A Work Project, presented as part of the requirements for the Award of a Master's degree in International Finance from the Nova School of Business and Economics.

ADVANCING FINANCIAL INCLUSION: A CREDIT RISK MODEL FOR CUSTOMERS WITH ADVERSE CREDIT HISTORY

Ricardo Andres Leon Gomez

Work project carried out under the supervision of:

Melissa Prado

17/05/2023

Abstract

This thesis addresses the challenge of effectively assessing credit risk for customers with unfavorable financial histories. The goal is to develop a model that can segment these customers based on various factors such as payment history, income, debt-to-income ratio, and behavioral patterns, in order to determine the likelihood of defaulting on new credit card payments. Machine learning algorithms, including decision trees, random forests, and neural networks, are employed to construct the model.

The results indicate that while the overall performance of the models in identifying good customers is initially low, logistic regression outperforms both random forest and XGBoost models. This unexpected finding can be attributed to the presence of a linear relationship between the independent variables and the logit of the dependent variable. Additionally, potential issues such as overfitting, class imbalance, and intricate feature interactions may have affected the performance of random forest and XGBoost models. Logistic regression, on the other hand, demonstrates simplicity and robustness, which makes it a viable option especially when dealing with limited datasets or fewer features.

Despite the models' suboptimal predictive accuracy, they offer valuable insights into the customer base and their behavioral patterns. These insights can inform the refinement of segmentation strategies, thereby enhancing risk management and decision-making processes. It is important to view the models' results within the broader context of business objectives and risk tolerance. This perspective ensures that the machine learning models serve not only as predictive tools but also as strategic assets in effectively managing risk while maximizing returns.

Key Words: Machine learning model, segment customers, financial information, behavioral patterns, demographics, default, loans, Re-banking, trust, financial product, credit cards, payment habit, credit score.

- Machine learning model: A mathematical algorithm that learns patterns in data and makes predictions or decisions based on that learning. In the context of this thesis, I will use different machine learning models to analyze customer data and determine which customers should receive credit cards to help them rebuild their credit scores.
- Customer segmentation: Separating customers into different groups based on similar characteristics. Customers will be segmented based on their behavioral patterns, demographics, and available financial data.
- Financial information: Financial information refers to data related to a customer's financial situation, such as income, debt, and credit score.
- Behavioral patterns: Behavioral patterns refer to how a customer uses credit, including how frequently they make payments, the amount of debt they carry, and their payment history. The payment behavior will be used to build the target variable, differentiating clients with good payment habits from those with bad or non-payment habits.
- Demographics: Demographics refer to a customer's age, gender, education, income level, and other demographic factors that can impact their creditworthiness. These factors will also be considered in the segmentation process.
- Default: Default occurs when a borrower fails to repay a loan as agreed. In the context of this thesis, customers have defaulted on their loans and are seeking to rebuild their credit scores.
- Re-banking: Re-banking refers to helping customers whom traditional financial institutions have rejected regain trust and access to financial products. In this thesis, QNT is a Re-banking company that has purchased portfolios of defaulted loans and

seeks to issue credit cards to help these customers rebuild their credit scores.

- Payment habit: Refers to a customer's history of making payments on time and in full. Developing good payment habits is critical for customers seeking to rebuild their credit scores.
- Credit score: A credit score is a numerical value representing a customer's creditworthiness based on their credit history, payment habits, and other financial factors. Improving a customer's credit score is the goal of the credit card issuance program in this thesis.

Introduction

The global financial system is a complex web of transactions and interactions between players such as individuals, businesses, banks, and financial institutions. One critical factor in maintaining these systems' integrity is the ability to assess and manage credit risk. Credit risk refers to the risk that a borrower will default on their loan or credit obligations, leading to financial losses for the lender.

Historically, credit risk assessment has been based on financial information such as credit scores, income, and employment history. However, there are situations where little financial information is known, and borrowers have an unfavorable financial footprint, such as those who have defaulted on loans. In such cases, it can be challenging for financial institutions to assess credit risk accurately and make informed lending decisions.

Recently, interest in developing models to predict credit risk and improve credit scoring has increased. Credit risk assessment is a critical aspect of the financial industry, and it helps to identify the likelihood that a borrower will default on their loan. Many financial institutions use credit scoring models to assess the risk of lending money to individuals. These models

typically use demographic and financial information to generate a credit score. However, traditional credit scoring models may be ineffective for customers with little financial information or an adverse credit history.

This is where QNT, a Re-banking company, comes in. QNT has bought portfolios of clients who have defaulted on their loans and are now rejected by most financial institutions. The company aims to help these clients regain trust in the financial system by issuing credit cards and helping them improve their credit scores. QNT needs to develop a model to segment clients based on their behavioral patterns, demographics, and available financial information to achieve this goal.

This work focuses on building a model to segment customers with unfavorable financial histories. The model will analyze various features such as payment history, income, debt-to-income ratio, and other behavioral patterns to determine the likelihood of a customer defaulting on the payments if a new credit card is to be issued. The goal is to issue credit cards to these customers so that they can rebuild trust in the financial system and improve their credit scores.

I will use machine learning algorithms to build this model, including decision trees, random forests, and neural networks. These algorithms have shown effectiveness in credit scoring and default prediction tasks (Nonato Jr et al., 2017; Acquisto & Fabbri, 2020). I will also follow best practices for credit scoring model development outlined in "Credit Scoring in R" by Chris De Brabanter (De Brabanter, 2017) and "The Handbook of Credit Risk Management" by Sylvain Bouteille (Bouteille, 2013).

Literature Review

Segmenting clients with default loans is critical for companies operating in the fintech industry. By effectively distinguishing between clients who have a higher chance of recovering from default and those who pose a greater risk of further defaulting, companies can optimize resource allocation and implement targeted strategies for debt management. In this literature review, we explore existing studies that have approached similar scenarios, highlighting the methodologies used, the significance of client segmentation, and the potential benefits of further research.

Data-driven approaches have been widely employed in the fintech industry to develop predictive models for client segmentation. Li et al. (2019) used a random forest algorithm to segment clients based on credit history, income, employment status, and demographic information. Their study demonstrated the model's effectiveness in identifying clients with a higher likelihood of loan recovery (Li et al., 2019). Similarly, Zhang et al. (2020) focused on credit-related features such as credit scores, loan history, payment behavior, and debt-to-income ratios to improve the accuracy of their segmentation model. These studies highlight the importance of leveraging available data to identify relevant features for effective client segmentation (Zhang et al., 2020).

Various machine learning algorithms and statistical models have been employed to tackle client segmentation in the fintech industry. Chen et al. (2018) applied a support vector machine algorithm to segment clients based on transactional data, achieving high accuracy in identifying clients at risk of further defaulting (Chen et al., 2018). These studies demonstrate the diverse range of models available for this task, and the choice of model depends on the specific requirements of the problem and the available data.

Validation and evaluation techniques are crucial in ensuring the reliability and generalizability of segmentation models. Wang et al. (2017) emphasized the importance of rigorous validation to ensure the robustness of their client segmentation model (Wang et al., 2017). Cross-validation, hold-out validation, and other methodologies are commonly used to assess model performance. Evaluation metrics such as accuracy, precision, recall, and AUC-ROC are used to compare different models. These validation and evaluation techniques provide insights into the performance and effectiveness of client segmentation models.

Client segmentation in the fintech industry holds significant practical implications. Firstly, it enables companies to optimize resource allocation by directing limited resources toward clients with a higher chance of recovery. This allows for more efficient debt management and improved operational efficiency. Secondly, effective client segmentation aids in risk management by identifying clients with a higher risk of defaulting again. This enables companies to implement tailored strategies, such as customized repayment plans or targeted collection efforts, to mitigate this risk. Thirdly, client segmentation helps optimize collection efforts by focusing on clients more likely to make payments or engage in rehabilitation programs. This increases the chances of successful debt recovery and reduces the overall default rate.

The proposed thesis aims to contribute to the existing research on client segmentation for companies that purchase default loans in the fintech industry. This research seeks to provide new insights into effective client segmentation strategies by implementing novel methodologies and leveraging unique datasets. The practical implications of this study, such as improved resource allocation and enhanced loan recovery rates, will directly benefit companies in managing default loans more efficiently. Additionally, the thesis will explore

the interpretability and explainability of the segmentation models, addressing the need for transparency and actionable insights for decision-making and strategic planning.

Methodology

The methodology of this work follows the guidelines stated by Yung-Chia Chang, Kuei-Hu Chang, and Guan-Jhih Wuis in their paper Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions (Chang, Chang, & Wu, 2018). The first steps involves the gathering and consolidation of data. The data is processed in the second step, and the most critical variables are selected. The third step involves applying machine-learning models to construct classification forecasting models of consumer credit risk. Figure 1 is a flowchart of the mythology previously just mentioned.

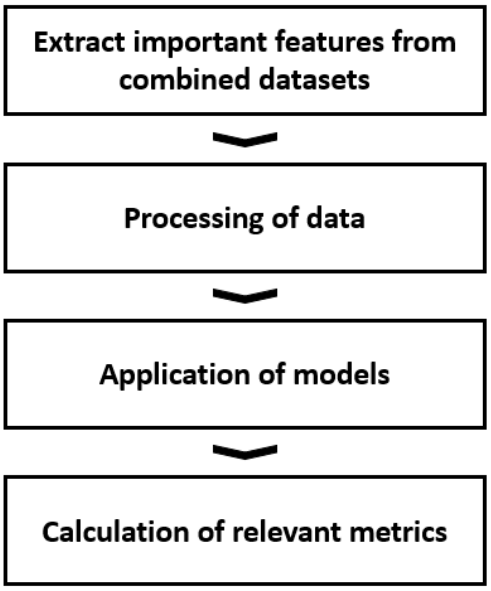


Figure 1 Methodology (Chang, Chang, & Wu, 2018)

Extract important features from combined datasets:

During this phase, the data source and the variables utilized in the model are identified. This step involves extracting the pertinent data from the data source(s) for the analysis. The data was gathered from historical data corresponding to the period of 2018 to 2023 of the loan books of QNT, a Re-banking financial company that has purchased portfolios of customers

who have defaulted on their loans. The data includes various aspects such as customer demographic information, payment history, income, credit bureau, and other related factors for individual clients.

The database was constructed using information from various sources, including QNT's data lake, bureau, and Mareigua. The original dataset contained 446,845 client records and 54 variables.

Table 1 presents a subset of 52 variables, comprising 41 financial variables and 11 non-financial variables. For a detailed understanding of each column's significance in the study's context, please refer to Appendix 1, which provides a comprehensive overview of each variable.

Financial variables	**Non-financial variables**
Debt ratio over incomes	Age
Debt ratio	City
Total Active Loans	Gender
Acierta Plus Score	Job type (employed, unemployed, pensioner, independent)
Number of active banks	
Debt Ratio	Income
Number of debts in real sector	Last Contact Date
Loans outstanding	Start date Customer
Restructured debt	Number Sources of Income

Table 1 Variables

Preprocessing of data:

This study will use a structured approach to develop a model that can segment clients with unfavorable financial history into two categories for credit card issuance. The methodology comprises four critical stages: Data preprocessing, Feature selection, Model training, and

Evaluation.

To ensure the adoption of best practices, the guidelines provided in "Credit Scoring in R" by Chris De Brabanter and "The Handbook of Credit Risk Management" by Sylvain Bouteille will be followed. These references offer comprehensive guidance for developing credit scoring models, covering all critical aspects of the process.

The initial stage of the model development process involves data preprocessing, which is crucial for ensuring the accuracy and dependability of the model. This step involves cleaning, transforming, and encoding the data for effective analysis. The following guidelines are essential in data preprocessing:

Managing missing values:

Missing data is a significant issue in the dataset, as it can lead to biased or incomplete analysis. Various techniques, such as mean imputation, regression imputation, or advanced methods like multiple imputations, can be employed to address this issue. Figure 2 illustrates the location of missing values in the dataset, highlighting the need to address this issue through appropriate techniques.

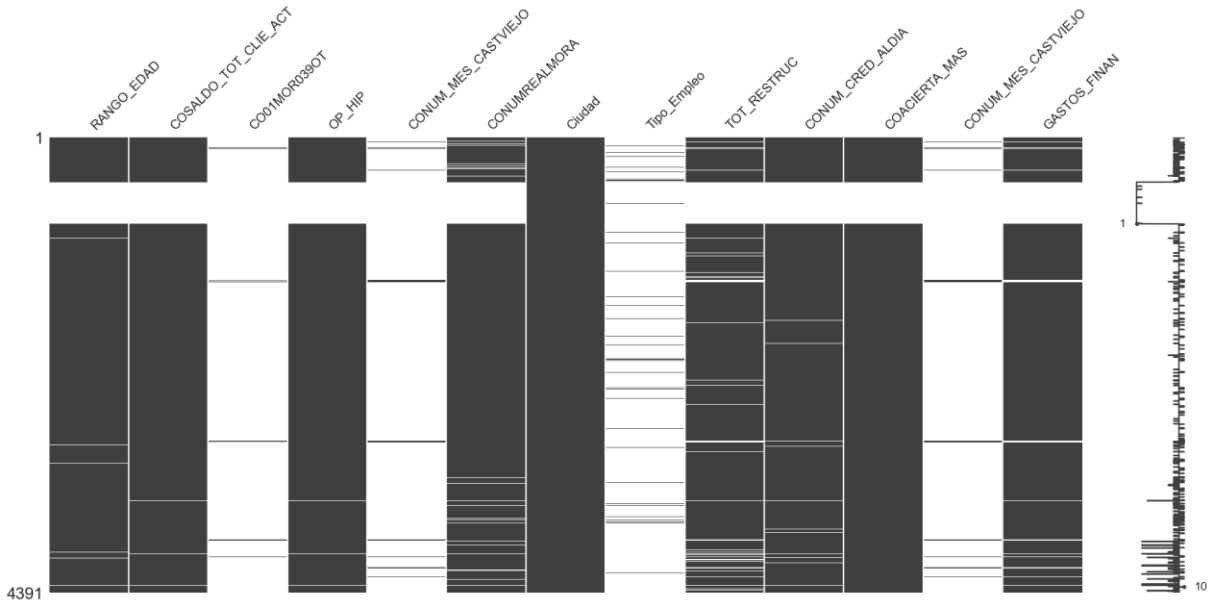


Figure 2 Sample of Missing Values

After thorough preprocessing and data cleaning, the entries with missing entries were removed from the dataset due to the high percentage of null values (higher than 30% of nulls), leaving only the most relevant and informative columns for the analysis. Additionally, the records with more than 50% of the variables in null were removed. As a result, 86,654 observations and 33 variables are collected, including 3,970 defaults and 75,196 non-default cases.

Data Name	No. of observations (Defaults/non-defaults)	Ratio of defaults to non-defaults	No. of Variables
Dataset after preprocessing null values	86,654 (3,970 / 75,196)	1: 18.94	33

Table 2 Summary of dataset

After removing columns with a high percentage of missing values, the next step involves filling in the null values for the remaining columns. Little and Rubin (2019) and Enders (2010) provide comprehensive overviews of missing data analysis techniques.

According to these sources, mean imputation techniques may vary according to the variable type. For numeric variables, missing values can be replaced with the mean value of the variable or filled with -99, a value that does not occur in the data. For categorical variables, missing values can be filled with the mode of the variable or a new category corresponding to -99.

Data transformation:

This stage involves constructing new variables using the existing ones with a proper structure. The first variable created is the Target Variable since it will define what is perceived as a good potential client from a bad one. The Target Variable is defined as all current QNT clients who are paying their loan, who have at least done three payments, and who have not default on any of these payments since they started their relationship with the company. Once the Target Variable is defined, normalization and aggregation operations are performed to transform the data and create a more uniform and symmetric distribution, thus improving the model's performance. Since complex models can fail to generalize well and lead to overfitting when dealing with categorical variables that contain too many categories.

To illustrate this transformation process, Table 3 provides some examples. For more details on this process, please refer to Appendix 2. The categorical variables were transformed into binary variables, also known as dummy variables. This process involves creating a new binary variable for each unique value in the original categorical variable. The new binary variables take a value of 1 if the original categorical variable has that value and 0 otherwise.

Variable	**Transformation**
Age	Group by deciles
City	Keep cities that group 80% of data and add new category with others

Incomes	Group by deciles
Housing Loans	Yes: Has housing loans No: Do not have housing loans
Restructured Loans	Yes: Has Restructured loans No: Do not have restructured loans
Default Loans	Yes: Has loans in default No: Do not have loans in default

Table 3 Transformations

Standardization was performed to account for the differences in scale among numeric features. This involves transforming the variables to have a mean of 0 and a standard deviation of 1, thus ensuring that each variable contributes equally to the analysis. This ensures that all features contribute equally to the analysis and helps avoid numerical instabilities during the model training process.

Feature Selection:

After completing the data preprocessing stage, the next step is identifying the key features for segmenting clients. Feature selection is a crucial step to ensure that the model utilizes only the most relevant features for analysis. Various methods for feature selection are available, including filter methods, wrapper methods, and embedded methods (Guyon & Elisseeff, 2003).

Filter methods involve ranking features based on statistical measures such as correlation and mutual information. On the other hand, wrapper methods select features based on their predictive power using a machine learning model.

Figure 3 shows the remaining columns' correlation matrix showing the features' correlation. To ensure the accuracy and reliability of the statistical model, it is essential to address the issue of correlated variables and multicollinearity. It represents a problem because it “undermines the statistical significance of an independent variable” Plenum Press (1997)

since the standard error of the coefficients will increase.

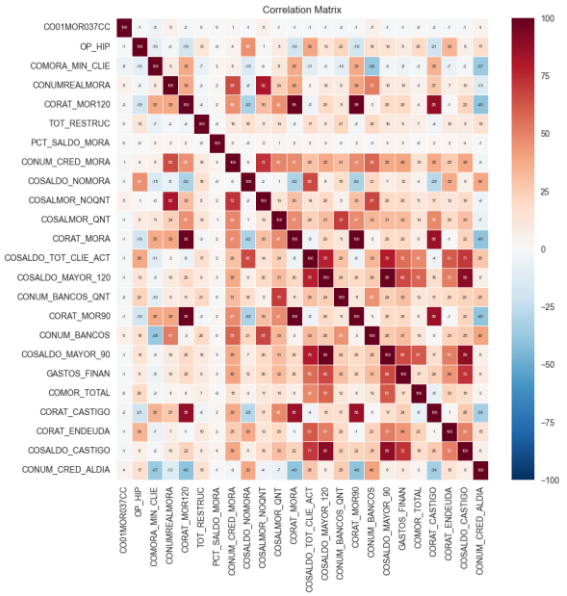


Table 4 Correlation of variables

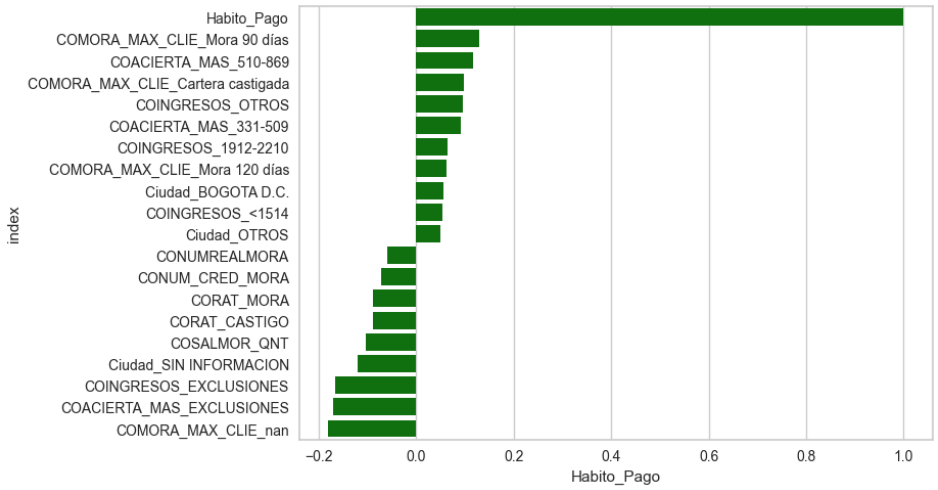
One practical approach to tackle this challenge is to employ the variance inflation factor (VIF), a measure of collinearity among predictor variables within a multiple regression to eliminate redundant features. So, as we can see in the formula, the greater the value of the R^2 is, the greater the VIF. Therefore, any variable that exhibits a high VIF value indicates that it is strongly correlated with other predictors and thus can be dropped from the dataset.

$$VIF = \frac{1}{1 - R^2}$$

To determine which variables to remove, experts in the field often recommend a threshold value of 10, such as Scott Menard (Menard, 2001), since it indicates a high correlation with other independent features. By applying this methodology, we can streamline the model and enhance its predictive power, ensuring that nonredundant variables are utilized.

After addressing multicollinearity issues among predictor variables, it is essential to analyze the relationship between the features and the target variable in the model context. To accomplish this, a univariate analysis is conducted, which examines the relationship between

each individual feature and the target variable. This approach helps identify significant predictors by independently assessing their impact on the outcome variable. By understanding the impact of individual features on the target variable, we can refine our model further and improve its overall predictive accuracy. Therefore, univariate analysis is crucial in developing a robust and dependable statistical model, as seen in Figure 4.



The degree of correlation between each predictor variable and the target variable provides valuable insight into their respective predictability. Liu, Li, and Zeng (2017) have shown that correlation analysis based on Pearson, Spearman, and Kendall coefficients can provide helpful information for financial data. When a predictor variable has a correlation coefficient close to zero, it implies little to no relationship between that variable and the target variable. Karami, Zare, and Mehri (2017) have evaluated several methods to deal with multicollinearity in logistic regression analysis, emphasizing the importance of eliminating variables that do not contribute to the model's predictive accuracy. Therefore, variables that do not exhibit any predictive power are also removed from the model. By eliminating variables that do not provide any predictive power, the model is streamlined to improve its interpretability. This approach ensures that the model is efficient and effective and provides accurate predictions,

critical for making informed decisions in various fields.

Application of models

I have employed a range of machine learning algorithms to train the models to compare their performance, such as logistic regression, ridge regression, random forests, and gradient boost. The effectiveness of these algorithms in credit scoring and default prediction tasks has been established by prior research, including Nonato Jr et al. (2017) and Acquisto & Fabbri (2020), among others. These algorithms were chosen based on their ability to handle large datasets, their robustness to overfitting, and their proven track record in similar applications.

Firstly, to ensure the validity and fairness of our machine learning models, we addressed the issue of class imbalance in our dataset, which is a common challenge in the field (He & Garcia, 2009). To tackle this problem, we employed Synthetic Minority Over-sampling Technique (SMOTE), a widely adopted method for oversampling the minority class in imbalanced datasets (Chawla et al., 2002). SMOTE has been shown to improve the performance of classifiers on imbalanced datasets (Barua et al., 2014) and is particularly effective when combined with cross-validation (Galar et al., 2012). We performed SMOTE on the input data to the train data using the fit_resample() method provided by the learn library (Lemaître et al., 2017).

1) Logistic Regression

Logistic regression is used to model binary classification and seeks to explain the influence of the explanatory variables on the occurrence of the dichotomous phenomenon. Logistic regression estimates the probability of a binary outcome based on input features and is often used for credit scoring and default prediction tasks. This probability is calculated through:

$$p(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1}}$$

In practical applications, a threshold probability value is typically chosen to dichotomize a given continuous variable. Specifically, observations with probabilities exceeding the threshold are conventionally assigned a value of 1, while those below the threshold are assigned a value of 0.

$$p(x_i) \geq Threshold \rightarrow y_i = 1$$

$$p(x_i) < Threshold \rightarrow y_i = 0$$

The logistic regression model was implemented using the LogisticRegression() class provided by the scikit-learn library. By default, this class employs a regularization strength of 1.0 to balance the bias-variance trade-off and prevent overfitting. The model utilizes the sag solver, a highly efficient algorithm for solving large-scale nonlinear optimization problems.

2) Ridge Regression:

Ridge regression is a linear regression variant used to prevent overfitting. Overfitting take place when the models complexity makes it fit the training data too closely, leading to poor generalization to new data. Ridge regression adds a regularization term to the loss function to penalize large coefficients and reduce model complexity. This technique is often used in credit scoring to help prevent overfitting and improve the model's accuracy. This was implemented using the RidgeClassifier() class provided by sci-kit-learn, which uses a default regularization strength of 1.0 and the squared hinge loss function.

3) Random Forests

The random forest algorithm is an ensemble method that is comprised of multiple decision trees. Each tree in the ensemble is constructed using a random subset of the training data, drawn with replacement, and a random subset of the input features. This process, known as

bootstrapping, helps create a diverse set of trees that can capture a wide range of relationships between the input and target variables. The final prediction of the random forest is obtained by combining the outputs of all the trees in the ensemble, typically through a majority voting mechanism. Figure 7 illustrates this process.

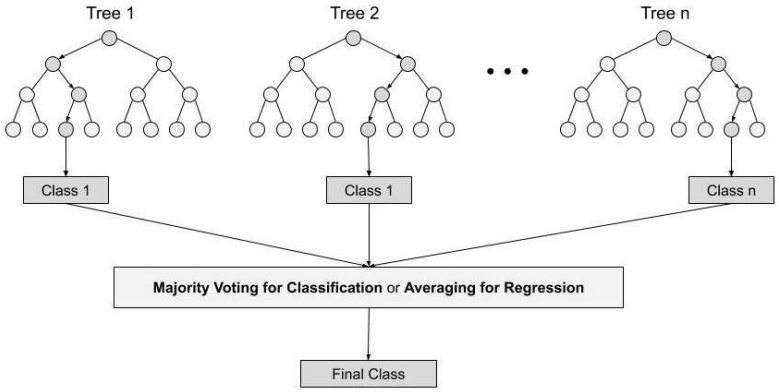


Table 5 Random Forest Algorithm. The Science of Machine Learning (2020).

By using random subsets of the training data, the random forest algorithm reduces the risk of overfitting and improves the model's ability to generalize to new, unseen data. Random forests are effective in credit scoring because they can handle nonlinear relationships between features and provide highly accurate predictions.

Random Forest was implemented using the RandomForestClassifier () class provided by the scikit-learn library, with 100 decision trees, Gini as the criterion for splitting nodes, and a maximum depth of 10 for each decision tree. The max_features parameter was set to sqrt, meaning each decision tree randomly selects a subset of features to split on.

4) Gradient Boosting:

The gradient boosting algorithm is an ensemble learning technique that combines multiple weak models to form a more robust model. In gradient boosting, each weak model is trained on the residual errors of the previous model, which helps improve the final model's accuracy.

This iterative process involves adding decision trees to the model, with each new tree being constructed to reduce the errors of the previous trees, as illustrated in Figure 9.

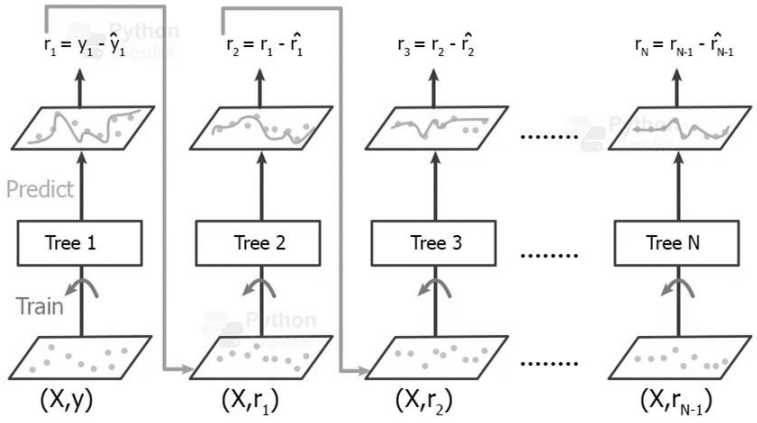
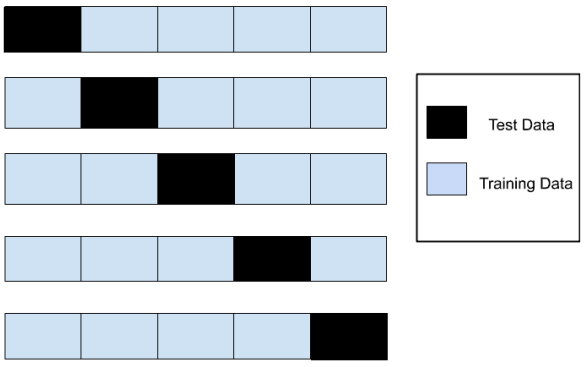


Table 6 Gradient Boosting Algorithm (Datacamp, 2023)

Gradient boosting is a widespread technique in credit risk due to its ability to handle large datasets and provide highly accurate predictions. As seen before, the algorithm is also robust to outliers and missing data, which are common in credit risk datasets.

Gradient Boosting was implemented using the GradientBoostingClassifier() class provided by scikit-learn, with 100 estimators (i.e., weak learners), a learning rate of 0.1, and a maximum depth of 3 for each decision tree.

To ensure that the models can provide reliable predictions and avoid overfitting, they are evaluated using cross-validation techniques. I used 5-fold cross-validation, a widely used machine learning technique for estimating models' performance on unseen data (Hastie et al., 2009). The goal of cross-validation is to divide the data into training and validation sets and assess the models using the validation set (Figure 3 Cross Validation), which helps to measure their accuracy and identify issues of underfitting or overfitting



(Nonato Jr et al., 2017; Acquisto & Fabbri, 2020). This approach is particularly useful for small datasets, providing a more reliable estimate of model performance than traditional train-test splits (Kohavi, 1995).

Figure 3 Cross Validation

Results:

Table 7Confusion Matrix

Confusion Matrix								
		1	0		precision	recall	f1-score	accuracy
Logistic Regression	1	724	467		0.09	0.61	0.15	0.66
	0	7661	14898		0.97	0.66	0.79	
Ridge Regression	1	724	467		0.09	0.61	0.15	0.66
	0	7661	14898		0.97	0.66	0.79	
Random Forest	1	110	1081		0.04	0.09	0.05	0.84
	0	2712	19847		0.95	0.88	0.91	
XGBoost	1	400	791		0.1	0.34	0.16	0.82
	0	3506	19053		0.96	0.84	0.9	

Cross-Validation Scores		
	Mean-Accuracy	Standard Deviation
Logistic Regression	0.6454	0.0043
Ridge Regression	0.6454	0.0043
Random Forest	0.8932	0.0028
XGBoost	0.8533	0.0038

At first glance the overall performance of the models is very low, when predicting which are the good customers. However, the regression models in this case are outperforming the Random Forest and the XGBoost model. The unexpected superior performance of the Logistic Regression model over the Random Forest and XGBoost models in this analysis is intriguing. Yet, it can be explained by delving deeper into the data and the characteristics of these models. Firstly, Logistic Regression excels when a linear relationship exists between the independent variables and the logit of the dependent variable. If such a linear structure is inherent in our data. It is not out of the roam of possibilities that the Logistic and Ridge Regression models

outperform Random Forest and XGBoost, which are optimized to handle intricate, non-linear relationships.

Secondly, we must consider the potential for overfitting. While the flexibility of Random Forest and XGBoost models allows them to capture complex patterns in data, it might lead to overfitting wherein the models learn the training data too well, failing to perform optimally on unseen data. The presence of noise in the data could exacerbate this issue, leading to a surge in false positive predictions.

The Random Forest and XGBoost performance are also highly dependent on the tuning of hyperparameters. Suboptimal settings for parameters such as the number of trees, maximum depth of trees, and learning rate may impair the performance of these models.

Class imbalance in the data is another critical factor to consider. If the data is skewed towards one class over the other, the Random Forest and XGBoost models might struggle to correctly classify the minority class, leading to a higher false positive rate. Logistic Regression, however, is less sensitive to such class imbalances.

Feature interactions are another area of consideration. While Logistic Regression assumes independence among predictors and can perform better if features are not highly correlated, Random Forest and XGBoost can model interactions between variables but may struggle if these interactions are overly complex.

Finally, the simplicity and robustness of Logistic Regression should not be overlooked. Despite being a simpler model, Logistic Regression can often perform quite well with fewer

computational resources. This simplicity can occasionally lead to more robust performance, especially when the dataset or the number of features is limited. As seen in Table 7 the two linear models can generalize better and therefore they are able to predict truer positive than the Random Forest and XGBoost models. Both the Logistic and Ridge Regression have almost the same prediction because during preprocessing all unnecessary variables were dropped and the Ridge Regression is not able to find any insignificant values to lower their weight. Therefore, both models are running with variables and almost identical weights. On the cross-validation models, the Random Forest and the XGBoost present a high degree of overfitting causing high mean-accuracy.

While the predictive models did not perform as expected in identifying potential good customers, it's important to understand that this doesn't necessarily indicate a failure in client segmentation. Instead, these results can be harnessed as a guide for refining and improving our segmentation strategies.

Each model, even if not perfect in its predictive accuracy, contributes to our understanding of the client base and their behavior patterns. This insight is particularly valuable when it comes to aligning our machine-learning models with specific business objectives.

To further investigate the predictive power of the dependent variable, we employed one well-known test that is commonly used in credit risk analysis: the Kolmogorov-Smirnov (KS) test. This method allows for a more thorough evaluation of the models' predictive capabilities and can provide additional insights into the performance of the models.

The Kolmogorov-Smirnov (KS) test is a statistical test that is commonly used to evaluate the divergence between the probability distributions of two independent samples (Cieslak et al., 2007). It is particularly useful for analyzing continuous variables, such as the probability of a binary outcome. By comparing the distributions of two samples, the KS test can help determine if there is a significant difference between them. A KS close to 1 indicates that the model can separate the two groups well, while a low value indicates that the model is not effective at distinguishing between the two groups.

As indicated in

Model Construction Method	Training Data	Test Data
	KS	KS
Logistic regression	0.27	0.26
Ridge	0.27	0.26
Gradient Boosting	0.42	0.2
Random Forest	0.92	0

Table 8, advanced analytics models such as random forest and gradient boosting exhibit higher KS values when trained with the available data. However, when evaluated with the test dataset, this indicator is not consistent. This suggests that these two models are overfitting the model with the training data, thereby limiting their effectiveness when applied to real-world scenarios. On the other hand, the logistic regression and Ridge models perform well on both training and test datasets, with a KS value of 0.26 that is consistent across both datasets. This indicates that these models can achieve a good balance between fit and generalization, making them more suitable for practical applications.

Model Construction Method	Training Data	Test Data
	KS	KS
Logistic regression	0.27	0.26

Ridge	0.27	0.26
Gradient Boosting	0.42	0.2
Random Forest	0.92	0

Table 8 KS Results

Moreover, for risk models, it is essential that the score is ordered based on the probability of risk. This ensures that the model can effectively differentiate between high-risk and low-risk individuals. As illustrated in Figure 4, the Ridge regression model achieves the desired behavior of the target variable (% Int M) exhibiting an ascending trend with the probability given by the model. This behavior is essential as it provides a clear and interpretable score that can be used to rank individuals based on their risk level. Overall, the Ridge regression model is effective in identifying high-risk individuals while minimizing false positives, making it a reliable tool for risk assessment in practical applications.

	min	max	Total	%_INT_T	%_AC_T	Buenos	%_INT_B	%_AC_B	Malos	%_INT_M	%_AC_M	KS
0	0.000%	26.874%	5542	10.00%	100.00%	5412	97.65%	100.00%	130	2.35%	100.00%	0.000000
1	26.876%	33.640%	5542	10.00%	90.00%	5390	97.26%	89.72%	152	2.74%	95.32%	0.056038
2	33.641%	37.059%	5541	10.00%	80.00%	5382	97.13%	79.48%	159	2.87%	89.85%	0.103742
3	37.059%	40.044%	5542	10.00%	70.00%	5379	97.06%	69.25%	163	2.94%	84.13%	0.148774
4	40.045%	42.979%	5541	10.00%	60.00%	5379	97.08%	59.03%	162	2.92%	78.27%	0.192310
5	42.981%	47.437%	5542	10.00%	50.00%	5350	96.54%	48.82%	192	3.46%	72.44%	0.236207
6	47.437%	52.868%	5541	10.00%	40.00%	5261	94.95%	38.65%	280	5.05%	65.53%	0.268756
7	52.869%	60.107%	5542	10.00%	30.00%	5146	92.85%	28.66%	396	7.15%	55.45%	0.267950
8	60.107%	68.169%	5541	10.00%	20.00%	5007	90.36%	18.88%	534	9.64%	41.20%	0.223216
9	68.171%	84.005%	5542	10.00%	10.00%	4931	88.98%	9.37%	611	11.02%	21.99%	0.126184

Figure 4 Validation Table

The next phase involves integrating the model into the broader context of our business problem. Here, the success of the model is not solely determined by its predictive accuracy but also by how it can be employed to manage risk and facilitate decision-making. This requires an in-depth analysis of the risk tolerance of the company, as well as an examination of the interplay between false positives and true positives.

False positives, in this context, refer to customers who are flagged as potential risks but who represent a good opportunity. Conversely, the true positives are customers correctly identified as potential clients. While a high number of false positives might initially seem detrimental, it's crucial to weigh this against the financial implications of accurately identifying true positives.

For instance, a model with a moderate number of false positives may still be beneficial if it can accurately flag a significant number of true positives, particularly if these true positives represent high-value opportunities. The key is to find an optimal balance between risk and reward, where the potential losses from false positives are outweighed by the gains made from correctly identifying true positives.

In conclusion, the models' results should not be viewed in isolation but rather interpreted within the broader context of our business objectives and risk tolerance. This approach ensures that our machine learning models serve not just as predictive tools but also as strategic assets in managing risk and maximizing returns. Moving forward, we should continually refine the models and segmentation strategies, leveraging the insights gained from each iteration to enhance the risk management and decision-making processes. As of now the recommended model to segment clients is the Logistic Regression because of its ability to generalize and predict more accurately than the other models and is also the simplest one in terms of computation power. However, if more information were to be available from other sources, the analysis could be run again to see if anything changes in terms of prediction capabilities.

References:

Acquisto, D., & Fabbri, M. (2020). Predicting credit default using machine learning techniques. International Journal of Engineering Business Management.

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE: Minority class weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering.

Bouteille, S. (2013). The handbook of credit risk management: Originating, assessing, and managing credit exposures. John Wiley & Sons.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research.

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Title of Journal

Datacamp. (2023). Gradient boosting (DB) Python. [Chart]. https://campus.datacamp.com/courses/machine-learning-with-tree-based-models-in-python/boosting?ex=5

De Brabanter, C. (2017). Credit scoring in R. Packt Publishing.

Enders, C. K. (2010). Applied missing data analysis. Guilford Press.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review

on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

Karami, A., Zare, H., & Mehri, A. (2017). Multicollinearity in logistic regression analysis: An evaluation of some methods. Journal of Applied Statistics, 44(8), 1391-1407.

Kim, S., & Yi, S. (2018). Feature selection using univariate and multivariate analysis in machine learning for predicting survival after surgery in breast cancer patients. Healthcare Informatics Research, 24(1), 38-45.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (pp. 1137-1143).

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1-5.

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.

Liu, X., Li, Y., & Zeng, X. (2017). Correlation analysis of financial data based on Pearson, Spearman and Kendall coefficients. Journal of Physics: Conference Series, 898(10), 102003.

Menard, S. (2001). Applied logistic regression analysis (2nd ed.). SAGE Publications,

Inc.

Nonato Jr, R., Silva, A., Magalhães, M., & Oliveira, H. (2017). Credit scoring with machine learning for default prediction: A comparative study. Expert Systems with Applications, 77, 114-123.

Plenum Press, New York. (1997). The problem of multicollinearity. In: Understanding regression analysis. Springer, Boston, MA. https://doi.org/10.1007/978-0-585-25657-3_37

Science of Machine Learning. (2020). Random forest models. [Chart]. https://www.ml-science.com/random-forest

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

Appendix 1

- MAX_DELINQ_RANGE: Maximum delinquency range of a client's active loans as reported to DataCredito. The delinquency range is an important indicator of a client's creditworthiness and financial stability, as it reflects their ability to make timely loan repayments.
- BAL_OVER120_RATIO: Ratio of balance over 120 days delinquent to the total balance in active loans reported to DataCredito. A higher ratio indicates a higher level of financial distress for the client, which may impact their credit risk profile.
- BAL_OVER90_RATIO: Ratio of balance over 90 days delinquent to the total balance in active loans reported to DataCredito. It serves as another metric for assessing the client's financial health.
- CHARGED_OFF_RATIO: Ratio of charged-off balance to total balance in active loans reported to DataCredito. Charged-off balances are loans that the lender has deemed uncollectible, and a higher ratio suggests a greater degree of financial difficulty for the client.
- DEBT_TO_INCOME_RATIO: Ratio of the total balance in active loans reported to DataCredito to the client's estimated income, as calculated by Quanto. This metric helps to evaluate the

client's debt-to-income ratio, which is a key factor in determining their overall credit risk.

- NON_DELINQ_BAL: Non-delinquent balance of a client's active loans reported to DataCredito. It provides insight into the client's ability to maintain a healthy credit standing by making timely loan repayments.
- TOTAL_DELINQ_BAL: Total delinquent balance of a client's active loans reported to DataCredito. It serves as a measure of the client's overall financial distress and credit risk.
- DELINQ_BAL_RATIO: Ratio of the delinquent balance to the total balance in active loans reported to DataCredito. It helps assess the client's credit risk by evaluating their propensity to default on loan repayments.
- EST_INCOME: Client's estimated income as calculated by Quanto DataCredito. Income is a crucial factor in determining a client's ability to repay loans and their overall credit risk.
- MORTGAGE_OWNERSHIP: Indicates whether the client has mortgage credit ownership. Clients with mortgage loans may have different credit risk profiles compared to those without such loans.
- TOTAL_BAL_ACTIVE_LOANS: Total balance of a client's active loans reported to DataCredito. It provides an overview of the client's outstanding debt, which is a key factor in determining credit risk.
- CHARGED_OFF_BAL: Charged-off balance in active loans reported to DataCredito. It helps to assess the client's credit risk by analyzing their history of defaulted loans.
- NUM_ACTIVE_BANKS: Number of banks with which the client has active loans.
- MIN_DELINQ_RANGE: Minimum delinquency range of the client's active loans reported to DataCredito.
- BAL_OVER90: Balance over 90 days delinquent in active loans reported to DataCredito.
- BAL_OVER120: Balance over 120 days delinquent in active loans reported to DataCredito.
- ACIERTA_PLUS_SCORE: Acierta Plus score, DataCredito’s model.
- NON_QNT_DELINQ_BANKS: Number of banks with delinquent balance different from those

in the network

- NUM_REAL_SECTOR_DELINQ: Number of entities with delinquency in the real sector
- QNT_DELINQ_BANKS: Number of banks with delinquent balance in the current Qnt network (Banks associated with Qnt)

Appendix 2

- Age_cat: The clients age is segmented the following way: <30, 30-40, 40-50, 50-60 and >60.
- City: Cities column is clean to create less values, since many of the cities were not being group due to different spelling. Also, cities with a low volume of clients are set to “Other.”
- INCOME_GROUPS: The client’s income is segmented into the following groups 1514, '1515-1911', '1912-2210','>2210', 'EXCLUSIONES'.
- MORTGAGE_STATUS: From the count of mortgage loans to 1 (has mortgage loans) and 0 (do not have mortgage loans).
- REAL_SECTOR_DEFAULT_LOANS: From the count of default loans in the real sector to Yes (has default loans in the real sector) and No (do not have default loans in the real sector).
- LOANS_OUTSIDE_QNT_BANKS: From the count loans with banks outside the QNT network, to 1 (has loans with banks outside the QNT network) and 0 (do not have loans with banks outside the QNT network).
- DEFAULT_LOANS: From the count default loans, to 1 (has default loans) and 0 (do not have default loans).
- HEALTHY_LOANS: From the count healthy loans, to 1 (has healthy loans) and 0 (do not have healthy loans).
- LOANS_INSIDE_QNT_BANKS: From the count loans with banks inside the QNT network, to 1 (has loans with banks inside the QNT network) and 0 (do not have loans with banks inside the QNT network).
- RESTRUCTURED_LOANS: From the count of total restructured loans, to 1 (has restructured loans) and 0 (do not have restructured loans).