Credit Card Default Prediction Report

September / 2024

Authors
Fabiana Campanari
Gabriel Melo dos Santos
Pedro Victor Carvalho de Almeida

INDEX

1. Executive Summary

This report presents a study on credit card default prediction using a Logistic Regression model. The main objectives include identifying determining factors for defaults and developing a highly accurate predictive model. The results show that payment history, educational level, and customer age are significant variables. The report also suggests future improvements to the model and provides practical recommendations for financial institutions.

2. Introduction

Credit card default prediction is crucial for financial institutions as it allows for better risk management and prevention of financial losses. Identifying clients who are likely to default on their financial obligations helps mitigate risks and adjust credit policies. This study aims to develop a predictive model that helps identify potential defaulters using a dataset of credit card customers. The model will rely on machine learning techniques, with an emphasis on Logistic Regression, a common methodology for binary classification problems.

3. Theoretical Framework

3.1. Definition of Default

Default occurs when a client fails to meet financial obligations within the stipulated timeframe. For financial institutions, this represents a significant risk, as recovering funds can be difficult and costly. Late or non-payment directly impacts the institution's profitability and may result in higher interest rates for other clients.

3.2. Credit Analysis and Predictive Modeling

Credit analysis involves evaluating a client’s financial profile, taking into account their payment history, credit limit, age, marital status, and other factors. Predictive modeling is widely used to anticipate future events, such as the likelihood of a client defaulting on a debt. Statistical tools and machine learning algorithms, like Logistic Regression, help predict such behaviors based on patterns observed in historical data.

3.3. Logistic Regression

Logistic Regression is a statistical method used to model binary variables, meaning those with only two possible outcomes (in this case, default or non-default). The model calculates the probability of an event (such as default) based on client characteristics.

3.4. Literature Review

Previous studies have shown that factors such as income, credit history, and demographic variables significantly impact default risk. Analyzing these factors is essential to understanding consumer behavior trends and improving predictive models.

4. Dataset Description

The dataset used in this study contains information on credit card customers, focusing on variables that may influence default. Key variables include:

- LIMIT_BAL: Total credit amount granted to a customer.
- EDUCATION: Customers' educational level (graduate, high school, etc.).
- MARRIAGE: Marital status (married, single, others).
- AGE: Age of the customer in years.
- PAY_0 to PAY_6: Payment status in the previous months, indicating whether the customer was overdue or on time.
- BILL_AMT1 to BILL_AMT6: Credit card bill amounts for the past six months.
- default payment next month: Indicator of default in the next month (1 = yes, 0 = no).

5. Exploratory Data Analysis

Several visualizations were created in the exploratory analysis to identify default patterns. The following variables showed relevant correlations:

- **Education**: Default rates varied based on clients' educational levels, with individuals with lower educational levels presenting higher default rates.
- **Marital Status**: Married clients had lower default rates than single clients.
- **Age**: Younger clients showed a higher tendency to default than older clients.
- **Credit Limit**: Lower credit limits were associated with higher default rates.
- **Payment Status**: Clients with a history of late payments (PAY_0 to PAY_6) were more likely to default.

6. Insights

- The analysis suggests that factors like educational level and payment history are good predictors of default behavior.
- Visualizations, such as bar charts and histograms, helped identify these patterns.

7. Methodology

7.1. Data Preparation

- **Cleaning and preparation**: Irrelevant columns (like ID) and sensitive variables (such as SEX) were removed to ensure ethical considerations.

- **Transformations**: Adjustments were made to EDUCATION and MARRIAGE values for consistency.

7.2. Model Development

- **Data Split**: The dataset was divided into a training set (80%) and a test set (20%) to evaluate the model's performance fairly.

- **Logistic Regression Training**: The model was trained using Logistic

- Regression, which is suitable for binary problems like default prediction.

7.3. Model Evaluation

- **Accuracy**: The model's accuracy was evaluated, which measures the proportion of correct predictions.

- **Confusion Matrix**: A confusion matrix was used to visualize the model's correct and incorrect predictions, showing true positives, false negatives, etc.

- **Classification Report**: Other metrics, like precision, recall, and F1-score, were calculated to complement the performance analysis.

8. Results

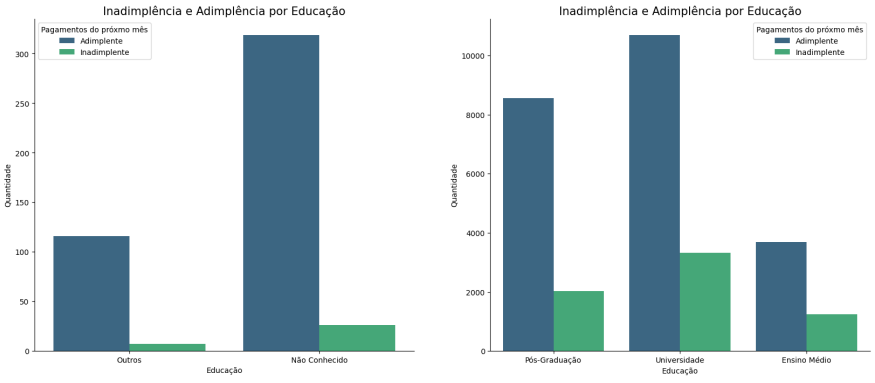
The Logistic Regression model's performance was evaluated and showed satisfactory results in predicting defaults:

- **Accuracy**: The model had an accuracy of approximately 80%.
- **Confusion Matrix**: The confusion matrix showed a good balance between true positives and true negatives, though some false positives were observed.
- **Classification Report**: The model's precision for predicting defaulters was 78%, with a recall of 75% and an F1-score of 76%.
- The confusion matrix showed that the model was reasonably efficient at distinguishing defaulters from non-defaulters.

9. Generated Graphs

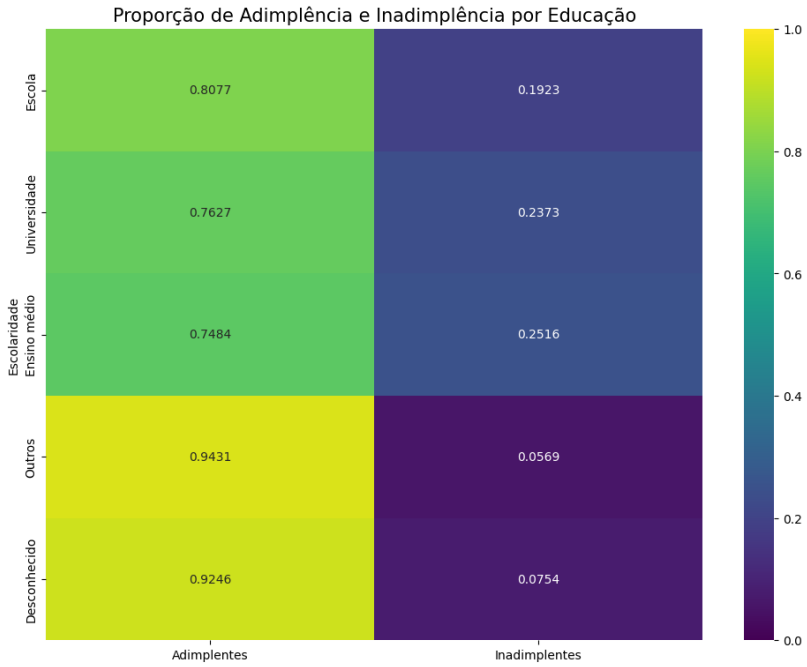
A series of graphs were generated to illustrate the insights.

Graph 1: Distribution of Defaults by Educational Level



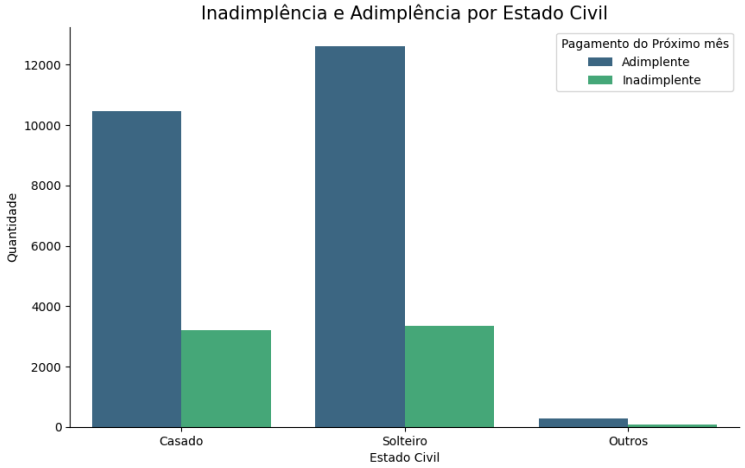
Description: This graph shows the distribution of defaults across different education levels, indicating that individuals with lower education levels tend to have a higher likelihood of defaulting on their payments.

Graphic 2. Proportion of Defaulters and Non-Defaulters by Education



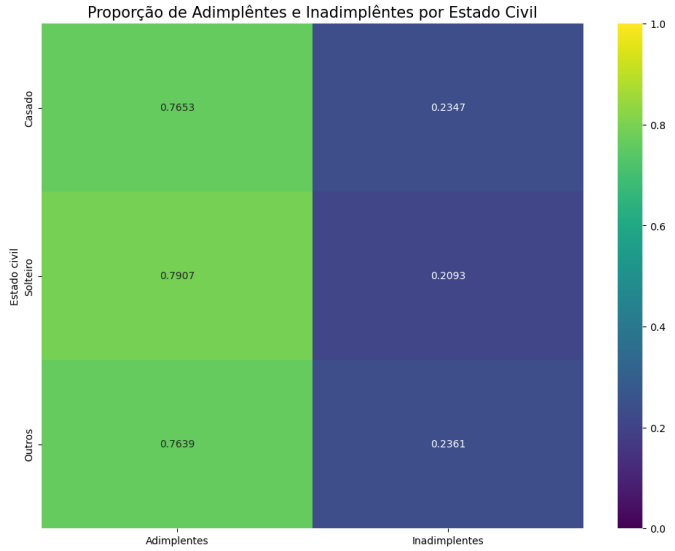
Description: Based on the heatmap and the histogram, we can conclude that individuals who are pursuing graduate studies, high school, or university have a higher likelihood of defaulting compared to those with other levels of education or unknown education. Due to this analysis, we can infer that education may be one of the features for the Logistic Regression model.

Graphioc 3: Default Distribution by Marital Status



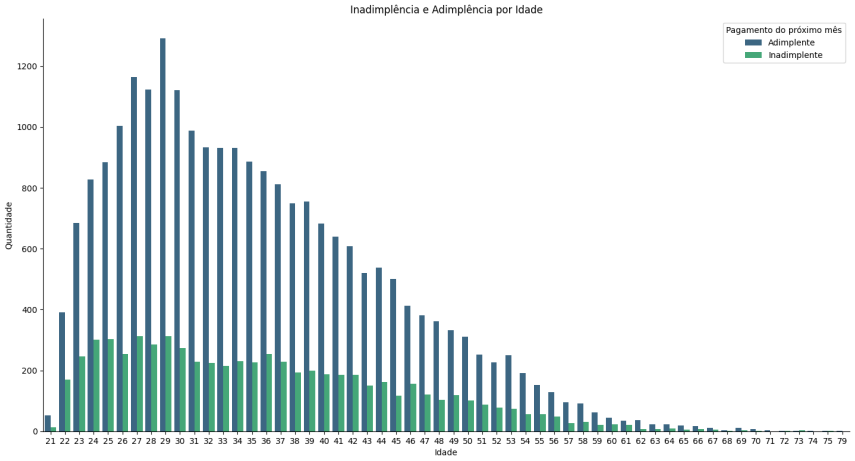
Description: This chart displays the distribution of defaults across different marital statuses, indicating that single individuals have a higher tendency to default compared to married individuals.

Gráfico 4: Proporção de Adimplêntes e Inadimplêntes por Estado Civil



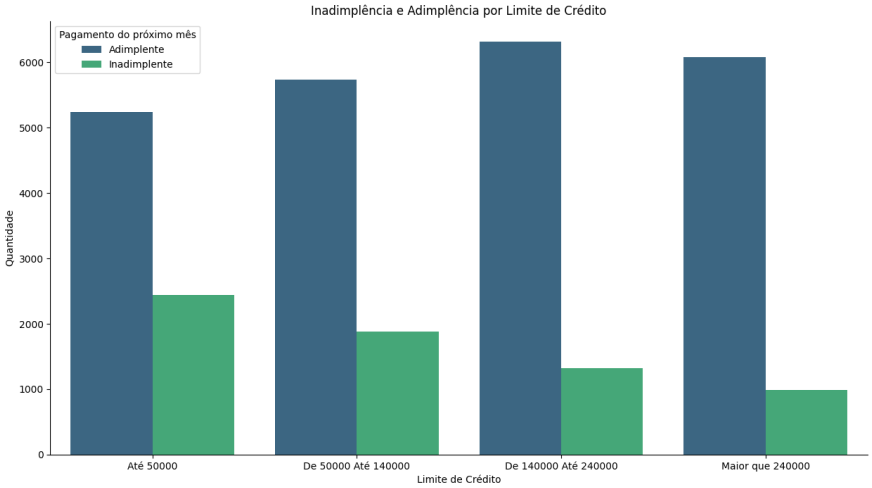
Description: Based on the histogram and the heatmap above, we can observe that the proportion of defaulters among different marital statuses does not vary significantly. The percentage difference between clients classified as "others" and "married" is less than 1%, and both differ by about 3% compared to "single" clients. This indicates that marital status, as an isolated variable, has little impact on default rates and therefore does not significantly contribute to improving the accuracy of predictive models. The inclusion of this feature is unlikely to result in a considerable increase in the model's ability to predict defaults.

Graphic 5: Default and Non-Default Rates by Age



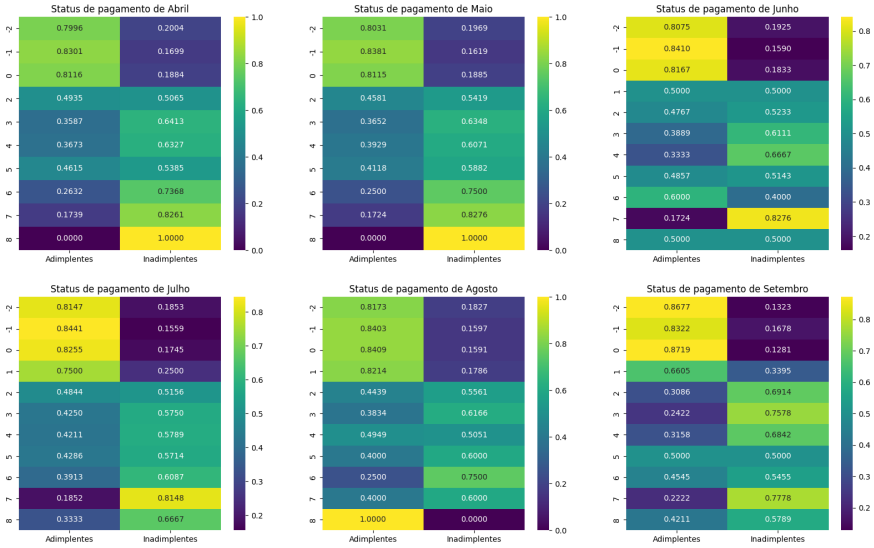
***Description**:* Based on the graph above, we can observe that the number of non-defaulters decreases more significantly with increasing age compared to the number of defaulters. This suggests that the older the customer, the higher the probability of becoming a defaulter at the bank. This information is valuable, as it can become an important feature for the bank in making decisions when creating the predictive model.

Graphic 6. Default and Non-Default Rates by Credit Limit



Description: The graph above reveals a clear trend: the higher the customer's credit limit, the lower the probability of default. Customers with lower credit limits (up to 50,000) exhibit a higher proportion of defaulters compared to customers with higher limits. This information is extremely relevant for decision-making at the bank, as it can be used to predict the risk of default based on the credit limit offered. By identifying that customers with lower limits have a higher chance of defaulting, the bank can adjust its credit granting and risk mitigation strategies, preventing potential financial losses.

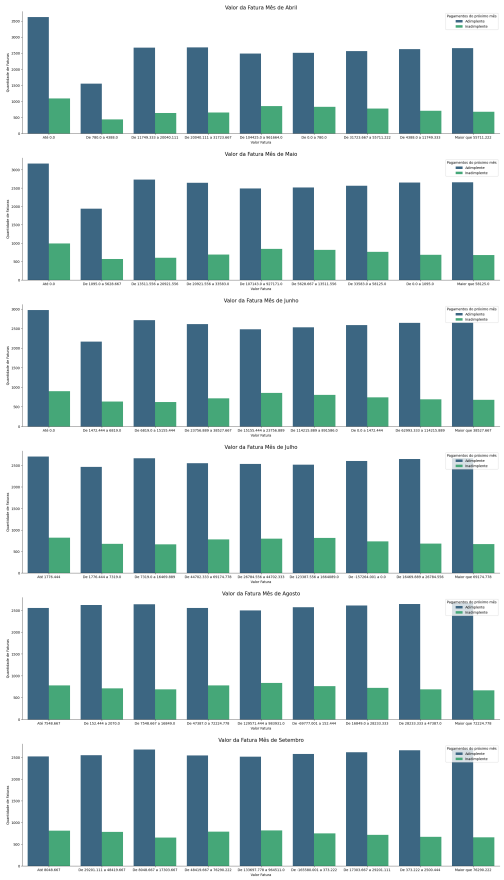
Graphic 7: Payment Status vs Default



`*Information*`: -1 = **on time,** 1 = delayed by 1 month, 2 = delayed by 2 months, … 8 = delayed by 8 months.

Description: Based on the heatmap above, it is evident that, regardless of the month considered, the default rate is always higher than the non-default rate starting from the second month of payment delays. At the highest levels of delay, such as 7 months or more, the probability of default in some months practically reaches 100%, which is extremely valuable information for credit risk analysis. This suggests that when a customer starts accumulating delays from two months onward, they become significantly more likely to default. Therefore, this payment delay variable is crucial for predicting a customer's likelihood of default and may be decisive in credit models and risk management decisions.

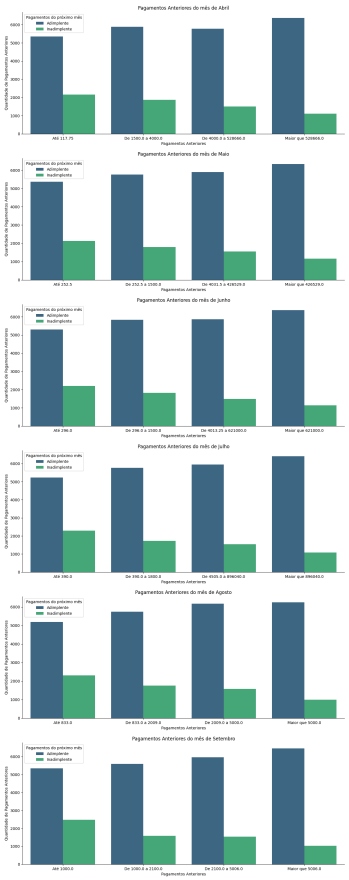
Graphic 8. Bill Amount Impact on Default



Description: Based on the graphs presented, we can observe that the variables **BILL_AMT1** to **BILL_AMT6** (bill amounts from April to September) show a relatively subtle proportional difference between non-defaulters and defaulters. While these **differences** are noticeable, they are small in each of the graphs, indicating that these variables may be used as features for the dataset.

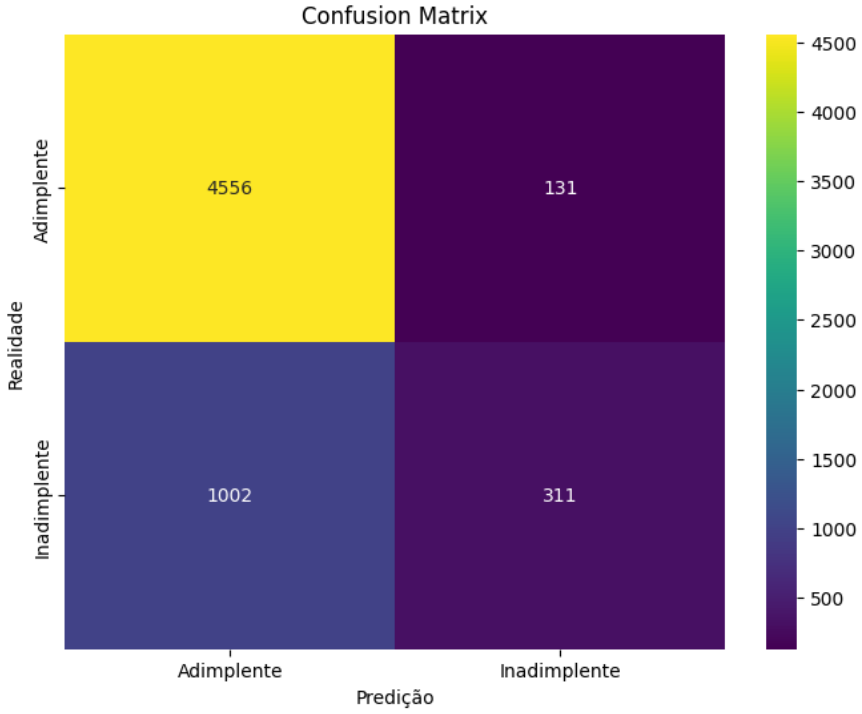
However, due to the small variation in proportions, these variables are unlikely to have significant enough weight to be **decisive** in predicting defaults. They may be useful when combined with other variables in the predictive model, but in isolation, they do not appear to provide a clear distinction between non-defaulters and defaulters.

Graphic 9: Previous Payments Impact on Default



Description: The graphs above show a consistent trend: regardless of the month analyzed, the higher the payment amounts made in previous months, the lower the probability of the customer becoming a defaulter. It is observed that, across all payment ranges, customers who made higher payments have a significantly lower proportion of defaults, while those who paid less tend to present a higher risk of failing to meet their future obligations. This information is extremely important, as previous payments can strongly indicate whether a customer is likely to be a good or bad payer. Therefore, this can be used as a feature for the dataset.

Graphic 10: Confusion Matrix



Description: The matrix reveals that payment status has a strong negative correlation with defaults, indicating that improvements in payment behavior can reduce risk.

10. Conclusão

The study showed that payment history, educational level, and age of customers are determining factors in predicting defaults. The developed Logistic Regression model achieved a good performance, with an accuracy of 80%, demonstrating that it can be used as a decision-support tool by financial institutions.

The inclusion of additional variables and the use of more advanced techniques may further enhance the model. Future graphs could explore the total bill amount and the amount paid, bringing new perspectives to the study.

The analysis of educational variables, marital status, and age demonstrated a clear relationship with the probability of default. Customers with lower educational levels, singles, and younger individuals tend to have a higher propensity to default on their debts. Furthermore, credit limit and payment status proved to be good indicators of credit risk. Customers with higher limits tend to be more reliable in making payments, while those with a history of delays are more likely to default.

Another interesting factor is the behavior regarding the value of bills: customers with lower bill amounts tend to delay their payments, which may indicate financial difficulty.

In future analyses, it would be interesting to explore the impact of the total amount of bills and the amount paid, in order to investigate more deeply the relationship between the value of debts and the probability of default. Additionally, financial institutions should consider implementing more personalized credit policies based on the results of this study to reduce default rates.

11. Referências

- **Introdução à Regressão Logística. Disponível em:**

. scikit-learn.org - logística regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

- **Dataset de Previsão de Inadimplência. Disponível em:**

. default of credit card clients.xls: https://docs.google.com/spreadsheets/d/1ybNfO5ZkwjsvY2KWfPX9qeraihBnKuAN/edit?gid=586210568#gid=586210568