**Loan Default Prediction based on the domain.**

Willer Pinheiro Oliveira - https://www.linkedin.com/in/willeroliveira

# Executive Summary

Credit risk assessment and loan management are critical global issues that deeply affect financial institutions. Many banks have tried to automate the loan approval process using data science and machine learning to make it more efficient and reduce human bias. Using predictive modeling techniques can streamline the loan origination process, improve credit recovery processes, and reduce bias in financing decisions. To contribute to this important process of change, we use several statistical tools and machine learning algorithms to predict, as accurately as possible, the probability that a borrower will default on real estate loans. We tested three different Machine Learning models (Logistic Regression, Decision Tree, and Random Forest), comparing them through several performance metrics, to define which, one would be the most suitable for this task. In this report we will present all the discoveries we made about the proposed data, the problems we found in the search for the best model and the way we solved these problems. We will also do a brief analysis of the likely costs of putting a project like this into production, the parties involved and the benefits of implementing this type of project. Finally, we will give our opinion on which solution was chosen and the justification for it, as well as everything that still needs to be done so that the project is ready to be put into production.

## Most important findings from the analysis

The insights below were organized according to the level of importance assigned by the random forest model but were also observed in the bivariate analysis.

* There appears to be a strong relationship between the number of non-performing credit lines and the percentage of non-performing customers. As the number of delinquent accounts increases, the percentage of delinquent customers also increases significantly, suggesting that individuals who are late on their loan payments are more likely to be classified as high risk and have their loans denied. Here, the DT and RF models seem to have agreed and considered this variable as the most important among all the others, as it occupies the 1st position in their rankings, for the classification of good and bad payers.

Chart, line chart

Description automatically generated

* In second place in the opinion of both models is the debt/revenue ratio. We can see that the percentage of commitment of income to debt payments increases, the greater the proportion of customers classified as defaulters. Therefore, a high debt-to-income ratio may indicate that a customer is already heavily committed to other debt, which may increase the risk of default.

Chart, histogram

Description automatically generated

The most interesting thing is to note that in the analyzed dataset the interval of this indicator varies between 0 and 200%. This means that some customers continued to use their credit even when they had 100% of their income committed to paying off debts. They are completely profligate and have doubled their ability to get into debt. My grandmother used to say: for some, at the bottom of the well there is still a trapdoor and a long ladder to the abyss.

* There is a very clear trend that, as the number of derogatory reports increases, the percentage of defaulting customers also increases, and the percentage of good payers decreases proportionally. This suggests that the count of derogatory reports can be a good indicator of credit risk, and lending companies may want to consider the number of derogatory reports when assessing the risk of lending money to an individual. This variable was ranked 8th and 3rd by the DT and RF models, respectively. Once again, I believe that the RF model was more assertive in its classification.

Chart, line chart

Description automatically generated

* There are some interesting observations about the age of the oldest credit line. Credit lines up to seem one year to have a higher percentage of defaulting customers, but after one year the trend reverses and the percentage of good payers increases as credit lines become longer. However, there are exceptions to this trend, and the relationship between the age of the credit line and the status of the customer (good payer or defaulter) is not well defined. Both models consider this variable to be one of the most important: DT considered it in 3rd place and RF in 4th place.

Chart, histogram

Description automatically generated

* There is a tendency that, as the number of recent credit queries increases, the percentage of defaulting customers also increases. However, after a certain level (approximately 7 to 12 consultations), the trend becomes less clear, and the proportion of good and bad payers fluctuates. This could indicate that many recent credit inquiries could be an indicator of a financially risky situation. The DT model attributed to this variable the 9th position in its ranking of importance, while the RF model considered it the most important, assigning it the 5th place. In fact, many credit inquiries could mean that this person may be seeking credit from several different sources and not being successful, and this calls for caution. Therefore, I am once again inclined to think that the RF model was more assertive.

Chart, line chart

Description automatically generated

* Loan value, property value, and outstanding mortgage value are important in credit risk scoring as they provide relevant clues about the likelihood of default occurring: For example, a loan that exceeds the property value or a mortgage debt that is high relative to the property's value may pose a greater credit risk, as the borrower may have less incentive to repay the debt if he believes his property is worth less than the amount owed on his mortgage.
* There does not seem to be a well-defined trend in the relationship between the number of existing credit lines and the customer's situation (delinquent or good payer). The models considered this variable with a low level of importance, in 10th place for the DT model and 9th place for RF.

A picture containing chart

Description automatically generated

* There seems to be a slight trend indicating that as the number of years a client has worked at their current job increases, the percentage of defaulting clients decreases, while the percentage of good payers increases. However, this trend does not seem to be very significant, as there are exceptions and the influence of job stability on defaults may not be strong. Our DT model gave slightly more importance (4th) to this variable than the RF model (10th). Considering that the RF model obtained better metrics, I believe that this may, in fact, be a less significant attribute for the credit risk analysis process.

Chart

Description automatically generated

Another insight that intrigued me, despite not being representative of the data set studied, was the number of cases with debts greater than their own property values. There are 51 rows in this condition, which is only 0.01% of the total rows in the dataset. See the graph below, that approximately ¼ of them are in default. 25% is a very high percentage. I think these people were not very motivated to pay off their mortgage, knowing that their property is worth less than their debt. Toughness!

A pie chart with a red and green circle

Description automatically generated with low confidence

Now that we have observed these “morphological” aspects, which are the main insights about the data, let's take a look at their “anatomy” and see what other insights they brought us during our study.

* **Dataset size:** The proposed dataset is small to train machine learning models with maximum efficiency. With only 5960 rows and 13 columns, it may have been responsible for the model overfitting problems to the training data that we highlighted. Models trained on small datasets may fit well on training data but may perform poorly on test data or new data, as they may learn from noise and random patterns rather than true, relevant patterns.
* **Missing values:** To handle missing values in the dataset, several measures were taken, such as excluding rows that did not contain values for certain columns, filling in missing values with calculations, and filling in missing values in other columns with the mode or median. This was done to ensure that the data is complete and ready to be used in machine learning models.
* **Data distribution:** The data distribution in most columns is asymmetric and skewed to the right. This may have implied bias in the mean, a concentration of low values, the presence of outliers, lower variability, and the need for transformations in the data to normalize the distribution.
* **Data imbalance:** lines referring to defaulters represent 80% while defaulters account for the remaining 20%. This is an important factor to consider in classification problems, as the model may tend to bias towards the majority class. We try to minimize this problem by over-sampling (SMOTE) and under-sampling (Tomek Links) to deal with unbalanced datasets. Even with the mitigation of this problem, the data imbalance resulted in a bias in the model, which we observed in the metrics, which showed better performances for the majority class. This resulted in underrepresentation in the predictions and the model had difficulty generalizing correctly to this class when subjected to new data.
* **Presence of many outliers:** In 48.6% of the rows, there is at least one outlier in one of the columns. Removing these lines would negatively affect the ability of the model to generalize to new data, so this option was not considered a viable option. To try to minimize the problem, I performed data transformations on two variables, as the other transformations used had no effect. But, even after making the transformations, 39.6% of the lines still had outliers. See the reduction in the number of rows after transforming the data.

**Before data transformation After data transformation**

A picture containing text, screenshot, font, diagram

Description automatically generatedA picture containing text, screenshot, diagram, font

Description automatically generated

A total of 530 rows with outliers were data transformed and no longer identified as containing outliers. In addition to data transformations, we tried using the DBSCAN algorithm, but without success for our dataset.

## Final specifications in the proposed model

The model selected was the Random Forest, for which we assigned a score of 8/10. Below you can see the comparison of the metrics of the three models with the test dataset, performed after adjusting the hyperparameters.

A screenshot of a graph

Description automatically generated with low confidence

Random Forest stood out with superior overall performance, showing the highest accuracy, precision, recall and F1-score among the three models. It demonstrated an excellent ability to reduce false negatives, indicating a superior ability to correctly identify positive cases. In addition, it showed a significant reduction in false positives, which represents an improvement in the identification of negative cases. However, there were still a small number of false negatives, albeit at a considerably lower rate compared to the other models. Given the critical importance of minimizing false negatives, the score given is high and there is still room for improvement.

Based on this analysis, despite the Random Forest model having stood out from the others, with a superior global performance, I do not recommend that it be put into production, because despite greatly reducing the cases of false negatives, it would not be good for the bank to admit a model that still struggles to correctly identify defaulters without mistaking them for good payers.

The main factor that I consider having contributed to the random forest model not better identifying false negatives was the reduced size of the data set. Generally, the larger the dataset, the better the model will perform, provided the model is properly trained and validated. Machine learning models need a large dataset to learn patterns and generalize well to new data. With a smaller dataset, the model can end up learning patterns that are very specific to the dataset, which can result in overfitting. And when it comes to Random Forest, which is a tree-set decision-making algorithm, the size of the dataset can affect model performance in several ways:

* **Variance:** The size of the dataset can affect model variance, which refers to the model's sensitivity to fluctuations in the training data. The smaller the dataset, the greater the model variance, which can lead to overfitting.
* **Bias:** The size of the dataset can also affect model bias, which refers to the tendency of the model to consistently predict values other than the true value. When the dataset is small, the model may not be able to capture the full complexity of the problem, which can result in a biased model.
* **Training time:** Another factor to be considered is the model's training time. Generally, Random Forest models are computationally intensive and can take a long time to train on large datasets.

In general, it is important to have a large enough dataset so that the model can successfully learn patterns and generalize well to new data. The exact size of the dataset needed can vary depending on the specific problem and the machine learning algorithm used.

I believe that two other factors, previously mentioned, also influenced the random forest model to not present a better result in relation to False Negatives:

* **Data imbalance:** 1/5 of the dataset lines only belong to class 1. Thus, it is natural that the metrics referring to the defaulting class had a lower performance. In our unbalanced dataset, where the minority class has a significantly smaller representation in relation to the majority class (class 0), the model created biases for the majority class during training, even after applying sampling techniques such as over-sampling (minority class increase) or under-sampling (majority class decrease).

In the balancing process, the algorithms made copies of class 1 records at random and reduced the amount of class 0 records, but the model was exposed to the same amount of data after balancing, the majority class records were more varied, while those of the minority class were partly duplicate samples. As a result, the model can learn to classify most samples as belonging to the majority class, achieving a high global accuracy due to the bias in favor of the majority class, but presenting a low detection rate (recall) for the minority class. This can result in false negatives, where minority class samples are incorrectly classified as belonging to the majority class, leading to an underestimation of the minority class presence by the model. The solution to this problem would be to train the model with more balanced data.

* **Presence of many outliers:** Even a Random Forest model, which is generally considered more robust than methods such as Decision Trees and Logistic Regression, can be affected by outliers present in about 39% of the rows of the dataset, even after data transformations. The presence of outliers in such a significant proportion affected the model's ability to correctly learn the underlying pattern in the data and hindered its ability to make accurate predictions.

For all these reasons, I believe that you need to address these issues before putting the model into production and even, as a last resort, consider training the model with a larger and more homogeneous dataset.

# Problem and Solution Summary

## The problem

The essential issue we must consider is the importance and necessity of carrying out proper credit risk assessment and loan management in financial institutions, especially given the impact of significant losses caused by defaulting borrowers. The challenge in solving this problem is to automate the loan approval process, using data science and machine learning, to make it more efficient and bias free, preventing the machine from learning the previous biases that have arisen due to the process. of human approval.

## Reasons for the proposed project and how it would affect the problem/business.

The decision to replace the credit risk analysis process, carried out by humans, with an automated system based on a machine learning model, has as its main reason the fact that if this is not done properly or if it is neglected, it can lead to significant losses for financial institutions. In addition to this, I can list some other good reasons why financial institutions should be concerned with this matter, and I will list them below in four main groups:

* **Risk reduction and quality improvement:** Automating the credit risk assessment process can significantly reduce the risks associated with lending by allowing the use of sophisticated algorithms and advanced statistical models for accurate and detailed assessment. Furthermore, by utilizing data science and machine learning techniques, it is possible to build more objective and bias-free models, minimizing the chances of misjudgments or inappropriate loan approvals.
* **Efficiency, productivity, and reduction of operating costs:** Automation can make the credit risk assessment process more efficient and productive, speeding up the loan approval process, reducing response time to customers, and increasing loan application processing capacity. Automation can also help reduce operational costs associated with the manual process, such as labor, training, and supplies expenses, and minimize losses from bad loans, resulting in cash savings.
* **Consistency, standardization, and regulatory compliance:** Automation ensures that the credit risk assessment process is consistent and standardized, avoiding discrepancies and variations in assessment criteria. This can help financial institutions meet regulatory requirements more accurately, ensuring decisions are based on objective and transparent criteria, which can reduce the risk of legal action, fines, and penalties.
* **Improved customer experience:** Automation can result in a faster and more convenient experience for customers applying for loans, with faster and less bureaucratic approval processes, consequently, it can lead to a higher level of customer satisfaction and loyalty to the financial institution.

# Recommendations for implementations

As we saw earlier, the selected model is robust and versatile, but still needs improvements and refinements until it can be put into production. For this reason, here are some of the main recommendations for implementing this Random Forest model:

* **Use a larger dataset:** Random Forest models tend to perform better with larger datasets because they rely on multiple decision trees to make predictions. A larger dataset allows the model to learn patterns more effectively and generalize well to new data, reducing the risk of overfitting.
* **Dealing with class imbalance:** If the dataset used to train the Random Forest model has imbalanced classes, meaning that some classes have significantly fewer examples than others, it is important to address this issue. Class imbalance can affect model performance, especially in terms of accuracy, recall and F1 score. Techniques such as oversampling, undersampling, or using weighted loss functions can be used to deal with class imbalance and improve model performance.
* **Tuning hyperparameters:** Random Forest models have several hyperparameters that can be tuned to optimize their performance. Some important hyperparameters to consider in tuning include the number of trees in the forest, the maximum depth of trees, and the minimum number of samples needed to split a node. Experimenting with different values of these hyperparameters can help fine-tune the model and improve its performance.
* **Feature Selection:** Random Forest models can handle many features, but not all features may be equally important for making accurate predictions. Feature selection techniques, such as using feature importance scores provided by the Random Forest model or using techniques such as recursive feature elimination (RFE), can help identify the most important features for the model's predictions. Removing irrelevant or redundant features can reduce noise and improve model accuracy.
* **Cross-validation:** It is important to evaluate the performance of the Random Forest model using cross-validation to obtain a more robust estimate of its performance. Cross-validation involves partitioning the dataset into multiple folds and training the model on different subsets of the data. This helps reduce the risk of overfitting and provides a more reliable estimate of model performance.
* **Joint learning:** Random Forest models are based on the concept of joint learning, which combines the predictions from multiple decision trees to make a final prediction. Experimenting with different ensemble techniques, such as bagging and boosting, can further improve the performance of the Random Forest model.
* **Interpretability:** Random Forest models are known for their ability to provide feature importance scores, which can help interpret model predictions. Analyzing and interpreting feature importance scores can provide insight into the most important features to make accurate predictions and help understand underlying patterns in the data.
* **Performance Metrics:** Carefully evaluate the performance of the Random Forest model using appropriate performance metrics such as accuracy, precision, recall, F1-score and AUC-ROC, considering the specific project requirements and objectives. Choose the performance metrics most relevant to the problem at hand and align them with the project objectives.
* **Try and Iterate:** Random Forest models offer flexibility and can be fine-tuned through experimentation and iteration. It is important to experiment with different hyperparameter configurations, feature selection techniques, and assembly techniques and iterate on the model to continually improve its performance.
* **Monitor model performance:** Once the Random Forest model is deployed, it is important to monitor its performance on new data and regularly update the model as needed. Track model performance metrics and continually evaluate its accuracy and reliability to ensure it continues to perform well in production.

## Main Stakeholders

Below, I've listed stakeholders who, in my opinion, would benefit most from the development and implementation of a project like this. They are organized, in my point of view, from the most benefited to the least benefited.

* **Banks and financial institutions:** The project model, particularly the Random Forest model, has the potential to help banks and financial institutions to identify defaulters and minimize false negatives, which can result in significant financial losses relatively accurately. By deploying an improved model with a larger and more diverse dataset, with better refined hyperparameters, and more appropriate and well-applied pre-processing techniques, banks and financial institutions can make more informed decisions about lending, risk assessment and management. portfolio, leading to better profitability and risk management.
* **Lenders and Investors:** Lenders and investors who have financial interests in loans and loan portfolios can benefit from implementing an accurate and reliable project model. The model can help lenders and investors assess the risks associated with loans and loan portfolios, make informed investment decisions, and mitigate potential financial losses.
* **Clients and borrowers:** Speed in the response of the application to borrow and the guarantee of not having any type of judgment/discrimination, due to worldviews and biases, offering equal chances to all who seek a loan at the institution. In addition, by avoiding undue defaults, the bank preserves the credibility of its customers, as well as its own.
* **Regulators and policy makers:** Financial sector regulators and policy makers can benefit from implementing an effective project model as it can help improve compliance with regulatory requirements such as risk assessment and credibility assessment. Finally, the precise identification of defaulters can also contribute to the stability of the financial system.
* **Scientists and data analysts:** these can benefit from the analysis of the model's performance and recommendations for improvement, in terms of expanding the team's experience and expertise in the bank's processes and the technologies used. Insights gained from analytics, such as the need for a larger and more diverse dataset, hyperparameter tuning, pre-processing techniques, and model selection, can help guide additional research and development efforts, leading to refinement and optimization. of the model for better performance and the development of the team and discovery of new techniques to carry out the tasks more accurately and quickly. As we saw earlier, the model selected for the best performance still needs refinements until it can minimize and even eliminate cases, still persistent, of false negatives. Thus, the team is taken out of its comfort zone to seek better answers to the bank's problems, and at the same time specializes and enriches itself in the process.

In summary, implementation of the design model, particularly with improvements based on the analysis and recommendations provided, can benefit a range of stakeholders, including banks and financial institutions, lenders and investors, customers and borrowers, regulators and policymakers, scientists and data analysts. Accurately identifying defaulters and minimizing false negatives can lead to better risk management, investment decision making, client protection, regulatory compliance, and other advances in the field of data science and machine learning.

## Expected benefits and costs

With the aim of presenting the benefits and cost expectations of implementing an automated credit risk assessment solution for loan management, I report below a simple fictitious case, which I believe will serve very well to illustrate this subject in a very didactic way. And in this way we will better understand the impact that the use of data science can bring to a financial institution in the way it solves the problems common to its business.

**Charles Bronson Bank case:**

The Charles Bronson Bank (CBB) is a financial institution located in the United States, which faces challenges related to credit risk analysis in its mortgage lending process. It was founded in 1935 by Mr. Walter Buchinsky, who left it to his son, actor Charles Dennis Buchinsky, better known by his stage name Charles Bronson (famous for the “Death Wish” franchise), who lent his name to this small but reliable family financial institution.

The financial world has changed a lot since Mr. Charles Bronson left us in 2003. Today bad debt makes the CBB business much less profitable than it was after the mid-1930s, when he founded the CBB. The bank benefited greatly from the series of policies known as the New Deal, implemented by the United States Government, as well as the financial regulation efforts that followed, such as the Glass-Steagall Act of 1933 and the creation of the Securities and Exchange Commission in 1934. In the Great Depression of '29, there was a boom in bankruptcies, people lost their jobs and stopped paying their mortgages, it all happened like a series of dominoes falling on top of each other. From 1935 onwards, default rates fell again, and banks benefited greatly from this more stable recovery scenario. Today we have yet another default epidemic spread across the country. God in heaven!

Currently, the bank uses traditional credit analysis methods, which involve a manual process carried out by a team of 20 credit analysts. However, this process has proven to be inefficient, resulting in high operational costs, delays in credit decisions and an increase in the rate of loan defaults.

The costs associated with the old CBB structure are as follows:

* **Credit Analyst Salaries:** The bank has a staff of 20 credit analysts, with a total salary cost of $1,770,000 per year, assuming an average salary of $88,500 per credit analyst.
* **Training and capacity building costs:** CBB also conducts periodic training to train its staff of credit analysts, at an estimated cost of $100,000 per year.
* **Operating costs:** In addition to salaries and training, the bank has ongoing operating costs, such as energy, communication, and other resources, estimated at $100,000 per year.
* **Bad debt loss:** Based on historical data, the bank's estimated loss due to bad debt is $3,000,000 per year.

Therefore, the total annual cost of the old CBB structure, including salaries, training, operating costs and bad debt, is $4,970,000 per year.

Against this backdrop, CBB's management board is considering implementing a data science-based solution to improve its credit risk analysis. After a detailed analysis, the bank's management team identified the following key components for implementing the data science project:

* **Investment in technology and infrastructure:** The bank will need to acquire data analysis software and IT infrastructure to support the implementation of the project. The total cost of this portion of the project is estimated to be approximately $500,000.
* **Skilled workforce:** The bank will need to hire a data science team comprised of professionals with varying levels of experience. The estimated annual charges and costs are as follows:
* Senior Data Scientist (1): $250,000
* Data Scientist (2): $150,000 each
* Data Engineer (1): $150,000
* Data Analyst (1): $100,000

The estimated annual cost of the data science team, considering average salaries in the US market, is $800,000.

* **Training and capacity building costs:** It will be necessary to invest in training and capacity building for data science professionals so that they are able to develop and implement the necessary analytical models. The estimated cost of training is $150,000.
* **Operating Costs:** There will be ongoing operating costs associated with maintaining the IT infrastructure and spending on power, communications, and other resources. These costs are estimated to be approximately $100,000 per year.

Therefore, the total annual cost of the data science team IT investment, training, and operating costs is $1,550,000.

Based on this information, the CBB management team performed a detailed analysis of the potential benefits of this data science project:

* Reduced labor and staff training costs would result in savings of $920,000 per year.
* **Delinquency rate reduction:** With the implementation of advanced analytical models, the bank expects a significant reduction in the mortgage loan delinquency rate. Based on conservative projections, it is estimated that reducing the default rate could reach 10%, which would result in savings of $5,000,000 per year in avoided losses.
* **Increased Operational Efficiency:** Automating and streamlining the credit analysis process using advanced analytical models should also result in a significant reduction in the time and effort required to analyze each loan application. It is estimated that the bank could increase its analytics capacity by up to 50%, which would save $1,000,000 per year in operating costs.
* **Improved customer experience:** With more efficient processes and faster credit decisions, CBB customers will have a better experience in the mortgage loan process, which can result in greater customer satisfaction and loyalty, as well as potential referrals for new businesses.

Based on the identified benefits, the Charles Bronson Bank management team performed a return on investment (ROI) analysis of the data science project. Considering a time horizon of 3 years, the ROI was estimated at approximately 120%, which indicates a significant return on the investment made.

As a result, CBB achieved a significant return on investment, with an estimated $6,000,000 increase in annual revenue and an improved customer experience. The data science project was considered a success and served as an example of how implementing advanced technologies can bring tangible benefits to financial institutions. Oh! Mr. Charles Bronson would have been very happy with the results if he hadn't left us in 2003! Longing Mr. Bronson!

This brief fictitious history could well be the same as that of any financial institution that thought of making this important change in its structure, for the sake of the survival of its business. Furthermore, this fictitious case can be the reality for a project like ours.

## Potential design risks or challenges of the proposed solution

Below are some of the risks that I identified, more specifically on the model of our study in this project, as well as more broadly, on the risk analysis process for real estate credit, in a broader sense.

**Model accuracy risk:** Although random forest and decision tree models have been used to classify customers as good or bad payers based on various variables, there is still a risk that these models may have limited accuracy due to the quality of training data and algorithms used.

**Risk of missing variables:** While the analysis identified some important variables for predicting loan defaults, there may be other relevant variables that were not considered in the model, such as detailed financial information of customers (history of payment of utility bills, employment and income history, among others), which may influence the probability of default and were not included in the model, some of them because they are not available in the analyzed dataset.

**Risk of data bias:** The results obtained by the model depend on the quality and representativeness of the training data used. If the data used is biased or does not adequately represent the target population, the model results may be skewed, which highlights the importance of ensuring that the data used is of high quality, free of bias, and adequately represents the target population of customers. question.

**Risk of incorrect interpretation of results:** The interpretation of the results obtained by the models can be complex and subject to misinterpretation, which highlights the need to ensure that the results are correctly understood and interpreted to avoid inappropriate decision-making based on the results of the project. Most of the time, models can indicate risk of default, but the decision to approve or deny loans is in human hands.

**Regulatory compliance risk:** The forecast of loan defaults is subject to specific laws and regulations in different jurisdictions. It is important to ensure that the project complies with all applicable regulations and laws, such as customer data protection and compliance with credit regulations, to avoid legal and financial consequences for the project and the organization.

**Risk of changes in the economic and market environment:** The economic environment can change over time, affecting customers' ability to repay their loans. Changes in interest rates, the supply of jobs, the general economic situation in the country or in specific sectors can have a direct impact on loan defaults. It is important to closely monitor the economic environment and regularly update the default prediction model to ensure its continued accuracy.

**Fraud risk:** Fraudsters can falsify personal information, documents, and proof of income to fraudulently obtain credit. This can cause financial harm to the financial institution and affect the trust of legitimate customers.

**Operational risk:** Operational risk is related to failures in internal processes, systems, people, and external events that may result in financial losses. In the context of granting credit, this may include processing errors, failures in information technology systems, inadequacy of policies and procedures, deficiencies in corporate governance, among others.

In summary, the granting of credit presents several risks, many of which are not mentioned here. Effective risk management when granting credit is essential for the sustainability and success of financial institutions, ensuring informed decision-making, protecting the interests of stakeholders, and minimizing the negative impacts of any credit problems.

## Additional analysis for unresolved issues

Previously, we detected three problems that we detected in our dataset that contributed to the performance of the selected model still being classified incorrectly, defaulters as if they were good payers (False Negative):

1) The imbalance of the data.

2) The large amount of outiliers.

3) The size of the dataset.

Therefore, let's see below what other measures can be taken to try to solve these problems.

**Problem 1:** I tried to apply over-sampling and under-sampling technique (SMOTETomek). Compare below random forest model metrics on test data (last test) before using SMOTETomek vs. after using SMOTETomek:

Before SMOTETomek After SMOTETomek

A screenshot of a graph

Description automatically generated with low confidence

A picture containing text, screenshot, font, number

Description automatically generated

Although the change was tenuous, we see an improvement in Cross-Vall, Class 1O recall and Class 1 F1-Score, indicating an improvement in the model's performance of correctly identifying the Defaulter Class. Despite the improvement, SMOTETomek's algorithm has not yet taken the model to the level of excellence we want, since there are still many false negatives. However, there are still other techniques to explore, such as:

* **ADASYN (Adaptive Synthetic Sampling):** It is a variation of SMOTE that adapts the generation rate of synthetic examples based on the density of the instances. This means that the generation of synthetic examples is more intense in regions where the minority class is scarcer, making it more suitable for unbalanced data with complex distributions.
* **Cost-sensitive Learning:** It is possible to assign different weights to classes during model training, giving more importance to the minority class. This can be done by adjusting class weights or using cost optimization methods, where the cost of misclassifying the minority class is weighted differently than the cost of misclassifying the majority class.
* **Modification of machine learning algorithms:** Some machine learning algorithms can be modified to deal directly with imbalanced data. For example, the SMOTEBoost algorithm combines the SMOTE technique with the boosting algorithm to create an ensemble of classifiers that is more robust against data imbalance.
* **Transfer Learning:** Transferring knowledge of a related problem can be useful to improve the performance of models on imbalanced data. For example, pre-training a model on balanced data from a related task and then tuning it for unbalanced data can help improve the generalizability of the model.

**Problem 2:** Two ways were tried: the first was to do some data transformations. This technique only worked with two variables, and the reduction in outliers was not significant. The second was clustering through DBSCAN, but DBSCAN was not able to find any clusters in the dataset.

Unfortunately, there is no "silver bullet" or one-size-fits-all solution when it comes to handling data outliers. The most appropriate approach may vary depending on the context of the problem, the nature of the data and the objectives of the analysis.

The data transformations, exponential, square root, Box-Cox that I used, can be useful to reduce the influence of outliers in some situations, especially when the data is asymmetrically distributed, which is the case in the dataset we are working with. However, these transformations are not always efficient on all fields in a dataset, as I have observed. In addition, some transformations can change the interpretation of the data or even corrupt the data set, as in the case of inverse transformations, which resulted in the insertion of "imp" in some lines of the dataset, which of course, generated an error when I was create the machine learning models.

Other approaches to handling outliers can include:

* Direct removal of points identified as outliers.
* The replacement of these points by imputed values.

I didn't try these two as I was hopeful that the transformations would give a better result.

* Application of robust statistical methods that are less sensitive to outliers, such as robust regression or robust parameter estimation methods.

However, it is important to be careful when removing or replacing outliers, as this can affect the representativeness and completeness of the data.

**Problem 3:** Unfortunately, if none of these yet-to-be-realized approaches work, you'll need to train the model with a larger dataset, not just in number of records, but also in number of variables. This new dataset will also need to be more homogeneous, containing a greater number of class 1 records. This way, the random forest model will “learn” better, with much more data to explore every detail that identifies classes 0 and 1.