



Credit Card Default Prediction

Project Group 5

1. Kiran Joseph
2. Amit Patra
3. Paul Ittoopunny

1. Introduction

Credit card default prediction is a critical task in the financial services industry. By identifying customers at risk of defaulting on their credit payments, banks and financial institutions can take proactive measures to mitigate financial losses and improve credit risk management.

In this project, we used the “Default of Credit Card Clients” dataset from the UCI Machine Learning Repository. The dataset includes demographic, payment, and bill-related data for 30,000 clients over a 6-month period. The objective is to develop a machine learning model that accurately predicts the likelihood of a client defaulting in the next month.

The project workflow involved:

- Data cleaning and preparation
- Exploratory Data Analysis (EDA)
- Model selection and training
- Model evaluation and interpretation
- Final recommendations

2. Data Cleaning and Preparation

Data cleaning and preparation is a critical phase in any data science pipeline, as it ensures that the data is structured, reliable, and ready for analysis. The raw dataset obtained from the UCI repository required several transformations to prepare it for exploratory analysis and modeling.

2.1 Loading and Initial Processing

- The dataset, obtained from the UCI Machine Learning Repository, was structurally adjusted to ensure proper alignment of column headers.
- The first row of the file (containing column headers) was misaligned and corrected by setting it as the column header and resetting the Data frame.
- The ID column was dropped as it holds no predictive relevance.
- The target column default payment next month was renamed to default for clarity.

2.2 Inspection of Structure and Data Types

Initial inspection focused on understanding the shape and quality of the dataset.

Checks Performed:

- Checking for Missing Values: No missing or null values were found in any of the columns.
- Reviewing Data Types: All features were confirmed to have appropriate data types, either categorical or numerical.
- Basic Statistical Review: Descriptive statistics such as mean, median, and standard deviation were reviewed to identify any anomalies or extreme values in features like credit limit, age, billing amounts, and past payments.
- Category Verification: Categorical fields such as gender, education level, and marital status were checked to ensure their values were within expected ranges.

2.3 Data Cleaning

Several refinements were made to improve data quality:

- Checked and removed duplicates (none were found).
- Converted any necessary string-encoded numerical fields to numeric.
- Reviewed distributions of demographic and financial features for anomalies or out-of-range values.

2.4 Class Balance and Distribution

The target variable default was examined to understand class imbalance—an important consideration for model performance.

Interpretation:

- ~77% of clients did not default.
- ~23% of clients defaulted, confirming a class imbalance.

This imbalance has significant implications for model performance and evaluation. As a result, the class imbalance was later addressed using weighted learning approaches during model training.

A bar chart visualization of this distribution was generated and included in the analysis to communicate this skew effectively.

2.5 Output of This Phase

At the conclusion of the cleaning and preparation phase:

- The dataset was fully structured and free of missing or duplicate data.
- Column names and types were standardized.
- A cleaned version of the dataset was saved for subsequent stages of exploratory data analysis and model training.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to better understand the structure, trends, and relationships within the dataset. This phase helped in identifying important patterns, potential issues like outliers, and provided insights that directly informed feature engineering and model design.

3.1 Univariate Analysis

The distribution of individual features was assessed to identify skewness, outliers, and typical value ranges.

- **LIMIT_BAL (Credit Limit):**
Displayed a right-skewed distribution. The majority of clients had lower credit limits, while a small portion had very high limits.
- **AGE:**
Followed an approximately normal distribution, centered around late 20s to early 30s, indicating a relatively young client base.
- **BILL_AMT1 (September Billing Amount):**
Exhibited a heavily right-skewed distribution. Most clients had low billing amounts, but a few had extremely high values, indicating the presence of outliers.

These distributions were visualized using histograms, enabling the identification of skewed features that may benefit from transformation or scaling during modeling.

3.2 Bivariate Analysis

Relationships between predictors and the target variable (default) were explored.

- **LIMIT_BAL vs Default:**
Clients with higher credit limits showed lower default rates. This suggests that credit limit could be a strong predictor of financial stability and repayment capability.

- AGE vs Default:

Slightly higher default rates were observed among younger clients, particularly those in their 20s and early 30s. However, the trend was not sharply pronounced across all age groups.

Bivariate analysis helped in assessing the discriminative power of continuous variables in predicting default behavior.

3.3 Correlation Analysis

To identify redundancy and multicollinearity among numerical features, a correlation matrix was created.

- Strong positive correlations were observed between:
 - Monthly billing amounts (BILL_AMT1 to BILL_AMT6)
 - Monthly repayment amounts (PAY_AMT1 to PAY_AMT6)

These correlations indicate temporal consistency in client behavior, i.e., those who had high bills or payments in one month tended to exhibit similar patterns across others. This insight was crucial for designing derived features (e.g., averages or trends) that summarize client behavior over time.

3.4 Outlier Detection

Outliers were primarily found in:

- Billing Amounts (BILL_AMT1-6)
- Payment Amounts (PAY_AMT1-6)

These extreme values were visualized and assessed for potential impact. Rather than removing them, they were retained due to their potential relevance in capturing high-risk financial behaviour. However, their influence was carefully monitored during model training and evaluation.

3.5 Summary of Insights for Modeling

- Key predictive features such as credit limit, recent payment history, and bill amounts demonstrated strong relevance.
- Skewed distributions and outliers were acknowledged and addressed through scaling techniques.
- Temporal consistency suggested that aggregations (like averages or ratios) could improve model performance.

4. Model Selection

The goal of this phase was to identify the most effective predictive model for classifying whether a credit card client would default on payment. Two supervised classification algorithms were selected: Logistic Regression (as a baseline model) and Random Forest Classifier (as a non-linear ensemble model).

The cleaned dataset was used to generate the input matrix X (predictors) and target vector y (default status). A stratified train-test split (80:20) ensured that the distribution of default cases was consistent across training and evaluation sets.

4.1 Handling Class Imbalance

Initial exploratory analysis revealed that only about 23% of the clients had defaulted, resulting in a class imbalance problem. To mitigate this during training, class weights were adjusted inversely proportional to class frequencies. This ensured the model would not be biased toward the majority class.

4.2 Baseline Model: Logistic Regression

Logistic Regression was chosen as a baseline due to its interpretability and efficiency in binary classification problems. Key steps included:

- Applied liblinear solver with L1/L2 regularization for stability.
- Increased the number of iterations to ensure convergence.
- Evaluated performance using 5-fold cross-validation with ROC-AUC as the primary metric.

This model served as a benchmark to evaluate improvements from more complex algorithms.

4.3 Advanced Model: Random Forest Classifier

Random Forest was used to capture non-linear interactions among features. The model is robust to overfitting and handles high-dimensional data well.

- Built with 100 trees (estimators) and automatic feature selection.

- Also trained using stratified sampling and class weights.
- Evaluated using 5-fold cross-validation, with ROC-AUC as the primary metric.

Initial results showed that Random Forest consistently outperformed Logistic Regression in terms of ROC-AUC scores, indicating better discriminatory power.

4.4 Hyperparameter Optimization

Both models were further optimized using Grid Search Cross-Validation to identify the best combination of parameters.

Logistic Regression Grid Search:

- Tuned over different values of regularization strength C
- Tested both L1 and L2 penalties
- Used ROC-AUC as the scoring metric


Random Forest Grid Search:

- Tuned n_estimators (number of trees), max_depth, and min_samples_split
- The best model showed improved performance over the default configuration

The final optimized models were saved for evaluation and comparison in the next phase.

4.5 Summary of Model Selection

Model	Method	ROC-AUC Score	Remarks
Logistic Regression	Grid Search	Moderate	Baseline model. Simple and interpretable
Random Forest	Grid Search	High	Best performance. Captures non-linear patterns



The Random Forest Classifier was selected as the final model due to its superior performance across validation folds. It was saved and passed on to the model analysis phase for further evaluation on unseen test data.

5. Model Analysis

This phase focused on evaluating the performance of the trained models on unseen test data. The purpose was to determine which model generalizes best and is most suitable for deployment in real-world scenarios. Two models were considered: Logistic Regression and Random Forest, both optimized through grid search during the model selection phase.

5.1 Model Evaluation on Test Data

The final evaluation was performed using the test subset (20% of the original data). The primary metric used was ROC-AUC (Receiver Operating Characteristic – Area Under Curve), which measures a model's ability to distinguish between the default and non-default classes.

Both models were reloaded and evaluated using a consistent framework to ensure fairness. The Random Forest model demonstrated better classification performance and was therefore selected as the final model.

5.2 ROC Curve Analysis

A ROC curve was plotted for the selected model to visualize its performance across various classification thresholds.

- The curve for the selected model (Random Forest) was plotted against the baseline diagonal (random classifier).
- The AUC (Area Under Curve) score was approximately 0.80, indicating strong discriminatory power.
- The model effectively balances sensitivity (true positive rate) and specificity (true negative rate), making it suitable for financial risk prediction.

This analysis confirmed that the model was capable of capturing underlying patterns in the data relevant to default prediction.

5.3 Feature Importance Analysis

Using the Random Forest model, the top 10 most important features contributing to the prediction were extracted and visualized. These features provide insight into the factors most associated with default behavior.

Top Predictive Features:

- LIMIT_BAL – Total credit limit assigned
- PAY_1, PAY_2 – Most recent repayment statuses
- Avg_Bill, Avg_Payment – Aggregated spending and repayment behavior
- Bill vs Payment Ratios – Financial discipline indicators


These findings aligned with insights from EDA and confirmed the relevance of the selected features.

5.4 Key Findings and Interpretability

- Random Forest was selected as the final model based on its higher accuracy and ROC-AUC.
- Feature interpretability was enhanced through importance plots, helping stakeholders understand which financial behaviors are most predictive of default.
- Logistic Regression remains available as a fallback model due to its simplicity and explainability.

5.5 Summary of Outcomes

- Conducted comprehensive evaluation of both baseline and advanced models on a holdout test set.
- Confirmed Random Forest as the best-performing model based on ROC-AUC and classification accuracy.
- Validated the model's ability to distinguish between defaulters and non-defaulters with an AUC score of approximately 0.80.
- Identified key drivers of default through feature importance analysis, enhancing model transparency and decision-making support.

- 
- Established a reliable, data-driven framework for credit risk prediction using real-world financial behavior indicators.

5.6 Future Recommendations

- Explore advanced algorithms like XGBoost for further performance gains.
- Implement K-S charts and probability calibration for better risk score reliability.
- Integrate the model into a decision support system or API for operational use.
- Use SHAP or LIME for model interpretability in production environments.

6. Conclusion and Recommendations

6.1 Conclusion

This project successfully developed a machine learning pipeline to predict credit card default based on historical client data from the UCI repository. The process included data preprocessing, exploratory data analysis, model training, and evaluation.

Key accomplishments include:

- Transforming raw, unstructured data into a clean and structured format ready for analysis.
- Uncovering valuable insights into financial behavior through EDA, such as temporal patterns in billing and repayment.
- Comparing baseline (Logistic Regression) and advanced (Random Forest) models using cross-validation and ROC-AUC scores.
- Selecting Random Forest as the final model due to its superior performance and feature interpretability.
- Achieving an AUC score of approximately 0.80, indicating strong model effectiveness in identifying potential defaulters.

Overall, the model presents a reliable and scalable approach for credit risk classification and has practical implications for use in credit scoring systems.

6.2 Recommendations

Based on the analysis and results, the following recommendations are proposed for future development and deployment:

Model Improvements

- Evaluate advanced ensemble models such as XGBoost, LightGBM, or CatBoost to further boost performance.
- Implement probability calibration (e.g., Platt scaling or isotonic regression) for better threshold optimization in real-world classification.

Interpretability and Fairness

- Integrate SHAP or LIME for model interpretability, especially important in financial contexts where transparency is required.
- Assess model fairness across demographic segments (e.g., gender, age) to ensure ethical application.

Deployment and Monitoring

- Develop a scoring API or dashboard for integration with financial systems.
- Implement continuous monitoring to detect data drift and model degradation over time.
- Use a fallback model like Logistic Regression when simpler, explainable results are needed for regulatory purposes.

Business Application

- Use predictions to flag high-risk customers for additional review or intervention.
- Combine model outputs with internal policy rules to improve approval workflows and reduce defaults.