# Université Jean Monnet - Saint-Etienne


# CREDIT RISK ANALYSIS

- Author :

  Mohamed Islem BEGGARI

- Lead teacher :

  Fabrice MUHLENBACH

30/07/2023

# TABLE OF Contents

# TABLE OF FIGURES

# 1. Introduction

## 1.1 What is Credit Analysis?

Credit analysis is the systematic process of evaluating the creditworthiness of individuals, businesses, or entities seeking to borrow funds from financial institutions or lenders. It involves a comprehensive assessment of the borrower's financial health, repayment capacity, and risk profile to determi+ne whether granting credit is feasible and at what terms.

The primary goal of credit analysis is to provide lenders with insights into the borrower's ability and willingness to fulfill their financial obligations. By analyzing various financial and non-financial factors, credit analysts aim to make informed decisions about extending credit, setting interest rates, and establishing loan terms. This process helps lenders mitigate the potential risks associated with defaults and loan delinquencies.

- **How is Credit Analysis Conducted?**

Credit professionals analyzing a prospective borrower will employ a variety of qualitative and quantitative techniques.

**Qualitative** techniques include trying to understand risks in the external environment, like where interest rates are heading and the state of the broader economy, among others. A framework like **PESTEL** is often employed.

For commercial lenders, specifically, they will also want to understand business characteristics – like the borrower's competitive advantage(s) and industry trends (using frameworks like **SWOT** and **Porter's five Forces**.

**Quantitative** elements of the analysis include assessing financial ratios using risk models, understanding financial projections, employing sensitivity analysis, and evaluating the strength of any physical collateral that could serve as security against the credit exposure.

- **What are those techniques?**

**PESTEL Analysis:**

PESTEL (Political, Economic, Sociocultural, Technological, Environmental, Legal) analysis is a framework used to assess the external macro-environmental factors that could impact a borrower's creditworthiness and a lender's decision-making process.it helps to identify potential risks and opportunities associated with a borrower's economic environment, regulatory landscape, societal trends, technological advancements, and more.

**SWOT Analysis:**

SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis is a strategic framework used to evaluate both internal and external factors affecting a borrower's creditworthiness. It assesses a borrower's financial health, business operations, industry trends, and competitive positioning to identify strengths that could lead to successful loan repayment, weaknesses that could pose challenges, opportunities for growth, and potential threats.

**Five Cs of Credit:**

The five Cs underpin the component parts of most risk rating and loan pricing models. The five Cs are:

**Character**: Assesses the borrower's trustworthiness, integrity, and credit history.

**Capacity**: Evaluates the borrower's ability to repay the loan based on their income, expenses, and financial obligations.

**Capital**: Examines the borrower's financial reserves, investments, and net worth.

**Collateral**: Considers the assets that the borrower pledges as security for the loan.

**Conditions**: Analyzes the economic, industry, and market conditions that could affect the borrower's ability to repay the loan.

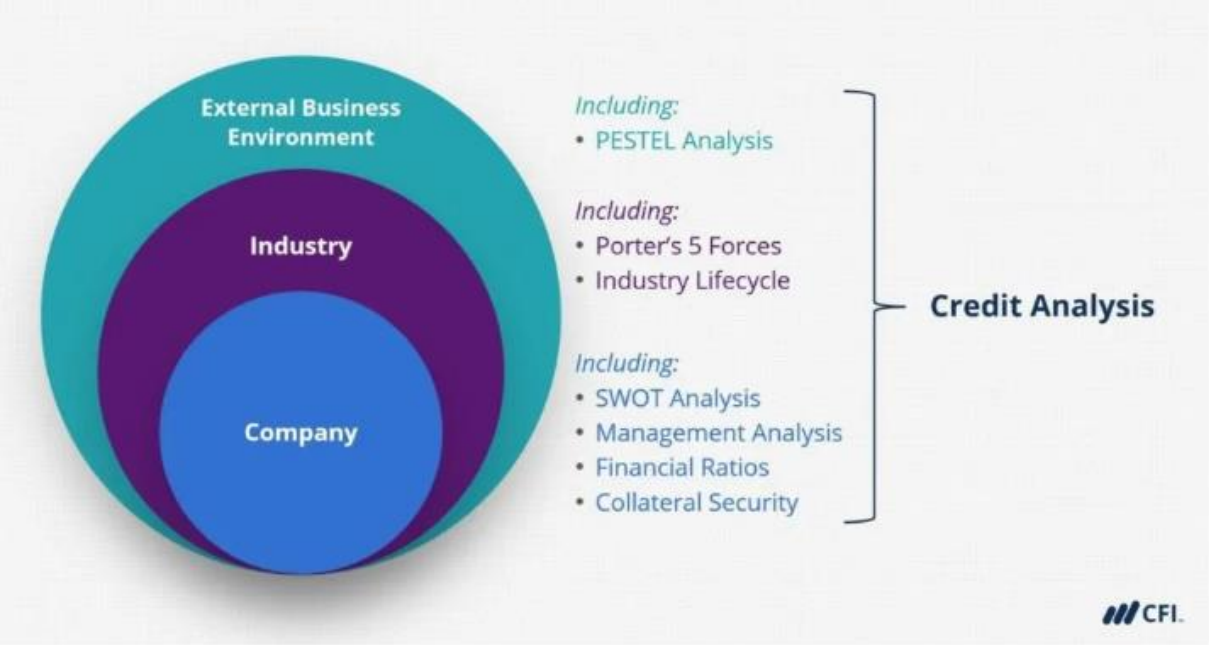


*Figure 1: Steps to follow in CREDIT ANALYSIS*
*Kyle Peterdy « credit analysis » corporatefinanceinstitute.com (January 09, 2020).*

## 1.2 What is Credit Risk Analysis?

While credit scoring helps paint an important picture of a customer's creditworthiness based on their financial history, it does not tell you much about their probability of default. Those with low credit scores may be at a higher risk for nonpayment, based on their history of default or other financial issues, but a good credit score does not necessarily mean a customer is a low risk.
Even with a stellar credit history, any business or individual faced with significant or unexpected economic hardships is at risk of default. That's why refining your credit-scoring technique is an important part of enhancing your credit risk analysis.
Credit risk analysis goes beyond the score of credit assessment and involves a strategic process through which lenders make informed decisions by evaluating the trade-offs between the advantages and drawbacks associated with assuming credit risk.

By balancing the costs and benefits of granting credit, lenders measure, analyze and manage risks their business is willing to accept.
The creditworthiness of the borrower, derived from the credit analysis process, is not the only risk lenders face. When granting credit, lenders also consider potential losses from non-performance, such as missed payments and potential bad debt. With such risks come costs, so lenders weigh them against anticipated benefits such as risk-adjusted return on capital (RAROC).

## 1.3 What are the goals and objectives of my research?

**Comprehensive Understanding of Credit Risk Analysis**

- To provide an in-depth overview of the concepts and principles of credit risk analysis, including its significance in the financial industry and its role in lending decisions.
- To review the existing literature and research on credit risk analysis techniques, methodologies, and models, highlighting their strengths and limitations.

**Development of Credit Risk Models**

- To develop and implement advanced credit risk prediction models that incorporate a wide range of relevant features, including financial and non-financial factors, to enhance the accuracy of risk assessment.
- To explore and compare various machine-learning algorithms, such as random forests, support vector machines, and neural networks, in the context of credit risk prediction.

**Handling Data Imbalance and Enhancing Model Generalization**

- To address the challenge of imbalanced data in credit risk analysis by employing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to balance the distribution of target classes.

**Performance Evaluation and Comparison**

- To conduct a comprehensive evaluation of the developed credit risk models using appropriate performance metrics, including accuracy, precision, F1-score, and ROC-AUC.
- To compare the performance of the developed models with benchmark models and assess their effectiveness in real-world credit risk assessment.

**Development of a Streamlined Web Application**

- To create an intuitive and user-friendly web application using Streamlit that allows users, including financial professionals and borrowers, to input relevant information and receive instant credit risk predictions.

# 2. Literature Review

The field of credit risk analysis has seen significant advancements over the years, with researchers and practitioners exploring various techniques to accurately assess and manage credit risk in lending and financial decision-making. This review highlights key credit risk analysis techniques, their strengths, limitations, and potential alternatives.

## 2.1 Existing literature on credit risk analysis techniques:

- **Traditional Credit Scoring Models (FICO):**

A credit score is a 3-digit number that reflects the likelihood that a consumer will repay his debts. With so many scoring methods used to determine your credit score, the variety of models means your score can vary several points, depending on whose model is used and what type of business is asking for it (department store, car dealership, bank).

**How is it used in real world?**

Traditional credit scoring models such as the FICO score. Statistical algorithms that use a borrower's credit history, payment behavior, and other relevant data to calculate a credit score. This score is a numerical representation of the borrower's creditworthiness and risk profile. This technique is Simple, well established and widely used by financial institutions. Lenders often set cutoff points on the credit score scale to classify borrowers into different grades, such as 'Excellent,' 'Good,' 'Fair,' and 'Poor.'


*Figure 2: FICO score interpretations*
*Listingsfor1 « How to Increase Your Credit Score» Listingsfor1.com (January 21, 2022).*

**How it is calculated?**

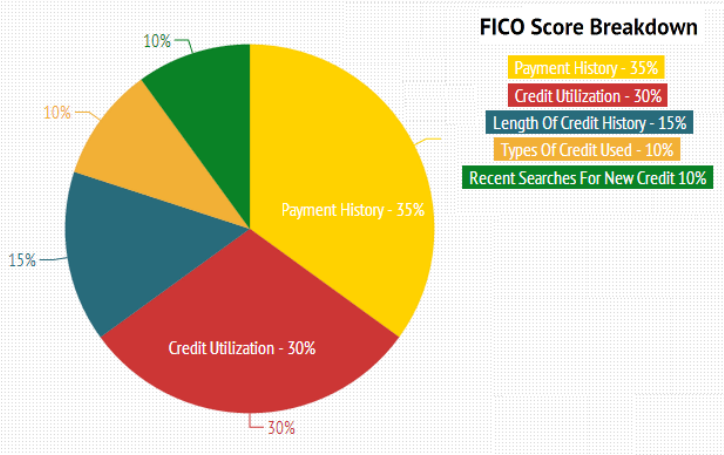


*Figure 3: How the FICO score is calculated*
*Doctor of credit « How FICO Score is calculated » doctorofcredit.com (October 22, 2013).*

- **Machine Learning Algorithms:**

Advanced machine learning algorithms, such as random forests, support vector machines, and gradient boosting, are increasingly used in credit risk analysis. These algorithms analyze a wide range of borrower attributes and historical data to predict creditworthiness and assign grades.

## 2.2 Different machine learning models and techniques used for our credit risk assessment.

In credit risk assessment, machine-learning models play a crucial role in analyzing large amounts of data to predict the likelihood of borrowers defaulting on their loans. Various machine-learning techniques are employed to build predictive models that help financial institutions make informed lending decisions.

Those are the most commonly used machine learning algorithms and we're about to use them in this project

- **Logistic Regression:**

Logistic regression is a fundamental classification algorithm used in credit risk analysis. It models the probability of a binary outcome (default or non-default) based on input features. Their interpretability is a key advantage, as the model provides coefficients that indicate the impact of each feature on the outcome.

- **Decision Trees:**

Decision trees are used to partition the data into subsets based on the values of input features. They are intuitive to understand and can capture non-linear relationships in the data.

- **Ensemble methods like Random Forests:**

They combine multiple decision trees to improve predictive performance.

- **Gradient Boosting Algorithms (XGBoost):**

XGBoost is a robust machine-learning algorithm that can help us understand the data and make better decisions by creating a strong predictive model.
They are known for their high predictive accuracy and robustness to noisy data, handling imbalanced datasets well and can capture complex interactions between features.

- **Support Vector Machines (SVM):**

SVM is a powerful classification algorithm that finds a hyperplane that best separates different classes. SVM maps training examples to points in space to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

- **Neural Networks:**

A neural network is a simplified model of the way the human brain processes information. Particularly deep learning models can capture intricate patterns in data.
They are capable of automatic feature extraction and can learn complex non-linear relationships.

# 3. Data Exploration

## 3.1 DATA OVERVIEW

The dataset used in this analysis contains information about various borrowers, including their age, income, loan intent, loan amount, and previous credit history. Additionally, it includes the loan grade, which indicates the level of risk associated with each loan application (ranging from "A" for low risk to "G" for high risk) and many more features.

| feature | description |
|---|---|
| person_age | The person's age in years |
| person_income | The person's annual income. |
| person_home_ownership | The type of home ownership (RENT, OWN, MORTGAGE, OTHER) |
| person_emp_length | the person's employment length in years. |
| loan_intent | the person's intent for the loan (PERSONAL, EDUCATION, MEDICAL, VENTURE, HOMEEMPROVEMENT, DEBTCONSOLIDATION). |
| loan_grade | the of risk on the loan(A,B,C,D,E,F,G)(A-> not riscky \| G-> very riscky |
| loan_amnt | the loan amount. |
| loan_int_rate | the loan interest rate (between 6% and 21%) |
| loan_status | Shows wether the loan is currently in default with 1 being default and 0 being non-default. |
| loan_percent_income | The percentage of person's income dedicated for the mortgage. |
| cb_person_default_on_file | If the person has a default history (YES , NO). |
| cb_person_cred_hist_length | The person's credit history. |

*Figure 4: description of the dataset's features*

In our database the loan grades are represented using a set of letters, 'A,' 'B,' 'C,' 'D,' 'E,' 'F,' and 'G,' where 'A' signifies the lowest credit risk and 'G' indicates the highest credit risk.

**A Grade:** Borrowers with an 'A' grade are typically considered to have excellent credit and pose a low credit risk. They are likely to have high credit scores and a history of responsible credit management.

**B Grade:** 'B' grade borrowers have good credit and are considered to have a lower credit risk compared to average borrowers. They may have a few minor issues in their credit history but overall demonstrate responsible credit behavior.

**C Grade:** Borrowers with a 'C' grade are often considered to have average or fair credit. Their credit risk may be moderate, and they may have a mix of positive and negative credit history.

**D Grade:** 'D' grade borrowers may have below-average credit and may be considered subprime. They may have a history of late payments, higher credit utilization, or other credit issues.

**E Grade:** Borrowers with an 'E' grade are likely to have poor credit and may have significant credit issues or a history of delinquencies.

**F Grade:** 'F' grade borrowers are considered high credit risk and may have a substantial history of late payments, defaults, or other serious credit problems.

**G Grade:** Borrowers with a 'G' grade are often considered the highest credit risk. They may have a history of multiple defaults, bankruptcies, or other severe credit issues.

## 3.2 BASIC INFORMATION

The dataset consists of **32,581 entries** and encompasses **12 columns**, reflecting various attributes related to credit risk assessment as you can see in the figure below. Notably, the **'person_emp_length'** and **'loan_int_rate'** columns exhibit missing values. The data types of the columns vary, with **3 columns** being of **float** type, **5 columns** being of **int** type, and **4 columns** being of **object** type.

When exploring the dataset through summary statistics, we find that:

- The **person_age** column has an **average age** of approximately **27.73 years**, with a **standard deviation** of **6.35 years**.
- The **person_income** column, reflecting the person's income, has a **mean income** of around **66,074$** and a **standard deviation** of **$61,983**.
- The **person_emp_length** column, representing the person's employment length, displays a **mean** value of approximately **4.79 years**.
- The **loan_amnt** column, indicating the loan amount, has an **average loan** value of **9,589.37$**, with a **standard deviation** of **6,322.09$**.
- The **loan_int_rate** column, representing the loan interest rate, has a **mean rate** of around **11.01%**.
- The **loan_percent_income** column, which represents the ratio of loan amount to income, has a **mean** value of **0.17**.
- The **cb_person_cred_hist_length** column, reflecting the person's credit history length, has a **mean** value of approximately **5.80 years.**
-

This table contain more statistics of the data:

| | person_age | person_income | person_emp_length | loan_amnt | loan_int_rate | loan_status | loan_percent_income | cb_person_cred_hist_length |
|---|---|---|---|---|---|---|---|---|
| count | 32581.000000 | 3.258100e+04 | 31686.000000 | 32581.000000 | 29465.000000 | 32581.000000 | 32581.000000 | 32581.000000 |
| mean | 27.734600 | 6.607485e+04 | 4.789686 | 9589.371106 | 11.011695 | 0.218164 | 0.170203 | 5.804211 |
| std | 6.348078 | 6.198312e+04 | 4.142630 | 6322.086646 | 3.240459 | 0.413006 | 0.106782 | 4.055001 |
| min | 20.000000 | 4.000000e+03 | 0.000000 | 500.000000 | 5.420000 | 0.000000 | 0.000000 | 2.000000 |
| 25% | 23.000000 | 3.850000e+04 | 2.000000 | 5000.000000 | 7.900000 | 0.000000 | 0.090000 | 3.000000 |
| 50% | 26.000000 | 5.500000e+04 | 4.000000 | 8000.000000 | 10.990000 | 0.000000 | 0.150000 | 4.000000 |
| 75% | 30.000000 | 7.920000e+04 | 7.000000 | 12200.000000 | 13.470000 | 0.000000 | 0.230000 | 8.000000 |
| max | 144.000000 | 6.000000e+06 | 123.000000 | 35000.000000 | 23.220000 | 1.000000 | 0.830000 | 30.000000 |

*Figure 5: Some statistics about our dataset*

# 4. DATA CLEANING

Data cleaning is a crucial and foundational step in the process of preparing and refining datasets for analysis. It involves a series of systematic procedures aimed at identifying, rectifying, and enhancing the quality of data. Raw data collected from various sources often contains imperfections, inconsistencies, and irregularities that can undermine the accuracy and reliability of subsequent analyses. Data cleaning addresses these issues, ensuring that the dataset is accurate, complete, and ready for meaningful insights and decision-making.

## 4.1 Data cleaning plan:

- **Checking / removing duplicates:** our dataset is full of duplicates, and we know that duplicates can skew analysis results and inflate counts, leading to misinterpretations. So are dealing with them :

```
In [7]: ## Checking for Duplicates
        dups = df.duplicated()
        dups.value_counts()

Out[7]: False    32416
        True       165
        dtype: int64
```

*Figure 6: Dealing with dupliate values*

As we can see we have **165 duplicates** in our dataset, we are going simply to remove them

.

- **Feature selection:** Feature selection involves identifying the most relevant attributes for analysis, streamlining the dataset and focusing efforts on the most impactful variables.

  For our case, we are removing the **'interest rate'** because it is a critical factor in loan repayment, as it directly affects the cost of borrowing for borrowers. However, interest rates are not standardized and can vary significantly from one lender to another, based on various factors such as creditworthiness, market conditions, and the lender's internal policies.

  The goal is to develop models that provide clear and interpretable insights into the factors influencing the likelihood of default. Including a variable like interest rate, which is subject to external and unpredictable factors, could make the model's interpretation more challenging and less actionable.

- **Removing outliers based on data knowledge & observation:** Outliers can introduce noise and disrupt statistical analyses, also they can lead to overfitting, where the model captures noise rather than true patterns making their identification and handling critical.

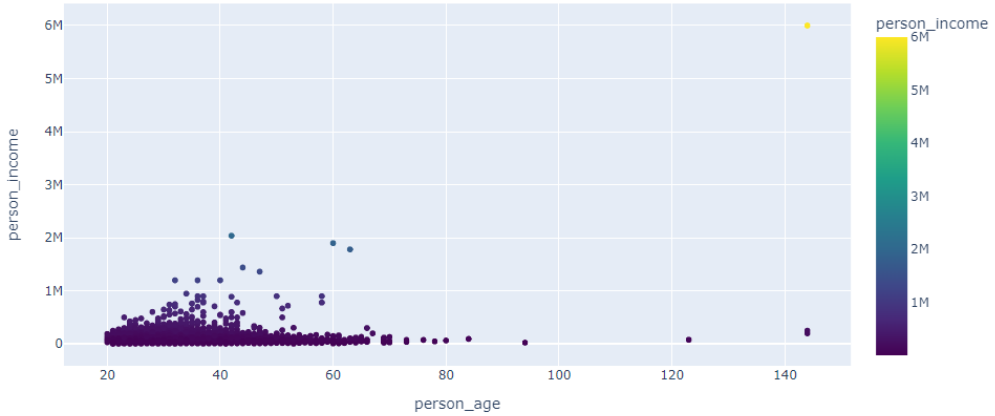


*Figure 7: scatter plot of age with respect to person's income*

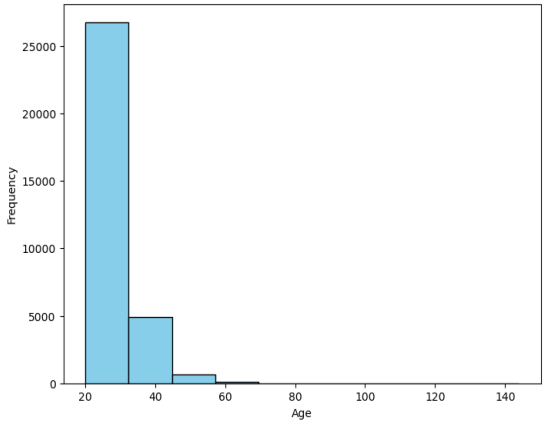


*Figure 8: distribution of person Age feature*

As we can see, we have persons with more than 100 years old in our dataset.

Following the distribution of age in the dataset and logically think that it is impossible to give a loan to someone older than 100 years old

we will consider those clients as outliers and remove them.

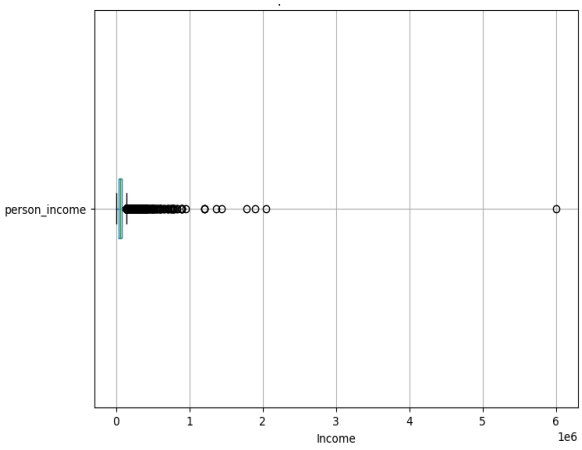


*Figure 9: Box plot of person_income feature*

Relying on this visualization, we can consider persons gaining more than 2M as outliers

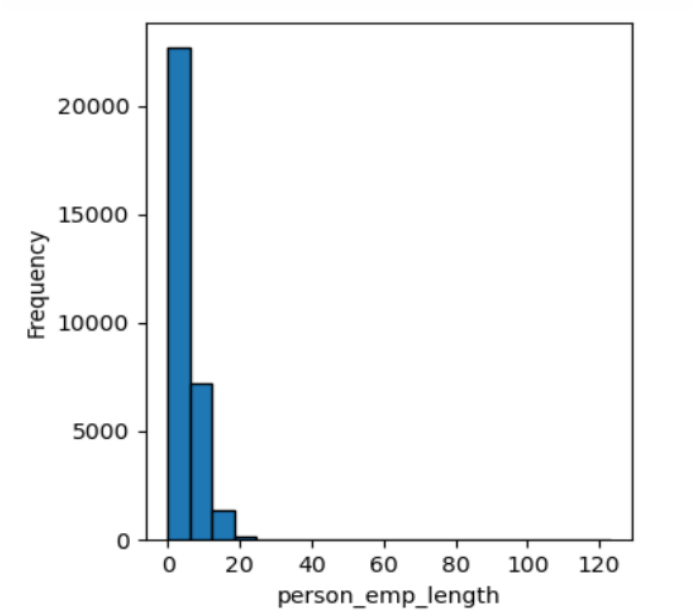


*Figure 10: distribution of person_emp_length feature*

We can also consider all the persons having more than **60 years** of work experience as outliers.

This threshold is assumed based on an upper bound of average employment duration and by considering the distribution of the feature **'person_emp_length'** in our dataset

- **Checking for Missing Data**: missing data can lead to biased conclusions and hinder the effectiveness of predictive models

  Missing data, or missing values, occur when you do not have data stored for certain variables or participants. Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons.

  There are typically three types of missing values:
  1. Missing completely at random (MCAR).
  2. Missing at random (MAR)
  3. Missing not at random (MNAR)

  Missing data are problematic because, depending on the type, they can sometimes cause sampling bias. This means our results may not be generalizable outside of our study because our data come from an unrepresentative sample.

  To deal with missing data we are going to use the **'msno library'** to identify the extent of missing data in our dataset through visualizations like the matrix plot and the bar chart.

  **Missingno** library offers a very nice way to visualize the distribution of NaN values. It is a Python library and compatible with Pandas.

  We will use the Bar chart, which will give us an idea about how many missing values are there in each column.

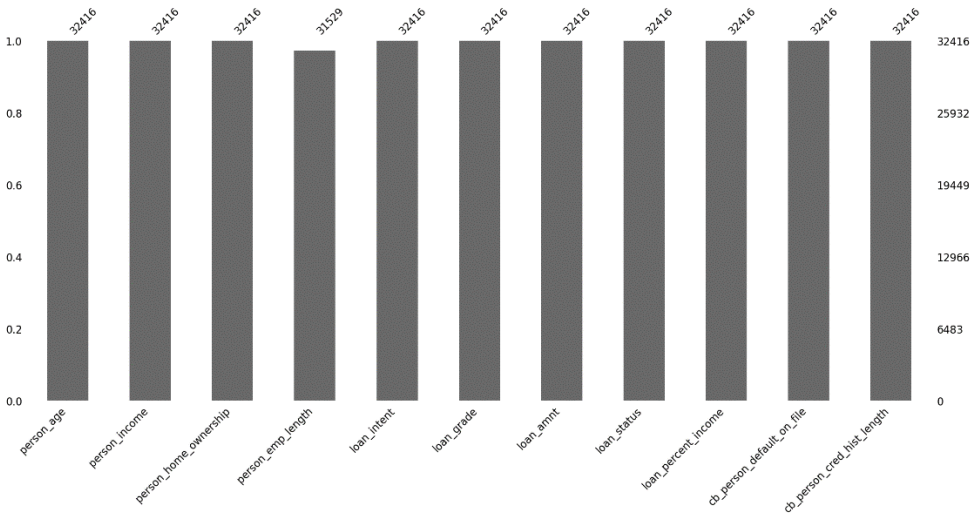


*Figure 11: Detecting the missing values in the features*

- **Dealing with Missing Data**:

  To deal with the missing data in the **'person_emp_length'** feature we are going to use the **'The IterativeImputer'** which is a powerful technique used to handle missing data in a dataset by imputing (replacing) missing values based on a predictive model.

  Unlike basic imputation methods that rely only on summary statistics, the IterativeImputer takes advantage of the relationships between variables in the dataset. It uses the other features in the dataset to predict the missing values, capturing the underlying patterns and correlations present in the data.

  It does so through an iterated round-robin fashion: at each step, a feature column is designated as output y and the other feature columns are treated as inputs X.
  A regressor is fit on (X, y) for known y. Then, the regressor is used to predict the missing values of y. This is done for each feature in an iterative fashion, and then is repeated for max_iter imputation rounds. The results of the final imputation round are returned.

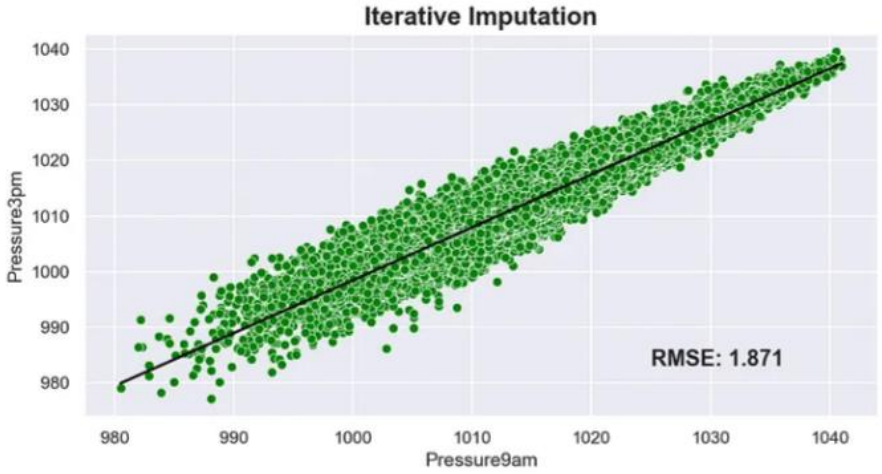


*Figure 12: iterative imputer*
*T.J Kyner « Iterative imputer » towardsdatascience.com (August 09, 2021).*

# 5. DATA VISUALIZATION & EXPLORATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

## 5.1 Analysing Categorical features:

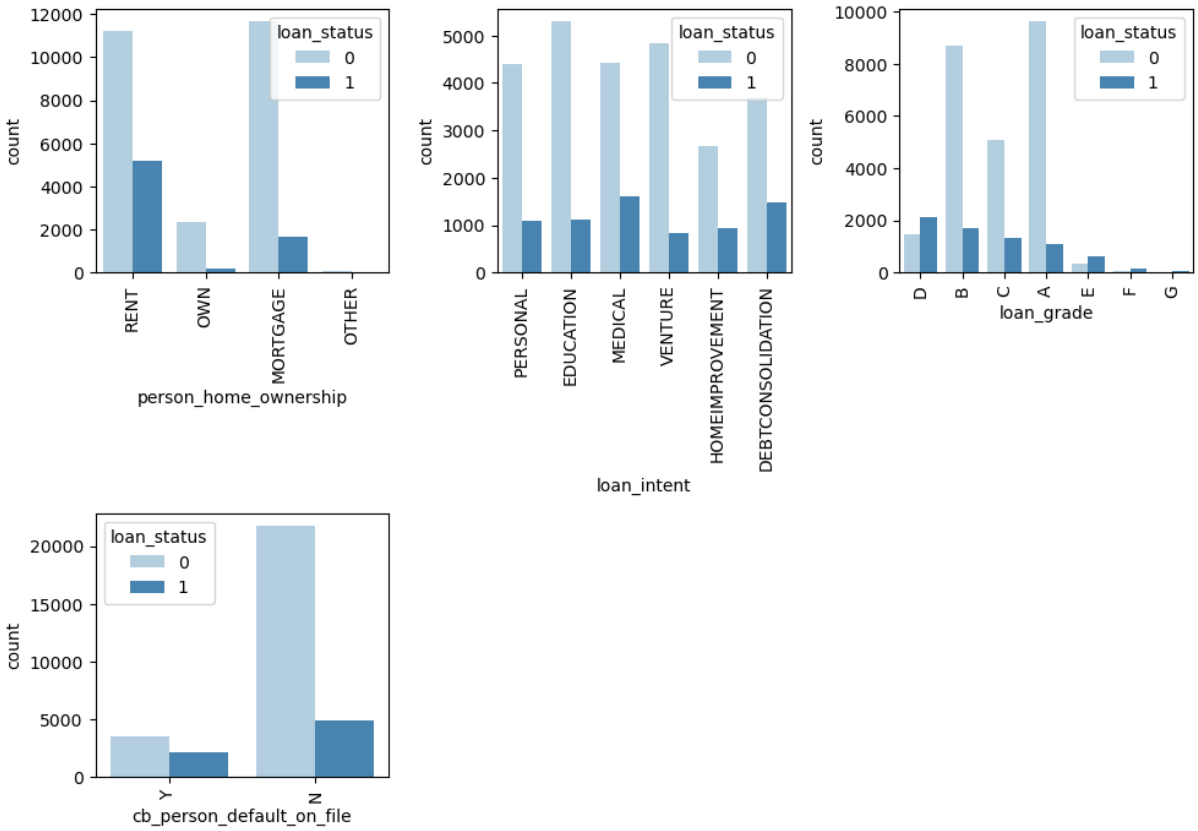


*Figure 13: A plot showing the influence of each categorical feature on loan_status*

In this visualization, we can make these remarks:

1- Borrowers that own their homes are very unlikely to default
2- Borrowers with RENT 'homeownership' are every likely to default (nearly half of the borrowers make defaults.
3- MORRAGE ownership is proportionally the most safe one
4- The loan intent does not affect the loan status very much because they have nearly the same distribution
5- Loan grade 'D' , 'E' , 'F' , 'G' will certainly cause defaults
6- Borrowers who have already made defaults will most likely do it again

All these patterns are going to be learnt by our machine learning models.
Here we are only trying to understand our data and the behavior of borrowers

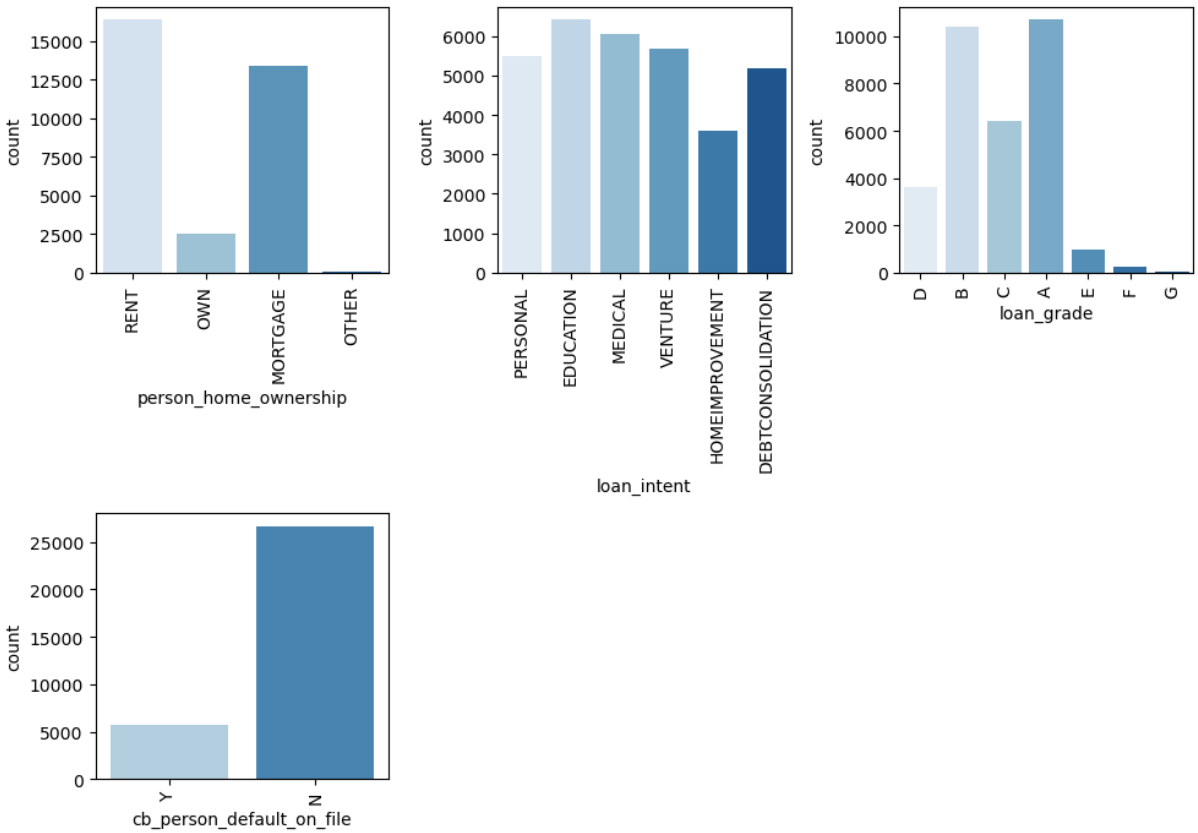


*Figure 14: distribution of all the categorical features*

In this visualization, we can make these remarks:

1- A very few owners are asking for loans
2- The loan-intent HOMEIMPROVEMENT is the least used for requesting a loan
3- The majority of borrowers are classed in 'A', 'B', 'C' grades which mean that they will most likely have an acceptance.
4- Borrowers having already defaults in payment are not too much ask for loan again
5- the people who make the most money are those who have mortgages

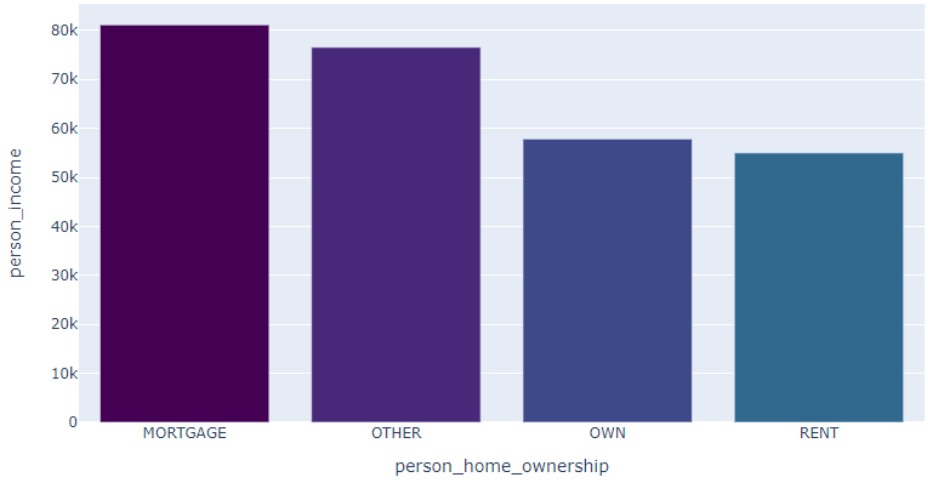


*Figure 15: Mean Person Income BY Homeownership*

## 5.2 Analysing numerical features:

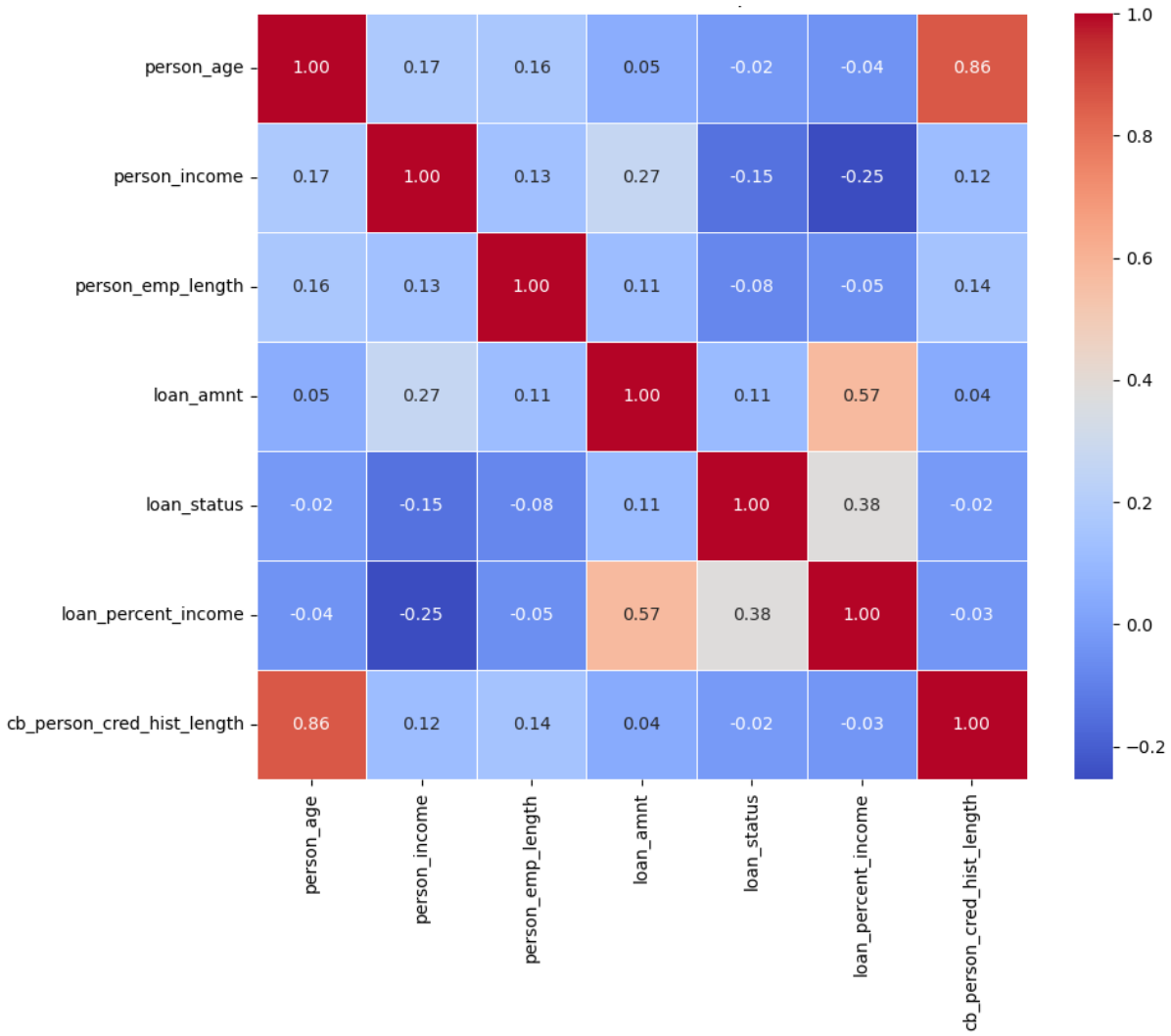


*Figure 16: Correlation Heatmap showing the relationships between numerical features*

**person_age -> cb_person_cred_hist_length:** A strong positive correlation between a person's age and length of credit history may indicate that older people tend to have longer credit histories. This is usually expected because older people have had more time to establish their credit history.

**loan_amount -> loan_percent_income:** The strong positive correlation between the amount of the loan and the percentage of income allocated to the loan suggests that the amounts of loans granted generally increase as the percentage of income allocated to loan repayment increases. This may indicate that lenders give higher loan amounts to those who spend more of their income on repayment.

**loan_amount -> person_income:** The strong positive correlation between the amount of the loan and the person's income indicates that people with higher incomes tend to obtain higher loan amounts. This is usually expected, as higher income may be associated with greater repayment capacity.

**loan_status -> loan_percent_income:** The strong positive correlation between loan status and the percentage of income allocated to the loan suggests that loans with higher income percentages may have higher odds of defaulting.

**loan_status -> loan_int_rate:** The strong positive correlation between loan status and interest rate indicates that loans with higher interest rates may have higher chances of defaulting.

**person_income -> loan_percent_income:** The strong negative correlation between person income and the percentage of income allocated to the loan indicates that people with higher income generally allocate a smaller portion of their income to loan repayment.

## 5.3 Analyzing target feature:

The target feature **loan_status** in the dataset represents the binary outcome of whether a loan was successfully paid off or not. It indicates whether the borrower met their obligation to repay the loan within the specified terms. This feature plays a crucial role in credit risk analysis, as it forms the basis for assessing the performance of loans and evaluating the creditworthiness of borrowers.

In the context of our dataset, the **loan_status** feature typically takes on two values:

- **0:** Indicates that the loan was successfully repaid within the terms, implying a positive outcome.
- **1:** Indicates that the borrower failed to repay the loan within the terms, indicating a negative outcome such as default or non-payment.

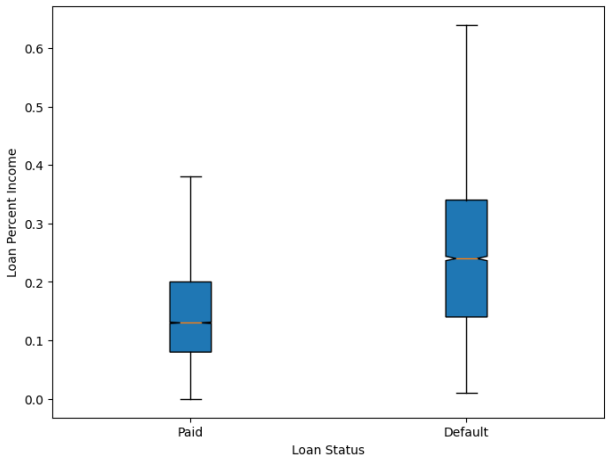


*Figure 17: Box Plot of Loan_status wrt loan_percent_income*

- We notice that unpaid loans tend to take bigger percentage of income

- We can assume that if the percentage of income dedicated to the loan is greater than 40% the borrowers will certainly not pay their loans within the terms

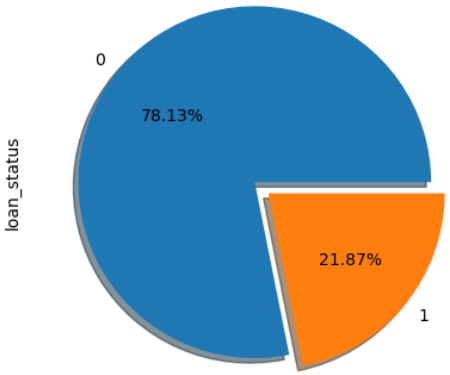


*Figure 18: loan status imbalance*

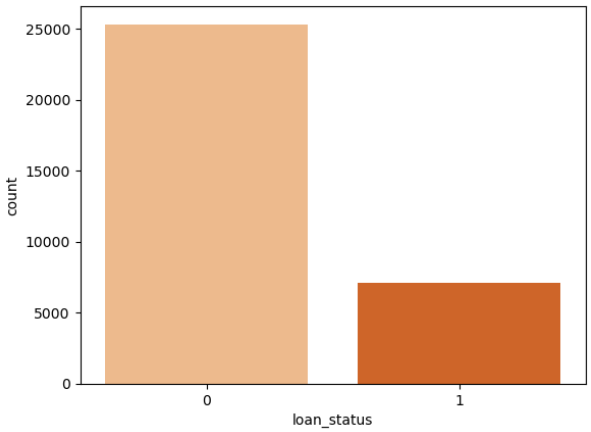


*Figure 19: Distribution of loan_status feature*

The Data is **highly IMBALANCED**. We will deal with oversampling techniques like KNN-SMOTE to solve this issue.

# 6. Data Preprocessing:

Every preprocessing technique is done only on the train-set. So splitting is mandatory before outlier removal, missing values handling, oversampling, scaling …etc. Therefore, we use the train_test_spit function of scikitlearn to spit our data to 80%training & 20% test

## 6.1 Creating the main pipeline

Using a pipeline of preprocessing is very important and have many advantages. Pipelines provide an organized workflow, making it easy to reproduce preprocessing steps consistently across different datasets or during model retraining pipeline is to assemble several steps that can be cross-validated together while setting different parameters which reduces the chances of errors and ensures uniformity.
When combining pipelines with techniques like cross-validation and hyper-parameter tuning, we ensure that preprocessing steps are consistently applied during each fold of cross-validation, leading to more reliable model performance estimates. The most important and the reason why we are using pipelines of preprocessing in this project is the Ease of Deployment because using a pipeline simplifies the process of deploying machine learning models into production. The same preprocessing steps applied during training can be easily replicated when new data arrives for prediction.

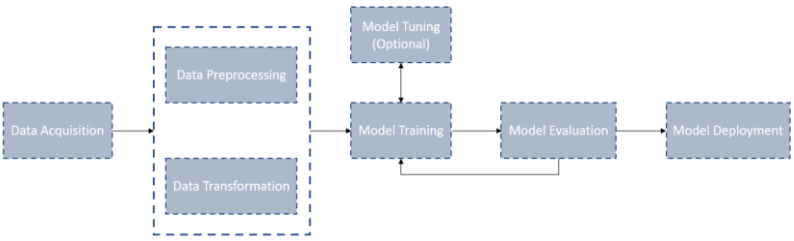


*Figure 20: How piplines work*
*David Hurley « How to use Pipelines to standardize data preprocessing » towardsdatascience.com (July 02, 2020).*

**The Main Pipeline will be made of two parts:**

- **Preprocessing for NUMERICAL VARIABLES:**

1- Iterative imputer - To handle missing values
We have already explained how we dealt with missing data and explain how the iterative imputer works

2- Scaling - To maintain the scale among features.
Here we scaled our data because many machine-learning algorithms, such as gradient-based optimization methods used in linear regression or neural networks converge faster and more reliably when features are on a similar scale.
Scaling improves the interpretability of coefficients and model outputs. In addition, prevents certain features from dominating the optimization process due to their larger scale, which can lead to more stable and efficient training.

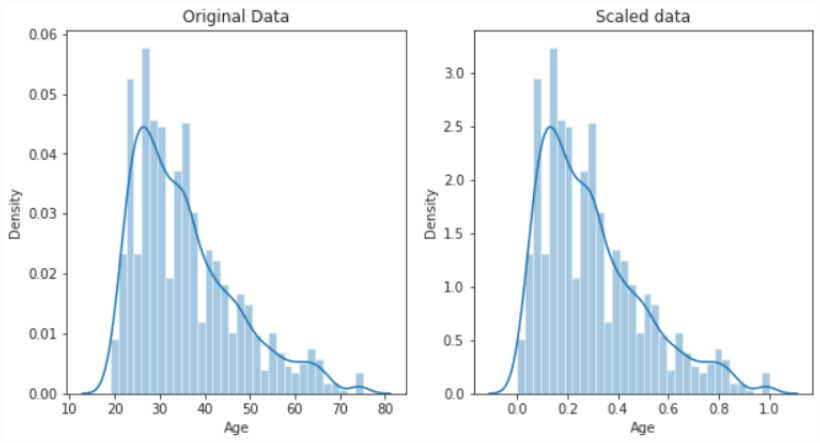


Figure 21: What do the scaling process do to our data
*Lawrence Alaso Krukrubo* « Scaling vs. Normalizing Data » towardsai.net (January 10, 2021).

- **Preprocessing for CATEGORICAL VARIABLES:**
1- One Hot Encoder - To encode each category for model interpretability

One-hot encoding is a process by which categorical data are converted into numerical features of a dataset, where each category is represented as a separate binary column. This is a required preprocessing step since machine-learning models require numerical data.
The reason why we used the One-hot-encoding technique instead of other techniques like “label encoding” for example is to Avoid Ordinal Bias. Some machine learning algorithms may incorrectly assume an ordinal relationship between categories if the categorical variable is encoded numerically (**e.g** using label encoding). One Hot Encoding eliminates this potential bias, as no numerical order is imposed on the categories.

## 6.2 Handling Data Imbalance:

Data imbalance occurs when the distribution of classes in a dataset is highly skewed, meaning that one class has significantly more instances than another class. This imbalance can lead to several problems like:

**Biased Model Training:** Imbalanced data can bias the model's learning process toward the majority class, making it less sensitive to the minority class. As a result, the model may have lower accuracy, precision, and recall for the minority class.

**Poor Generalization:** Models trained on imbalanced data may have difficulty generalizing to new, unseen data, especially when the minority class is of interest. The model may make biased predictions in favor of the majority class.

**Loss of Information:** Imbalanced data can lead to the loss of valuable information from the minority class. The model may not capture important patterns or characteristics of the minority class due to the scarcity of data.

**Evaluation Metrics Bias:** Traditional evaluation metrics like accuracy can be misleading in the presence of imbalanced data. A high accuracy score may not necessarily indicate a good model, as it could be driven by the majority class predictions.

One way of solving this issue is to under-sample the majority class. That is to say, we would exclude rows corresponding to the majority class such that there are roughly the same amount of rows for both the majority and minority classes. However, in doing so, we miss many data that could be used to train our model thus improving its accuracy (e.g. higher bias). Another option is to over-sample the minority class. In other words, we randomly duplicate observations of the minority class. The problem with this approach is that it leads to overfitting because the model learns from the same examples. Therefore, the solution is to use an over-sampling technique well implemented to deal with data imbalance without causing high bias or overfitting.

**SMOTE (Synthetic Minority Over-sampling Technique)** is an oversampling technique designed to address the issues caused by data imbalance. It generates synthetic samples for the minority class by creating new instances that are similar to existing minority class instances.

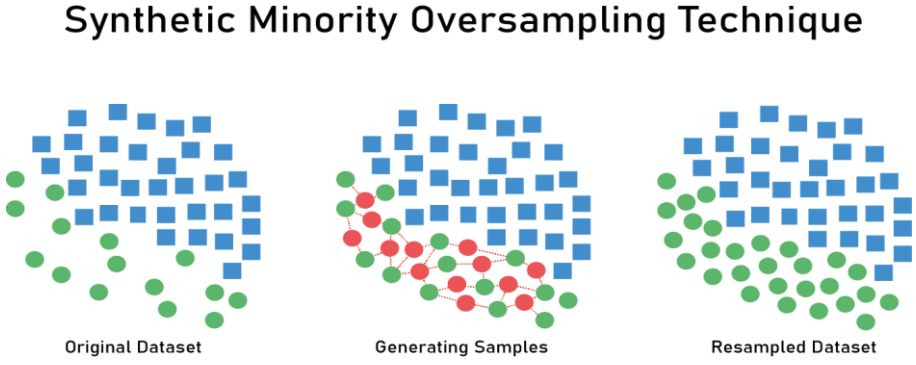


*Figure 22: SMOTE technique*
*Emilia Orellana « SMOTE » medium.com (December 10, 2020).*

The SMOTE algorithm can be described as follows:

- Take difference between a sample and its nearest neighbor
- Multiply the difference by a random number between 0 and 1
- Add this difference to the sample to generate a new synthetic example in feature space
- Continue on with next nearest neighbor up to user-defined number

SMOTE effectively increases the representation of the minority class by generating synthetic samples. This helps the model better learn the patterns and characteristics of the minority class which will allow us to achieve higher accuracy, precision, recall, and F1-score for both classes and avoid overfitting.

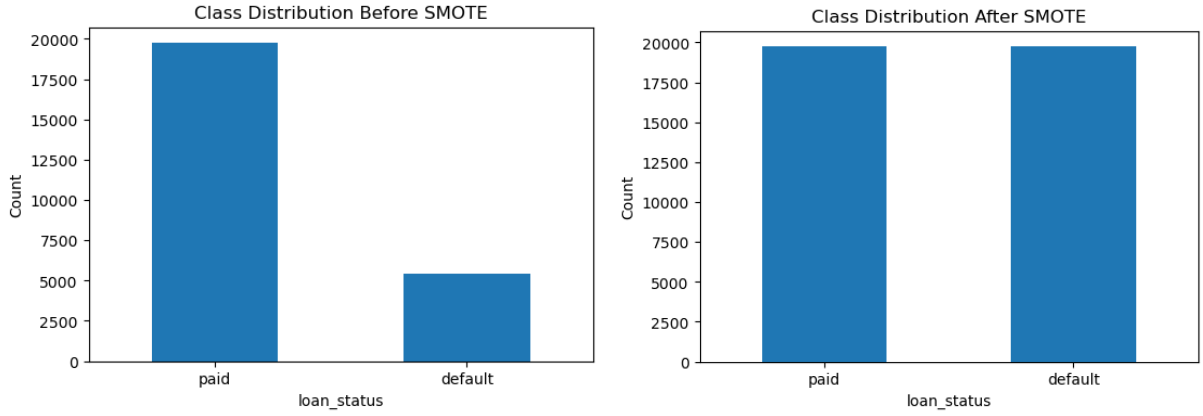


*Figure 23: the effect of SMOTE on our data*

Now that our dataset is balanced, we can proceed with training machine-learning models to predict credit risk. This step involves selecting appropriate algorithms, training the models, and evaluating their performance.

# 7. Model Selection and Hyper-parameter Tuning:

## 7.1 Model selection:

After an extensive review of similar projects and a thorough investigation into credit risk analysis, it became evident that certain machine learning models were somehow the preferred choices. These models, namely **Logistic Regression, Random Forests, Gradient Boosting Algorithms (such as XGBoost), Support Vector Machines (SVM), and Neural Networks**, have emerged as staples in credit risk assessment tasks.

One common thread among these models is their ability to effectively handle complex and multi-dimensional data inherent to credit risk scenarios. Their versatility in capturing both linear and non-linear relationships ensures the exploration of intricate patterns that may influence credit risk outcomes. Moreover, they give a very clear insight and adaptability, which make them promising candidates for credit risk analysis.

By using these models, we can dig deep into the data and discover hidden patterns, leading to predictions that are more accurate. Their collective strengths lie in their capability to handle feature-rich datasets, identify crucial risk factors, and offer a balance between interpretability and predictive performance.

As a result, these chosen models serve as a strong foundation for our credit risk analysis project.

## 7.2 Hyper-parameter tuning:

After completing the data cleaning and preparation process, the next step was to identify suitable machine learning models for the credit risk analysis task. Initial attempts to train these models on the prepared data resulted in suboptimal performance, with low accuracy scores on both the training and test sets. In order to enhance the model's predictive capabilities, an extensive hyper-parameter tuning process was conducted using **GridSearchCV**.

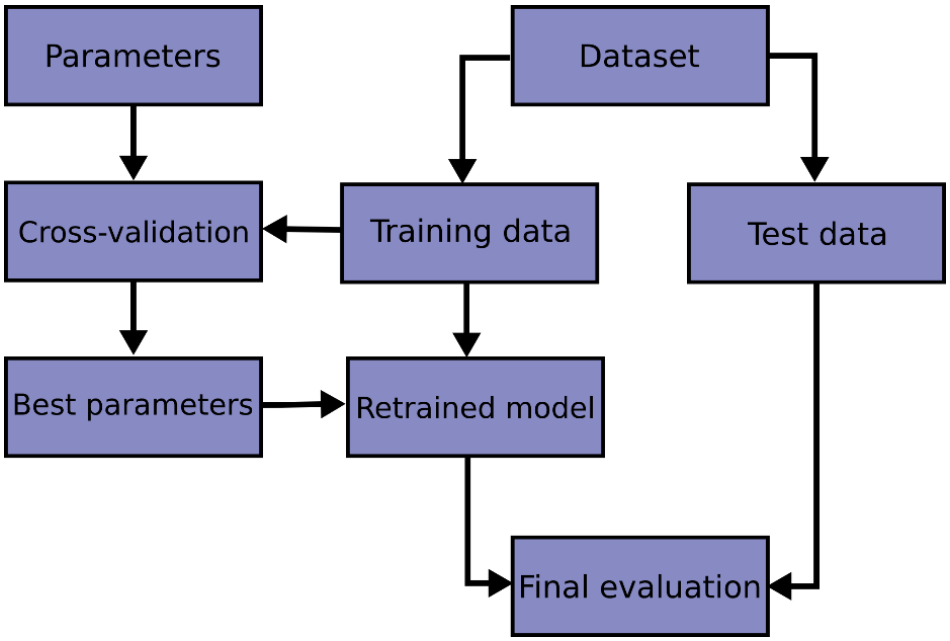


*Figure 24: Hyper-parameter Tuning process steps*
*Ojala and Garriga« Cross-validation: evaluating estimator performance» scikit-learn.org .*

This approach systematically explored various combinations of hyper-parameters, optimizing the model's configuration.

As a result of the hyper-parameter tuning, a significant improvement in performance was achieved. The final model exhibited a remarkable 95% accuracy on the training data and a robust 88% accuracy on the previously unseen test data. This outcome was deemed satisfactory and provided a solid foundation for proceeding to the deployment phase of the project.

# 8. Model Evaluation:

The model evaluation permits us to evaluate the performance of a model, and compare different models, to choose the best one to send into production. Its role is to systematically analyze how well the trained models generalize to new, unseen data and how accurately they fulfill the project's objectives.

During this phase, various evaluation metrics are used to measure the models' performance, such as accuracy, precision, recall, F1 score, and others that align with the specific goals of the project.

Using these metrics, we will be able to compare our models and evaluate them to identify which one performs the best It helps in avoiding overfitting (model performs well on training data but poorly on new data) and underfitting (model is too simplistic to capture patterns).

The model evaluation phase and hyperparameter-tuning phase work in tandem to iteratively assess and enhance the performance of the models. Throughout this iterative process, models are trained, evaluated, and fine-tuned multiple times to ensure that the final chosen model is optimized for the given task. This iterative approach allows for a comprehensive understanding of each model's behavior and its potential to achieve accurate predictions.

## 8.1 Metrics used to evaluate model performance:

- **Cross-Validation Score (Cross-Val Score):**

Cross-validation is a robust technique used to assess model performance by partitioning the dataset into multiple subsets (folds) for training and validation. It helps mitigate potential bias and provides a more accurate estimate of a model's generalization ability. The Cross-Validation Score quantifies how well a model performs across different subsets of the data. By averaging the scores from each fold, we obtain a more reliable measure of the model's overall performance.

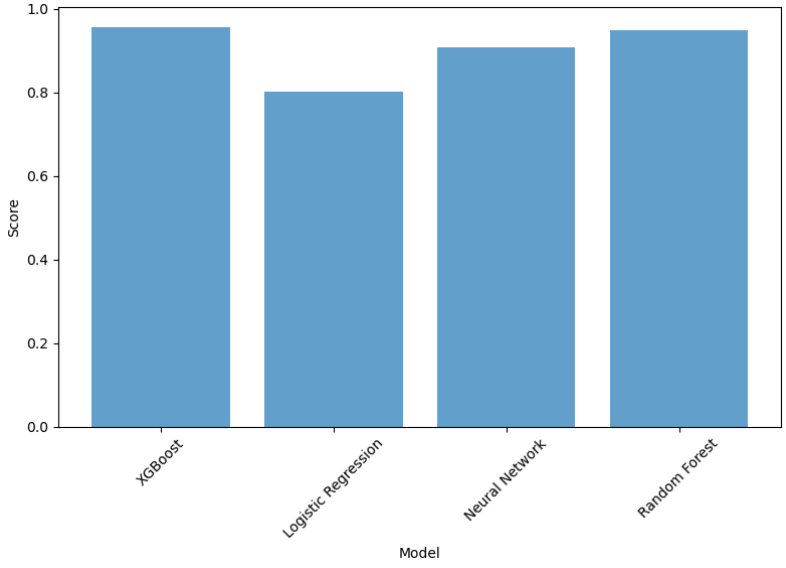


*Figure 25: Model evaluation Metric: Cross-Val Score*

As we can see our models are making a very good performance on our data, especially for XGBOOST and Random Forest, we will see with other metrics to have a better idea about the behavior of our models on new data.

- **Accuracy:**

Accuracy is a fundamental metric that measures the proportion of correct predictions made by the model. It is calculated as the ratio of correctly predicted instances to the total number of instances. While easy to interpret, accuracy can be misleading in imbalanced datasets where the classes are not equally represented.

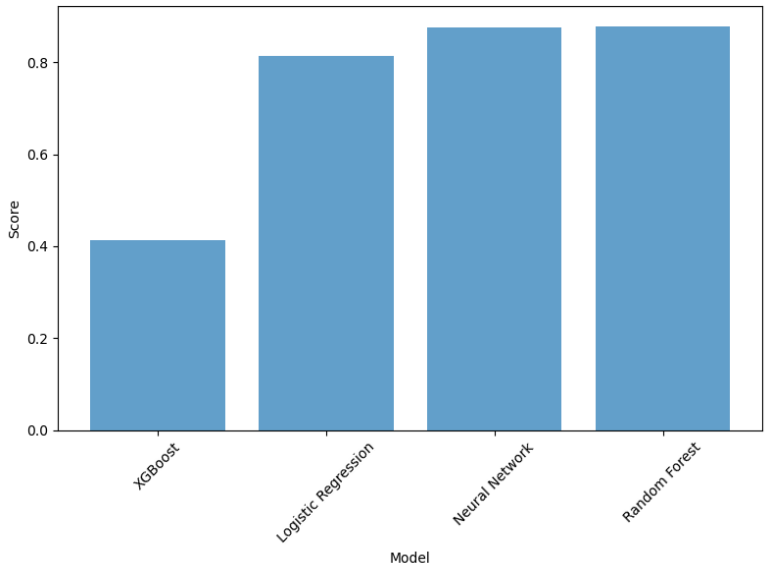


*Figure 26: Model Evaluation Metric: Accuracy*

The accuracy of **XGBOOST** is bad and that is mean that the model had **overfit** the training data and he is not able to generalize on unseen data.
For the other models and especially **Random Forest**, they are making a very good score.

- **F1 Score:**

The F1 Score is a balanced metric that considers both precision and recall. Precision represents the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive instances. It harmonizes these two metrics, providing a robust evaluation of the model's ability to correctly classify positive instances while minimizing false positives and false negatives.

The f1-score is the best metric for our case because it is very powerful and meaningful when dealing with imbalanced datasets.

We can see that Random Forest Neural Network and Logistic regression are scoring a very high F1-score, which indicates that the models achieve a good balance between precision and recall. Which mean that the model is effective at correctly identifying positive instances (high recall) while also minimizing the rate of false positives (high precision).

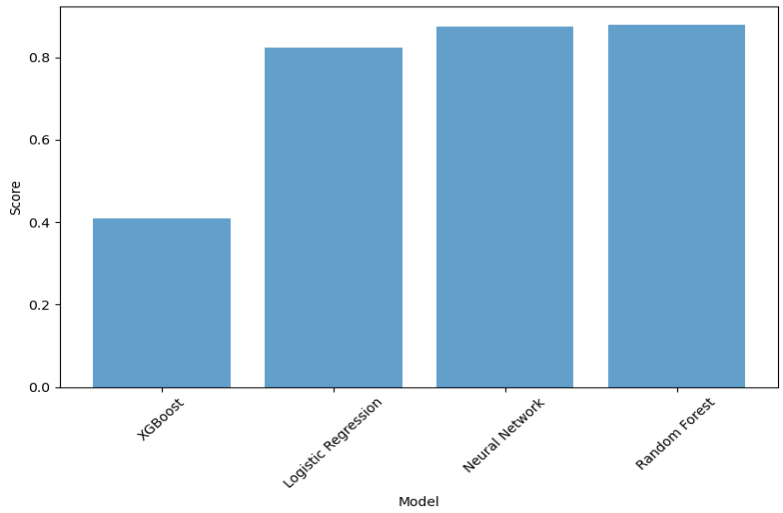


*Figure 27: Model Evaluation Metric: F1 Score*

For the XGBOOST model, it is scoring a very low F1 score, which mean that the model is struggling to strike a balance between precision and recall. This could be due to a higher rate of false positives, false negatives, or both. I suppose that the XGBOOST model is dealing well with the synthetically generated samples that we generated using SMOTE oversampling technique.

- **Mean Squared Root Error (MSRE):**

RMSE metric measures the average squared difference between the predicted and actual values. RMSE assess the model's predictive accuracy by quantifying the magnitude of prediction errors. Lower RMSE values indicate better model performance in minimizing prediction errors.

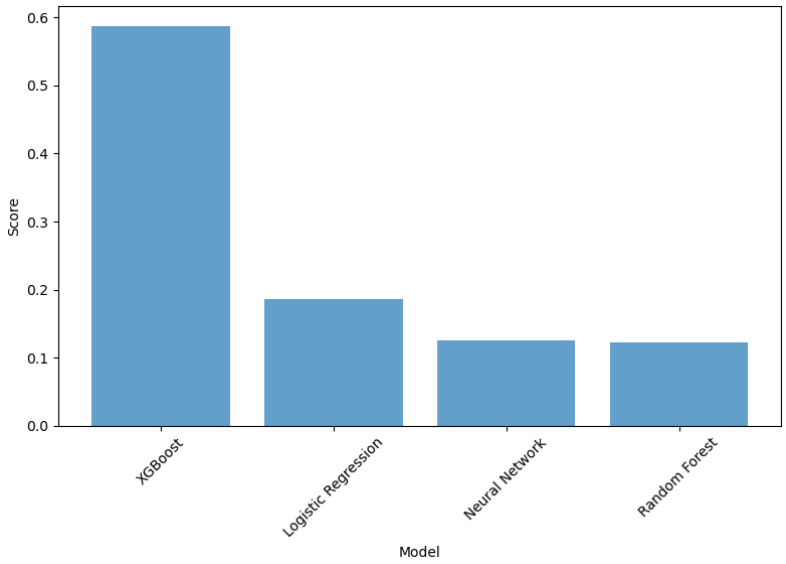


*Figure 28: Model Evaluation Metric: MSRE*

As we can see the XGBOOST unlike all of the other models make many mistakes and cannot minimize the prediction error, as it should.

- **The confusion matrix:**

To take a deeper look on our model performance and have a very clear idea about their weaknesses we are going to see the confusion matrix of each model.
The confusion matrix is a fundamental tool in model evaluation for classification problems. It provides a visual representation of the performance of a classification model by summarizing the predicted classes against the actual classes in a tabular format.

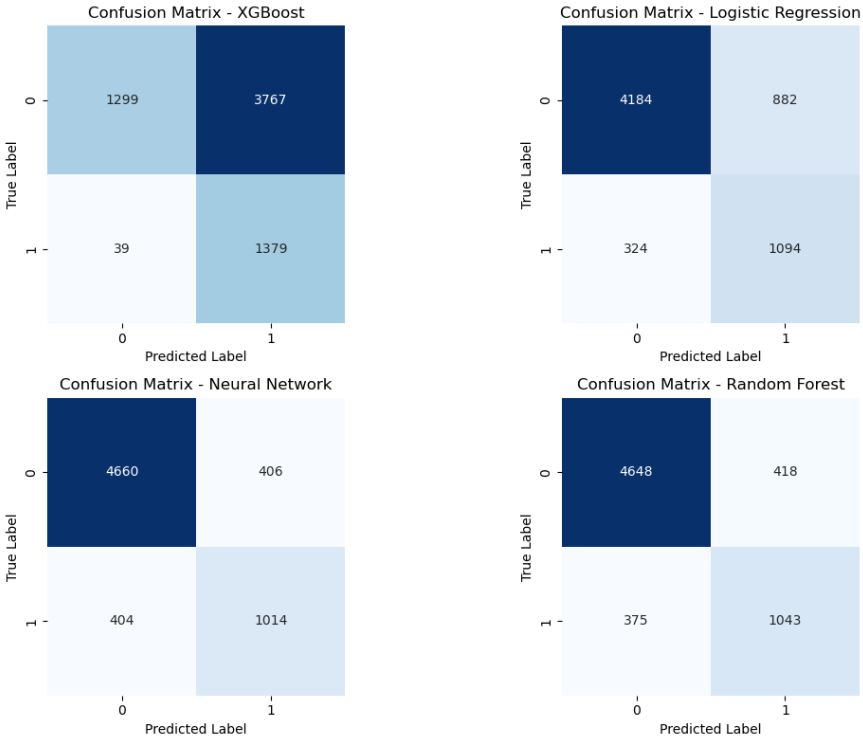


*Figure 29: confusion matrix of our trained models performences*

Each cell in the confusion matrix represents a specific outcome:

**True Positive (TP):** The model correctly predicted a positive class instance as positive.

**False Positive (FP):** The model incorrectly predicted a negative class instance as positive.

**True Negative (TN):** The model correctly predicted a negative class instance as negative.

**False Negative (FN):** The model incorrectly predicted a positive class instance as negative.

| | Correct classifications | Wrong classifications |
|---|---|---|
| **XGBOOST** | **2678** | **3806** |
| **Random Forest** | **5691** | **793** |
| **Logistic regression** | **5278** | **1206** |
| **Neural Network** | **5674** | **810** |

Our decision is clearly made, with the selection of the Random Forest model as our preferred choice. However, we are committed to further enhancing our model's performance by investigating its learning curve. Through this analysis, we aim to determine whether augmenting our dataset with additional data could potentially lead to improvements in the model's predictive capabilities. This approach aligns with our pursuit of optimizing the model's performance to ensure its robustness and accuracy in credit risk analysis.

- **Learning Curve:**

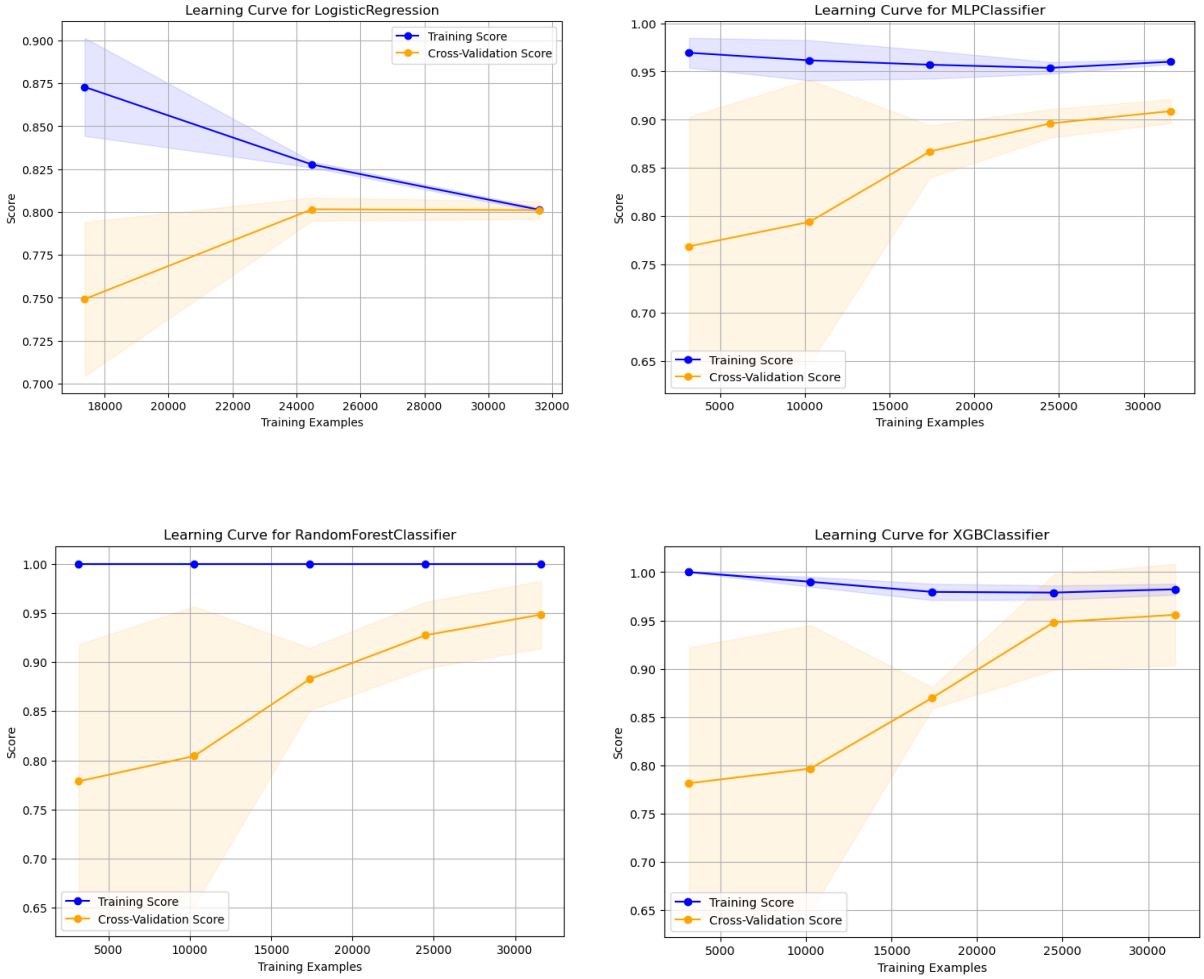


*Figure 30: Learning Curves of our trained models*

In these plots and except for the Logistic Regression one. The training score remains high regardless of the size of the training set. On the other hand, the test score, which measures the model's performance on new, unseen data, demonstrates a clear upward trajectory as the training dataset size increases. Indeed, this improvement continues without showing signs of leveling off.

Observing such situation is an indication that it might be useful to acquire new data to train the model because the increasing test score signifies that the model's generalization performance has room for enhancement. Leveraging a larger and more diverse training dataset could likely lead to further improvements in the model's ability to make accurate predictions on unseen data.

# 9. Deployment of the web application using Streamlit:

## 9.1 What is Streamlit?


Streamlit is an open-source Python library used for building and deploying data science and machine learning web applications. It provides a simple and intuitive way to create interactive dashboards and web apps for data visualization, machine learning, and other data-related tasks.

## 9.2 Why Streamlit?

Streamlit simplifies the process of creating web applications that display data analysis, visualization, and machine learning models.
With Streamlit, we can create custom web applications with minimal coding effort. It provides an easy-to-use syntax for creating interactive widgets, charts, and other visualizations that can be easily deployed and shared with others.

## 9.3 THE WEB APP:

The development and deployment of the web application using Streamlit marked a pivotal phase in our project, as it brought our credit risk analysis to life, making it accessible and interactive for end-users.

During the development process, we leveraged Streamlit's intuitive APIs to create a visually appealing and user-friendly interface. We incorporated various widgets and components provided by Streamlit to create input forms for users to input their personal and financial details. These widgets, such as sliders, text inputs, and select boxes, enabled users to provide essential information necessary for the credit risk assessment. As part of the development, we also embedded explanatory text and headers to enhance user understanding and engagement.

The integration of our pre-trained credit risk model into the Streamlit application was a seamless process. We saved the pipeline of pre-processing and the best model trained which is

the **'Random Forest'**, after that we used Streamlit's Python integration capabilities to load the saved model and preprocess user inputs before making predictions.

To apply the necessary transformations to user inputs and pass them through the model for prediction. We had to pass them by the pre-saved pipeline of preprocessing that we used to pre-process our original data, to have the same kind of data for the model to interpret.

We saved our best model and the preprocessing pipeline thanks to the **'pickle'** module of python.

In summary, the development and deployment of the web application using Streamlit provided an impactful way to deliver the fruits of our credit risk analysis to end-users.

By harnessing Streamlit's capabilities, we successfully created an engaging and interactive platform that empowers users to make informed financial decisions based on our trained credit risk model.

## 9.4 Illustration of the web app:

Those are some screenshots taken from our web-app:

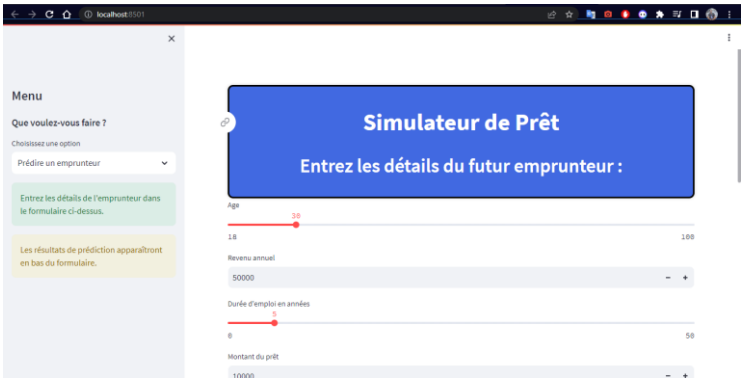


*Figure 31: The HOME page of the web app*

Here user can put his financial and personal information to get treated


*Figure 32: Coryright Page*

There is no specific functionality on this page

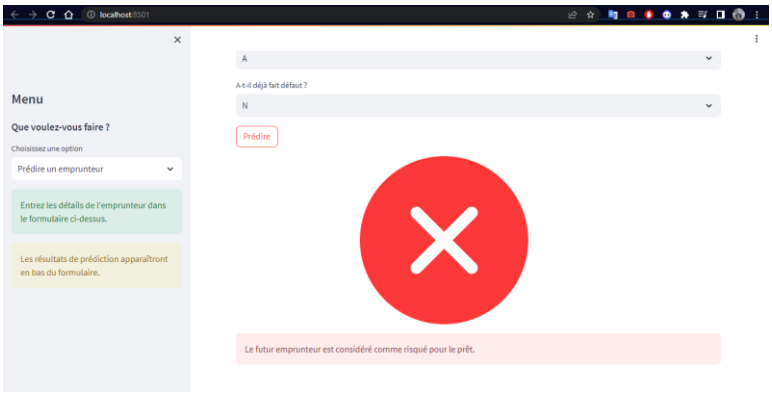


*Figure 33: The disapproved case*

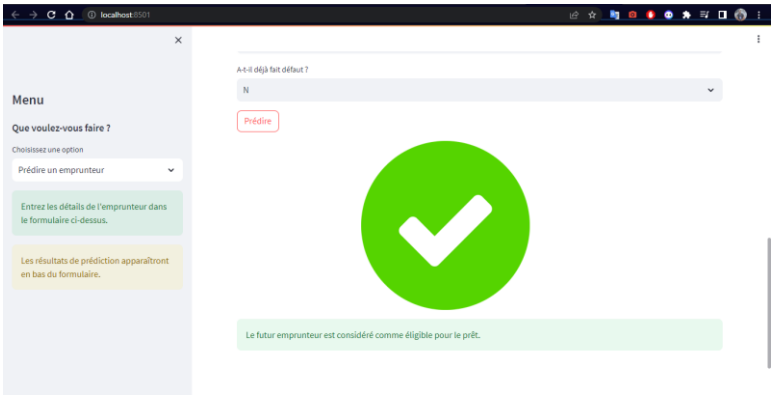


*Figure 34: The approved case*

# 10. Conclusion:

In conclusion, in this research we discovered the world of credit risk analysis, presenting a long learning journey through various stages of data preprocessing, model development, and evaluation. By using a range of machine learning techniques and methods, this study has demonstrated the potential of predictive models in assessing creditworthiness accurately.

In this project, we have been exposed to many concepts like:

--> **Building a pipeline**
--> **Hyperparameter tuning**
--> **Evaluating models**
--> **Building our first Streamlit application**
--> **Deploying it.**

The chosen ensemble of models, including Logistic Regression, Random Forests, XGBoost, and Neural Networks, was grounded in their practicality and effectiveness in credit risk analysis. These models, with their diverse strengths, were carefully evaluated and fine-tuned to achieve optimal performance.

The web application deployment marked a significant step towards the practical application of the developed models, making credit risk assessment more accessible and user-friendly. By bridging the gap between theoretical insights and real-world usability, this project highlights the potential of data science in addressing complex financial challenges.

In conclusion, this credit risk analysis project demonstrates the power of data science and machine learning in the financial industry.
This web app can serve as a valuable tool for financial institutions to assess credit risk, make informed lending decisions, and mitigate potential losses.

# 11. Future work:

However, as with any data science project, there are a few points to keep in mind:

**Model Robustness:** Although the achieved accuracy is excellent, it is essential to test the model's robustness on a wider range of scenarios and data distributions.
In addition, as we did see in the learning curves observations, we can add more data to improve the accuracy of our models. Also we can do some feature importance to understand better the features that count for the learning process.

**Continuous monitoring:** Credit risk is a dynamic domain, and models need regular updates to adapt to changing economic conditions and borrower behaviors.

**Model Deployment:** Deploying a machine-learning model in production involves careful considerations, such as scalability, security, and version control.

**Ethical Considerations:** Credit risk models must be fair and unbiased. Continuously monitor for any potential bias and ensure fairness in lending decisions.

# 12. BIBLIOGRAPHY:


Kyle, Peterdy «pestel analysis. » corporatefinanceinstitute.com « Link » (May 9, 2022).

Gabriel Lip «Credit Risk Anlysis. » corporatefinanceinstitute.com « Link » (October 09, 2019)

Kyle, Peterdy «Credit analysis. » corporatefinanceinstitute.com « Link » (January 09, 2020).

Allianz trade «How to improve your credit risk analysis process» « Link » (December 11, 2019)

ADAM, HAYES « What is a Financial Institution? » investopedia.com« Link » (May 25, 2023).

finhealthnetwork « What is Financial Health? » « Link » (April 28, 2022)

Bill Fay« What is a Credit Score & How is it Calculated? » debt.org «Link» (December 16, 2021).

Julie Sherrier « What Is a FICO Score? » lendingtree.com « Link » (December 30, 2022).

ibm « What is logistic regression? » ibm.com « Link »

ibm « What is Random Forest? » ibm.com « Link »

Simplilearn « What is XGBoost? » simplilearn.com « Link » (June 02, 2023)

ibm « The Neural Networks Model » ibm.com « Link »

wikipedia « Support vector machine» wikipedia.org « Link »

SujanDutta « Visualize missing values (NaN) values » debt.org « Link » (July 03, 2019).

scikit-learn.« Imputation of missing values» scikit-learn.org « Link »

tableau « What is Data Visualization ? » Tableau.com « Link »

Scikit-learn « Preprocessing » scikit-learn.org « Link »

Scikit-learn « Pipeline » scikit-learn.org « Link »

Scikit-learn « Importance of Feature Scaling» scikit-learn.org « Link »

datagy « One-hot-encoding» datagy.io « Link » (February 02, 2022).

Cory maklin «Synthetic Minority Over-sampling Technique» medium.com «Link»(May 14, 2022)

Ching (chingis) «Indtroduction to SMOTE» towardsdatascience.com «Link» (February 06, 2021)

Rahul Shah « Tune Hyperparameters with GridSearchCV» analyticsvidhya.com «Link» (June 23, 2021)

Scikit-learn « GridSearchCV » scikit-learn.org « Link »

Angelica Do Duca « Model Evaluation in Scikit-learn » towardsdatascience.com «Link» (May 17, 2022)

Jean-Christophe chouinard «Model Evaluation» jcchouinard.com «Link» (July 23, 2023)

Jason Brownlee « How to use Learning Curves to Diagnose Machine Learning Model Performance» machinelearningmastery.com «Link» (July 23, 2023)

Andy McDonald « Getting started with Streamlit» towardsdatascience.com «Link» (May 23, 2022)

# 13. ANNEXE

## DEFINITIONS:

CREDITWORTHINESS: The extent to which a person or company is considered suitable to receive financial credit, often based on their reliability in paying money back in the past.

Lender: A lender is an individual, a group (public or private), or a financial institution that makes funds available to a person or business with the expectation that the funds will be repaid

Borrower: A borrower is a person or business that receives money from a lender with the agreement to pay it back within a specified period.

Financial institution: A financial institution (FI) is a company engaged in the business of dealing with financial and monetary transactions such as deposits, loans, investments, and currency exchange.

Financial health: financial health is a composite measurement of an individual's financial life. Unlike narrow metrics such as credit scores, financial health assesses whether people are spending, saving, borrowing, and planning in ways that will enable them to be resilient and pursue opportunities.

Loan terms: Loan terms refer to the terms and conditions involved when borrowing money. This can include the loan's repayment period, the interest rate and fees associated with the loan, penalty fees borrowers might be charged, and any other special conditions that may apply.

Interest rate: The interest rate is the amount a lender charges a borrower and is a percentage of the principal—the amount loaned. The interest rate on a loan is typically noted on an annual basis known as the annual percentage rate (APR).

## Important links:

**Link to the dataset:** https://www.kaggle.com/datasets/laotse/credit-risk-dataset

**Link to the GitHub repository:** https://github.com/islem711/credit_risk_analysis