Analyzing Demographic and Behavioral Factors Associated with Credit Card Default Risk

Rafael F. de A. S. Lima
Centro Universitário IBMEC
IBMEC-RJ
Rio de Janeiro, Brazil
rafael.abreu.lima@outlook.com

Bernardo R. B. Loureiro
Centro Universitário IBMEC
IBMEC-RJ
Rio de Janeiro, Brazil
bernalourodri@gmail.com

Lucca Lanzellotti
Centro Universitário IBMEC
IBMEC-RJ
Rio de Janeiro, Brazil
luccalanzellotti@gmail.com

João P. Alencar
Centro Universitário IBMEC
IBMEC-RJ
Rio de Janeiro, Brazil
jpsa.alencar17@gmail.com

Thiago Silva de Souza
Centro Universitário IBMEC
IBMEC-RJ
Rio de Janeiro, Brazil
t.souza@ibmec.edu.br

Rigel P. Fernandes
Centro Universitário IBMEC
IBMEC-RJ
Rio de Janeiro, Brazil
rigelfernandes@gmail.com

***Abstract*—This study investigates key predictors of credit card default risk using demographic and behavioral data from credit card clients in Taiwan. Through exploratory data analysis and statistical modeling, we identify significant associations between default behavior and variables including payment history, credit limits, education level, marital status, and spending patterns. A logistic regression model was developed using oversampling to address class imbalance, achieving 68.5% accuracy and 71.3% AUC in predicting default status. Our findings highlight that recent payment delays (PAY_0), lower credit limits (LIMIT_BAL), and demographic characteristics are critical indicators of default risk.**

***Index Terms*—credit card default, demographic risk factors, payment behavior, logistic regression, financial risk management.**

I. Introduction

Credit risk management remains a critical challenge for financial institutions, particularly in credit card operations where defaults can trigger significant economic repercussions [1]. While predictive models exist, identifying behavioral and demographic factors associated with default continues to require deeper investigation.

This study examines key determinants of credit card default through analysis of real-world data encompassing:

- Demographic variables (gender, education, marital status, age).
- Financial behaviors (payment history, bill amounts, repayment patterns).
- Credit limits and utilization trends.

Addressing three core questions (association between demographic factors and default risk, feasibility of efficient predictive models, and relationship between payment history and future behavior), we employ exploratory data analysis (EDA), statistical testing (ANOVA, t-tests), and logistic regression modeling.

The significance of this research lies in:

- Identifying proactive risk indicators (e.g., recent delays - PAY_0).
- Proposing an approach to mitigate class imbalance challenges.
- Generating actionable insights for credit assessment strategies.

A. About credit card default risk

Credit risk represents a critical challenge for financial institutions [2], occurring when cardholders fail to meet minimum payment obligations. Traditional risk assessment models have relied primarily on credit scores and basic financial indicators, often failing to capture the nuanced behavioral patterns that precede default events. The evolution of data analytics and machine learning has enabled more sophisticated models that incorporate demographic characteristics, payment history, and behavioral indicators.

The economic impact extends beyond individual institutions, affecting overall consumer credit market stability. Understanding demographic patterns across age groups, educational backgrounds, marital statuses, and gender is essential for developing targeted risk management strategies and ensuring fair lending practices. Modern approaches require comprehensive analysis of how these demographic factors interact with financial behaviors to predict default risk accurately.

B. Interaction effects between payment history and demographic factors

While demographic factors and payment history have been studied independently, their interactions present significant potential for improving predictive accuracy. Payment history variables (PAY_0 through PAY_6 and PAY_AMT1 through PAY_AMT6) provide temporal insights into consumer financial behavior, revealing patterns of financial stress and payment prioritization strategies.

Demographic characteristics may moderate payment behavior relationships in important ways. Younger consumers may exhibit different payment patterns than older demographics under similar financial constraints, while educational background may influence responses to financial stress. Gender differences in financial behavior, combined with factors like marital status or age, create complex interaction effects that traditional additive models may overlook.

The temporal progression of payment behaviors may reveal different risk trajectories across demographic segments. Understanding these interaction effects is crucial for developing models that move beyond simple additive frameworks to capture the multiplicative relationships characterizing real-world financial behavior, ultimately enabling more accurate risk prediction and effective risk management strategies.

II. Literature Review

Credit risk prediction has emerged as a vital domain of study within the financial services industry, especially regarding the mitigation of losses related to credit card default. Traditional statistical methods such as logistic regression and discriminant analysis have long served as standard tools for credit scoring. However, these approaches often struggle to model complex, nonlinear relationships and perform poorly under class imbalance conditions commonly observed in credit default datasets [2].

To overcome these limitations, recent studies have explored the adoption of machine learning (ML) techniques and hybrid models. Chi et al. [2] proposed combining traditional statistical methods with artificial intelligence (AI) approaches, showing that hybrid models—such as logistic regression combined with multilayer perceptron (LR+MLP)—yielded superior predictive performance compared to standalone models. The study evaluated 16 such combinations across multiple datasets and demonstrated that models integrating LR with neural networks exhibited better generalizability and interpretability.

Further advancements have been made with the introduction of ensemble learning. Aruleba and Sun [8] employed ensemble classifiers including Random Forest, AdaBoost, XGBoost, and LightGBM, in conjunction with SMOTE-ENN for handling class imbalance. Their results indicated that XGBoost achieved the best recall on the German Credit dataset, while Random Forest outperformed others on the Australian dataset. Additionally, they integrated SHapley Additive exPlanations (SHAP) to provide feature attribution, enhancing the interpretability of the models.

Li [1] evaluated the performance of XGBoost against logistic regression for credit risk prediction and concluded that XGBoost significantly outperformed logistic regression in terms of classification accuracy. The study highlighted the importance of feature engineering and comprehensive data preprocessing in enhancing model performance.

Yu et al. [3] expanded the exploration of ensemble techniques by comparing LightGBM, XGBoost, and TabNet on a large-scale bank dataset. Their pipeline incorporated PCA for dimensionality reduction and SMOTEENN for class balance, ultimately demonstrating that LightGBM achieved the best trade-off between performance and computational efficiency.

Atiya [4] provided a comprehensive review of bankruptcy prediction with neural networks. The study emphasized the transition from traditional ratio-based methods to more sophisticated machine learning frameworks. Neural networks, particularly multilayer perceptrons, showed superior accuracy due to their capacity to model non-linear relationships among financial indicators.

In summary, the literature suggests a paradigm shift from classical statistical methods toward more flexible and powerful ML-based approaches for credit risk prediction. The integration of ensemble learning, hybrid modeling, and explainable AI has significantly enhanced both the predictive performance and transparency of credit scoring systems. This study builds upon such foundations by focusing on demographic and behavioral factors using logistic regression, while addressing challenges like data imbalance through oversampling techniques.

III. Methodology

A. Dataset Description

The dataset employed in this study is the ”Default of Credit Card Clients” dataset, sourced from the UCI Machine Learning Repository and available on Kaggle. It contains anonymized data from 30,000 credit card clients in Taiwan, with records collected in 2005. [1]

The dataset consists of 25 variables, comprising 24 predictor variables and one binary target variable, default.payment.next.month. This target variable indicates whether a client defaulted on their payment in the subsequent month (1 = yes, 0 = no). The predictor variables can be broadly categorized as follows [2]:

- **Client Demographics**: Includes variables such as SEX (gender), EDUCATION (level of education), and MARRIAGE (marital status).
- **Credit Limit**: The LIMIT_BAL variable, which specifies the amount of credit provided to the client.
- **History of Past Payment**: A critical set of variables (PAY_0, PAY_2 through PAY_6) that document the client's repayment status from April to September 2005.
- **Bill Statement Amounts**: The BILL_AMT1 through BILL_AMT6 variables, which detail the amount on the client's bill statement for the same six-month period.
- **Previous Payment Amounts**: Correspondingly, PAY_AMT1 through PAY_AMT6 record the amount of previous payments made by the client.

B. Data Preprocessing

The preprocessing phase involved several critical steps to prepare the UCI Credit Card dataset for analysis and modeling. The dataset initially contained 30,000 observations with 25 variables, including demographic information, payment history, bill amounts, and payment amounts for six consecutive months.

Initial data quality assessment revealed no missing values in the dataset, ensuring data completeness for subsequent analyses. However, the dataset contained 35 duplicate records, which were identified but retained in the analysis to maintain sample size and representativeness of real-world scenarios where duplicate entries may occur.

Categorical variables were transformed from numeric codes to meaningful factor variables to enhance interpretability. The SEX variable was recoded from numeric values (1, 2) to categorical labels (”Male”, ”Female”). The EDUCATION variable was transformed from seven numeric levels (0-6) to meaningful categories including ”Graduate School”, ”University”, ”High School”, and ”Others”, with unknown values appropriately labeled. Similarly, the MARRIAGE variable was recoded from numeric codes (0-3) to descriptive categories (”Married”, ”Single”, ”Others”, ”Unknown”). The target variable, default.payment.next.month, was converted from binary numeric (0, 1) to factor levels (”No”, ”Yes”) for clarity in model interpretation.

A new categorical variable, age_range, was created by binning the continuous AGE variable into six age groups using 10-year intervals: ”21-29”, ”30-39”, ”40-49”, ”50-59”, ”60-69”, and ”70-79”. This transformation facilitated better understanding of age-related patterns in credit default behavior and improved model interpretability.

The ID variable was removed from the dataset as it served only as a unique identifier and provided no predictive value for the analysis. This reduced the feature space from 25 to 24 variables, focusing on meaningful predictors for credit default prediction.

The preprocessed dataset was randomly partitioned into training and testing sets using a 70:30 split ratio. A random seed (123) was set to ensure reproducibility of results.

Analysis of the target variable revealed significant class imbalance, with the majority class (non-default) substantially outnumbering the minority class (default). To address this imbalance and improve model performance on minority class prediction, oversampling was applied to the training data using the ROSE package. The oversampling technique increased the representation of the minority class by generating synthetic examples, creating a more balanced training dataset with twice the number of non-default cases to ensure adequate representation of both classes during model training.

C. Exploratory Data Analysis (EDA)

To explore the characteristics of the dataset and identify potential predictors of default behavior, we conducted a series of visual analyses across key demographic and behavioral variables.

Figure 1 presents the distribution of default status segmented by age range. The majority of defaults occur in the younger segments (21–39), with a noticeable decline as age increases, suggesting a potential inverse relationship between age and credit risk.

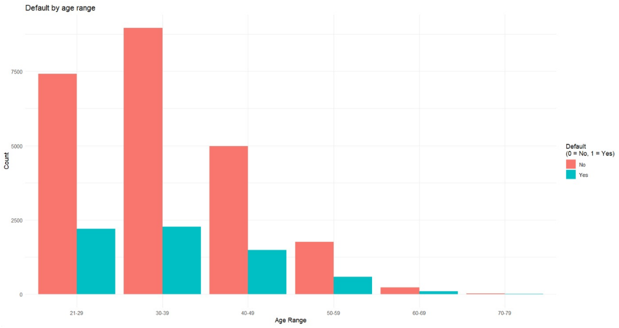

Fig. 1. Default by Age Range

Figure 2 illustrates the overall age distribution of credit card clients. The majority of clients fall between 21 and 50 years old, with a clear peak in the 30–39 age range.

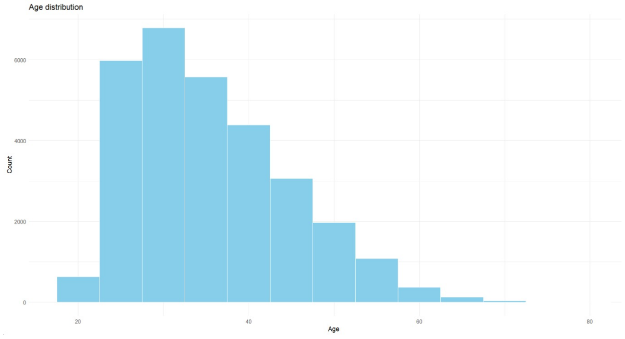

Fig. 2. Age Distribution

Figure 3 shows the default distribution across education levels. University and graduate-level clients represent the largest groups, and defaults are concentrated among university-level clients.

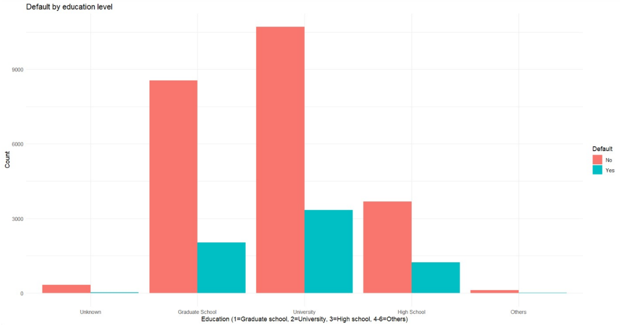

Fig. 3. Default by Education Level

In Figure 4, default rates are visualized by gender. Female clients outnumber males, and while both genders exhibit notable default rates, the proportion of female defaulters is slightly higher.

Figure 5 presents the overall distribution of the default variable. This plot confirms the class imbalance, with a significant majority of clients not defaulting.

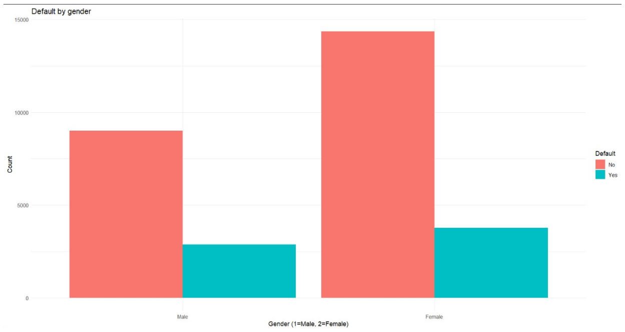

Fig. 4. Default by Gender

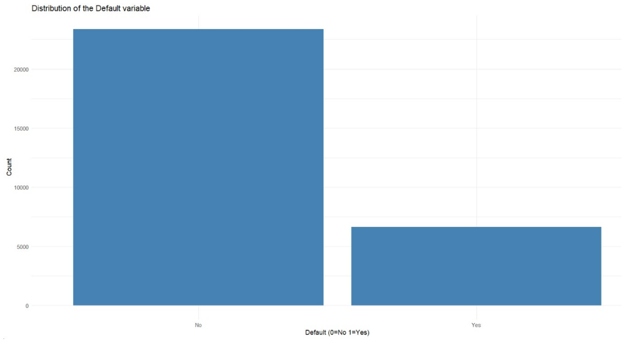

Fig. 5. Distribution of the Default Variable

Figure 6 displays the frequency of the PAY_0 variable, reflecting clients' most recent repayment status. A high concentration of values at 0 (no delay) is observed, with smaller peaks at -1 and 1, indicating a range of payment behaviors.

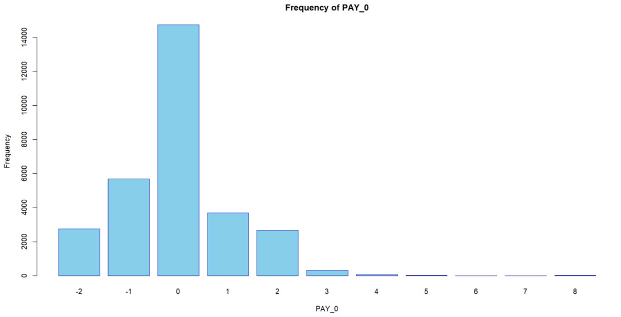

Fig. 6. Frequency of PAY_0

The distribution of marital status, shown in Figure 7, indicates that the majority of clients are single or married, with few labeled as "others" or "unknown".

Education level frequencies, as shown in Figure 8, confirm that most clients have university or graduate-level education.

Figure 9 shows the gender distribution, with a higher number of female clients in the dataset.

Lastly, Figure 10 presents the distribution of all numerical variables, emphasizing skewness and the presence of outliers in financial behavior indicators.

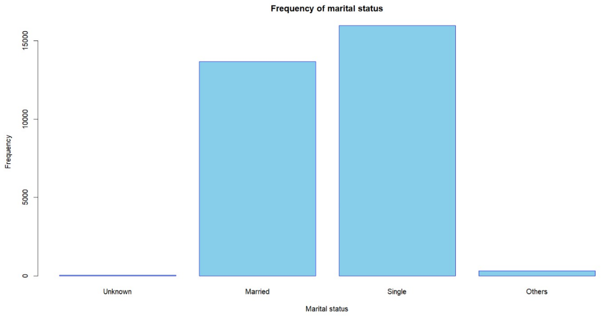

Fig. 7. Frequency of Marital Status

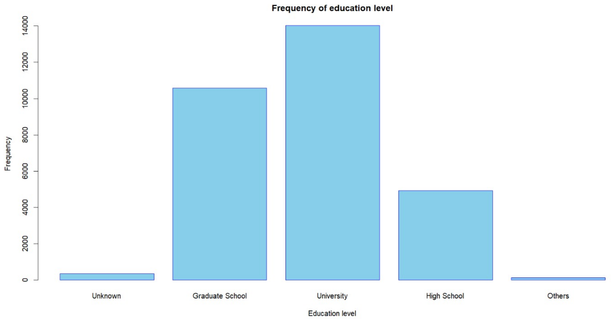

Fig. 8. Frequency of Education Level

Prior to model development, an Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset's structure, identify underlying patterns, and examine the relationships between variables. This process was essential for informing subsequent preprocessing and modeling decisions.

The analysis began by examining the distributions of key continuous variables (Figure **??**). Histograms were generated for AGE and LIMIT_BAL to understand their spread and central tendency. Furthermore, the distribution of the target variable, default.payment.next.month, was analyzed to assess the class balance between defaulting and non-defaulting clients.

To investigate the influence of demographic factors on default rates, bar charts were created. These visualizations compared the proportion of defaults across the different categories of SEX, EDUCATION, and MARRIAGE. The relationship between continuous predictors and the default outcome was also explored. For instance, density plots were used to compare the distribution of LIMIT_BAL for both defaulting and non-defaulting clients.

Finally, to assess the linear relationships and potential multicollinearity among the numeric predictor variables, a correlation matrix was computed and visualized using a correlogram. The insights derived from this EDA phase were instrumental in understanding the primary drivers of default and guiding the feature engineering process.

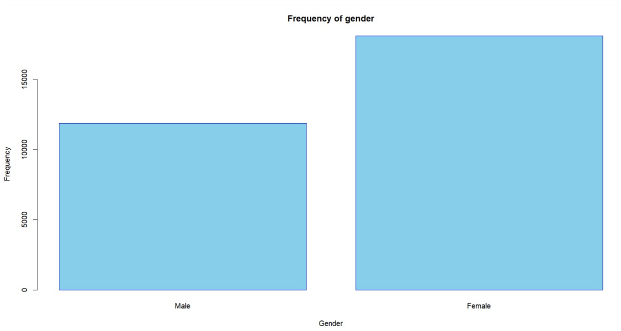

Fig. 9. Frequency of Gender

IV. Results and discussions

A. Exploratory Analysis Results

- **Descriptive Statistics and Data Distribution**: The exploratory data analysis revealed important characteristics of the credit card dataset. Age distribution showed customers ranging from a minimum of 21 years to a maximum of 79 years, with the majority concentrated in younger age groups. The dataset exhibited significant class imbalance in the target variable, with the majority of customers (approximately 77.88%) not defaulting on their payments, while only 22.12% experienced default in the subsequent month.
- **Normality Assessment**: Statistical testing using the Anderson-Darling [8] test revealed that none of the continuous variables in the dataset followed a normal distribution. This finding was consistent across all tested variables including LIMIT_BAL and all BILL_AMT and PAY_AMT variables, indicating the need for non-parametric approaches or data transformation techniques in subsequent analyses.
- **Correlation Analysis**: Correlation analysis revealed several noteworthy patterns among the variables. Strong positive correlations were observed among consecutive payment status variables (PAY_0 through PAY_6), suggesting consistent payment behavior patterns over time. Similarly, bill amount variables (BILL_AMT1 through BILL_AMT6) showed high intercorrelations, indicating stability in customers' outstanding balances across months. Payment amount variables (PAY_AMT1 through PAY_AMT6) also demonstrated positive correlations, reflecting consistent payment patterns among customers.
- **Demographic and Behavioral Patterns**: Analysis of demographic factors revealed distinct patterns in default behavior. Statistical testing showed significant differences in credit limits across different demographic groups. Customers with higher education levels demonstrated significantly different credit limit distributions compared to those with lower education levels. Similarly, marital status showed statistically significant associations with credit limits, with married and single customers exhibiting different credit profiles.
- **Distribution Characteristics**: The analysis of variable distributions showed that most numerical variables exhibited right-skewed distributions with positive skewness values, indicating the presence of outliers and non-normal data patterns. Kurtosis analysis revealed that several variables had high kurtosis values, suggesting heavy-tailed distributions with more extreme values than would be expected in normal distributions.

B. Predictive Model's Performance

- **Model Development and Training**: A logistic regression model [6] was developed using multiple predictor variables including demographic factors (LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE), payment history variables (PAY_0 through PAY_6), bill amounts (BILL_AMT1 through BILL_AMT6), and payment amounts (PAY_AMT1 through PAY_AMT6). The model was trained on the balanced dataset created through oversampling techniques to address the original class imbalance issue.
- **Model Evaluation Metrics**: The trained logistic regression model demonstrated moderate predictive performance on the test dataset. The model achieved an overall accuracy of 67.63%, indicating that approximately two-thirds of the predictions were correctly classified. While this accuracy level suggests reasonable predictive capability, it also highlights the inherent complexity of credit default prediction.
- **ROC Analysis and AUC Performance**: The model's discriminatory power was further evaluated using Receiver Operating Characteristic (ROC) analysis. The Area Under the Curve (AUC) value of 0.7127 (71.27%) indicates moderate discriminatory ability, falling into the acceptable range for binary classification problems. This AUC value suggests that the model can distinguish between defaulting and non-defaulting customers with reasonable effectiveness, though there remains room for improvement through feature engineering or alternative modeling approaches [7].
- **Confusion Matrix Analysis**: The confusion matrix analysis provided detailed insights into the model's classification performance, showing the distribution of true positives, true negatives, false positives, and false negatives. This analysis revealed the model's strengths and limitations in correctly identifying both defaulting and non-defaulting customers, providing a foundation for understanding the practical implications of the model's predictions in real-world credit risk assessment scenarios.

C. Discussions

- **Key Findings and Implications**: This study successfully identified critical factors associated with credit card default risk, addressing the primary research questions. The analysis revealed that demographic and behavioral factors play significant roles in determining default probability, with credit limit allocation emerging as a particularly

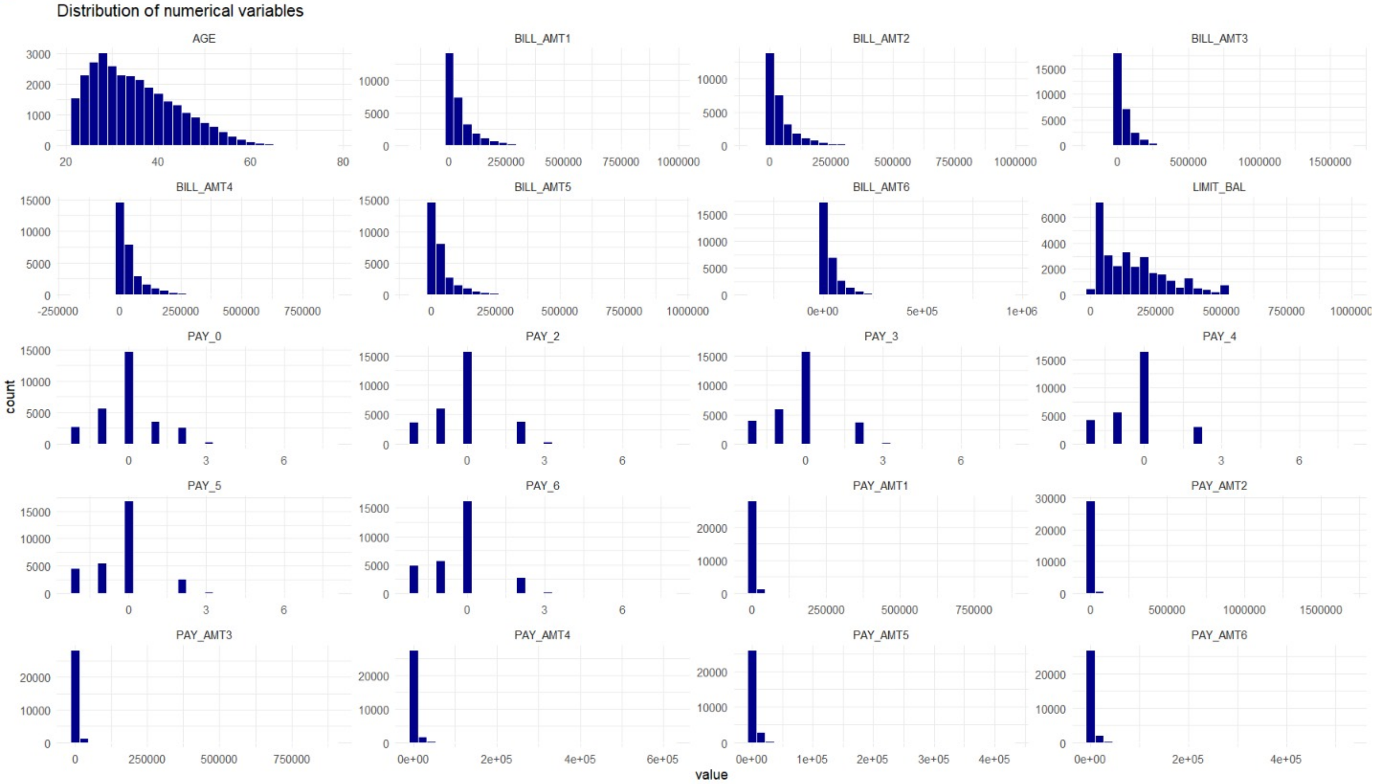

Fig. 10. Distribution of Numerical Variables

strong predictor. Customers who defaulted demonstrated substantially lower average credit limits compared to non-defaulting customers, suggesting that financial institutions already incorporate risk assessment in their credit allocation decisions. The payment history variables showed strong intercorrelations, indicating that past payment behavior serves as a reliable indicator of future payment patterns. This finding supports traditional credit scoring approaches that heavily weigh payment history in risk assessment models.

- **Model Performance and Practical Applications**: The logistic regression model achieved moderate predictive performance with an accuracy of 67.63% and an AUC of 71.27%. While these metrics indicate reasonable discriminatory power, they also highlight the inherent complexity of credit default prediction. The model could serve as a useful screening tool for financial institutions when combined with expert judgment and additional risk assessment procedures.

D. Web Application

To improve accessibility and replicability of the predictive model, we deployed it as a web application using R Shiny. The tool allows users to input demographic and financial variables and receive a prediction of default probability. The application is publicly available at: **https://rafaelf-lima.shinyapps.io/api_shiny/**.

V. Threats to Validity

This study relies on the Default of Credit Card Clients dataset, which was collected in 2005 and includes data from clients in Taiwan. This presents a potential threat to external validity, as the economic, cultural, and regulatory context of credit usage may differ significantly across countries and over time. Consequently, the generalization of the results to current populations or other geographic regions should be made with caution. Future research could explore more recent and diverse datasets to enhance the robustness and applicability of the findings.

Additionally, a threat to internal validity arises from the use of oversampling techniques to address class imbalance. While synthetic sampling methods like ROSE help balance the training data, they may introduce noise or artificial patterns that do not reflect the true underlying distribution of the population. This can lead to overfitting, where the model performs well on the training data but fails to generalize effectively to unseen cases. Future research should consider alternative resampling strategies, regularization techniques, or ensemble methods to mitigate this risk.

VI. Conclusion

This study provides valuable insights into credit card default prediction using machine learning techniques applied to the UCI Credit Card dataset. The research successfully identified key demographic and behavioral factors associated with default risk and demonstrated the feasibility of constructing predictive models for credit risk assessment. The analysis revealed that credit limit allocation, payment history patterns, and demographic characteristics serve as important predictors of default behavior. Future research should focus on exploring advanced machine learning techniques and incorporating additional data sources to further improve predictive performance in credit risk assessment applications.

References

[1] Addy, Wilhelmina and Ugochukwu, Chinonye and Oyewole, Adedoyin and Adeoye, Omotayo and Okoye, Chinwe. (2024). Predictive analytics in credit risk management for banks: A comprehensive review. GSC Advanced Research and Reviews. 18. 434-449. 10.30574/gscarr.2024.18.2.0077.

[2] L. Jantsch, J. L. Becker, P. Solana-González, and A. A. Vanti, “Analysis of default risk in credit card use / Análise do risco de inadimplência na utilização de cartões de crédito”, Braz. J. Develop., vol. 7, no. 6, pp. 62634–62656, Jun. 2021.

[3] I. Yeh. ”Default of Credit Card Clients” UCI Machine Learning Repository, 2009. [Online]. Available: https://doi.org/10.24432/C55S3H.

[4] Kaggle, “Default of Credit Card Clients Dataset,” [Online]. Available: http://kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset.

[5] NIST, “Anderson-Darling Test” [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm.

[6] D. Jurafsky and J. Martin, “Speech and Language Processing” [Online]. Available: https://web.stanford.edu/ jurafsky/slp3/5.pdf.

[7] Google, “Classification: ROC and AUC — Machine Learning” [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=pt-br.

[8] M. Merlo, “Default-of-Credit-Card-Clients-Dataset-Analisys” GitHub, [Online]. Available: https://github.com/MatteoM95/Default-of-Credit-Card-Clients-Dataset-Analisys. keywords: Predictive models;Credit cards;Data models;Decision trees;Boosting;Australia;Machine learning;imbalanced data;customer credit risk;credit card default model;interpretable model;gradient boosted decision tree.

[9] Y. Li et al., “A Study on the Calibration of Multiple Non-Overlapping Cameras for SLAM,” in *Proc. 2019 CHI Conf. on Human Factors in Computing Systems*, ACM, 2019.

[10] Z. Li et al., “Deception Detection via Multimodal Analysis,” in *Proc. 2019 CHI Conf. on Human Factors in Computing Systems*, ACM, 2019.