DEFAULT OF CREDIT CARD CLIENTS

Group 6

Yan Kevin ZE
Arnaud FOUBEUDA BOZAHBE
Olusola Timothy OGUNDEPO
Consolee NISINGIZWE

November, 2025

Overview

Introduction and Research Context

The Real-World Problem

- Credit card defaults pose significant financial risks to banks and lenders
- Predicting defaults helps institutions manage risk and make better lending decisions
- Early identification allows for proactive measures like credit limit adjustments

Our Research Questions

- Can we reliably predict which customers will default next month?
- What factors are associated with customers with high default risks?
- How can banks use this information in their risk management?

Dataset Overview

Data Source

- Credit card default data from Taiwan
- 30,000 customer records with 24 variables
- Historical payment data, demographic information, and billing amounts
- Target variable: Default payment next month (0 = No, 1 = Yes)

Key Variables

- **Payment History**: PAY_0 to PAY_6 (recent payment status)
- **Demographic**: SEX, EDUCATION, MARRIAGE, AGE
- **Financial**: LIMIT_BAL, BILL_AMT1-6, PAY_AMT1-6
- **Target**: Default payment next month

Data Quality and Preparation

Data Cleaning Steps

- Removed customer ID (Redundant)
- Checked for missing values (found none)
- Ensured proper factor encoding for the target variable (default)

Target Variable Distribution

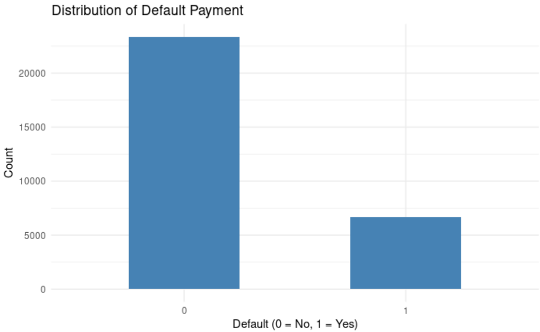

Figure 1: The Distribution of Default Payment

- Clear class imbalance: Majority of customers don't default
- Important for understanding model performance trade-offs

Key Relationships with Default

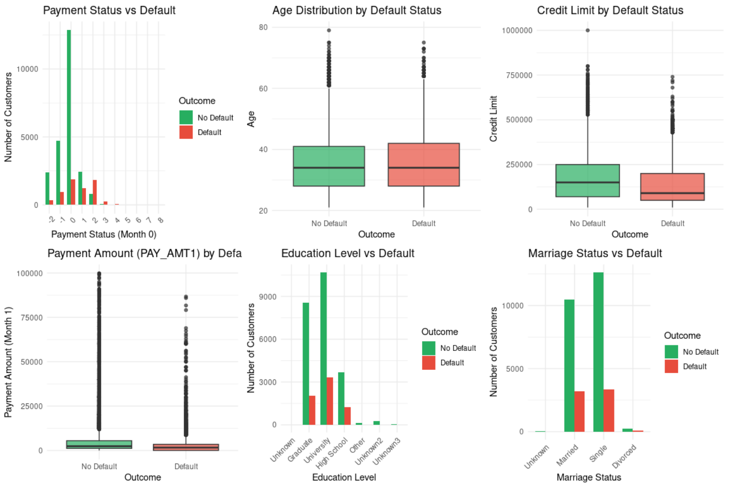

Figure 2: Relationships of Variables with Default Payment

Why Logistic Regression?

Theoretical Foundation

- Appropriate for binary classification problems
- Provides probability estimates, not just classifications
- Coefficients are interpretable as log-odds
- Well-established statistical properties

Practical Advantages

- Handles both continuous and categorical predictors
- Computationally efficient
- Results are easily explainable to non-technical stakeholders
- Robust to violations of some assumptions

Feature Selection Strategy

Forward Selection using AIC

- Started with best individual predictor (PAY_0)
- Sequentially added features that most improved model fit
- Used Akaike Information Criterion (AIC) for model comparison
- Lower AIC indicates better model balancing fit and complexity

Selection Results

- Initial model (PAY_0 only): AIC = 28,535.57
- Final selected model: AIC = 27,917.81
- Improvement: 617.76 AIC points
- Selected 18 out of 23 available features

AIC Progression - Diminishing Returns

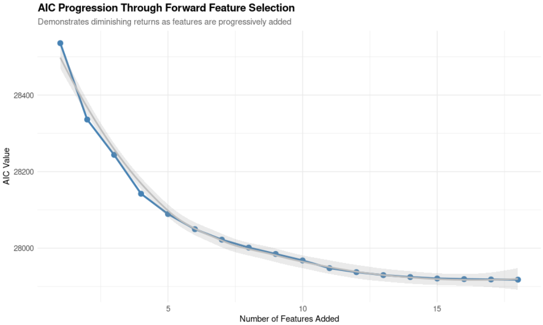

Figure 3: AIC Progression through Forward Feature Selection

- First few features provided largest improvements
- Later additions offered smaller gains
- A good features selection criteria is at 18

Key Risk Factors Identified

Strongest Predictors (Increases Default Risk)

- **PAY_0**: Recent payment status (OR = 1.78, 95% CI [1.71, 1.86])
- Each unit increase multiplies default odds by 1.78 (78% increase)
- Recent payment delays are the single biggest red flag

Protective Factors (Decreases Default Risk)

- **MARRIAGE**: Married customers (OR = 0.86, 14% reduction)
- **EDUCATION**: Higher education levels (OR = 0.90, 10% reduction)
- **Payment Amounts**: Larger payments reduce default probability

Coefficient Visualization with Confidence Intervals

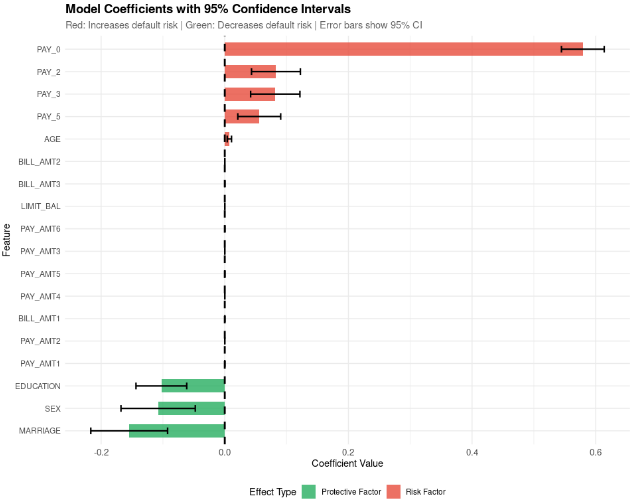

Figure 4: Model Coefficients with 95% Confidence Intervals

Experimental Design: Train-Test Split

Stratified Sampling

- 70% training data, 30% test data
- Maintained original class proportions in both sets
- Prevents accidental bias in test set composition
- Training: 21,001 observations
- Testing: 8,999 observations

Class Distribution Preservation

Dataset	Default Rate	Training	Test
Overall	22.12%	22.12%	22.11%

Performance Metrics

Metric	Value	Interpretation
Accuracy	81.20%	Overall correct predictions
Sensitivity	24.72%	Default detection rate
Specificity	97.23%	Non-default identification
Precision	71.72%	Prediction reliability
Balanced Accuracy	60.98%	Fair average performance
AUC	0.7250	Discrimination ability

Confidence in Results

- Accuracy 95% CI: [80.38%, 82.00%]
- Narrow interval indicates reliable estimates
- Results likely generalizable to new data

Confusion Matrix Analysis

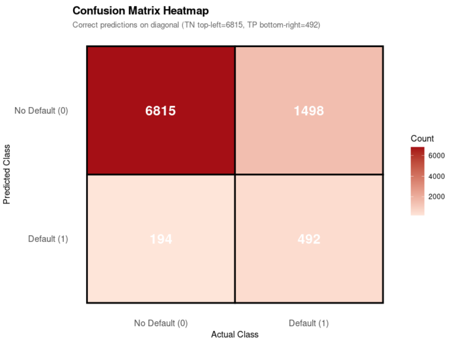

Figure 5: Confusion Matrix Heatmap

- **True Negatives**: 6,815 (correctly identified non-defaulters)
- **True Positives**: 492 (correctly identified defaulters)
- **False Negatives**: 1,498 (missed defaults - main weakness)
- **False Positives**: 195 (false alarms minimal)

ROC Curve and Model Discrimination

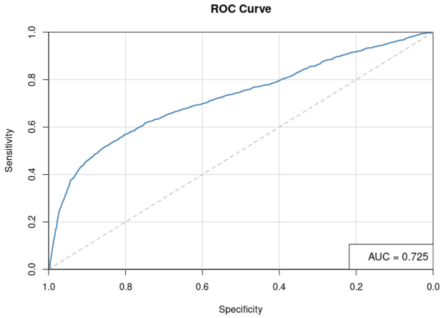

Figure 6: ROC Curve with AUC = 0.7250

- AUC = 0.7250 indicates reasonable discrimination ability
- Curve shows trade off between sensitivity and specificity

Summary of Findings

What We Learned

- **Payment history is crucial**: Recent payment behavior is the strongest predictor
- **Demographics matter**: Marriage and education are protective factors
- **Model is conservative**: Excellent at identifying safe customers, misses many risky ones
- **Trade-offs are inevitable**: High specificity comes at the cost of lower sensitivity

Statistical Confidence

- Most risk factors show strong statistical significance ($p < 0.001$)
- Confidence intervals are narrow, indicating precise estimates
- Odds ratios provide intuitive interpretation of effect sizes

Future Work and Improvements

Methodological Enhancements

- Try different classification thresholds to balance sensitivity/specificity
- Experiment with ensemble methods or neural networks
- Incorporate time-series analysis of payment patterns
- Address class imbalance with sampling techniques

Practical Extensions

- Develop early warning system with multiple risk thresholds
- Integrate with other data sources (income, employment history)
- Create dynamic models that update with new payment behavior
- Build customer segmentation based on risk profiles


Thank you
for Listening!