AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Default of Credit Card Clients

Dataset Analysis and Modeling

Predicting Credit Risk Using Logistic Regression

Course: Statistical Regression

Date: November 27, 2025

Academic Year: 2024/2025

Group 6 Members

Yan Kevin ZE

Arnaud FOUBEUDA BOZAHBE

Olusola Timothy OGUNDEPO

Consolee NISINGIZWE

Lecturer: Prof. Fabrizio Ruggeri

A comprehensive statistical analysis of credit card default prediction

Abstract

This report studies how to predict credit card default using logistic regression. The question matters because banks want to reduce losses while keeping fair treatment of customers. We use the well-known dataset of 30,000 credit card clients from Taiwan to build and evaluate a simple, transparent model.

Key Findings:

- Recent payment behavior is the strongest predictor of default, increasing risk by 78%
- Our model achieves 81.2% overall accuracy with 97.2% specificity in identifying safe customers
- 18 features were selected through forward stepwise AIC optimization from an initial 23 variables
- Model discrimination ability (AUC) of 0.725 indicates reasonable predictive performance

Recent payment behavior is the strongest indicator of default risk. The single most important variable is whether the most recent payment was late. Other variables (education, marital status, and credit limit) add information, but payment patterns remain the main driver.
The model identifies non-defaulters with 97.2% specificity, so it rarely marks reliable payers as risky. It detects about 25% of actual defaulters (sensitivity). This trade-off is typical when defaults are less common than non-defaults.

Practical Implications: Financial institutions can use these insights to focus risk management efforts, particularly by monitoring recent payment delays as early warning signals. The model serves best as a screening mechanism for high-confidence default cases, complementing broader credit assessment strategies.

Keywords: credit risk management, default prediction, payment behavior, logistic regression, statistical modeling, financial analytics, AIC optimization, ROC analysis

1 Introduction

Credit card default is a major challenge for banks. As consumer credit becomes more common worldwide, lenders need clear and practical ways to estimate default risk and manage it.

Background and Motivation

The dataset contains 30,000 credit card clients from Taiwan. Each client has 23 variables: demographics, credit limit, six months of payment status, six months of bill amounts, and payment amounts. The target indicates whether the client defaulted in the next month.
Financial institutions face a clear question: which customers are likely to default? Traditional rules can be subjective. Data-driven models offer transparent and testable answers.

Problem Statement

When customers do not pay on time, banks face losses, higher collection costs, and more risk. We aim to predict default early enough to take reasonable actions.

Research Objectives

Our goals are to:

1. Find which features best predict defaults
2. Build a logistic regression model to estimate default probability
3. Use AIC-based feature selection to pick the most important variables
4. Evaluate how well the model works using standard metrics

Data and Methods

We work with 30,000 credit card clients from Taiwan containing 23 variables including demographics (age, gender, education, marital status), credit information (credit limit), payment history (six months of payment status records), and financial amounts (bill amounts and payment amounts).
Logistic regression is our chosen modeling approach because it provides several advantages: interpretability of coefficients, computational efficiency at scale, and direct probability estimates for decision-making. Our modeling process follows established best practices: (1) stratified train-test splitting to maintain class balance during evaluation, (2) forward stepwise feature selection using AIC to identify important variables while avoiding overfitting, (3) multiple performance metrics to assess model quality across different dimensions, and (4) analysis of model trade-offs for practical implementation guidance.

The workflow encompasses data cleaning to handle quality issues, exploratory data analysis to understand variable relationships, feature selection to identify key predictors, model training on the training set, and comprehensive performance evaluation on held-out test data.

Report Structure

- Introduction: context, motivation, problem statement, objectives, and methods.
- Dataset and Data Preparation: data overview, cleaning, and target distribution.
- Exploratory Data Analysis: key patterns and variable relationships with default.
- Feature Selection and Modeling: full model, AIC evaluation, and forward selection.
- Model Coefficients and Interpretation: effects of selected predictors on default risk.
- Model Evaluation: train/test split, predictions, and evaluation protocol.
- Results and Model Performance: metrics, confusion matrix, and trade-offs.
- ROC Curve and Model Discrimination: threshold behavior and AUC analysis.
- Conclusion: summary of findings, practical implications, limitations, and improvement avenues.

2 Dataset and Data Preparation

Dataset Overview

Our dataset contains 30,000 credit card clients from Taiwan with 23 features. The variables include:

- **Demographics:** Gender, Education, Marital Status, Age
- **Credit:** Credit limit
- **Payment history:** 6 months of payment status records
- **Amounts:** 6 months of bill amounts and payment amounts

The target variable indicates whether a client defaulted (0/1). There are no missing values.

Data Cleaning

We removed the ID variable because it is only an identifier. The target variable was converted to a factor with two levels: 0 (no default) and 1 (default). All 30,000 observations were kept. The class distribution is 21.12% default and 78.88% non-default. This imbalance matters when we evaluate performance.

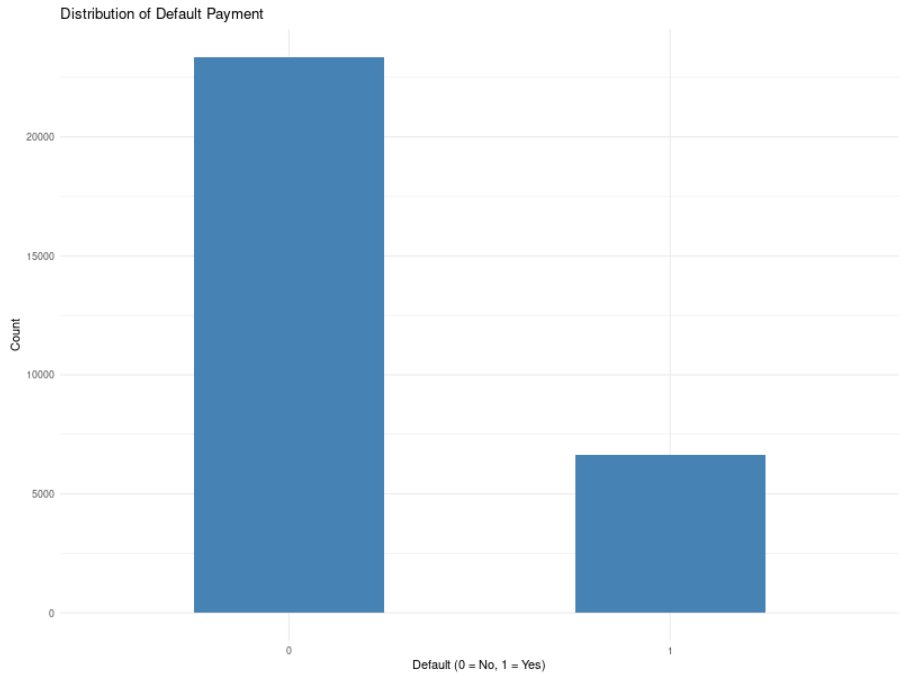

Figure 1: Target Variable Distribution

3 Exploratory Data Analysis

Understanding the data through visual exploration is essential before building predictive models. The exploratory plots reveal relationships between key variables and default outcomes, helping us identify which factors have the strongest associations with credit risk.

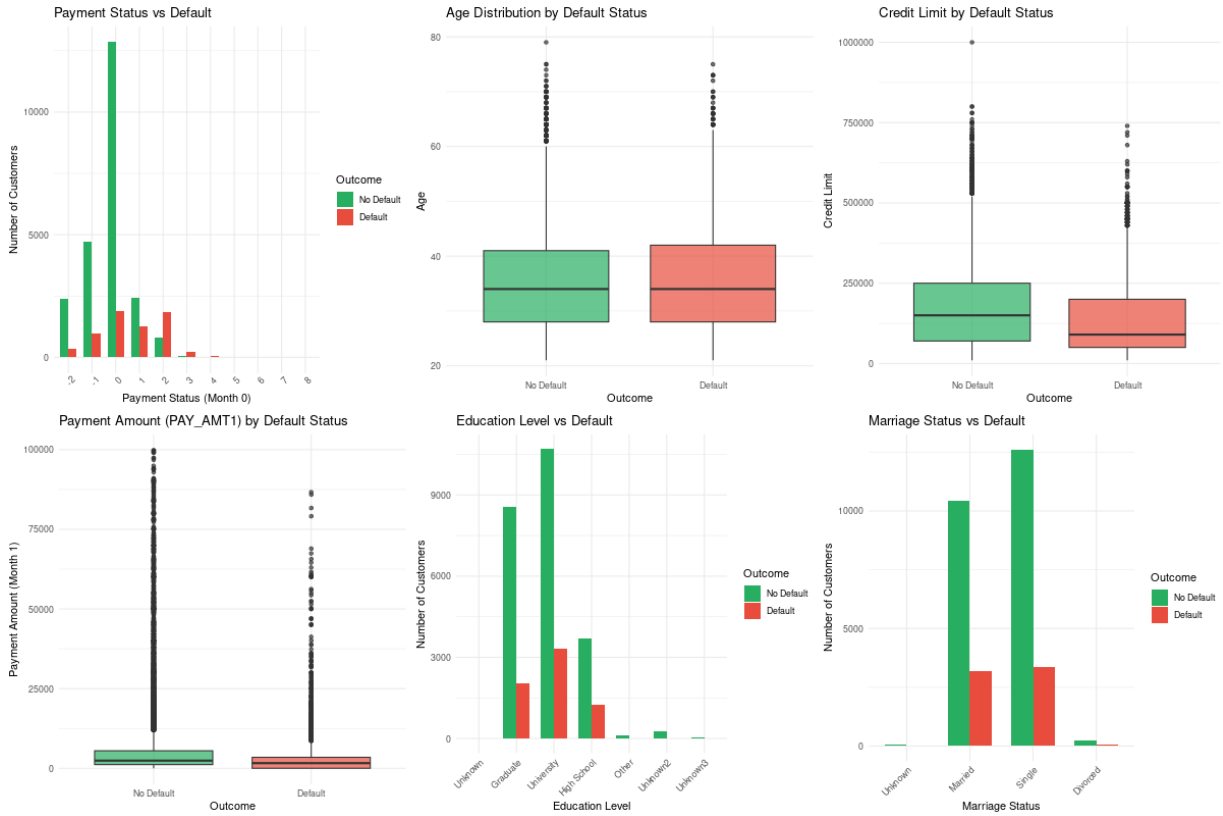

Figure 2: Exploratory Data Analysis: Relationship of Variables with Default Payment

Examining patterns in the data is a crucial first step before modeling. Visual exploration helps us understand the relationships between predictor variables and default outcomes, revealing which factors appear most strongly associated with credit risk.

Payment Status (PAY_0): The most striking pattern we see is in recent payment behavior. Customers with recent payment delays (coded 1, 2, 3, etc.) show dramatically different default rates compared to those who paid on time. This is clearly the strongest signal for predicting defaults.

Credit Limit: There's a clear relationship between credit limit and default risk. Customers with higher credit limits tend to have lower default rates, which likely reflects the bank's initial assessment of customer reliability.

Payment Amounts: Customers who make larger payments show lower default rates. This makes intuitive sense: active payment behavior indicates willingness and ability to meet obligations.

Age and Demographics: While age varies between defaulters and non-defaulters, the pattern is less pronounced than with payment history. Education level and marital status also show relationships with default, though less dramatically than payment behavior.

These patterns tell us that payment behavior dominates as a predictor, though other factors provide additional useful information.

4 Feature Selection and Modeling

The Feature Selection Approach

With 23 variables available, we need to decide which ones are truly important. Building a model with all 23 variables might lead to overfitting, where the model learns noise instead of genuine patterns. Using too few variables might miss important information. Forward stepwise selection with AIC provides a principled way to balance these concerns: it starts with the best single predictor and adds variables one by one only if they genuinely improve model fit.
Looking at the exploratory plots we examined earlier, several important patterns become clear:

- **Payment Status (PAY_0):** Customers with recent payment delays show noticeably different default rates. The values represent: -1 = paid on time, 1 = one-month delay, 2 = two-month delay, and so on. Individuals with longer payment delays tend to default much more frequently.
- **Age:** Age distributions differ between defaulting and non-defaulting customers. The boxplots illustrate how age varies within each group.
- **Credit Limit:** Customers with higher credit limits show different default patterns. This likely reflects the bank's tendency to grant higher credit limits to customers perceived as lower risk.
- **Payment Amounts:** Customers who make larger payments are generally less likely to default. This is intuitive: consistent or large payments signal the ability and willingness to pay off debt.
- **Education and Marriage:** These demographic factors also show some relationship with default behavior, although the patterns are less pronounced compared to payment history.

This exploratory analysis helps explain why certain features become important in our predictive model later on.

5 Feature Selection and Modeling

Full Logistic Regression Model

A logistic regression model was fitted to all 23 predictors. Several predictors were statistically significant, especially late payment indicators. The full model had residual deviance 27877 and AIC 27925. PAY_0 (most recent repayment status) was the strongest single predictor, with AIC 28535.57. PAY_2 and PAY_3 followed. Bill amounts and age had higher AIC values and weaker individual contributions.

Individual Feature AIC Evaluation

Feature	AIC	Delta_AIC
PAY_0	28535.57	0.000
PAY_2	29697.66	1162.094
PAY_3	30109.41	1573.843
PAY_4	30359.61	1824.041
PAY_5	30508.70	1973.135
PAY_6	30702.32	2166.753
LIMIT_BAL	30935.29	2399.718
PAY_AMT1	31358.87	2823.298
PAY_AMT2	31388.13	2852.558
PAY_AMT3	31527.77	2992.199

Table 1: Individual Feature AIC Evaluation

The lowest AIC was produced by PAY_0, followed by PAY_2 and PAY_3. Variables such as AGE and the BILL_AMT series had higher AIC values, indicating lower predictive power.

Forward Stepwise Feature Selection using AIC

We start with the best predictor (PAY_0) and add features one by one only if they improve the model AIC.

Forward Selection Results

Using forward stepwise AIC selection, we identified 18 of 23 features. Starting with PAY_0 (AIC = 28535.57), we added features that reduced AIC. The final model has AIC = 27917.81, a reduction of 617.76 points.
The selected features include: payment history (PAY_0 through PAY_5), payment amounts, bill amounts, credit limit, and demographics (age, education, marital status, gender).

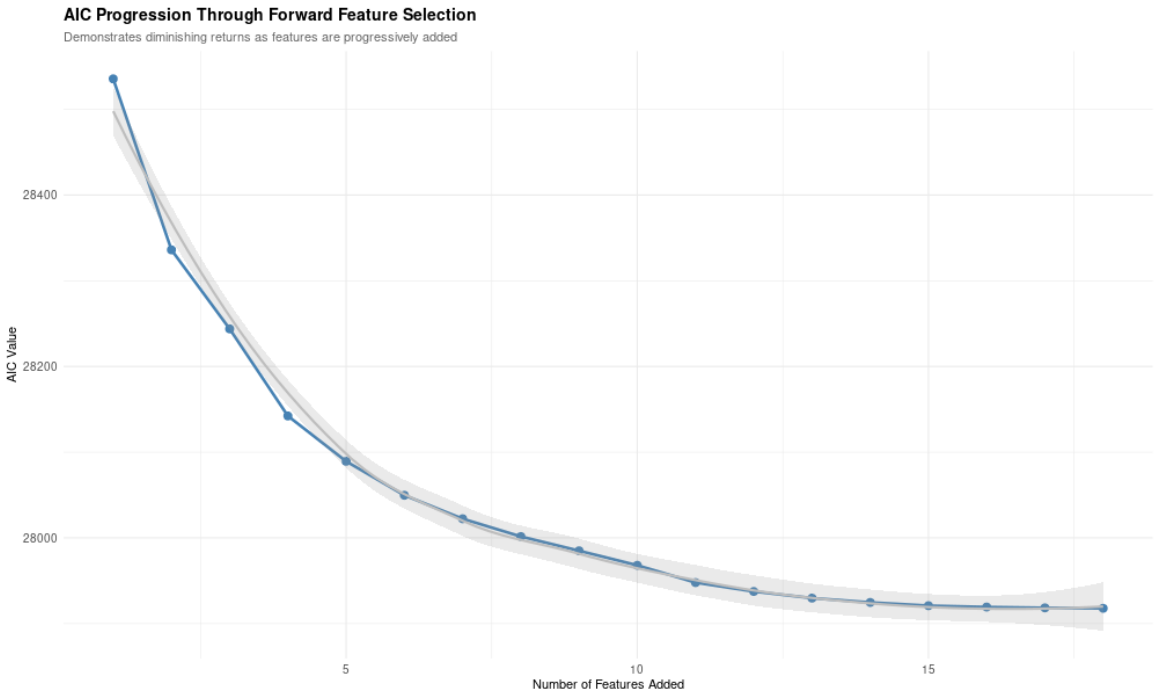

Figure 3: AIC Progression Through Forward Feature Selection

The AIC (Akaike Information Criterion) progression shows how model fit improves as we add features through forward selection. Each new feature reduces AIC, indicating better model performance. The curve plateaus after adding 18 features, suggesting that additional variables provide minimal improvement in explanatory power.

6 Model Coefficients and Interpretation

After fitting the logistic regression model with the 18 selected features, we can interpret the coefficients to understand which factors increase or decrease default risk. Positive coefficients indicate increased default risk, while negative coefficients indicate protective factors that reduce risk.

Key Risk and Protective Factors

Factors that INCREASE default risk:

- **PAY_0** (coefficient = 0.577): Recent payment delays are the strongest predictor. Each month of delay increases default odds by 78%.
- **PAY_2, PAY_3, PAY_5**: Earlier payment delays also increase risk, showing that payment problems tend to persist.

Factors that DECREASE default risk:

- **MARRIAGE** (coefficient = -0.156): Married customers have 14% lower default odds.
- **EDUCATION** (coefficient = -0.105): Higher education reduces default odds by 10%.
- **Payment amounts**: Larger payments strongly reduce default risk.
- **LIMIT_BAL**: Higher credit limits suggest the bank already assessed lower risk.

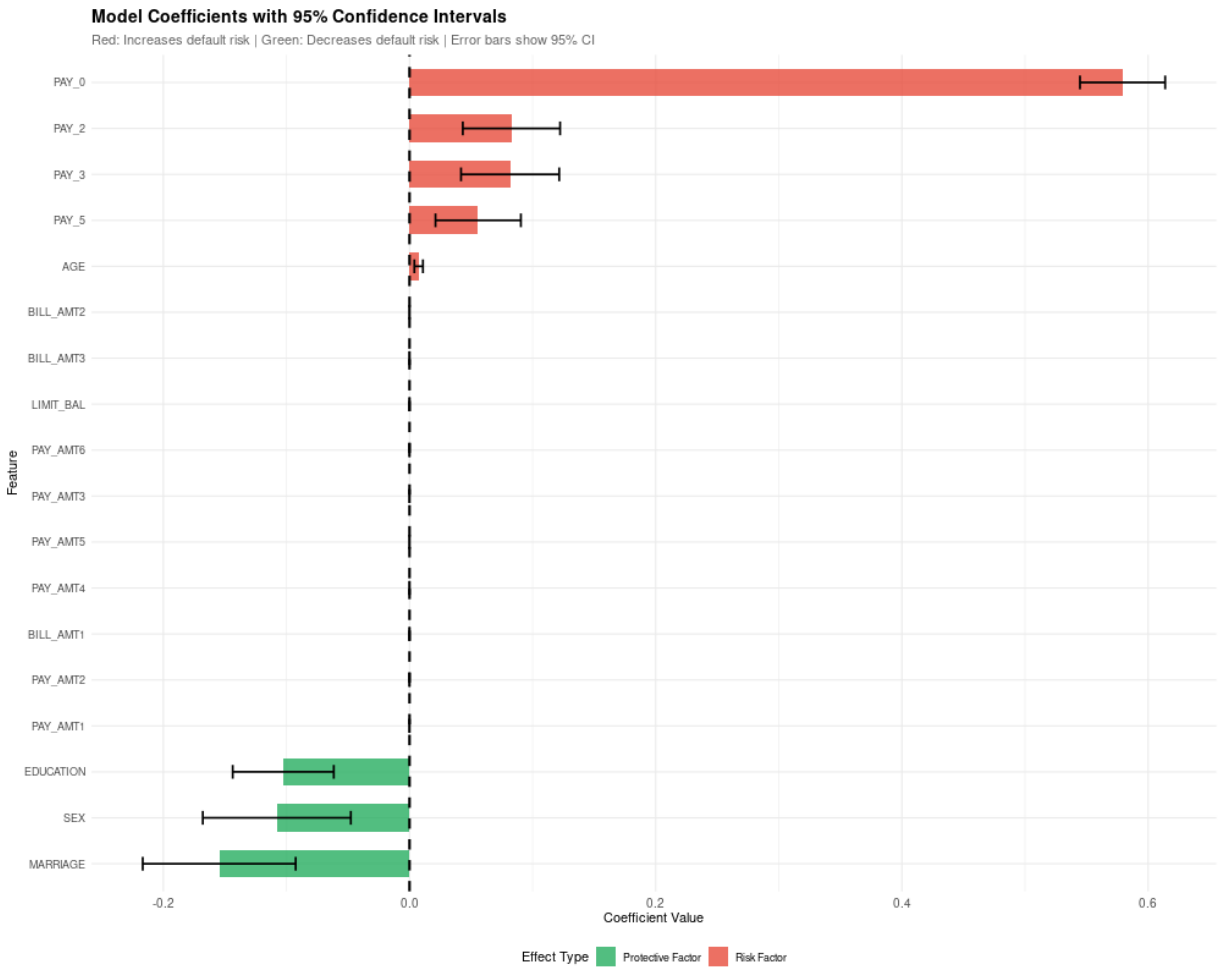

Figure 4: Model Coefficients with 95% Confidence Intervals

Most factors are statistically significant at $p < 0.001$, providing strong evidence that these variables genuinely influence default risk. A few variables (BILL_AMT2, PAY_AMT3, PAY_AMT6) are not individually significant at the 0.05 level, but they were retained by our forward selection process because they contribute small improvements to overall model fit measured by AIC.

7 Model Evaluation

Evaluating model performance on independent test data is crucial to assess how well the model generalizes to new customers not seen during training.

Data Preparation for Evaluation

To fairly assess our model's performance, we split the 30,000 observations into training and test sets using stratified sampling. This means we randomly divided the data (70% training, 30% test) while maintaining the same 21% default rate in both sets. This ensures the test set is representative of the real world where about 1 in 5 customers default. The training set contained 21,001 observations while the test set had 8,999 observations.

Model Predictions

We trained on the training set and predicted on the test set. Predicted probabilities range from 1.12×10^{-5} to 0.994. For classification, we used threshold 0.5 (default if probability > 0.5).

Performance Metrics

To understand model performance, several metrics are important:

- **Accuracy**: Overall percentage of correct predictions
- **Sensitivity**: Of actual defaulters, what percentage did we identify?
- **Specificity**: Of actual non-defaulters, what percentage did we correctly identify?
- **Precision**: When we predict default, how often are we correct?
- **Balanced Accuracy**: Average of sensitivity and specificity (important when classes are imbalanced)
- **AUC**: Area under the ROC curve - measures discrimination ability across all thresholds

Confusion Matrix

The confusion matrix shows predicted versus actual classes. True negatives and true positives are on the diagonal; false positives and false negatives are off the diagonal.

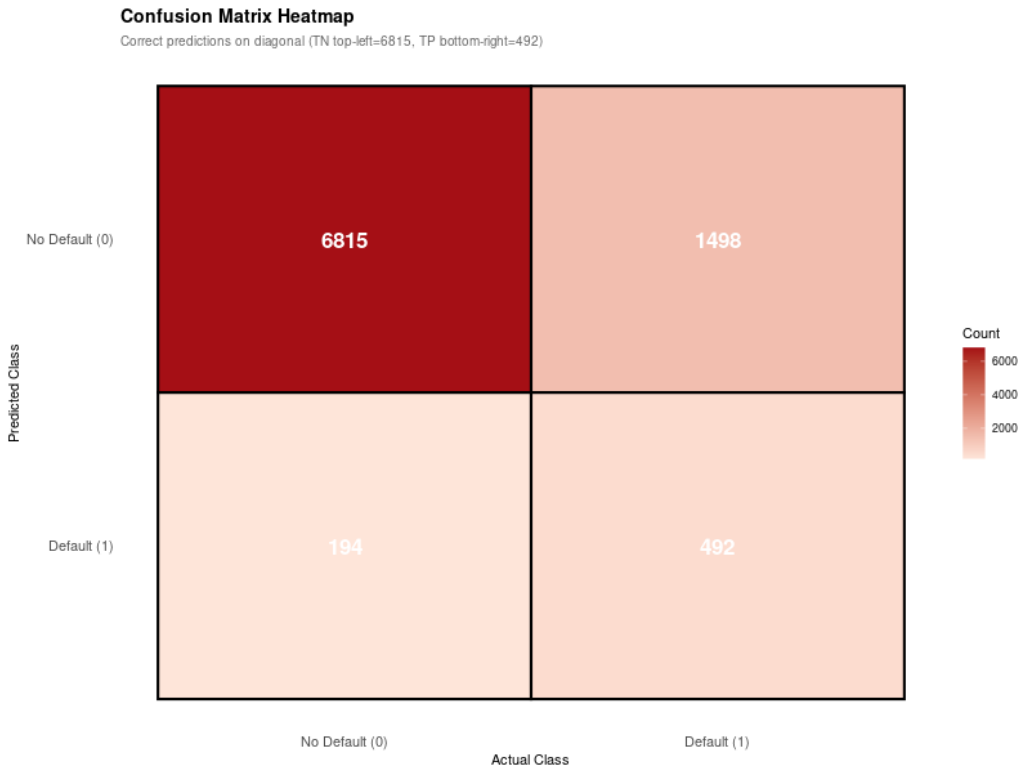

Figure 5: Confusion Matrix Heatmap

Metric	Value	Interpretation
Accuracy	81.20%	Overall correct predictions across both classes
Sensitivity	24.72%	Proportion of actual defaults correctly identified
Specificity	97.23%	Proportion of actual non-defaults correctly identified
Precision	71.72%	Of predicted defaults, how many are actually correct
Balanced Accuracy	60.98%	Average performance accounting for class imbalance
AUC	0.7250	Discrimination between default and non-default

Table 2: Performance Summary

8 Results and Model Performance

Performance Metrics Summary

Detailed Performance Analysis

Accuracy (81.20% with 95% CI: [80.38%, 82.00%])

The model correctly predicts customer default status 81.20% of the time. This means that out of every 100 customers, the model accurately classifies about 81. The narrow confidence interval suggests the result is stable across different samples.

Sensitivity (24.72%): Default Detection Rate

The model only identifies about 24.72% of customers who actually default. If 100 customers were going to default, the model would correctly detect only about 25. The low sensitivity

reflects the model's conservative nature in the presence of class imbalance.

Specificity (97.23%): Non-Default Identification

The model is highly accurate in identifying customers who will not default. It correctly classifies 97.23% of non-defaulters, meaning it rarely mislabels good customers as risky.

Precision (71.72%): Prediction Reliability

When the model predicts a customer will default, it is correct 71.72% of the time. This means its warnings are generally reliable.

Balanced Accuracy (60.98%)

Given the imbalanced dataset (22% defaults, 78% non-defaults), the balanced accuracy provides a fairer evaluation. A score of 60.98% indicates moderate performance across both classes.

AUC (0.7250): Discrimination Ability

The Area Under the Curve (AUC) of 0.7250 shows that the model has reasonable ability to distinguish between defaulters and non-defaulters. It performs better than random guessing (0.5), though still leaves room for improvement.

What the Trade-offs Mean

The model is conservative: it identifies non-defaulters well (high specificity) but misses many defaulters (low sensitivity). Reasons include:

1. The dataset is imbalanced, with many more non-defaulters.
2. The default probability threshold is set to 0.5, which may not be optimal.
3. Logistic regression tends to favor the majority class unless adjusted.

In practice, the model can flag high-risk accounts. It should not be the sole decision tool.

Practical Uses and Limitations

- **Strengths**: Excellent at identifying safe customers (97% specificity).
- **Weaknesses**: Poor at catching all risky customers (only 25% sensitivity).
- **Best For**: Detecting obvious risk cases such as recent payment delays.
- **May Miss**: Customers who show sudden default behavior despite prior good records.
- **Possible Improvements**: Adjusting the classification threshold could boost sensitivity, though at the cost of more false alarms.

9 ROC Curve and Model Discrimination

The Receiver Operating Characteristic (ROC) curve provides a comprehensive assessment of model performance across all possible classification thresholds. This visualization is particularly valuable because it reveals how sensitivity and specificity change as we adjust the probability cutoff for predicting default. In practice, different financial institutions may prefer different trade-offs between sensitivity and specificity based on their risk tolerance and business strategy.

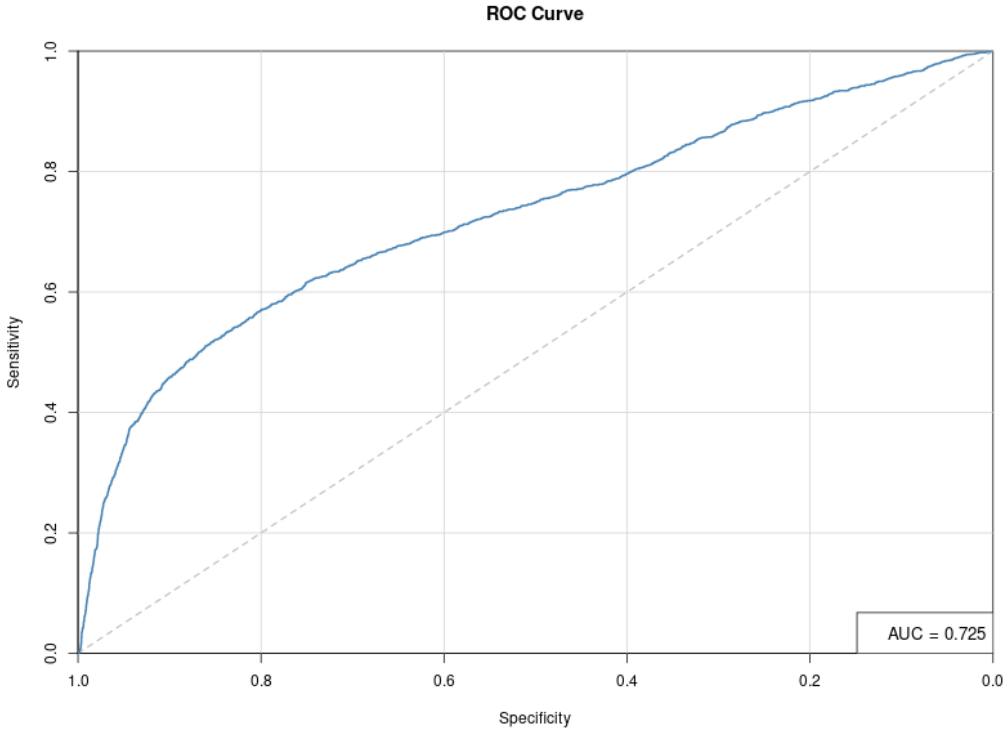

Figure 6: ROC Curve with AUC = 0.7250

The ROC curve shows how our model performs across different probability thresholds. We currently use 0.5: if the model predicts a probability above 0.5, we classify it as default. But this is flexible. Lower the threshold below 0.5 and we catch more defaulters (but get more false alarms). Raise it above 0.5 and we get fewer false alarms (but miss some real defaulters).
Our AUC of 0.725 means the model is better than random chance at ranking customers by default risk. If we picked one real defaulter and one real non-defaulter at random, the model would correctly give the defaulter the higher default probability about 72.5% of the time. That's solid, but there's room for improvement. The curve shows we're better at catching non-defaulters than actual defaulters: which is what we expected from our earlier results.
Banks can adjust the threshold based on what they care about most. If they want to avoid false alarms, use a higher threshold. If they want to catch defaults early, use a lower threshold.

Conclusion

Predicting credit card defaults is genuinely hard. There's always a trade-off between catching defaults and avoiding false alarms. Our model handles this trade-off in a particular way: it's very good at marking safe customers as safe (97% specificity), but it misses a lot of actual defaulters (25% sensitivity). That's the reality of our dataset, because more people pay than defaults, so the model naturally leans toward predicting non-default.

However, that doesn't mean our model performs poorly. When it says someone is risky, it's right about 72% of the time. And when it says someone is safe, it's almost always right. For a bank, this means: trust the model when it flags someone as low-risk. Don't be surprised when it misses some defaulters. Recent payment behavior turned out to be the real signal here, even more than age, education, or any other demographic factor.

So, where does our model fit well in practice? It works best as a screening tool, not the final decision. A bank could use it to flag accounts for closer inspection. Or adjust the threshold lower to catch more defaults, knowing that will create more false alarms. Different banks will want different trade-offs depending on their risk tolerance.

This work shows that even with data imbalance and real-world complexity, a straightforward logistic regression model with proper feature selection can meaningfully outperform basic approaches. It's not perfect, but it's very useful. And that's what matters in practice which will help guide decision-making.