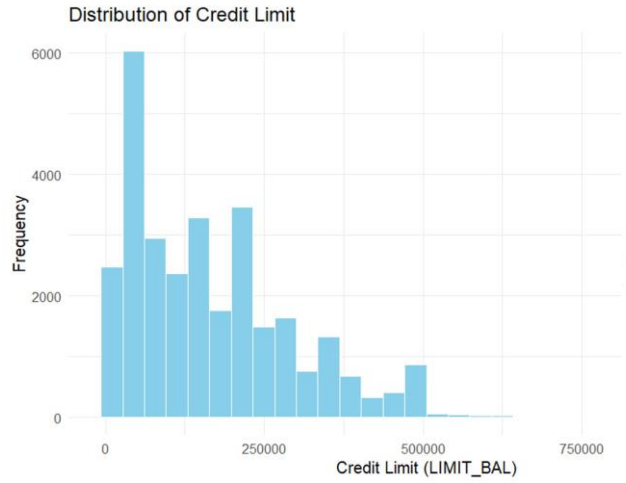
Distribution of Default Outcomes

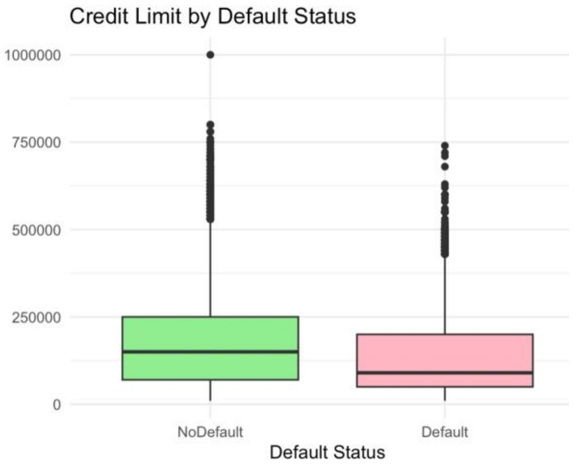
The bar chart of the target variable shows a clear class imbalance, with approximately 78% of clients labeled NoDefault and 22% labeled Default. This imbalance indicates that most customers remain current on their payments and highlights the need for balanced training methods during model development.



Credit Limit Patterns

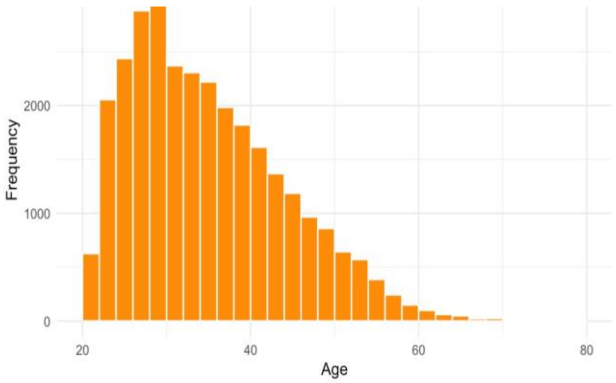
The distribution of LIMIT_BAL is strongly right skewed, indicating that while most customers have relatively modest credit limits, a small subset holds extremely high limits. A boxplot comparison between defaulting and non-defaulting groups shows that defaulting clients tend to have lower average credit limits, suggesting financial capacity may play a role in repayment behavior.

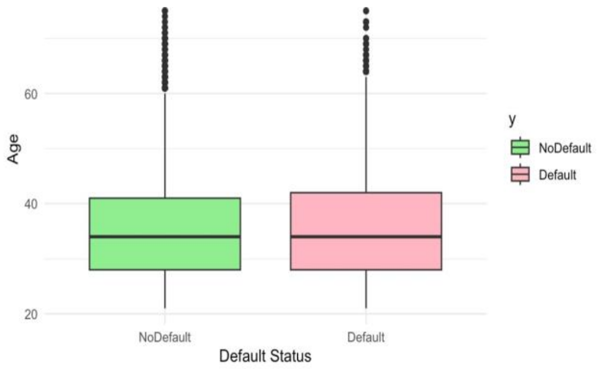




Age Characteristics

The age distribution indicates that the majority of clients fall between 25 and 40 years old, with a median near the mid-thirties. When segmented by default status, age differences become minimal; both groups display similar central tendencies and spread. Thus, age alone is not a strong differentiator in default behavior.



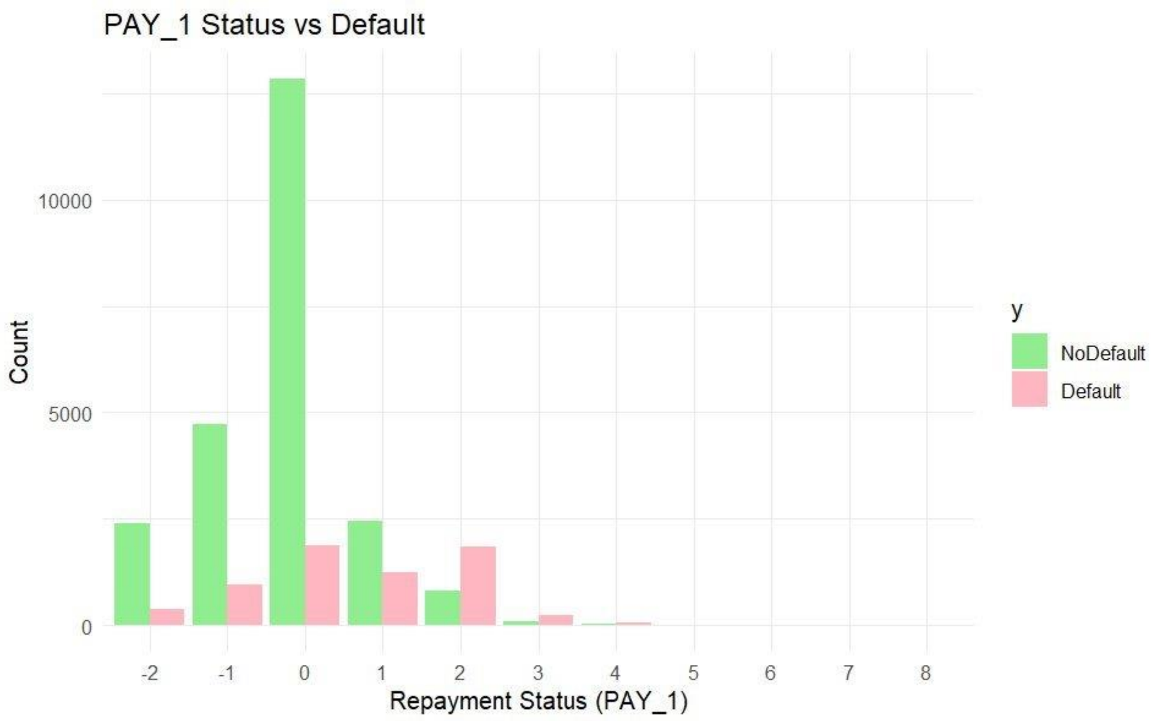


Repayment History (PAY Variables)

Repayment history is one of the most informative predictors in the dataset. The PAY_0 variable, which reflects the most recent month's repayment status, exhibits a clear relationship with default:

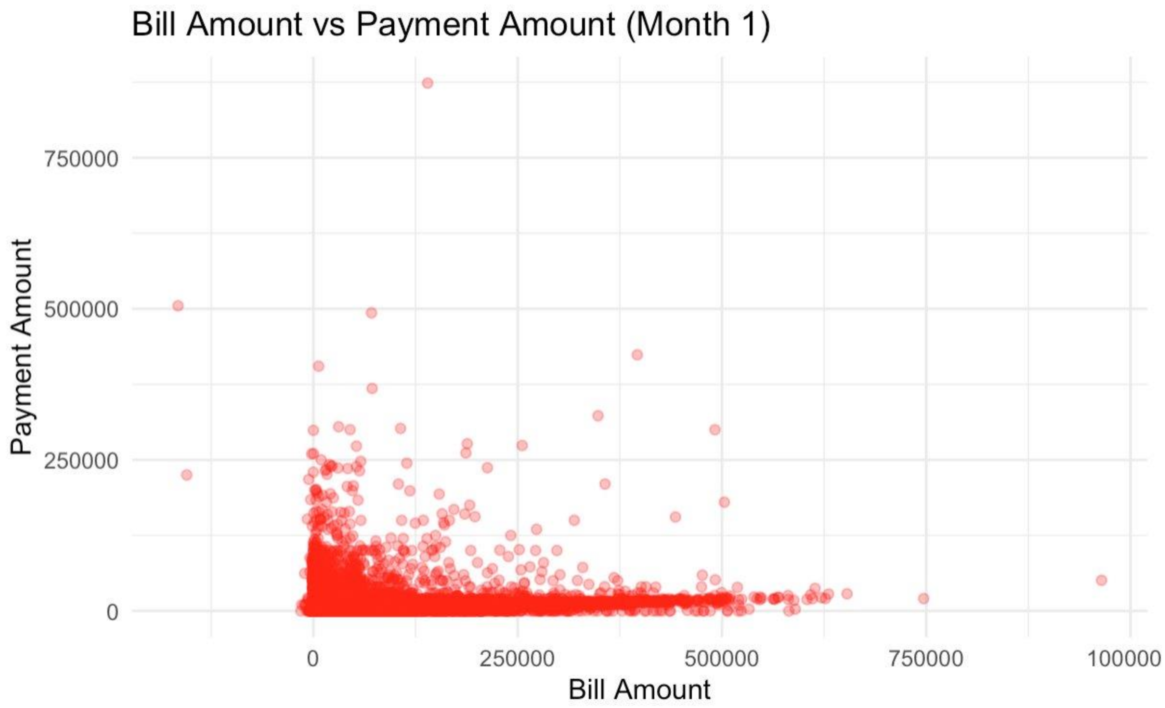
PAY_1 values of -2, -1, or 0 → low default frequency
PAY_1 values of 1 or higher → large increase in default probability

This pattern indicates that clients with recent delays are at substantially higher risk, aligning with findings in the literature and reflected in the splits of the Decision Tree model.



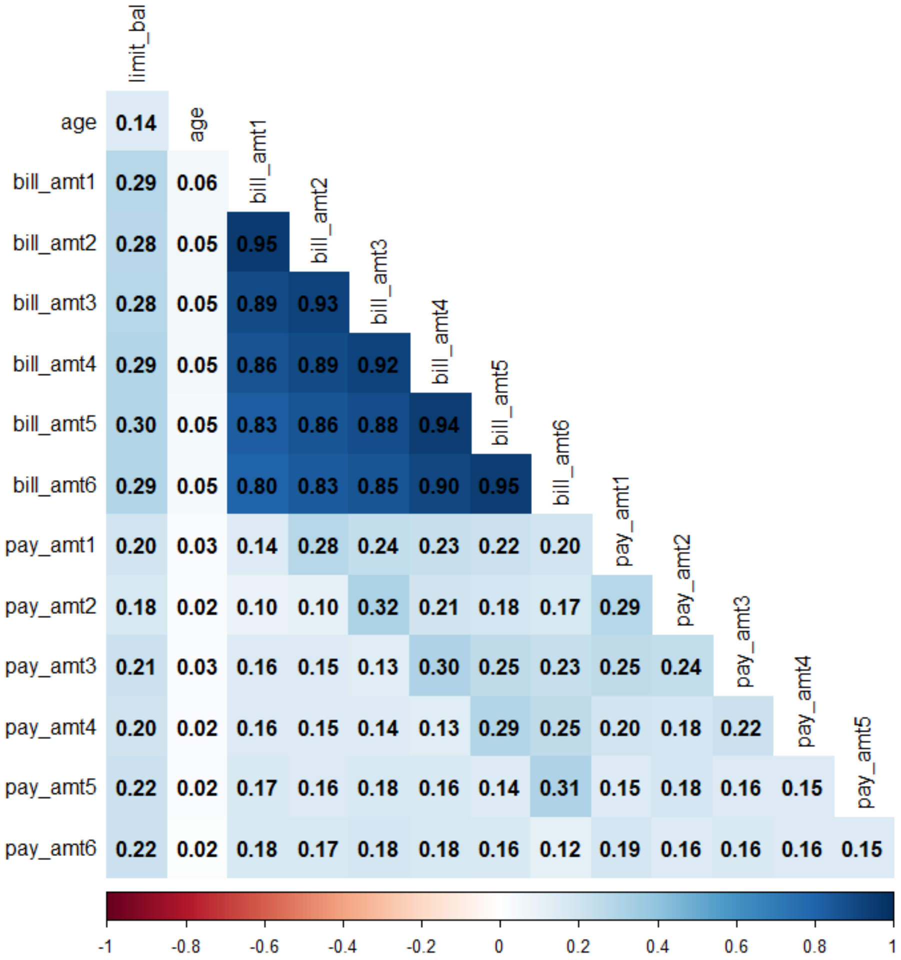
Financial Behavior Across Billing Cycles

A scatter plot comparing BILL_AMT1 to PAY_AMT1 reveals a positive trend—higher bill statements are typically accompanied by higher payment amounts. However, the plot also shows substantial variability and the presence of extreme outliers, indicating diverse financial behaviors among customers.



Correlation Heatmap

The correlation heatmap highlights clear structural patterns among the numeric variables. The BILL_AMT1–BILL_AMT6 variables display very strong positive correlations, forming a tight cluster that reflects the sequential and cumulative nature of monthly billing cycles. In contrast, the PAY_AMT variables show only weak-to-moderate correlations with one another, indicating greater variability in monthly repayment behaviors. The figure also shows that LIMIT_BAL has only modest correlation with bill amounts and repayment variables, suggesting that while credit limit influences financial capacity, it does not strongly mirror short-term billing or payment patterns. Given the exceptionally high inter-correlations among the BILL_AMT variables, several of these predictors were removed during preprocessing to reduce multicollinearity, prevent redundant information from inflating model complexity, and improve overall model stability.



	limit_bal	age	bill_amt1	bill_amt2	bill_amt3	bill_amt4	bill_amt5	bill_amt6	pay_amt1	pay_amt2	pay_amt3	pay_amt4	pay_amt5
age	0.14												
bill_amt1	0.29	0.06											
bill_amt2	0.28	0.05	0.95										
bill_amt3	0.28	0.05	0.89	0.93									
bill_amt4	0.29	0.05	0.86	0.89	0.92								
bill_amt5	0.30	0.05	0.83	0.86	0.88	0.94							
bill_amt6	0.29	0.05	0.80	0.83	0.85	0.90	0.95						
pay_amt1	0.20	0.03	0.14	0.28	0.24	0.23	0.22	0.20					
pay_amt2	0.18	0.02	0.10	0.10	0.32	0.21	0.18	0.17	0.29				
pay_amt3	0.21	0.03	0.16	0.15	0.13	0.30	0.25	0.23	0.25	0.24			
pay_amt4	0.20	0.02	0.16	0.15	0.14	0.13	0.29	0.25	0.20	0.18	0.22		
pay_amt5	0.22	0.02	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	
pay_amt6	0.22	0.02	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15

Models

We discovered in our exploratory data analysis that there was an imbalance between our two target classes, No Default and Default. About 78% of the instances had a class of No Default, and about 22% of the instances had a class of Default. To handle this imbalance, we applied oversampling to the training dataset and trained the models on both imbalanced and balanced training sets.

Since our models are working with imbalanced data, we prioritized the model's sensitivity scores in addition to their accuracy metrics. While it is important that the models make accurate classifications, the accuracy score can be biased by imbalanced data. Additionally, the model's sensitivity score determines how well the model classifies a customer defaulting on their loan payment. If a model has a low sensitivity, they may predict a customer as not defaulting on their loan when they really do default, leading to the bank losing money.

Logistic Regression

The logistic regression models were built using all the attributes that were not highly correlated since the model performed significantly worse when additional attributes were removed. This could be because while some attributes may not have been statistically significant, they may have informed the model on some underlying relationships within the data.

Imbalanced: Statistically Significant Attributes

- limit_bal
- sexFemale
- educationOthers
- marriageSingle
- age
- bill_amt1
- pay_amt1
- pay_amt2
- pay_amt3

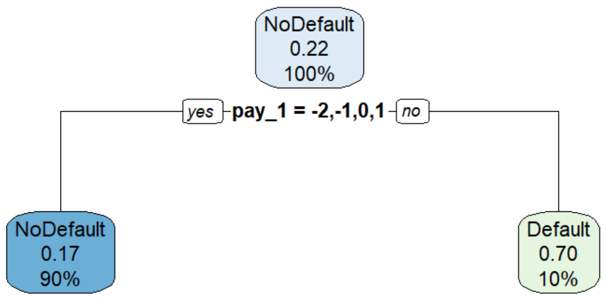
Balanced: Statistically Significant Attributes

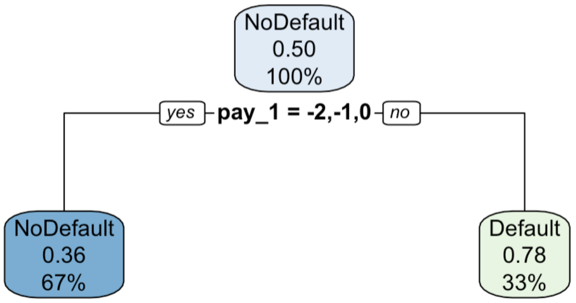
- limit_bal
- sexFemale
- educationOthers
- marriageSingle
- age
- bill_amt1
- pay_amt1
- pay_amt2
- pay_amt3
- pay_amt4
- pay_amt5
- pay_amt6

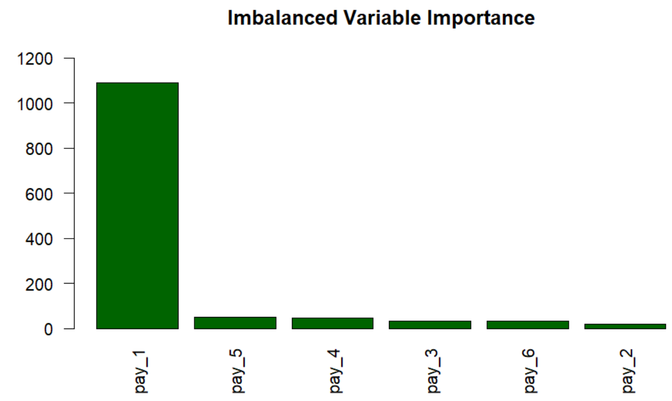
Both the imbalanced and balanced logistic regression models found the attributes limit_bal, sexFemale, educationOthers, marriageSingle, age, bill_amt1, pay_amt1, pay_amt2, and pay_amt3 to be statistically significant. However, the balanced model differs by also finding attributes pay_amt4, pay_amt5, and pay_amt6 statistically significant as well.

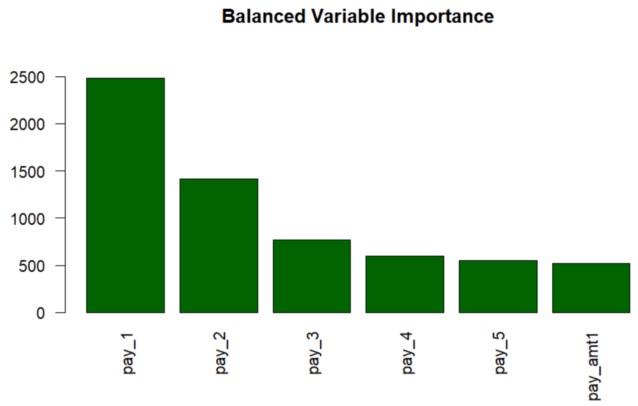
CART Decision Tree

We developed a CART decision tree model because a decision tree with the CART algorithm applied can work with mixed data types (numeric and categorical), is robust to outliers, and is easily interpretable. The complexity parameter (cp) was set to 0.01 to prevent the model overfitting the data. This means that any potential split would only go through if it increased the model's performance by at least 0.01.


Imbalanced Decision Tree Model


Balanced Decision Tree Model

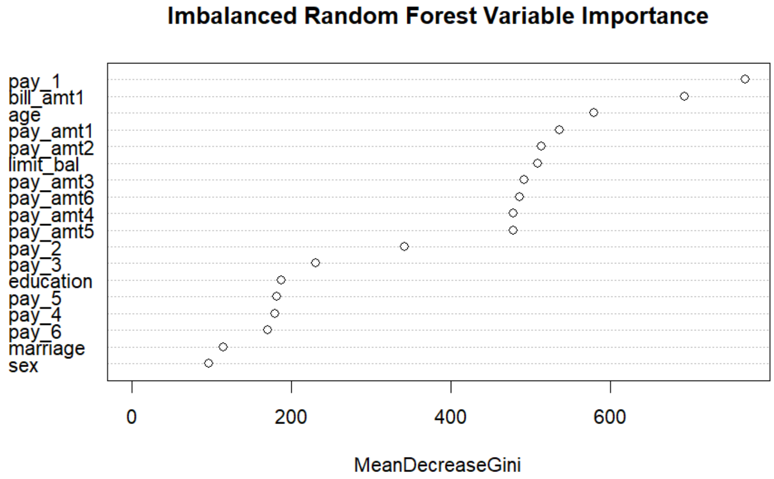


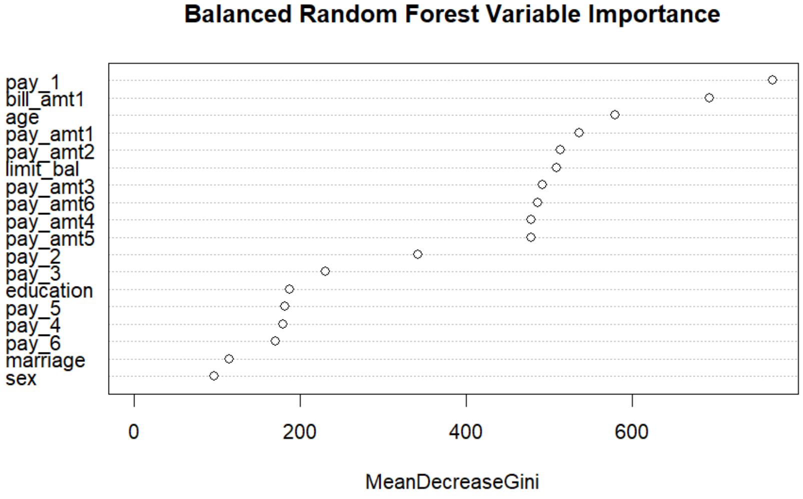


The balanced and imbalanced decision tree models both only selected one splitting point, pay_1, to classify the data. Both models also ranked pay_1 with the highest variable importantce, though the balanced model places moderate importance on other payment variables.

Random Forest

The random forest models were created with 500 trees and set to try a random subset of 4 variables at each split. This process helps to reduce the correlation between the individual decision trees within the random forest, decreasing the model's variability and increasing its predictive power.

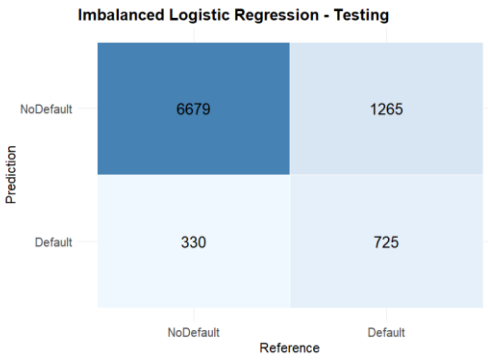


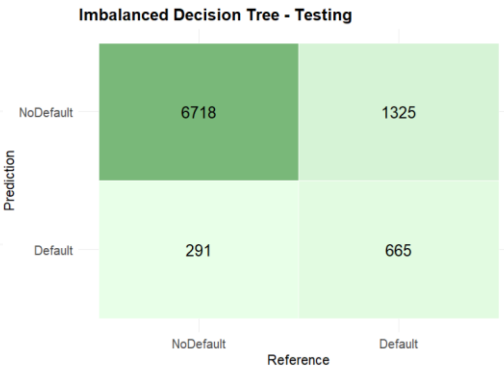


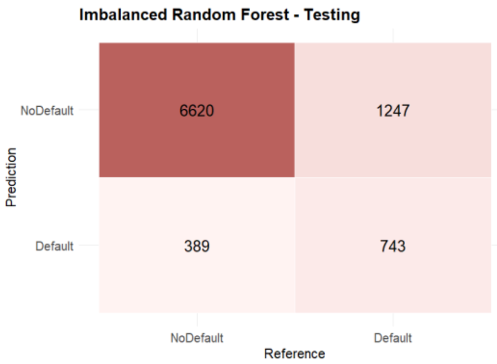
The imbalanced and balanced random forest models ranked variable importance similarly, with pay_1 causing the highest reduction in Gini impurity within the model. Additional attributes of note are bill_amt1, age, pay_amt1, and pay_amt2.

Results - Imbalanced Training Set

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



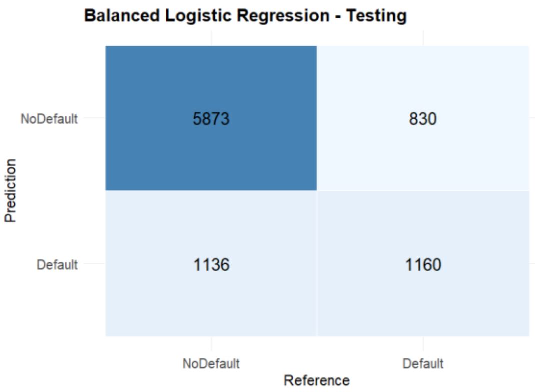


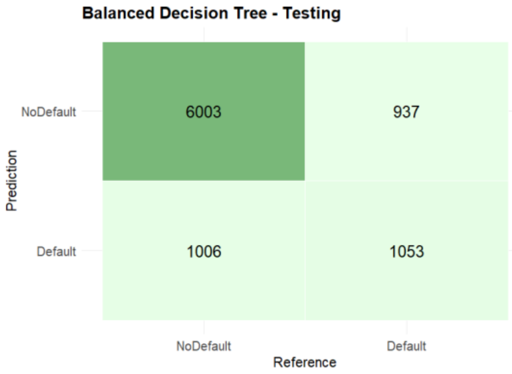


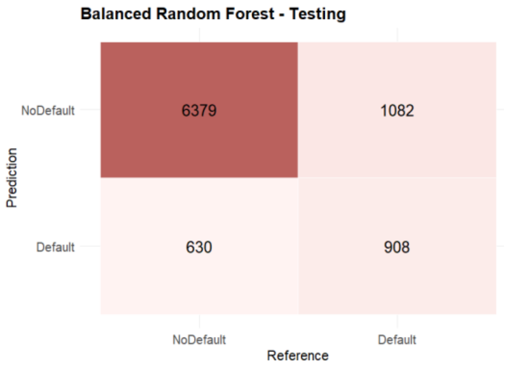
Although all three models had relatively high accuracy scores, they had significantly low sensitivity scores, indicated that they primarily predicted the negative class, No Default, for the majority of test instances.

Results - Balanced Training Set

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

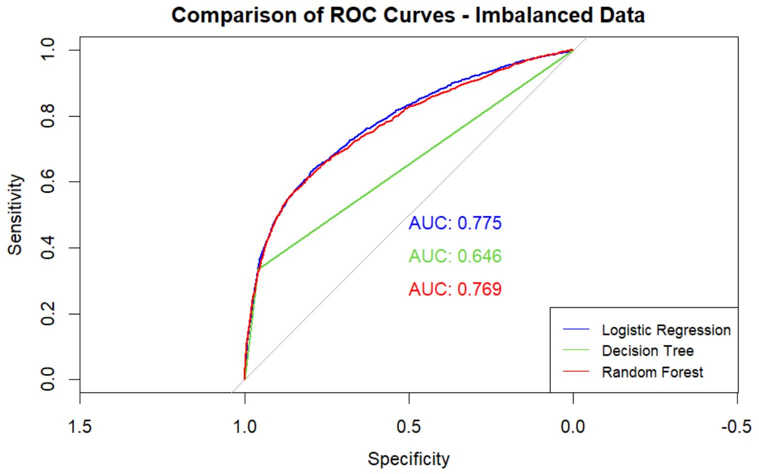


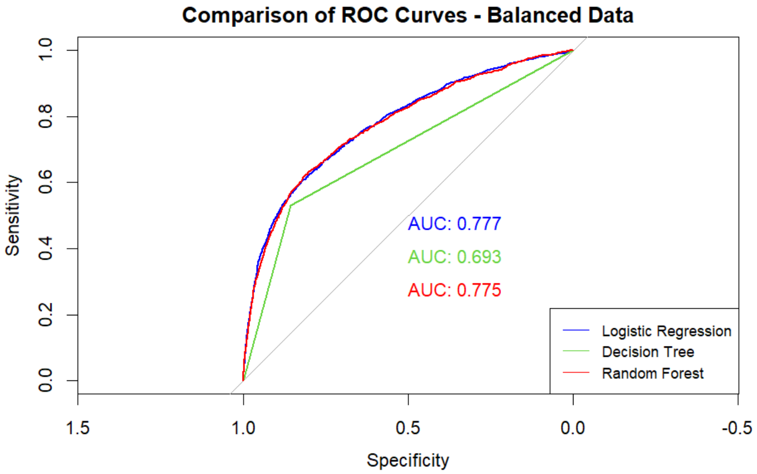




The test accuracy for all three models decreased, however their sensitivity scores all increased. The random forest model is overfitted to the training data, but it still had the highest test accuracy at 0.810. Overall, the models are more able to accurately predict both classes, however, it is of note that although the sensitivity scores increased, the highest score is still only 0.583. This indicates that these models still struggle with predicting the positive class, Default.

ROC Curves





The ROC curves were generally similar for all three models for both the imbalanced and balanced training sets. The decision tree model had the largest AUC change, increasing from 0.646 to 0.693. The logistic regression and random forest models had similar AUC metrics.

Conclusions

This project evaluated multiple data-mining models to predict credit card default using a large real-world dataset collected from a major bank in Taiwan. Through extensive preprocessing, exploratory analysis, and comparison of several modeling approaches, repayment history (the PAY variables) consistently emerged as the strongest and most reliable indicator of default risk. Logistic Regression provided an effective and interpretable baseline and achieved the highest sensitivity among the balanced models, making it especially useful for correctly identifying clients at risk of default. Decision Trees offered clear interpretability through intuitive repayment-based splits but were limited in capturing more complex patterns. Random Forest models achieved the highest overall predictive performance, producing the strongest accuracy and AUC under both imbalanced and balanced conditions, though with slightly lower sensitivity.

The results also show that addressing class imbalance significantly improves fairness and model performance. Balanced versions of the models—particularly Logistic Regression—detected substantially more default cases while maintaining stability and interpretability. Overall, while Random Forest delivers the best predictive power, Balanced Logistic Regression offers the most reliable and fair approach for early detection of defaulting clients. These findings underscore the importance of combining accuracy, sensitivity, and interpretability when developing credit-risk models for financial institutions.

Future Work

Future work can build on this analysis in several meaningful directions. Incorporating additional financial attributes—such as income level, spending habits, or long-term transaction patterns—may enhance predictive performance by capturing a more complete picture of customer behavior. Exploring more advanced machine-learning methods, including XGBoost, Gradient Boosting, and Neural Networks, could uncover nonlinear relationships not captured by traditional models. More sophisticated resampling strategies such as SMOTE or hybrid approaches may further improve sensitivity in the presence of class imbalance.

Feature selection or dimensionality-reduction techniques could help reduce multicollinearity and simplify model structure, improving stability and interpretability. For complex ensemble methods, explainability tools such as SHAP values could provide deeper insights into model decisions—an important requirement in financial risk management. Finally, validating results with k-fold cross-validation or time-based splits would strengthen the generalizability of the models and better reflect operational conditions in real banking environments.

References

Yeh, I-Cheng. 2016. “UCI Machine Learning Repository.” Archive.ics.uci.edu. January 25, 2016. https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients.