Analysis of the "Loan Default" dataset

1. Introduction

In the contemporary economic scenario, financial loan services play a crucial role in the financial dynamics of both individuals and companies. Whether for acquiring consumer goods, investing in education, expanding businesses, or other financial needs, loans play a fundamental role by providing access to capital that can drive economic growth and social well-being. However, while these services are vital for stimulating economic activity, default control emerges as a significant challenge for financial institutions that offer these loans. From this perspective, it is imperative to highlight the importance of using data analysis to understand and control default in this type of service, as it not only strengthens the capacity of financial institutions to manage risks but also promotes financial sustainability and operational efficiency.

In the present study, a dataset obtained from the "Kaggle" online platform was used, named "Loan Default Prediction Dataset" (https://www.kaggle.com/datasets/nikhil1e9/loan-default). The dataset comprises 255,347 rows and 18 columns, with one column for identification, 16 columns for features, and one column for the target. Table 1 shows the description of each column.

Table 1. Description of columns in the "Loan Default Prediction Dataset" dataset.

Column name	Description
Loan ID	A unique identifier for each loan
Age	The age of the borrower
Income	The annual income of the borrower
Loan Amount	The amount of Money being borrowed
Credit Score	The credit score of the borrower, indicating their creditworthiness
Months Employed	The number of months the borrower has been employed
Num Credit Lines	The number of credit lines the borrower has open
Interest Rate	The interest rate (annual) for the loan
Loan Term	The term length of the loan in months
DTI Ratio	The Debt-to-income-ratio, indicating the borrower's debt compared to their income
Education	The highest level of education attained by the borrower (PhD, Master's, Bachelor's, High School)
Employment Type	The employment current status of the borrower (full-time, part-time, self-employed, unemployed)
Marital Status	The marital status of the borrower (single, married, divorced)
Has Mortgage	Whether the borrower has a mortgage (yes or no)
Has Dependents	Whether the borrower has dependents (yes or no)
Loan Purpose	The purpose of the loan (home, auto, education, business, other)
Has Cosigner	Whether the loan has a co-signer (yes or no)
Default	The binary target variable indicating whether the loan defaulted (1) or not (0)

The objective of the study was to carry out a descriptive and exploratory analysis of the data, as well as predictive modeling to predict the probability of default based on the borrower's profile. To develop the analyses, the following tools were used: DBeaver (SQL language), PowerBI, and Jupyter Lab (Python language).

2. Descriptive and exploratory analysis of quantitative variables

The table 2 presents the number of observations (count), the values of mean, standard deviation (std), minimum value (min), 25th percentile (25%), 50th percentile (50%), 75th percentile (75%) and maximum value (max) for all quantitative variables in the database.

Table 2. Descriptive analysis of the quantitative variables data (Age, Income, Loan Amount, Months Employed, Interest Rate, and DTI Ratio).

	Age	Income	LoanAmount	CreditScore	MonthsEmployed	InterestRate	DTIRatio
count	255347.00	255347.00	255347.00	255347.00	255347.00	255347.00	255347.00
mean	43.50	82499.30	127578.87	574.26	59.54	13.49	0.50
std	14.99	38963.01	70840.71	158.90	34.64	6.64	0.23
min	18.00	15000.00	5000.00	300.00	0.00	2.00	0.10
25%	31.00	48825.50	66156.00	437.00	30.00	7.77	0.30
50%	43.00	82466.00	127556.00	574.00	60.00	13.46	0.50
75%	56.00	116219.00	188985.00	712.00	90.00	19.25	0.70
max	69.00	149999.00	249999.00	849.00	119.00	25.00	0.90

We can observe that the minimum, mean, and maximum age of the borrowers are 18, 43, and 69 years, respectively. The minimum value of Income (annual salary) is $15,000, while the maximum value is approximately $149,999, around 10 times the minimum value. Additionally, 50% of the borrowers have an Income greater than $82,466. The credit score ranges from 300 to 849, and the loan amount ranges from $5,000 to $249,999, with an average of approximately $127,600. Regarding the employment length (Months Employed), the database includes borrowers who are unemployed/retired (0 months) to borrowers who have been employed for more than 9 years (119 months). On average, the borrowers have been employed for approximately 5 years. The Interest rate can vary from 2 to 25 % per year (simple interest). On average, the rate is 13.5 %. We can also add that only 25 % of the borrowers pay an annual interest rate greater than 19.25 %.

The data of the variable representing the ratio between the loan amount and the client's annual salary (DTI Ratio) has a minimum value of 0.10, a mean of 0.50, and a maximum of 0.90. This means that, on average, the borrowers have an annual salary twice as high as the loan amount. Additionally, approximately 25 % of the borrowers have a DTI Ratio greater than 0.70, meaning that for these borrowers, the loan amount represents 70% of their annual salary (Income).

Finally, it is interesting to highlight that in all the data analyzed in Table 2, the disparity between mean and median values is negligible. This indicates that these data exhibit a balanced distribution.

Distribution graphs were obtained for each quantitative variable. These graphs revealed that all variables were evenly distributed. In other words, the data from all quantitative variables had the same distribution function shape. To illustrate, Figure 1 presents the distribution graph of the “Income” variable. Additionally, using Spearman's correlation coefficient, no correlation was evidenced between these variables. In Figure 2, the

coefficient values are presented in the form of a heatmap. The closer the value is to 1 or -1, the stronger the correlation between the variables. However, as observed in Figure 2, the coefficient values obtained are very close to zero, indicating that there is no significant correlation between the variables.

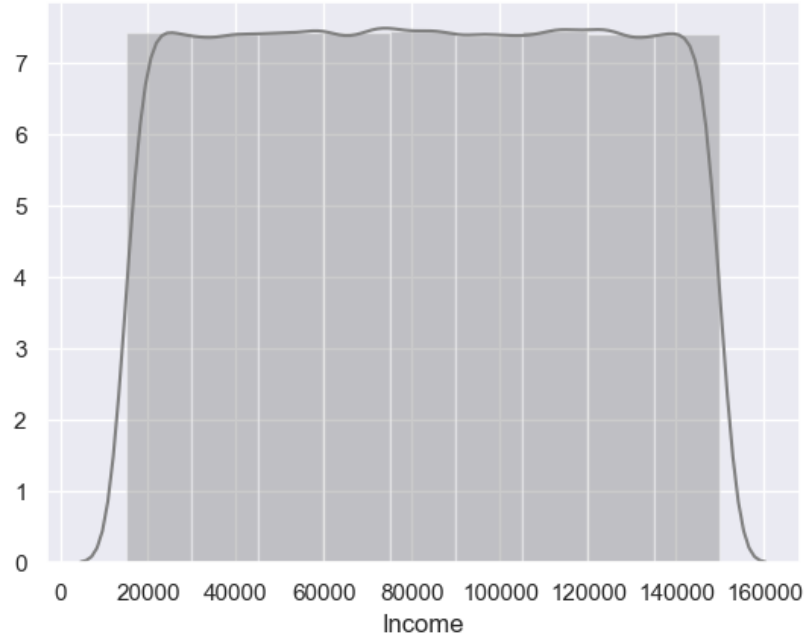


Figure 1. Graph of the distribution function of the “Income” variable.

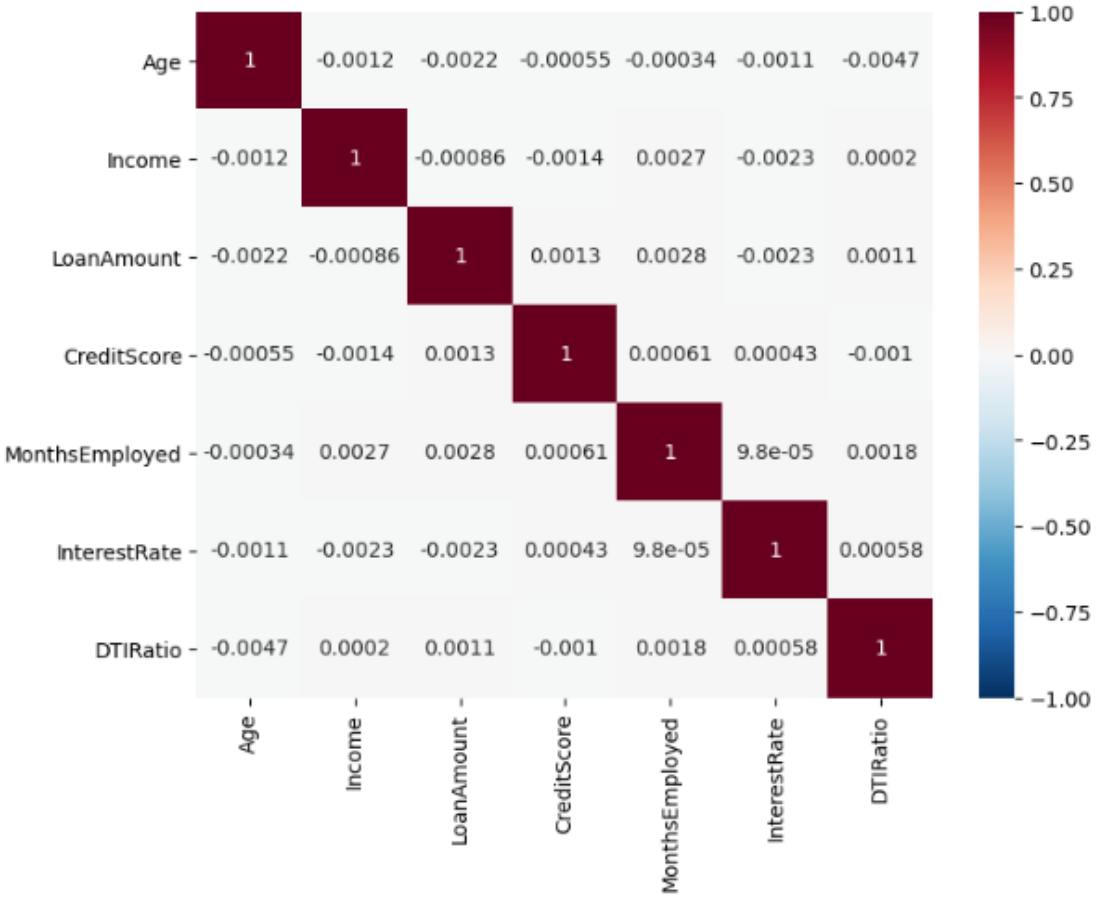


Figure 2. Heatmap of Spearman's Correlation Coefficients for all quantitative variables.

Figure 3 shows the loan amount and the percentages of loan amount in relation to default (Default). The total percentage of default in the dataset was 11.6 %. The total loan amount granted was $32,576,880,572, with approximately 13 % of this amount, or $4,285,312,531, not being repaid (Default = 1).

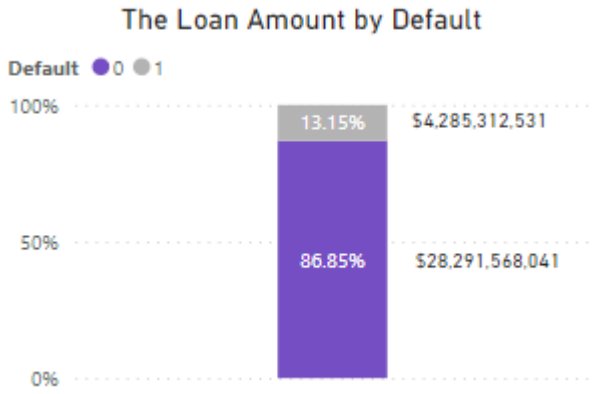
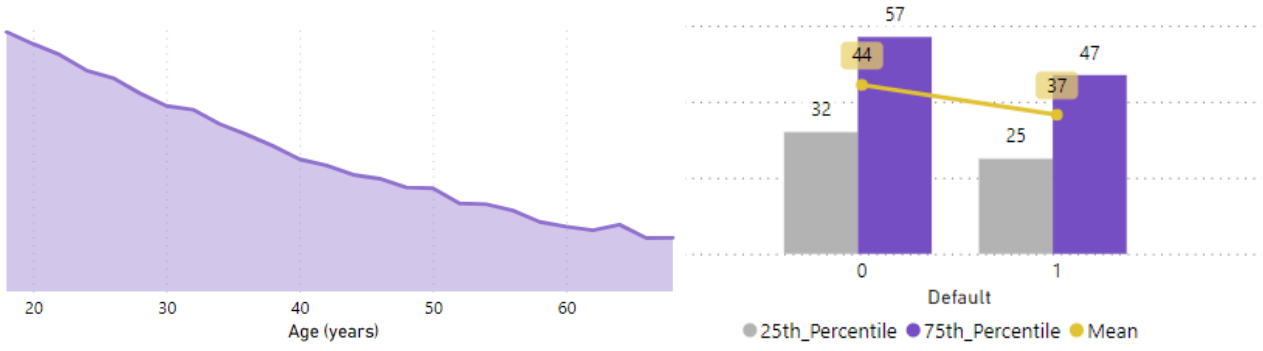


Figure 3. Loan amount repaid (Default = 0) and not repaid (Default = 1).

2.1 Distribution of the quantitative variables by Default

Figure 4 - (a) presents the distribution curve of the age data of the defaulting borrowers. We can observe that the area under the density curve continuously and sharply increases as age decreases, reaching a maximum at the age of 18. Figure 4 - (b) displays a bar graph with the mean values and the 25th and 75th percentiles of the age of defaulting (Default = 1) and non-defaulting borrowers (Default = 0). It can be concluded that 75 % of the defaulting borrowers are aged up to 47 years, exactly 10 years less than 75 % of the non-defaulting borrowers. Additionally, the average age of the defaulting borrowers is 25 years, while for the non-defaulters it is 32 years. Therefore, considering the observations above, we conclude that the defaulting borrowers tend to be younger than the non-defaulting borrowers.

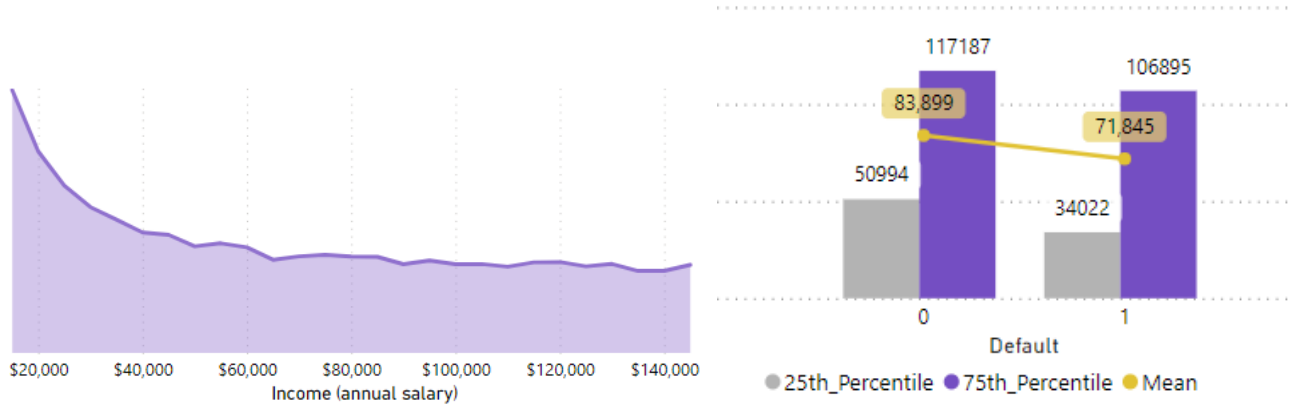


(a) (b)

Figure 4. Distribution of the "Age" data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the "Age" data by default (b).

The graph with the distribution curve of the Income data of the defaulting borrowers is presented in Figure 5 - (a). In this graph, we observe that the data follows a uniform distribution when the values of income are above $60,000. The area under the density curve begins to increase smoothly until reaching an income of $40,000

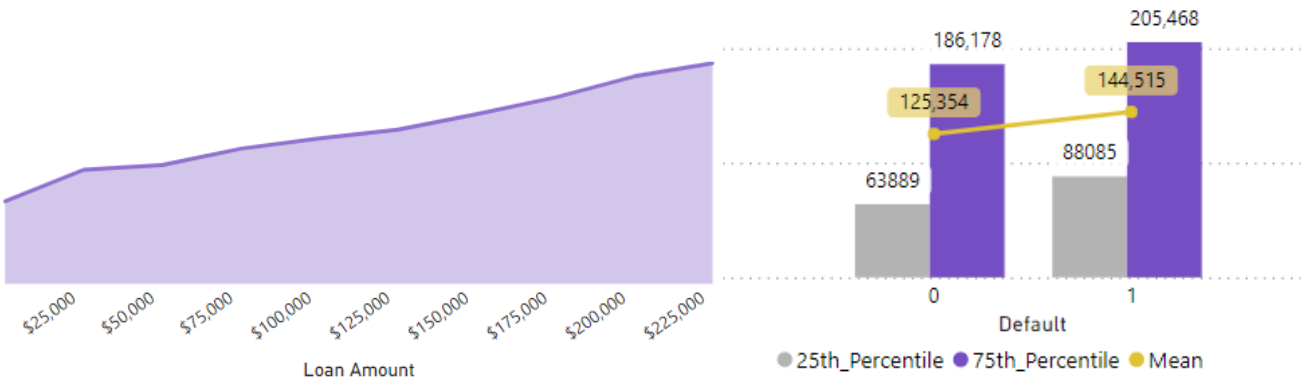
and undergoes an exponential increase as the values of income decreases, reaching a maximum value at $15,000. Figure 5 - (b) shows a bar graph with the mean values and the 25th and 75th percentiles of Income for the defaulting (Default = 1) and non-defaulting (Default = 0) borrowers. The income of the defaulting borrowers has an average of $71,845, approximately $12,000 less than the non-defaulting borrowers. The percentile values are also lower for the defaulting borrowers, with a difference of 33 % between the 25th percentiles and 8.8 % between the 75th percentiles. These results suggest that defaulting borrowers tend to receive a lower annual salary (income) compared to non-defaulting borrowers.



(a) (b)

Figure 5. Distribution of "Income" data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the "Income" data by default (b).

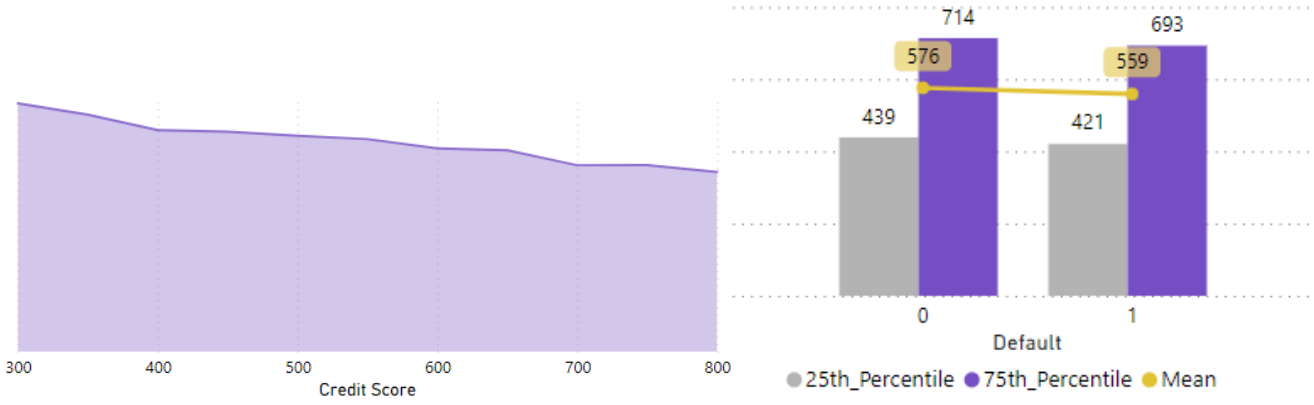
Figure 6 presents the distribution curve graph of the Loan Amount variable of the defaulting borrowers (a) and the bar graph with the mean and 25th and 75th percentiles of the Loan Amount data by default (b). Analyzing the data distribution curves (Fig. 6 - a), it is observed that the area under the density curve continuously increases as the loan amount increase. Observing Figure 6 - (b), we verify that the highest mean and percentile values belong to the group of defaulting borrowers. The difference between the 25th percentiles is 37.9 %, between the means is 15.3 %, and between the 75th percentiles is 10.4 %.



(a) (b)

Figure 6. Distribution of “Loan Amount” data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the " Loan Amount " data by default (b).

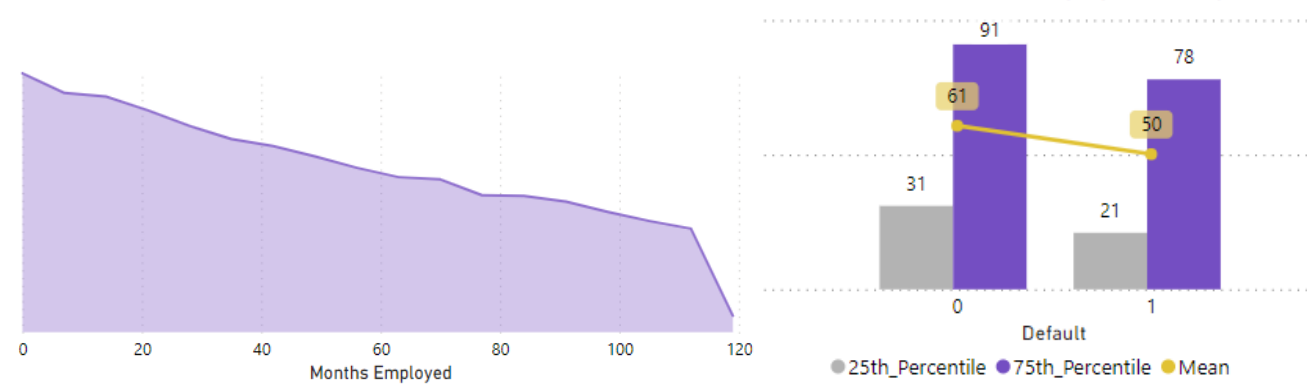
Figure 7 presents the distribution curve graph of the Credit Score variable of the defaulting borrowers (a) and the bar graph of the mean and 25th and 75th percentiles of the Credit Score data by default (b). The area under the data distribution curve undergoes a slight reduction as the Credit Score value increases, with a maximum value at Credit Score 300 and a minimum value at Credit Score 800. Observing Figure 7 - (b), we verify that defaulting borrowers present slightly lower Credit Score values compared to the non-defaulting group. However, the differences between the means and percentiles are not significant, approximately 3 %.



(a) (b)

Figure 7. Distribution of "Credit Score" data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the " Credit Score " data by default (b).

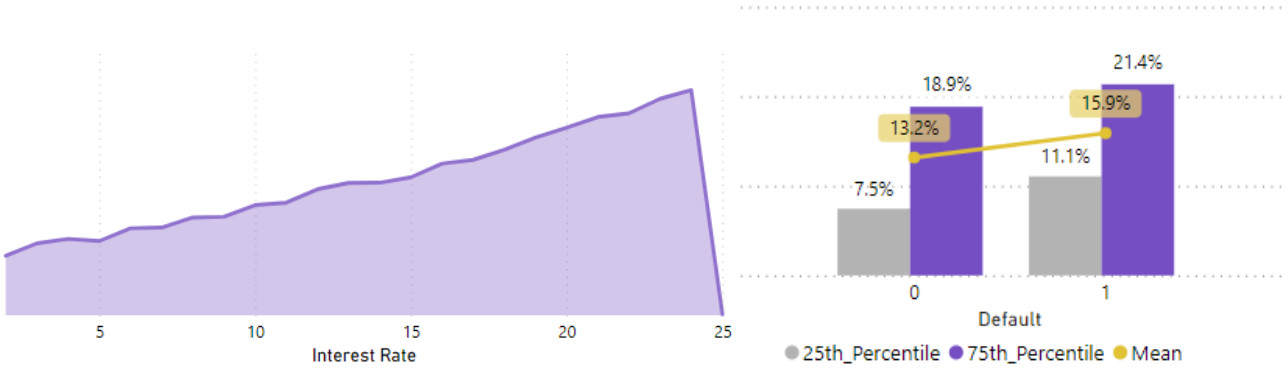
Figure 8 presents the distribution curve graph of the Months Employed variable of the defaulting borrowers (a) and the bar graph of the mean and 25th and 75th percentiles of the Months Employed data by default (b). We can observe that the curve area increases as the months of employment decrease, reaching a maximum value in the early months of employment. Observing Figure 8 - (b), we verify that the lowest mean and percentile values belong to the group of defaulting borrowers. The difference between the 25th percentiles is 10 months, between the means is 11 months, and between the 75th percentiles is 13 months. In summary, the results suggest that defaulting borrowers tend to be employed for a shorter period compared to the non-defaulting borrowers.



(a) (b)

Figure 8. Distribution of "Months Employed" data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the " Months Employed " data by default (b).

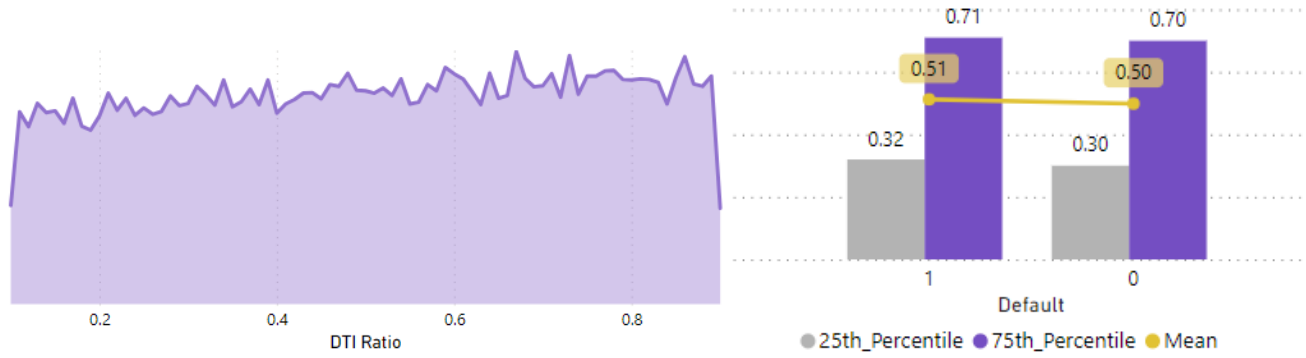
Figure 9 presents the distribution curve graph of the Interest Rate variable of the defaulting borrowers (a) and the bar graph of the mean and 25th and 75th percentiles of the Interest Rate data by default (b). Analyzing Figure 9 - (a), it can be observed that the curve of the data shows a sharp and continuous increase as the Interest Rate values increase, with a maximum concentration at 25 % of Interest Rate. Observing Figure 9 - (b), we verify that the mean and percentile values for defaulting borrowers (Default = 1) are higher than the values related to non-defaulting borrowers (Default = 0). The difference is 3.6 percentage points between the 25th percentiles, 2.7 percentage points between the means, and 2.5 percentage points between the 75th percentiles. Therefore, we can affirm that defaulting borrowers tend to have higher Interest Rate values compared to the non-defaulting borrowers.



(a) (b)

Figure 9. Distribution of "Interest Rate" data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the " Interest Rate " data by default (b).

Figure 10 presents the distribution curve graph of the DTI Ratio variable of the defaulting borrowers (a) and the bar graph of the mean and 25th and 75th percentiles of the DTI Ratio data by default (b). The distribution curve shows a uniform behavior. However, the curve of the data experiences a slight increase as the DTI Ratio also increases. Observing Figure 10 - (b), we verify that there is practically no difference between the mean and the percentile measures of the two groups of borrowers.



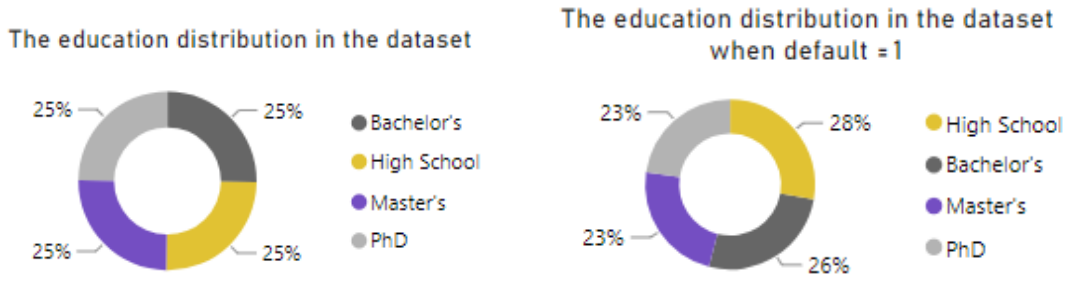
(a) (b)

Figure 10. Distribution of "DTI Ratio" data of the defaulting borrowers (a); Values of the 25th percentile, mean, and 75th percentile of the " DTI Ratio " data by default (b).

3. Descriptive and exploratory analysis of qualitative variables

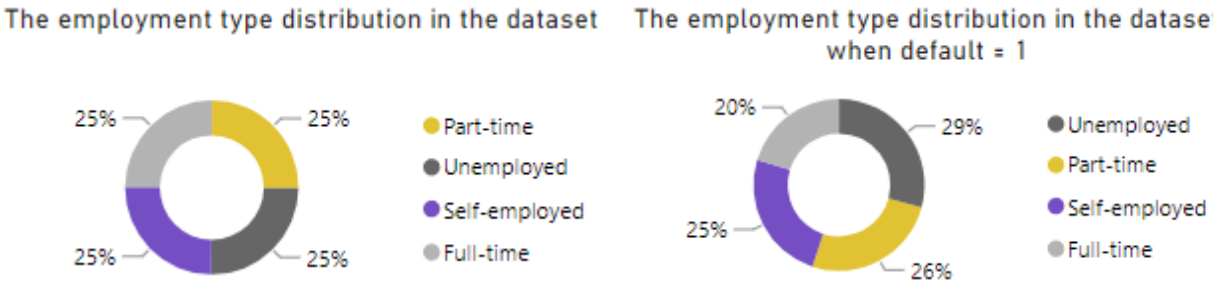
The graphs from Figures 11 to 16 - (a) show how the categories of the categorical variables are distributed in the dataset. It was observed that all categorical variables have a perfectly balanced distribution, meaning each category has the same number of observations as the other categories of the same variable.

The graphs from Figures 11 to 16 - (b) show how the categories of the categorical variables are distributed in the defaulting borrowers data (Default = 1). Comparing with the unfiltered conditions (a), we noticed that the distribution of categories was slightly altered. Nevertheless, it is important to highlight the categories that showed a small increase in concentration when Default = 1: "High School" and "Bachelor's" (Fig. 11), "Unemployed" (Fig. 12), "Divorced" (Fig. 13), and "4" in "Credit Lines" (Fig. 14). This suggests that borrowers belonging to these categories may have a greater tendency to default.



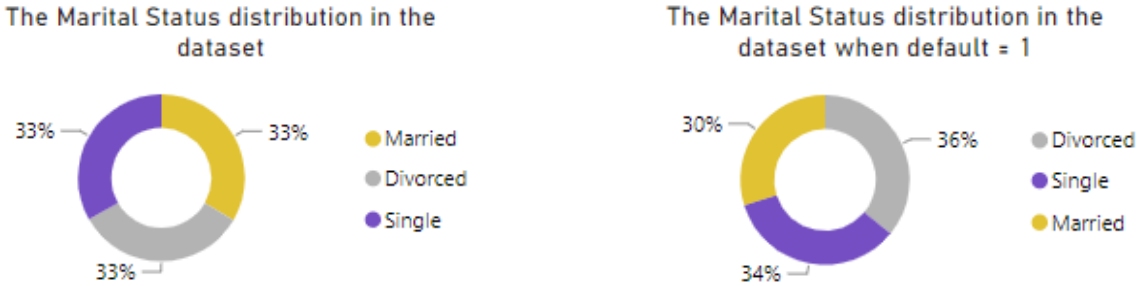
(a) (b)

Figure 11. Distribution of the Education categories in the database (a) and Distribution of the Education categories in the database with Default =1 (b).



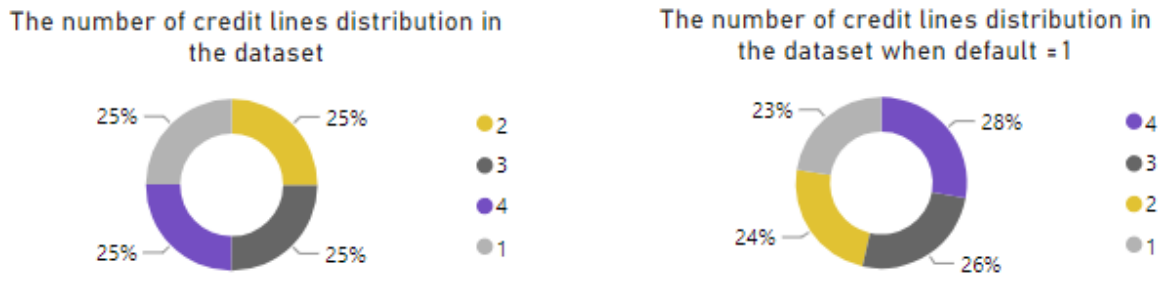
(a) (b)

Figure 12. Distribution of the employment types in the database (a) and Distribution of the employment types in the database with Default =1 (b).



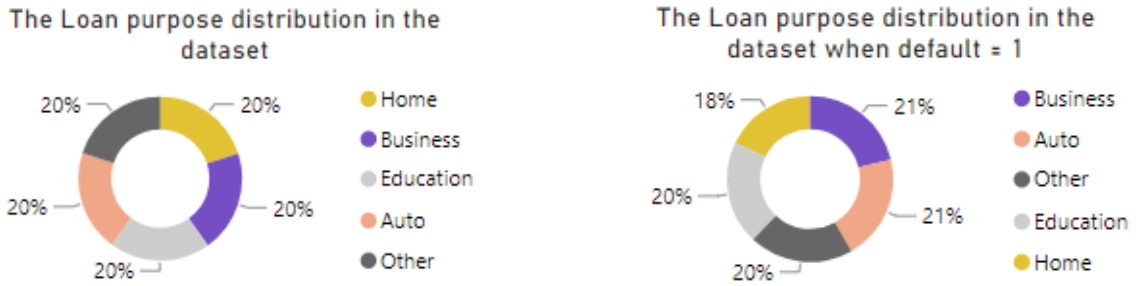
(a) (b)

Figure 13. Distribution of the marital status categories in the database (a) and Distribution of the marital status categories in the database with Default =1 (b).



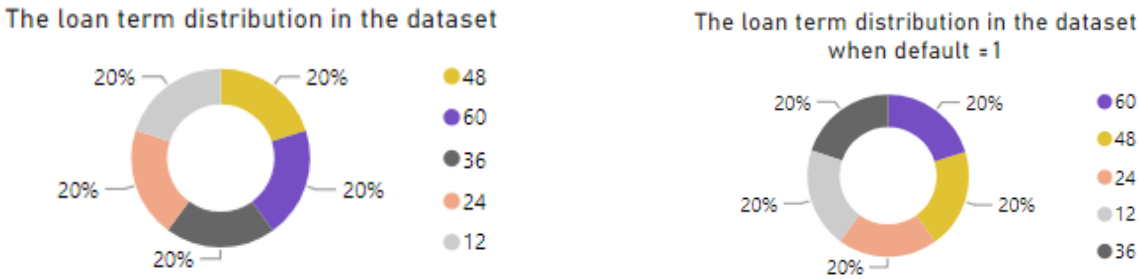
(a) (b)

Figure 14. Distribution of the credit lines categories in the database (a) and Distribution of the credit lines categories in the default customer data group (b).



(a) (b)

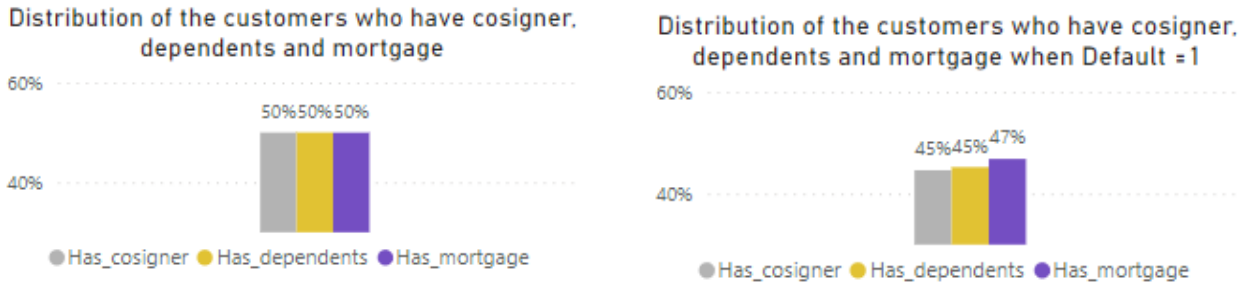
Figure 15. Distribution of the loan purpose categories in the database (a) and Distribution of the loan purpose categories in the database with Default =1 (b).



(a) (b)

Figure 16. Distribution of the loan term categories in the database (a) and Distribution of the loan term categories in the database with Default =1 (b).

The figure 17 - (a) presents the distribution graph of the binary variables "Has_cosigner," "Has_dependents," and "Has_mortgage," where we observe that these variables are also equally distributed. In other words, 50 % of the borrowers have a cosigner, 50 % have dependents, and 50 % have a mortgage. Figure 17 - (b) presents the distribution graph of the same binary variables, but in the dataset filtered for Default = 1. In this case, we noticed a slight reduction in the concentration of borrowers who have a cosigner (45 %), have dependents (45 %), and have a mortgage (47 %). This may indicate a higher tendency towards default for borrowers who do not have a cosigner, dependents, or a mortgage. However, we emphasize that this is a very small trend.



(a) (b)

Figure 17. Distribution of data for the binary variables "Has_cosigner", "Has_dependents" and "Has_mortgage" (a) Distribution of data for the binary variables "Has_cosigner", "Has_dependents" and "Has_mortgage" when Default =1 (b).

4. Evaluating the predictive power of each variable

4.1 Information Value (IV)

During this phase of the analysis, the Information Value (IV) was obtained, a metric that allows evaluating the impact (predictive power) of each variable on the target. In the case of the present analysis, the IV reveals how much a variable can predict a borrower's default (target variable).

The general formula for calculating the IV of a variable (feature) is as follows:

$$IV = \sum_{i=1}^{n} (p_i - q_i) \times \ln\left(\frac{p_i}{q_i}\right)$$

Where, *pi* is the proportion of the positive category (Default = 1) for the i-th category of the variable, *qi* is the proportion of the negative category (Default = 0) for the i-th category of the variable. The sum is taken over all categories of the same variable.

It is important to emphasized that, to calculate the IV of a variable, it must have categories. That is, in the case of a quantitative variable, it is necessary to first transform it into a categorical variable. This can be easily done by separating the values into categories of different ranges and similar probability of default.

Once the IV has been calculated for each variable, we use a classification rule to define the predictive power of each variable:

IV < 0.02: Very weak

0.02 ≤ IV < 0.1: Weak

0.1 ≤ IV < 0.3: Medium

0.3 ≤ IV < 0.5: Strong

The dataset was divided into training (70%) and testing (30%) sets. The Information Value (IV) values were calculated using **only the training set** to avoid leakage of information from the testing set to the prediction model.

Table 3 presents the IV values and the default prediction powers obtained for each variable. The variables that presented the greatest predictive powers were “Age”, “Interest Rate” and “Income”, which were classified with medium predictive power. The variables “Months Employed”, “Loan Amount” and “Employment Type” presented “weak” predictive powers, while the remaining variables presented “very weak” predictive powers.

Table 3. The Information Value (IV) and the impact power of each variable.

	Feature	Sum_IV	Impact
0	Age	0.279	medium
1	InterestRate	0.164	medium
2	Income	0.105	medium
3	MonthsEmployed	0.086	weak
4	LoanAmount	0.069	weak
5	EmploymentType	0.022	weak
6	HasCoSigner	0.016	very weak
7	HasDependents	0.013	very weak
8	CreditScore	0.011	very weak
9	NumCreditLines	0.008	very weak
10	MaritalStatus	0.008	very weak
11	Education	0.007	very weak
12	LoanPurpose	0.006	very weak
13	HasMortgage	0.005	very weak
14	DTIRatio	0.004	very weak
15	LoanTerm	0.000	very weak

4.2 Z test (hypothesis test)

The variables “Months Employed” and “Loan Amount” obtained IV values of 0.05 to 0.10, which classifies them as variables with weak predictive power. However, the previous exploratory analysis suggested that these variables have some relevance in predicting the default profile (figures 6 and 8). In order to include them in the training of the prediction models, a Z test (hypothesis test) was carried out to investigate whether the proportions of defaulting borrowers in the different categories (value ranges) were significantly different.

As observed in Figure 18, the categories of both variables presented p-values lower than 0.05 and very close to zero. This indicates that the default rates among the different categories (in the same variable) are statistically different. Therefore, we can infer that these features have a relevant predictive power, even though their initial IV values suggest weak power.

	Test_Z_stat	p-value
0	9.760650	1.660877e-22
1	10.507534	7.975177e-26
2	7.402782	1.333602e-13
3	7.556849	4.129501e-14

(a)

	Test_Z_stat	p-value
0	-7.412657	1.237936e-13
1	-7.018845	2.237100e-12
2	-8.431282	3.418982e-17
3	-9.076297	1.123312e-19

(b)

Figure 18. The Z test statistics and the p-values of the categories of the feature “Months Employed” (a) and the feature “Loan Amount” (b).

Therefore, to proceed with the analysis, only the variables that showed some relevant predictive power were considered. In other words, only the following five variables (features): "Age", "Interest Rate", "Income", "Months Employed", and "Loan Amount".

5. Evaluating the borrower profiles according to the probability of default

Figure 19 displays bar charts showing the default percentage for each category of the variables "Age" (a), "Interest Rate" (b), "Income" (c), "Months Employed" (d), and "Loan Amount" (e).

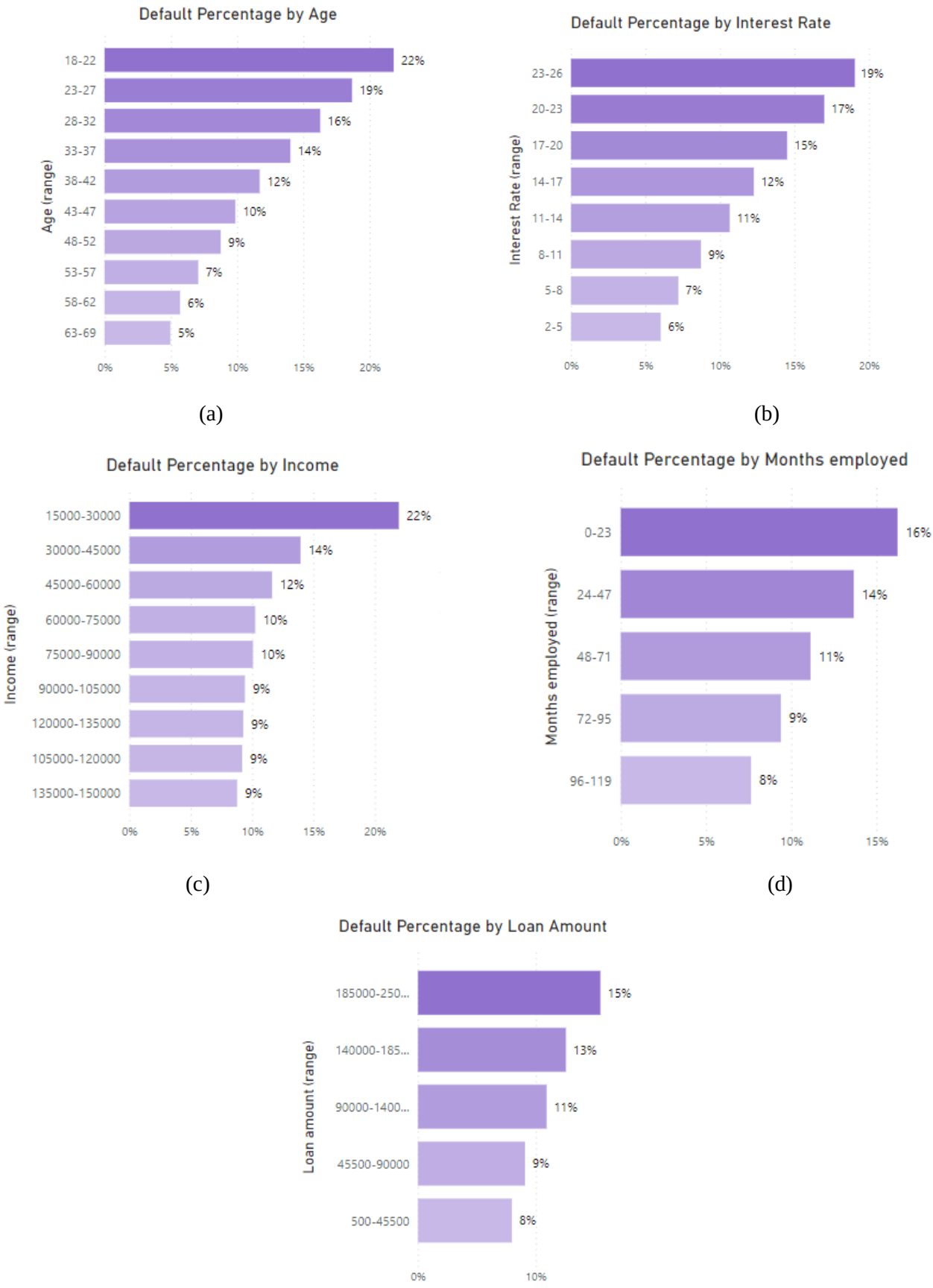


Figure 19. Default percentage by categories of the variable "Age" (a), "Interest Rate" (b), "Income" (c), "Months Employed" (d) and "Loan Amount" (e).

Observing the charts in Figure 19, we note that all the five features have categories with default percentages higher than the total default percentage of the dataset (11.6 %). We can identify the categories most inclined to default by selecting those with default probabilities equal to or greater than 15 %: Age from 18 to 32 years old, income from $15,000 to $30,000, Interest rate from 17 to 26 %, Months Employed from 0 to 23, and Loan Amount from $185,000 to $250,000.

Therefore, it is possible to infer that the default probability is higher for:

- Young customers;
- Customers with low income;
- Customers paying high interest rates;
- Customers who have been employed for a short time;
- Customers with a high loan amount.

Different combinations of features were tested to assess the default probability of different borrower profiles. Although there are five important variables to assess, the profiles tested were composed of only three variables. This is because the combination of a greater number of variables significantly reduces the number of observations per profile, which makes it impossible to use the frequentist theory to estimate the probability of default.

In chart 20, the 10 profiles of customers most likely to default are displayed. The profile most likely to default was borrowers aged between 18 and 22 years old, with an income of US$15,000 to US$30,000 and a loan value greater than US$185,000. This profile has 731 observations in the database and a 58 % probability of default.

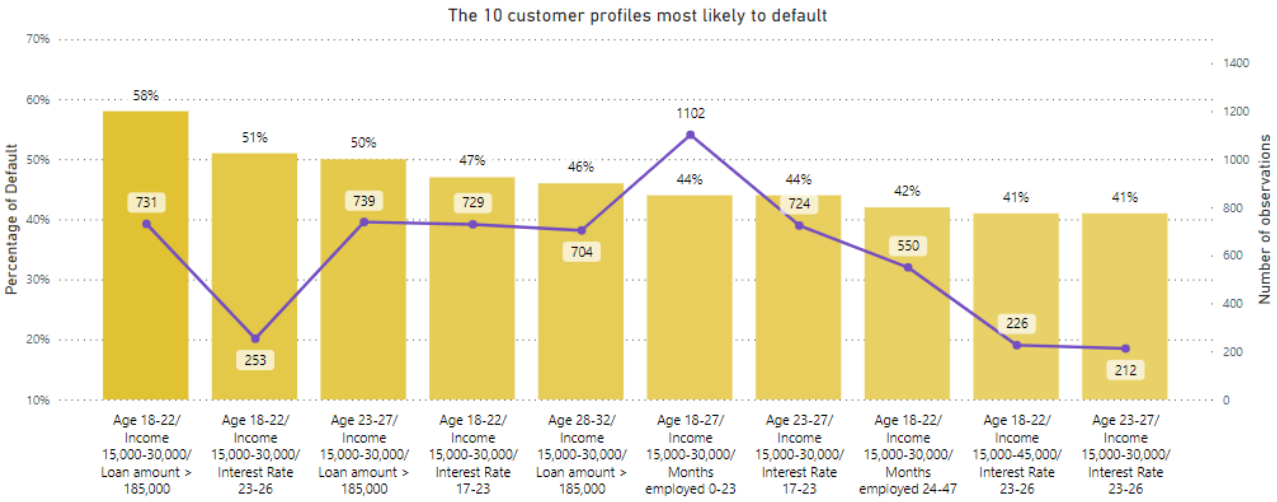


Figure 20. Top 10 borrower profiles most likely to default and the number of observations in the database for each profile.

6. Prediction Models

The training set was used to train the following prediction models: Bagging of Logistic Regression, Artificial Neural Networks, Random Forest, and Gradient Boosting Classifier. As already established, only the five most relevant features were considered: "Age", "Interest Rate", "Income", "Months Employed" and "Loan Amount." The Bayesian optimization technique was used to find the best hyperparameters for the models.

The prediction models were evaluated according to the metrics of Recall, AUROC (area under the ROC curve), KS, and accuracy, following this priority order. The Recall metric is crucial for this type of business since it represents the percentage of correct predictions in positive observations (Default = 1). In other words, Recall reveals the percentage of actual default cases that the model predicted. In this context, the neural network prediction model performed the worst, with only 59 % of Recall and 2 % of KS on the test set. Table 4 shows the performances in Recall, AUROC, KS, and accuracy resulting from the models training with the training set. The Random Forest model presented 67 % of Recall, while the Gradient Boosting model presented 66 % and the Random Forest model obtained 65 %. Despite achieving satisfactory Recall values, the models did not present good generalization. Which means that when applied to the test dataset, these models lost performance significantly.

Table 4. Performances on the test set and performances variation (between training and test sets) of the Random Forest, Bagging of Logistic Regression, Gradient Boosting Classifier and Artificial Neural Networks prediction models.

	Recall_Test	Recall_Var	AUROC_Test	AUROC_Var	KS_Test	KS_Var	Accur_Test	Accur_Var
Random_For	0.67	-0.11	0.73	-0.06	0.33	-0.21	0.66	-0.07
Gradient_Boost	0.66	-0.11	0.72	-0.08	0.33	-0.21	0.67	-0.06
Logist_Regres	0.65	-0.03	0.71	-0.01	0.31	-0.09	0.66	-0.01
Art_Neural_Net	0.57	0.00	0.74	-0.01	0.02	-0.95	0.75	0.12

With the intention of improving the performance and generalization of the models, a feature engineering process was carried out where the five most relevant features were summarized into a single feature called "score".

Firstly, the quantitative features were transformed into categorical features. To divide the features values into different groups (categories), the following criteria were used: The number of observations should be similar between categories, and the number of categories should be large enough to ensure a representative Default percentage for each category.

After the categorization of the features, the new feature "score" was obtained by summing the Default percentages corresponding to each category. In the case of categories where the variables showed the highest prediction impact ("Age," "Interest Rate," and "Income"), the proportion value was multiplied by 1.5.

Considering that the default among borrowers with high loan amounts represents a greater risk for the business, it was decided to calculate four different types of scores. Each one received a different weight for the

percentage of default related to the Loan Amount. In the calculation of Score_1, this percentage was multiplied by 1 (weight 1); in the calculation of Score_2, it was multiplied by 1.5 (weight 1.5); in Score_3 by 2.0 (weight 2); and in Score_4 by 2.5 (weight 2.5).

To illustrate, below is the calculation of Score_1 for observation ID I38PQUQS96 (table 5):

Score_1 (ID: I38PQUQS96) = (0.07 x 1.5) +(0.12 x 1.5) + (0.10 x 1.5) + 0.10 + 0.09 x (1)

Score_1 (ID: I38PQUQS96) = 0.625

Table 5. Customer profile of ID I38PQUQS96 and Default proportions by category.

Features (ID: I38PQUQS96)	Category of the features	Probability of Default
Age	Group 8	0.07
Interest Rate	Group 6	0.12
Income	Group 8	0.10
Months Employed	Group 7	0.10
Loan Amount	Group 2	0.09

In summary, four different “score” sets were used to train the previously mentioned models, except for the Neural Network model.

The following nomenclature was used for the models:

Grad_n = Gradient Boosting Classifier model;

Bagg_n = Bagging of Logistic Regression model;

Rand_n = Random Forest model;

Where "n" refers to the number of the training set used. For example, training set 1 refers to the feature "score_1", training set 2 refers to the feature "score_2", and so on.

Table 6 presents the performance results of the models trained with the four different training sets (four different "scores"). All models obtained similar performances, where the Recall values ranged from 66 to 68 %. The maximum Recall value was obtained by model 1 using the Gradient Boosting Classifier. The AUROC results ranged from 73 % to 74 %, the KS from 34 % to 35 % (considered a very good discrimination level), and the Accuracy between 66 % and 68 %. All trained models showed a satisfactory generalization, with a maximum variation between training and testing of 1 % in accuracy, 3 % in KS, 1 % in AUROC, and 2 % in Recall.

Table 6. Recall, AUROC, KS and Accuracy values of the prediction models: Gradient Boosting (Grad), Random Forest (Rand) and Bagging of Logistic Regression (Bagg).

	Recall_Test	Recall_Var	AUROC_Test	AUROC_Var	KS_Test	KS_Var	Accur_Test	Accur_Var
Grad_1	0.68	0.00	0.73	0.00	0.35	0.03	0.66	-0.01
Bagg_1	0.67	0.02	0.74	0.01	0.35	0.03	0.68	0.01
Rand_2	0.67	0.02	0.74	0.00	0.35	0.00	0.68	0.01
Grad_2	0.67	0.02	0.73	0.00	0.35	0.00	0.68	0.01
Bagg_4	0.67	0.02	0.73	0.00	0.34	0.03	0.67	0.00
Grad_4	0.67	0.02	0.73	0.00	0.34	0.00	0.67	0.00
Bagg_3	0.67	0.00	0.74	0.01	0.35	0.03	0.67	0.00
Bagg_2	0.67	0.00	0.74	0.01	0.35	0.00	0.67	0.00
Rand_1	0.66	0.02	0.73	0.00	0.35	0.03	0.68	0.01
Grad_3	0.66	0.02	0.73	0.00	0.35	0.03	0.68	0.01
Rand_3	0.66	0.02	0.73	0.00	0.35	0.03	0.68	0.01
Rand_4	0.66	0.02	0.73	0.00	0.34	0.03	0.68	0.01

7. Assessment of possible economic advantages

With the aim of evaluating the potential economic advantages obtained through the best prediction models, the invested values, the gross income values, the gross profit values, and the ROI values (percentage of return on investment) were obtained through the prediction of Gradient Boosting Classifier models (Grad_1, Grad_2 and Grad_3).

The invested values were calculated by summing the values of the "Loan Amount" feature, while the gross income values were obtained according to the following calculation:

Gross Income = (Sum of Loan Amount)* + (Sum of Loan Amount)* x (Interest Rate)/100 x (Loan Term)/12

**Considering only the values of Loan Amount when Default = 0*

Note: The annual interest rate was considered as a simple interest rate.

The Gross Profit was calculated by subtracting the total invested value from the total Gross Income.

The ROI values were obtained through the calculation: ROI = 100 x (Gross Income - invested value)/ invested value. The ROI represents the investment efficiency relative to the cost.

Additionally, these values were also obtained considering different threshold values for each model. The threshold defines the value of the probability of default from which the target must be considered as 1. That is, if a threshold of 0.5 is defined, it means that all borrowers who have an estimated default probability

greater than 0.5 should be considered as defaulters. In this situation, it is assumed that the loan will not be authorized for these customers. The thresholds of 0.45, 0.50, 0.55, 0.60, 0.65, and 0.70 were used.

Table 7 presents the "top 10" models in descending order of % ROI and their respective threshold values, gross income amount, invested amount, unpaid amount, default percentage, and gross profit amount. It can be observed that the models 2 and 3 with thresholds of 0.50 and 0.55 promoted the highest ROI values (ranging from 25.5 to 25.6 %) with a final Default of 4 or 5 %. That represents a reduction of seven percentage points in default compared to the original dataset (without the use of a prediction model).

Table 7. The top 10 models with the highest % ROI values and their respective threshold values, gross income amount, invested amount, unpaid amount, Default percentage and gross profit values.

	Threshold	Gross_income	Invested_amount	Unpaid_amount	% Final_Default	Gross Profit	% ROI
Grad_2_0.5	0.5	2.382893e+10	18972569429	1193035581	4.0	4.856356e+09	25.60
Grad_3_0.55	0.55	2.662440e+10	21200417904	1487174211	5.0	5.423977e+09	25.58
Grad_2_0.55	0.55	2.700862e+10	21508606846	1516926880	5.0	5.500009e+09	25.57
Grad_3_0.5	0.5	2.316945e+10	18462423814	1151386522	4.0	4.707023e+09	25.50
Grad_1_0.55	0.55	2.776961e+10	22130494468	1606999417	5.0	5.639114e+09	25.48
Grad_1_0.5	0.5	2.360141e+10	18809779623	1183674878	4.0	4.791634e+09	25.47
Grad_3_0.6	0.6	2.921066e+10	23288947712	1788865926	6.0	5.921711e+09	25.43
Grad_2_0.6	0.6	2.999824e+10	23935032905	1888888760	6.0	6.063212e+09	25.33
Grad_1_0.6	0.6	2.972960e+10	23722496257	1850345701	6.0	6.007102e+09	25.32
Grad_2_0.45	0.45	2.085894e+10	16649153359	953250617	3.0	4.209785e+09	25.29

Table 8 presents the "top 10" models in descending order of gross profit and their respective threshold values, gross income amount, invested amount, unpaid amount, default percentage, and % ROI. It was observed that none of the models promoted higher gross profit than the original one (without prediction model). Additionally, the order of the models in the table indicates that the higher the threshold, the higher the gross profit and the smaller the ROI value.

Table 8. The top 10 models with the highest gross profit values and their respective threshold values, gross income amount, invested amount, unpaid amount, Default percentage and ROI.

	Threshold	Gross_income	Invested_amount	Unpaid_amount	% Final_Default	Gross Profit	% ROI
original	-	3.945692e+10	32576880572	4285312531	12.0	6.880039e+09	21.12
Grad_2_0.7	0.7	3.530920e+10	28432163909	2814818284	8.0	6.877036e+09	24.19
Grad_1_0.7	0.7	3.531395e+10	28456591310	2825660994	8.0	6.857363e+09	24.10
Grad_3_0.7	0.7	3.429302e+10	27530261116	2589402520	8.0	6.762764e+09	24.56
Grad_2_0.65	0.65	3.278305e+10	26238451718	2298757843	7.0	6.544594e+09	24.94
Grad_3_0.65	0.65	3.219078e+10	25742479410	2210788742	7.0	6.448301e+09	25.05
Grad_1_0.65	0.65	3.225340e+10	25827615717	2239234992	7.0	6.425788e+09	24.88
Grad_2_0.6	0.6	2.999824e+10	23935032905	1888888760	6.0	6.063212e+09	25.33
Grad_1_0.6	0.6	2.972960e+10	23722496257	1850345701	6.0	6.007102e+09	25.32
Grad_3_0.6	0.6	2.921066e+10	23288947712	1788865926	6.0	5.921711e+09	25.43

Observing tables 7 and 8, we noticed that the models 1, 2, and 3 with a threshold of 0.60 are present in both tables. This means that these models can present a desirable "balance" between gross profit and ROI values. In figure 20, it is possible to compare the ROI and gross profit values obtained for the original dataset (without a prediction model) with the values obtained for the dataset considering the use of the model 2 and the thresholds of 0.70, 0.65, 0.60, 0.55 and 0.5.

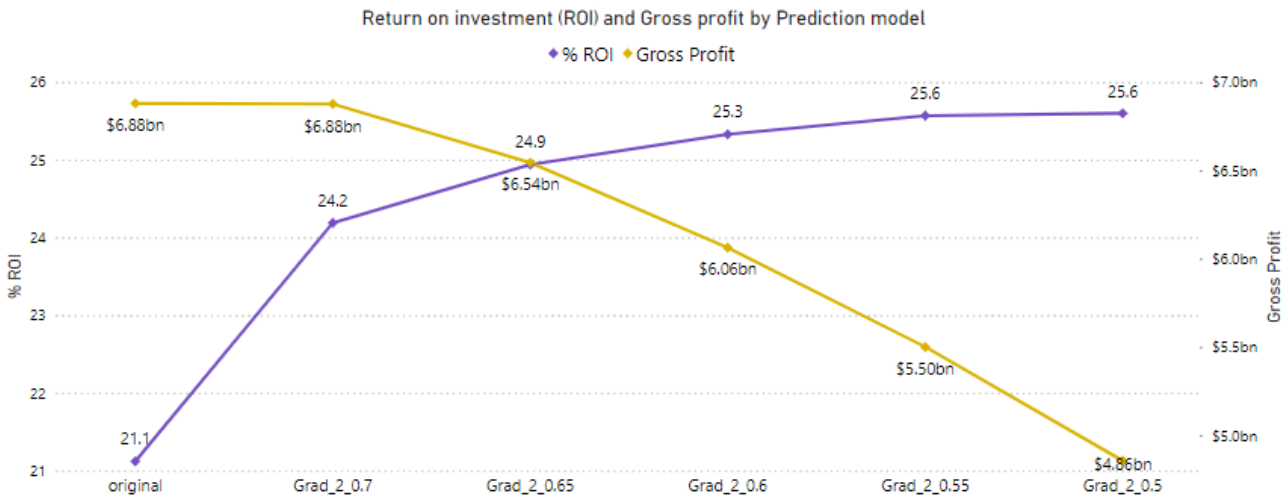


Figure 20. Return on investment (ROI) and Gross Profit by Prediction model and threshold.

The ROI and gross profit obtained from the original dataset (without a prediction model) are 21.12 % and $6.88 billion, respectively. The choice of the model and the threshold will depend on the business objectives. If an ROI of 24 % is considered desirable, the model 2 with a threshold of 0.70 would be a good choice, as it presents an ROI of 24.20 % and a gross profit value of $ 6.877 billion, slightly lower than the original gross profit. However, if the ROI is considered more important than gross profit, there are several model options

with an ROI higher than 25 %. In this case, the choice of model will depend on the amount of gross profit that can be "sacrificed" in order to obtain a higher ROI. The model 2 with a threshold of 0.65, for example, presents a good balance, with an ROI of 24.94 % and a gross profit of $6.54 billion.