

ECOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION



Final Project Report



INTEGRATION OF CLIMATE RISK IN INTERNAL CREDIT MODELS (PD)

Authors :

Mamadou DIALLO
Pape Tidiane GUEYE
William SAMBA
Habib TOURE

Supervisors:

Jordan BELOUCHAT
Asmynour YOUSOUUF

2022 - 2023

CONTENTS

1 Literature review	3
1.1 Physical risk	3
1.1.1 Relevant indicators for measuring physical risk	3
1.1.2 Integration of physical risk factors in financial risk assessment	4
1.2 Transition risk	5
1.2.1 Relevant indicators for measuring transition risk	5
1.2.2 Integration of transition risk factors in financial risk assessment	5
2 Data presentation and processing	7
2.1 Credit related data	7
2.1.1 Creation of the target variable	7
2.1.2 Variables preselection: first steps of selection	8
2.1.3 Database sampling	8
2.1.4 Variables selection	8
2.1.5 Handling of missing values	9
2.2 Climate related data	10
2.3 Discretization of quantitative variables and grouping of modalities for categorical variables	11
3 Description of the portfolio	14
3.1 Default rate of the portfolio	14
3.2 Triggers of counter-party default	16
4 Modeling framework	19
4.1 Modeling of the default using logistic regression	19
4.1.1 Theoretical background of Logistic Regression	19
4.1.2 Model estimation	21
4.1.3 Interpretations	23
4.1.4 Model performances	24

4.2	Projection of the default rate under 3 scenarios of temperature rise	25
4.3	Challenger models	26
4.3.1	Random Forest: "wisdom of crowds" (Chloé-Agathe Azencott, 2018)	26
4.3.2	XGBoost: (Tianqi Chen et al., 2016)	26
4.3.3	Neural network	27
4.3.4	Results of the machine learning models	28
4.3.5	Explanation with Shapley values	29
4.4	Correlation coefficient	IV
4.5	K-nearest neighbor algorithm (KNN)	IV
4.6	Correlation to the target variable	V
4.7	Descriptive charts	VI
4.7.1	Grade: univariate distribution	VI
4.7.2	Late fees received to date	VII
4.7.3	Fico score	VII
4.7.4	Loan term: number of payments of the loan	VIII
4.7.5	Monthly debt ratio	VIII
4.7.6	Average revenu per capita	IX
4.7.7	Number of trades opened in past 24 months	IX
4.7.8	Region	X
4.8	Perfomances on machine learning models without climate related variables	X

LIST OF TABLES

2.1	Name and label of variables	9
2.2	Name and label of climate variables	11
2.3	List of categorical variables after discretization	12
3.1	Distribution of the portfolio according to the default	14
4.1	Estimation with credit related variables only	22
4.2	Estimation with climate related variables	23
4.3	Results from Machine Learning models	29

Acknowledgments

We would like to particularly thank Mrs Asmynour YOUSSEUF and Mr Jordan BELOUCHAT for the knowledge they transmitted and for the quality of supervision we benefited from them during this project. Their pertinent remarks and their availability to answer our various questions were indispensable.

Our thanks also go to the Adway firm for the relevance of the subject proposed for the realisation this project.

Abstract

In a constantly evolving economic environment, risk management is an important part of financial institutions. In order to protect themselves, some banks are integrating new risks such as climate risk into their statistical models.

Our study focuses on a portfolio containing individuals living in the USA who have taken out a loan with a financial institution . The data contains information on the characteristics of the customer such as his rating, information about the loan (maturity, type of loan), as well as a target variable indicating the loan status of the customer.

The aim of this project is first to implement a methodology for integrating climate risk into credit models and then to evaluate the impact of climate indicators on the estimation of the probability of default of a borrower.

Several findings emerge from this study. First, the rating grade is a key indicator of a client's financial health. Clients with E as grade are 7 times more likely to default compared to those with A as grade, all else being equal. Then, other characteristics such as the number of trades opened in past 24 months or the number of payment on the loan are also key indicators of default. Concerning climate indicators, we find that individuals living in higher risk states are more likely to default. This effect is captured by the variable **insurance** which gives an idea of the insurance premiums paid by individuals. We also find that, despite the lack of granularity, climate data allows one to improve model performance for default modelling either with Logistic Regression or other Machine Learning.

Finally, we conclude this study by comparing the performance of the logistic regression model with three machine learning approaches: Random Forests, XGBoost and Neural Network. We find that these methods do significantly outperform our logistic regression model.

Keywords : Risk, credit, climate, logistic regression, Machine learning, default, probability.

INTRODUCTION

The bank's core business remains credit. It represents an important part of the income, but it also generates a very high cost in case of non-repayment. Therefore, there is a risk related to the non-repayment of the loan granted: this is the credit risk.

The credit risk is defined as the risk of loss on a debt instrument resulting from the failure of the borrower to make required payments to the debtor such as a bank. We generally distinguish two types of credit risk:

- Default risk, which arises when the borrower is not able to settle an obligation in full when due. An example is a mortgage loan where the borrower defaults on his payment.
- Downgrading risk, which concerns debt securities. This is the risk of a change in the debt market value caused by the increase of the obligor's credit risk.

Against this background, the Basel Committee on Banking Supervision (BCBS) requires banks to have sufficient capital and reserves to absorb losses in the event of default. Historically, to calculate the probability of default, banks use traditional risk drivers. However, today with climate change, new drivers have appeared in the calculation of the probability of default in internal models: these are the climate risk factors.

Climate risk, driven by floods, droughts or hurricanes, is the exposure of economic agents to climate impacts and their repercussions on their activities. These climatic phenomena are not new in the history of humanity, but they have increased in recent years to the point of being the focus of all concerns. There are three main types of climate risks:

- Physical risk, which can be divided in two types: "acute" and "chronic". The former corresponds to the direct losses associated with the damage caused by climatic hazards on economic actors. These effects can be direct as damage to property or reduction in productivity, or indirect, such as the disruption of supply chains. The chronic physical risk corresponds to losses caused by progressive change in climate.
- Transition risk, which corresponds to the economic consequences of implementing a low-carbon economic model. It may result, for example, from the rapid adoption of climate policies unfavorable to certain sectors of activity (fossil fuels, transport, etc.), from the acceleration of technological progress or change in customer preferences.
- Legal risk, which corresponds to the compensation that a legal entity could be required to pay in the event that it is deemed responsible for damage caused by the consequences of climate change.

In this study, we will first conduct a literature review to understand what has already been done around the integration of climate risk into internal credit models. This literature review is a critical assessment of research developments in the area of integrating climate risk into the calculation of the probability of default. It will allow us to better understand the issues at stake. Then, in a second step, we will integrate the physical risk in the calculation of the probability of default. We will conclude with an analysis of our modeling results and a final conclusion.

CHAPTER 1

LITERATURE REVIEW

The climate risk is at the heart of concerns of governments, companies and ordinary citizens. It is becoming essential to assess, anticipate and integrate it into any company strategy. Since 1988, the Intergovernmental Panel on Climate Change (IPCC) has been assessing the state of knowledge on global climate change, its impacts and the means to mitigate and adapt to them. The consequences of global warming do not spare any continent: Harvey, Sandy, Doria in the Caribbean, Danielle near the European coast, drought in Africa, flooding in Asia (nearly 1000 deaths in Pakistan in 2022)... The frequency of these extreme weather events has considerably increased in recent years. In this context, climate risk can be seen as a more predictable hazard by companies. Nevertheless, if these physical risks are the subject of increasingly complex and elaborate analysis models, their financial impacts are still poorly identified and difficult to anticipate (EY, 2022).

Through the literature, there is a large debate around the choice of climate risk indicators and the way of integrating them into the financial models, mainly the credit risk models. Two main approaches are used to integrating climate risk drivers into financial risk assessment : **direct approach** and **indirect approach**. The direct approach consists of integrating the climate variables directly into the computations of the credit risk parameters. The indirect approach consists of assessing the impact of climate risk through macroeconomic factors.

In the following, we will first address the physical risk and then the transition risk.

1.1 Physical risk

1.1.1 Relevant indicators for measuring physical risk

Different indicators are frequently used in the literature to quantify physical risk in credit models.

An initial relevant indicator that is widely used is **level of temperature**. The variation over time of the temperature in relation to its average over the period makes it possible to capture both the "acute" risk materialized by the peaks (very strong rise or very strong fall in relation to the average) and the "chronic" risk materialized by the increasing trend of the temperatures [1].

A second very important indicator is the **level of rainfall**. In their article [2], researchers Carolyn Kousky and al. show through a case study on the importance of this variable in measuring flood risk. A third indicator is the integration of **geospatial data** (topographic data, satellite data) in risk models to take into account geographic heterogeneity. As mentioned in Ulrich's article, since

climatic events depend on geographic diversity, financial losses related to acute or chronic risks will not be the same. To this end, we can consider variables such as the region, the province, or the postal code of the counterparty, depending on the granularity of the data we have at our disposal. However, the location of household assets as well as the location of a firm's production sites, suppliers and customers provides a first insight, but does not offer information on the resilience of these assets.

In addition, an **occurrence score of extreme event** is a variable that could be used. There are several types of extreme events such as floods, wildfires, cyclones, hurricanes, typhoons. Pierre Monnin [3] and Ulrich Nguemmo [1] mention for example the possibility to use the probability of occurrence of a flood. This type of information could be obtained using rating organizations such as "four twenty seven" or calculated by a specific method to the banking entity. As an example, R.M. Walles [4] quantifies the risk of a flood occurring in a given area using the proportion of insurance related to this catastrophe. However, it should be noted that the physical risk score may be biased compared to other risk scores due to a short historical depth and poor data availability on small statistical units.

Sometimes, after a disaster, we want to estimate the damage in order to assess the impact that it might have on a borrower's ability to pay. In this perspective Claire Zhang [5], builds a model to predict **property damage** due to a flood. She defines the property damage as the ratio of insurance claims over coverage and models it from the number of floors, the basement types and the flood zone.

Moreover, these indicators don't take into account efforts made by corporates in terms of adaptability and sustainability. To that purpose, Kyriakos Chatzitheodorou and al., propose a methodology allowing banks to assess corporate sustainability risks. The proposed methodology is a scoring approach on eight categories of indicators provided by the GRI¹. The indicators include the **physical risk** which covers three topics of sustainability: water sources significantly affected by withdrawal of water; percentage and total volume of water recycled and reused and total environmental protection expenditures and investments. The following part shows how to integrate those variables in financial risk models.

1.1.2 Integration of physical risk factors in financial risk assessment

In the context of physical risk, the direct approach is the most widely used approach in the literature. In Walles [4], to incorporate climate change risk into the prediction of Probability of Default (PD), flood risk was used in addition to the classical variables for physical risk. Then to estimate the differences in PD of individual loans the author used a logistic regression model that describes the linear relationship between the logarithmic ratio of default and non-default probabilities. Likewise, Carolyn Kousky and al. [2] models loan performance by directly integrating in a logistic regression flood damage with the loan characteristics. Loan characteristics in this model include the debt-to-income ratio, the minimum of borrower's and co-borrower's credit score, the number of borrowers, the occupancy type, the loan product type and the loan purpose.

The indirect approach is based on scenarios analysis. The formulation of physical risk scenarios differs from that of transition risk scenarios because it requires particular attention to the geographical

¹Global Reporting Initiative, founded in Boston 1997, the GRI is an independent international standard-setting organism for sustainability performance

location of the issues at stake, even more than to the sectoral dimension. An additional difficulty lies in the fact that the physical risk is manifested by extreme climatic episodes whose occurrence and impact on human and economic activities are practically impossible to predict. The scenarios used in the climate stress test exercises assume, first of all, that climate change will materialize in the form of an increase in the frequency and cost of climatic events over extreme weather events over the projection period considered. In the case of the ACPR's physical risk stress exercise [6], the physical risk assumptions are based on, among other things, the projections of the major contributors to the Intergovernmental Panel on Climate Change (IPCC) climate projections.

1.2 Transition risk

1.2.1 Relevant indicators for measuring transition risk

One of the main used indicator to capturing transition risk is the **Green House Gas (GHG)** (Novela and al. [7] ;Ulrich and al [1]; Andrea Cruz and al. [8]). This indicator shows to what extent the business plan of the company depends on GHG emission. There are different ways of using this indicator in transition risk modeling. One way of depicting a mapping of portfolio exposure to transition risk is to create homogeneous clusters of GHG emission (as done in Novela and al. [7]). In addition to GHG emission, there is the **Carbon footprint** that takes into account the revenue of the corporate and represents the emission per unit of income [7]. Carbon footprint is widely used by financial market participants to integrate climate change into their valuation models for transition risk [3].

However, current measures of corporate carbon footprints and green house gas emission often only include a company's direct emissions (scope 1) and emissions related to the energy it uses (scope 2), and therefore do not cover emissions from a company's entire value chain, including those of its suppliers and customers (scope 3) (Pierre Monnin, 2018). In addition, these indicators (GHG emissions and carbon footprint) do not deliver any information either about firm's possibilities to switch to low-carbon technologies, or its preparedness to do so and its ability to pass on higher costs to its customers. Therefore, the BCBS recommend integration of variable such as corporate's **investment plans toward green technology, qualitative information on corporate climate strategies, quantification of GHG reduction and its cost**[9].

These variables can be integrated in existing financial risk modeling through two different main approaches.

1.2.2 Integration of transition risk factors in financial risk assessment

While assessing the impact of transition risk in credit risk, a direct and an indirect approach can be used.

The first one has been used by Simone Novela and al., 2022 [7] with a sample that includes the constituents of the Euro Stoxx 50 Index, a market capitalization-weighted blue chip index designed to represent the 50 largest companies in the Eurozone. This sample is divided into homogeneous terciles based on the level of CO₂ emissions: the first tercile (Tercile 1) contains the companies with the lowest level of carbon emissions; the second (Tercile 2) contains the companies with a medium

level of carbon emissions; the third (Tercile 3) contains the companies with the highest level of carbon emissions. Then the author applied the method of Black & Cox Distance to Default.

The indirect approach is the main used approach as recommended by the BCBS². In the 2021 report on climate risk measurement guidelines, the BCBS highlighted three main steps in climate-credit risk assessment : translating climate risk drivers into economic risk factors to have climate-adjusted economic risk factors; linking climate-adjusted economic risk factors to exposure; and translating climate-adjusted economic risk into financial risk. This three stage modeling has become the cornerstone of several studies.

Based on these guidelines, Andrea Cruz and al. [8], provide a wide framework for assessing the impact of transition risk on credit risk. For example, transition scenarios encompassing policy, preference and technology change may cause : a direct cost of CO₂ emission depending on the GHG emission and the carbon price; an indirect cost of CO₂ emission depending on the supply chain; a capital expenditure resulting from required investment to reduce emission; change in income resulting in prices and consumer demand. According to their paper, these transmission channels will affect the corporate's default through their **cash flows, capital and collateral**. Therefore, these variables can be recalibrated according to the scenarios and reintroduced into a model in order to get the probability of default accounting for the transition paths.

In the same vein, a simulation based on a series of interactions was performed by Walles [4]. For each scenario, the author presents the evolution of specific economic variables for the future. For example, a carbon tax scenario will lead to an increase in the price of oil and gas, a change in consumer preferences will lead to layoffs through a slowdown in economic activity in a certain sector. This will ultimately result in a decline in disposable income. As a result, domestic demand will decline, resulting in lower economic growth for the entire economy [4].

After this crucial step of literature review, which is essential in any study on climate risk, as it allowed us to gather and analyze existing knowledge on a relatively new and constantly evolving field of research, with numerous publications and emerging research, we will now attack chapter 2 of our study, which will allow us to take our data in hand.

²Basel Committee on Banking Supervision

CHAPTER 2

DATA PRESENTATION AND PROCESSING

In order to shed light on integration of climate risk into credit risk models, high quality data are required. In this chapter, we present the data used in this study and the processing steps. These data can be divided into two main groups: **credit related data** and **climate related data**.

2.1 Credit related data

For this study, we relied on a database (**Lending Club Loan data**¹) provided by the Kaggle platform. This is an online platform dedicated to machine learning, data mining and data science. It allows users to find and share datasets, participate in data science competitions, collaborate with other users, and access educational resources. Our database consists of **160 variables**:

- Financial health indicator: for example the notation grade provided by the Lending Club, annual income, etc.
- Information about the loan: loan status, the issue date of the loan, the amount, the maturity, interest rate, etc.
- Borrower's indebtedness: for example the maximum current balance owed on all revolving accounts.

The scope of our study covers nearly **2 million** statistical units.

2.1.1 Creation of the target variable

The **loans status** variable was re-coded to serve as a target variable. This variable presents the current status of the loan. We consider a borrower defaults when at least one of the cases occurs:

- the bank raise a **charge off** on borrower's account, meaning the bank has create a specific provision in order to prepare for the default;
- the borrower's reimbursement are more than 31 days late.
- the borrower has already been declared in default by the bank.

¹<https://www.kaggle.com/datasets/wordsforthewise/lending-club>

The implicit hypothesis is that when a borrower is in these cases, it is probable that he will not meet his commitment, which can have serious financial and legal consequences. Once our target variable is created, we did few processing steps to select the explicatives ones.

2.1.2 Variables preselection: first steps of selection

The data processing stage and the final database constitution began with the elimination of variables with a high percentage of missing values or identical values (95% in this case). In fact, these variables do not provide viable information to correctly discriminate the target variable.

2.1.3 Database sampling

Once this pre-selection was done, we proceeded to the database sampling. One part of the database will be used to build the statistical models (train dataset) and another part to evaluate the performance of these models (test dataset). Our database was divided into two samples, training and test, with 70% of the observations in the training. This was done by a stratified random draw on the target variable which is the default indicator, the year the loan was granted and the address of the state where the loan was granted. These variables were chosen in order to maintain the same representativeness of the population in both samples.

2.1.4 Variables selection

Following the sampling, we conducted a selection of variables based on statistical tests on the training sample. These tests consisted of evaluating the relationship between the explanatory variables and the target variable (the default indicator). Two tests were performed: a **Cramer test** and a **Kruskal-Wallis test**.

- **Cramer's test**

To investigate the existence of a relationship between two categorical variables, a χ^2 test is used. The Cramer test measures the strength of the relationship between the two categorical variables. It is based on a test statistic called Cramer's V and is defined as follows:

$$V = \sqrt{\frac{\chi^2}{n \times \min(c - 1, r - 1)}}$$

With n the total sample size, r the number of rows and c the number of columns.

The closer the Cramer's V is to 1, the stronger the link between the two variables and conversely the closer it is to 0, the weaker the link.

- **The Kruskal-Wallis test**

The Kruskal-Wallis test is a non-parametric test often used as an alternative to the ANOVA to study the relationship between a qualitative variable and a quantitative variable. It allows to test if K samples come from the same population. In our context, it will allow for each of our quantitative variables, to compare the two populations in presence (the one with a default indicator of 1 and the one with a default indicator of 0) and thus to give an idea of

the relationship between the quantitative variables and the target variable. The test is based on the following statistic:

$$K = \frac{1}{N(N+1)} \sum_{i=1}^n [R_i^2 - 3(N+1)]$$

where n_i is the size of sample i, N is the sum of n_i , and R_i is the sum of the ranks of sample i among all samples.

Based on the Kruskal-Wallis test statistic, the first 15 quantitative variables most related to the target variable are selected. Then, based on a correlation study (calculation of Pearson and Spearman correlation coefficients), when there is a correlation between two or more variables (see in the appendix for the correlation tests used 4.4), we keep the variables most correlated to the target variable. As for the cramer V , we retain the variables whose cramer V is higher than 0.10.

The variables retained as a result of this selection are presented below:

Table 2.1: Name and label of variables

Variable	Label
<code>total_rec_late_fee</code>	Late fees received to date.
<code>last_fico_range_low</code>	The lower boundary range of the borrower's last FICO.
<code>fico_range_high</code>	The upper boundary range of the borrower's FICO.
<code>dti</code>	A ratio calculated using the borrower's total monthly debt payments.
<code>grade</code>	LC assigned loan grade.
<code>term</code>	The number of payments on the loan.
<code>last_pymnt_amnt</code>	Last total payment amount received.
<code>acc_open_past_24mths</code>	Number of trades opened in past 24 months.
<code>mort_acc</code>	Number of mortgage accounts.
<code>total_rec_prncp</code>	Principal received to date.
<code>income</code>	Average income per capita at last payment date.

2.1.5 Handling of missing values

After the selection of variables, we performed a treatment of the missing values present in the database. Among the variables present in the final database, some variables had missing data. Indeed, for the variable `dti` we simply deleted the 374 missing observations because this number is small compared to the number of observations in the database. Then, we chose to delete some observations that had missing data. Finally, the variable `last_pymnt_d` had 2313 missing values that we imputed using a K-nearest neighbor algorithm with K equal to 10 (see Appendix 4.5). The algorithm was trained on the complete observations and predictions were made on the observations containing the missing values. The variable `mort_acc` and `openacc` that had missing values were also imputed by the median. At the end of these different steps, we obtained a training sample of **20 variables** and **1 540 723 observations** and a test sample of 20 variables and **667 175**

observations.

In this section, we presented the credit data available in the database and then discussed all the processing and transformation steps that are necessary to have quality data.

In the next section, we will discuss the origin of the climate data that will be used in the rest of the project. Note that all the processing and transformations done for the credit variables will be applied for some climate variables.

2.2 Climate related data

We focused in this project on the integration of physical risk in credit risk. In this section, we present the climate related data used in this study.

One of the major problems in the use of climate data is the choice of granularity. The granularity of the data refers to its level of detail. The finer the granularity, the more detailed the data are and the more precise the analysis will be. The climate data used in this study are all at the state level. We were constrained to do so because the credit data did not contain a precise borrower's address.

Different categories of climate variables were selected. The first category concerns the number of occurrences of certain natural disasters. The number of occurrences of a disaster gives an idea of the occurrence risk of the disaster and thus allows a comparison in terms of risk in the different states of the USA. For each of the selected claims, we counted the number of occurrences of the disaster over the last three years preceding the date of the last payment made by the client. We made the assumption that beyond 3 years, the events would have very little effect on the probability of default of individuals. The related database comes from the **Federal Emergency Management Agency (FEMA)** website. The disasters selected are **wildfires**, **hurricanes** and **floods**. The choice of these disasters is explained by their impact on the daily lives of the populations affected. Indeed, being victim of such disasters could have an impact on the places of residence, the jobs and the means of travel of the populations, and thus have financial repercussions.

However, from the frequency of occurrence, we are not able to quantify the level of damage of a disaster. For this reason, we have added a second category of climate variables to the database: the damage caused by the different disasters. The damages are expressed in financial losses (in dollars). For example, in the case of floods, the variable associated with the damage due to this disaster is created by summing up all the damage listed in the state concerned over the last three years preceding the client's last payment.

We then added two other indicators to the base. The first indicator is the average home **insurance premium** by state over the various years. The higher the risk of natural disasters in a state, the higher the insurance premiums in that state. An individual in the database is therefore assigned the value of the average insurance premium paid in their state in the year corresponding to the year of their last payment. The insurance variable was then discretized following the same principle as the credit variables. We collected the data associated with insurance on the National

Association of Insurance Commissioners (NAIC) platform. The second indicator is the deviation of the average temperature of the year of the individual's last payment from the average temperature observed during the period 1895-1930. The choice of the period is explained firstly by the availability of data. Secondly, the period is far enough away to capture possible deviations of temperatures from the period of the loans. Regarding the source of these data, they come from the National Centers for Environmental Information (NCEI).

Finally, we added to the database a variable named region which gives the region in which the individual lives. Depending on the region of residence, an individual has more or less risk of suffering a natural disaster.

The variables retained as a result of this selection are presented below:

Table 2.2: Name and label of climate variables

Variable	Label
Nfire	number of fires in the three years prior to the last payment date
Nhurricane	number of hurricanes in the three years prior to the last payment date
Insurance	average insurance premium by state
Ecart_temp	difference between the year of the last payment and the average over the pre-industrial period
Region	client's region
Nflood	number of floods in the three years prior to the last payment date
damage_fire	sum of damages caused by a fire in the three years prior to the last payment date
damage_flood	sum of damages caused by a flood in the three years prior to the last payment date
damage_storm	sum of damages caused by a storm in the three years prior to the last payment date

Now that we have the climate variables, we will discretize all the variables (credit and climate) to limit the effects of outliers, among others.

2.3 Discretization of quantitative variables and grouping of modalities for categorical variables

Discretization of continuous variables is an operation that consists in transforming a continuous variable into a discrete variable using a limited number of values. There are several advantages to discretizing continuous variables in statistical models. First, it limits the impact of extreme values and takes into account non-linear effects or interactions (to improve the readability of the final score grid).

For the discretization of our continuous variables, we first proceed to split our variables into a large number of classes (20 classes). Then, we group the successive classes according to the proximity of the default rates and the consistency of the cases. We retain only transformations of 2 to 5 terms per variable whose terms are sufficiently separated in terms of risk (relative differences between default rates). Regarding the grouping of modalities for categorical variables, we applied the criteria of at least 5% of the population in each modality. After discretizing our continuous variables and grouping our modalities for the categorical variables, we took care to check the stability in volume and risk of the transformations thus selected and to calculate the statistical indicators Tschuprow

T and Cramer V for each transformation. The list of variables is presented in the following table:

Table 2.3: List of categorical variables after discretization

Variables	Modalities	Proportion
LC assigned loan grade	A B C D E	19.40 29.48 28.67 14.19 8.25
Late fees received to date	No fee fee >0	96.52 3.48
The lower boundary range the borrower's last FICO pulled belongs to	<570 [570 ; 620[[620 ; 660[>660	8.69 8.66 9.34 73.31
Last total payment amount received	<=90]90 ; 1200] >1200	4.2 61.46 34.34
Principal received to date	<=2400 [2400 ; 4430[[4430 ; 9500[>9500	19.60 14.82 25.35 40.12
The upper boundary range the borrower's FICO at loan origination belongs to	<=710 >710	68.11 31.89
The number of payments on the loan	36 monts 60 months	71.40 28.6
Ratio calculated using the borrower's total monthly debt payments	<= 20 >20 <=10000	58.97 41.03 39.79
The listed amount of the loan applied for by the borrower	[10000 ; 15000] >15000 <=40000	20.51 79.70 2.27
The borrower's income]40000 ; 55000] [55000 ; 65000] >65000 <=4	55.94 33.17 8.61 58.21
Number of trades opened in past 24 months	[4 ; 7] >7 0	27.01 14.77 43
Number of mortgage accounts]0 ; 2] >2 <=1000	31.96 25.04 23.38
Annual home insurance premiums paid by tenants or owners	[1000 ; 1700] >1700 Midwest	58.96 17.66 17.97
The borrower's region of residence	Northeast South West East North Cent East South Cent Middle Atlantic	20.70 35.87 25.45 13.06 4.32 15.43
The borrower's Division of residence	Mountain New England Pacific South Atlantic West North Cent West South Cent	7.87 5.27 17.58 20.42 4.92 11.13

A summary of the main steps in variable processing is presented below.

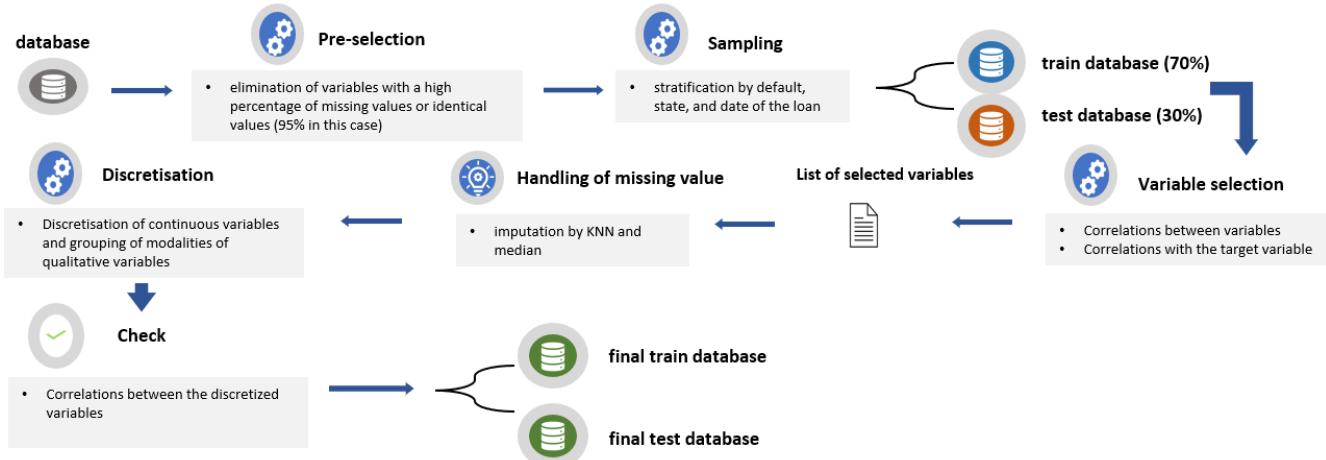


Figure 2.1: Main steps of data processing

In this chapter, we first discussed the origin of the data we use to carry out our study, and then we detailed the different steps that allowed us to have a database ready for modeling. We have taken care to document the entire variable selection phase, the treatment of missing values as well as the discretization of continuous variables and the grouping of certain modalities for qualitative variables. The diagram above is a summary that allows us to visualize all the essential steps to arrive at a database ready for modeling.

In the next chapter, we will present the characteristics of the customer portfolio in our database. We will see how the customers are distributed with respect to the target variable, and then we will try to understand the mechanism that triggers the counterparty default.

CHAPTER 3

DESCRIPTION OF THE PORTFOLIO

3.1 Default rate of the portfolio

The table below describes the distribution of the portfolio in terms of default occurrence. According to this table, our portfolio counts 266774 defaulting client, representing 12.08% of the database. This default rate is not steady over time as shown in figure below 3.1

Modalities	Proportion	Volume
Default	12.08	266774
Non-Default	87.92	1941124

Table 3.1: Distribution of the portfolio according to the default

In the graphic below 3.1, default rate is globally decreasing. Two main periods catch our attention. First in 2008-2009, as showed in the graphic , the default rate reached a peak of 42% in 2009, before decreasing progressively. This peak can correspond to the effect of the **subprime crisis**, which explains these high levels of default in 2008 and 2009. The volume of the portfolio for those years is quite small, reason why we can assume that this effect of **subprime crisis** will not affect our modeling. Second, between 2018 and 2019, the number of accepted loans increased by 136.65% and the default rate experienced a sharpe decrease (from 17.3% to 0.3%). This phenomenon can be explained by the fact that in 2019 the USA experienced the lowest unemployment rate (3.6%) in the half century¹. We can also note few some differences between states.

¹https://www.lemonde.fr/international/article/2019/05/03/etats-unis-le-taux-de-chomage-au-plus-bas-depuis-un-demi-siecle_5458021_3210.html

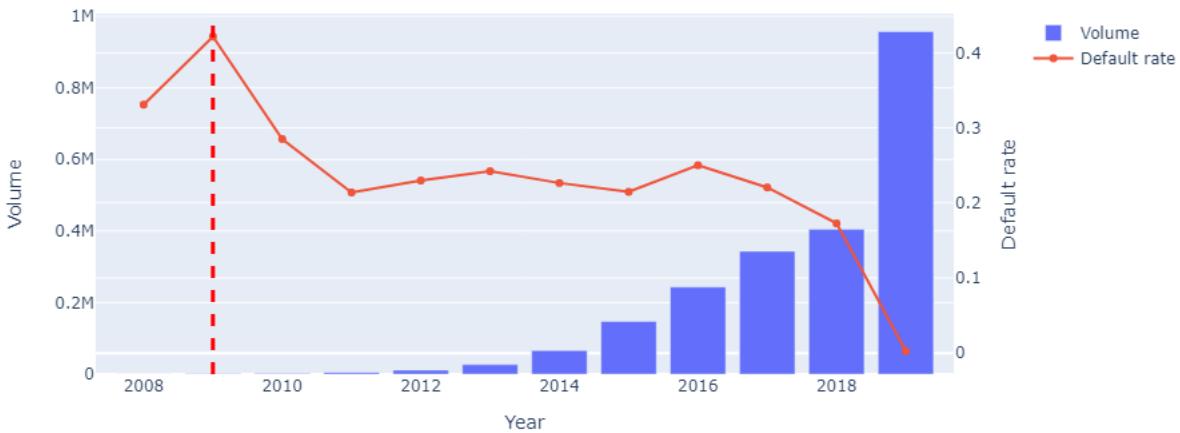


Figure 3.1: Default over time

In the graphic below we notice that the default rate is higher in Alabama (14.7%), Arkansas (14.4%) and Louisiana (14.2%). On the other side, default rate is lower in Maine (5.7%), Idaho (7.5%) and Vermont (7.6%).

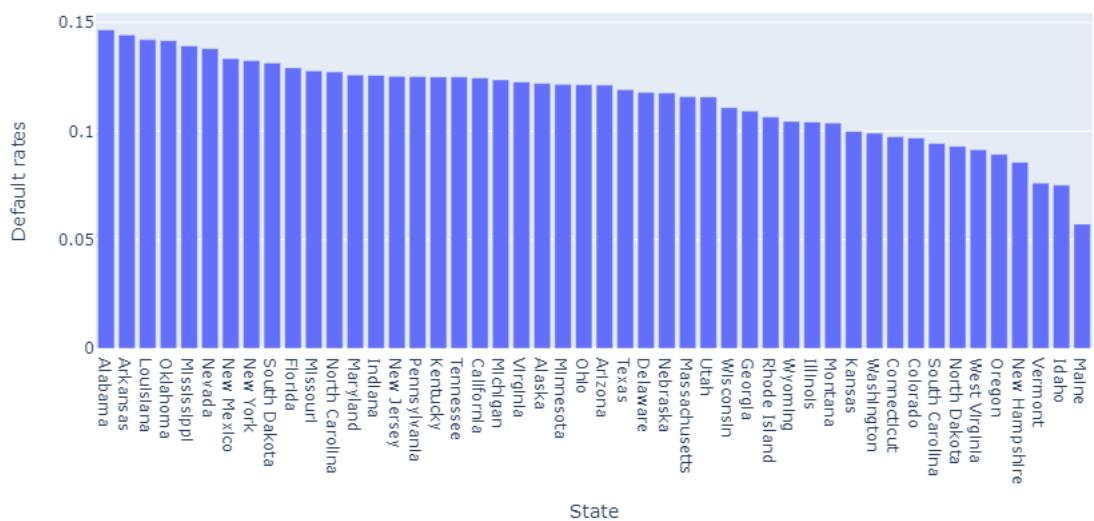


Figure 3.2: Default rate by state

To better understand what triggers the status of default, we can analyse the interaction of the default variable and the most related co-variables.

3.2 Triggers of counter-party default

Bivariate analysis made with most explicative variables allows to better grasp paths that trigger the default and avoids to use post default variables in our model. For example, variables like **the last payment received to date** and **the total payment amount** are post-loan variables and does not allow to estimate default at any time. After removing these variables, the **grade note** is the most correlated variable to the default indicator 4.12 among categorical variables. Counter-party in grade A has 3.3% default risk and the further you go from A to E, the more the defect rate increases and reaches a peak of 29.97% in grade E. A second very important variable is the **the average income per capita**. This variable shows how rich the state is and we can see that as income increases, default decreases. It also shows that the difference in default rates across states is not only explained by geographical location but rather by economic conditions.

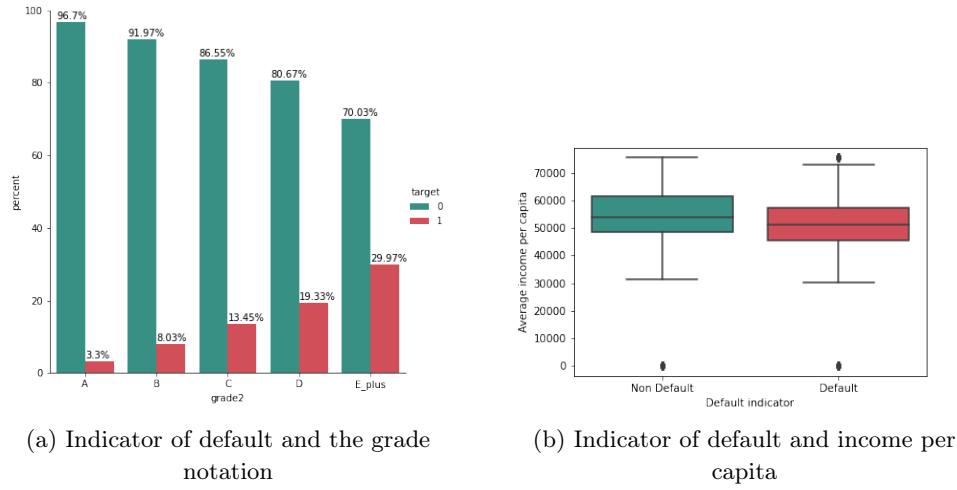


Figure 3.3: Default indicator vs most correlated variables

The purpose of the loans can also be an important variable for the business despite his quite low Cramer V score. As we can see in the graphic below, default rate differs slightly between purposes. We note that small businesses are the highest risk purpose with 19.03% followed by educational purpose (default risk of 16.5%). This quite logic because the USA is a capitalist economy with fierce competition between companies. Therefore, many small business may fall due to lack of funds.

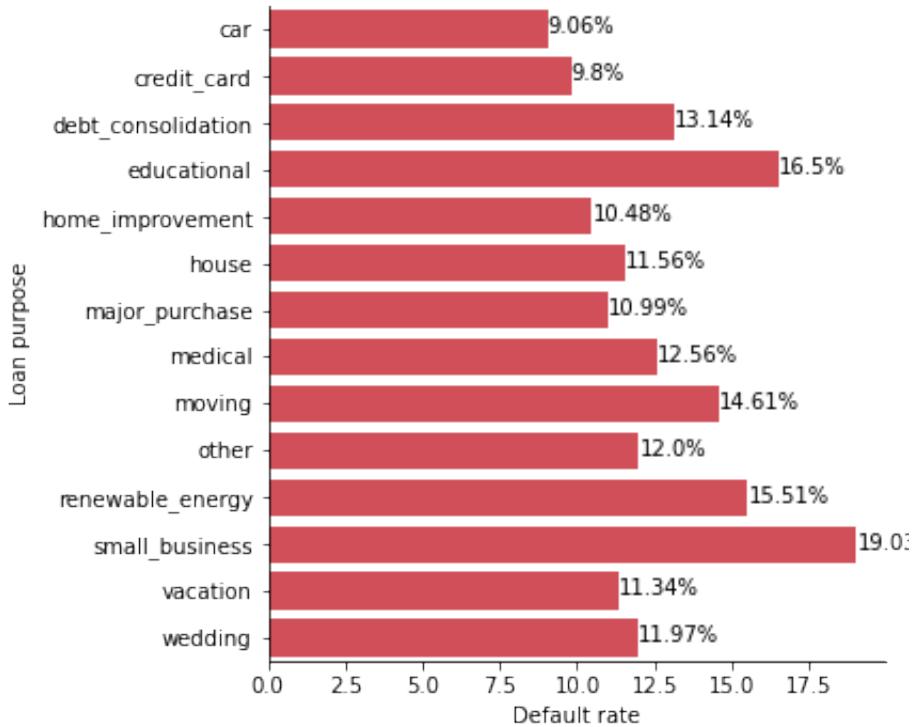


Figure 3.4: Default rate by purpose of loan

In addition to the economic and financial variable, climate variables can add more explanations to the default occurrence. We will add analyse with the geographical position, the number of floods, wildfires and hurricanes and the shift of temperature relatively to pre-industrial period.

As we can see in **figure (a)** 3.5, default is more frequent in the Mid-South and slightly in the West of USA. On the basis of the pre-industrial period, rise in temperature is more severe in East (particularly South-East) and West **figure (b)** 3.5. In the West, occurrence of high temperature is correlated with high level of default rate. The same correlation is remarked between the number of flood in last 3 years and the default, **figure (c)** 3.5. One can note that flood occurrence is more severe in the South of USA (for example in Texas where we count on average 3 declared floods) where default rate is higher too. For the hurricanes, **figure (e)** 3.5, they strike more in South-East but are nearly not present in the rest of the Country. On the other side, Wildfires occur in the East and remain quite missing in the rest of the country, **figure (d)** 3.5.

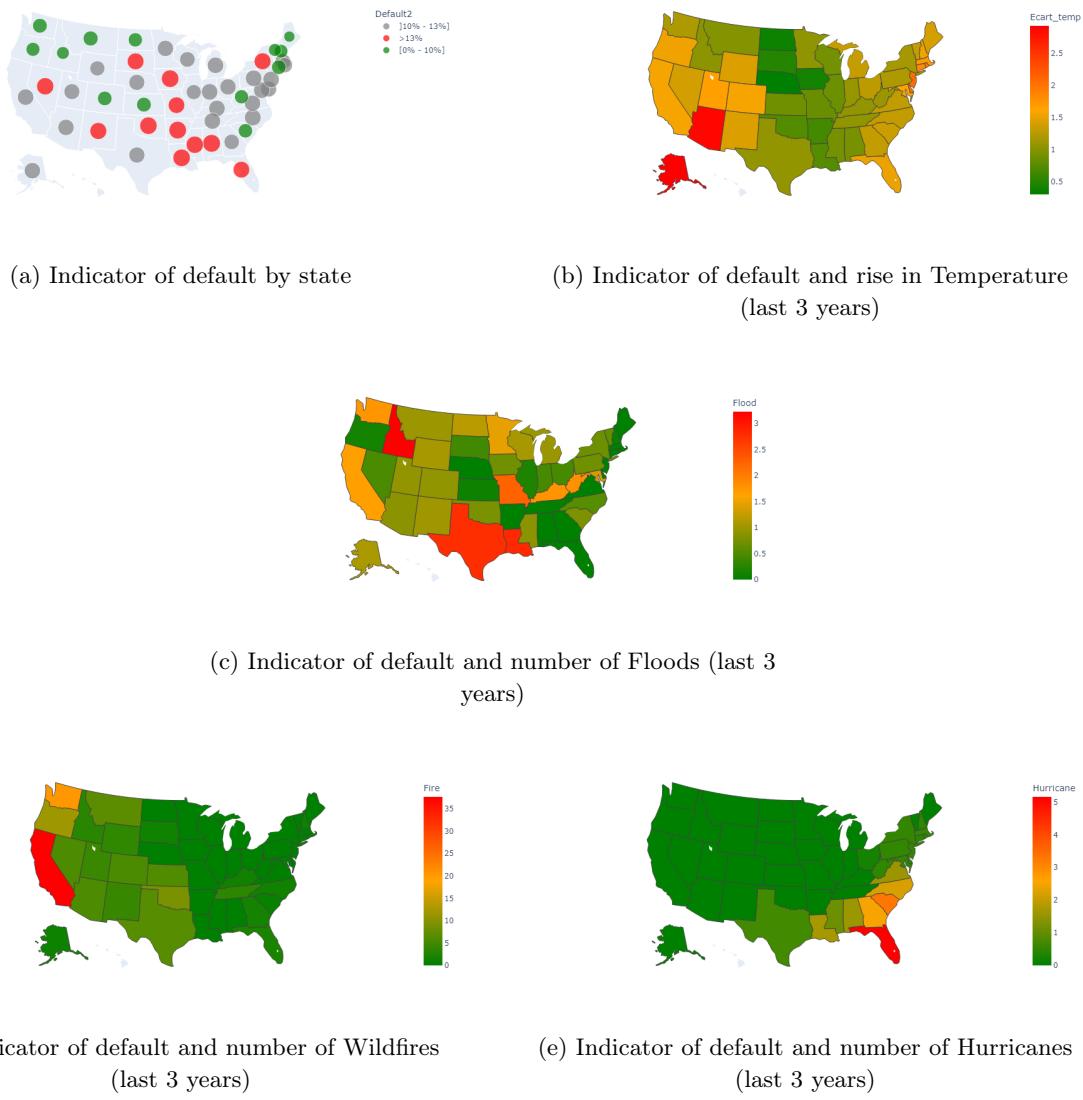


Figure 3.5: Integration of climate variable to bivariate analysis

After discovering the characteristics of our portfolio in this chapter and the triggering mechanism of a counterparty default, we will see in the next chapter how to model a customer default using logistic regression and machine learning models.

CHAPTER 4

MODELING FRAMEWORK

As presented above, our target variable is a binary one, therefore we encounter a **classification problem**. To integrate climate variables into counterparties default modeling, we have selected a direct approach (presented in literature review). Thus we have to deal with a problem of model specification, which means the way we represent the dependencies between the default and the covariates. The Logistic regression, our first model, is widely used for default modeling. It assumes a linear relation between the default probability and the covariate (as presented below). To encompass other types of dependencies, we used other machine learning algorithms such as Random Forest, Extreme Gradient Boosting (XGBoost) and Neural Network. Below the principles, advantages and limits of each of them are presented.

4.1 Modeling of the default using logistic regression

4.1.1 Theoretical background of Logistic Regression

Model specification

Logistic regression belongs to a family of models called Generalized Linear Models (GLM). The objective of a GLM model is to predict the conditional expectation of a categorical response variable Y conditional on a set of quantitative or qualitative variables $X = (X_1, X_2, \dots, X_p)$ called covariates. A GLM model is entirely determined by the knowledge of a link function (function linking X to Y) and a class of laws linking Y to X . In logistic regression, the logit link is used. It gives the following specification:

$$\ln\left[\frac{P(Y_i = 1|X_{i1} = x_{i1}, X_{i2} = x_{i2}, \dots, X_{ip} = x_{ip})}{1 - P(Y_i = 1|X_{i1} = x_{i1}, X_{i2} = x_{i2}, \dots, X_{ip} = x_{ip})}\right] = \beta_0 + \beta_1.X_{i1} + \beta_2.X_{i2} + \dots + \beta_p.X_{ip} \quad (4.1)$$

The different coefficients can be estimated through log-likelihood maximisation with the following program:

$$\max(\beta_0, \beta_1, \beta_2, \beta_p) l(\beta) : \sum_{i=1}^n Y_i(\beta_0 + \beta_1X_{i1} + \dots + \beta_pX_{ip}) - \sum_{i=1}^n \ln(1 + \exp(\beta_0 + \dots + \beta_pX_{ip}))$$

Overall significance of the model and individual coefficients

To determine whether the model is significant or not, the test statistic is computed as the difference between the null deviance and the residual deviance.

The deviance D is calculated from the log likelihoods :

$$D = -2L^{model}(\beta) - 2L^{saturated\ model}(\beta)$$

We therefore compare our model with the saturated model. In particular, we use the deviance test (or likelihood ratio test) to test the significance of the model.

For the individual significance of the coefficients the hypothesis tested is the nullity of the coefficient β_i , the alternative hypothesis being β_i non-null. Thus, if the resulting p-value of this test is less than 5%, the null hypothesis will be rejected and the coefficient can therefore be considered significant.

The statistic used for this test is the Wald statistic. It is defined as follows :

$$W = \frac{\hat{\beta}_i^2}{V(\hat{\beta}_i)} \sim \chi^2(1)$$

Evaluation of model performance

To get an idea of the predictive quality of the model, we use the AUC metric and the ROC curve. The ROC curve is constructed by plotting sensitivity versus (1- specificity) for different thresholds. Given a threshold s, the sensitivity, on the ordinate, can be interpreted as the rate of true positives, while the specificity represents the rate of true negatives.

- **Sensitivity (True Positive Rate) :** $TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$
- **Specificity (True Negative Rate) :** $TNR = \frac{True\ Negative}{True\ Negative + False\ Positive}$

The closer the ROC curve is to the first bisector, the more inoperative the model is. On the other hand, the further away the curve is from the bisector, the better the predictive power of the model. The AUC (Area Under Curve) metric therefore measures the area under the ROC curve. A totally random model will have an AUC of 0.5 while a perfect model will have an AUC of 1.

Gini index

The Gini index, invented by Corrado Gini in 1910, is an indicator that measures the association between the two subsamples of a variable X, restricted to the two groups of a binary qualitative variable Y. In the context of logistic regression, it is an indicator of the goodness of fit of the model. It varies between 0 and 1. The closer it is to 1, the better the model. It can be calculated from the area under the ROC curve (AUC).

$$I_G = 2 \times AUC - 1$$

One of the main advantage of this model is that the interpretability is quite straightforward. The output of this model shows directly the characteristics that have significant¹ influence on counterparty default. A second advantage is that logistic regression is not computationally expensive, thus the fit does not take too much time and does not require extreme computer performance.

Nevertheless it has some disadvantages. In the equation 4.1 above, the log-linearity assumed by the model is a strong hypothesis. The true dependencies between the default and the features may be more complex than that.

In the first section of this chapter, we laid the theoretical foundation for logistic regression. Mastering this theoretical framework is essential for understanding the estimation of the model, which we will discuss in the next section.

4.1.2 Model estimation

Choice of variables

After an initial selection of variables based on criteria such as correlations between explanatory variables and correlation with the target variable, we retain 17 credit-related variables and 6 climate-related variables. Then, we proceeded to a step-by-step selection, i.e. by adding the variables one by one in the model. The objective is to have a parsimonious model. A model will be as parsimonious as it will contain the least number of variables possible and able to give good performance. The idea is to reconcile parsimony and model performance. This can also help to reduce the risk of overfitting the model, by adding the variables one by one to the model, we eliminate the variables that do not have a significant contribution to the Gini index, or that present too great a difference in Gini between the train base and the test base.

A first one is performed with only the credit-related variables. In this first model, the variables `last_fico_range_low`, `total_rec_prncp` and `last_pymnt_amnt` were removed from the model because of the over-fitting of the model. The variable `mort_acc` was removed because it did not bring any gain in Gini. This model finally contains 9 variables.

¹A p-value indicating whether or not a variable is statistically significant in the model.

Table 4.1: Estimation with credit related variables only

Variable	Modalities	Estimation	Odds ratio	Wald 95% C.I
grade	B	0.7614***	2.141	2.091 2.193
	C	1.2313***	3.426	3.346 3.507
	D	1.5939***	4.923	4.801 5.047
	E	2.1060***	8.215	8.000 8.436
	A	0	.	.
Late fees received to date	fee >0	1.8081***	6.099	5.985 6.215
	No fee	0	.	.
The upper boundary range the borrower's FICO at loan origination belongs to	<=710	0.2758***	1.318	1.300 1.336
	>710	0	.	.
The number of payments on the loan	60 months	0.1024***	1.108	1.095 1.121
	36 months	0	.	.
A ratio calculated using the borrower's total monthly debt payments	>20	0.1477***	1.159	1.147 1.171
	<=20	0	.	.
Average income per capita	<=40000	1.9130***	6.774	6.504 7.054
]40000 ; 55000]	1.4020***	4.063	3.942 4.189
]55000 ; 65000]	1.2657***	3.545	3.437 3.657
	>65000	0	.	.
Number of trades opened in past 24 months	>7	0.3909***	1.478	1.458 1.499
]4 ; 7]	0.2342***	1.264	1.249 1.279
	<=4	0	.	.
Intercept		-4.9857***		

*** : pvalue <.0001

A second regression was then performed by adding the climate variables to the first model. The final model has 15 variables (9 credit-related and 6 climate-related). The results are presented below.

Table 4.2: Estimation with climate related variables

Variable	Modalities	Estimation	Odds ratio	Wald 95% C.I
grade	B	0.7365***	2.089	2.039 2.139
	C	1.2016***	3.325	3.247 3.405
	D	1.5476***	4.700	4.583 4.820
	E	2.0091***	7.456	7.259 7.660
	A	0		
Late fees received to date	fee >0	1.8497***	6.358	6.236 6.482
	No fee	0		
The upper boundary range the borrower's FICO at loan origination belongs to	<=710	0.2693***	1.309	1.291 1.327
	>710	0		
The number of payments on the loan	60 months	0.1814***	1.199	1.185 1.213
	36 months	0		
A ratio calculated using the borrower's total monthly debt payments	>20	0.1817***	1.199	1.187 1.212
	<=20	0		
Average income per capita	<=40000	2.9346***	18.813	17.971 19.695
]40000 ; 55000]	2.3493***	10.478	10.104 10.866
]55000 ; 65000]	1.9234***	6.844	6.624 7.072
	>65000	0		
Number of trades opened in past 24 months	>7	0.3836***	1.468	1.447 1.488
]4 ; 7]	0.2274***	1.255	1.240 1.271
	<=4	0		
Annual home insurance premiums paid by tenants or owners	>1700	0.7734***	2.167	2.118 2.218
]1000 ; 1700]	0.4143***	1.513	1.490 1.538
	<=1000	0		
The borrower's region of residence	Midwest	-0.1882***	0.828	0.811 0.846
	Northeast	0.4537***	1.574	1.538 1.611
	South	-0.2601***	0.771	0.753 0.789
Difference of temperature between the year of the last payment and the average over the pre-industrial period		0.1853***	1.204	1.197 1.210
Number of fires in the three years prior to the last payment date		-0.00469***	0.995	0.995 0.996
Number of floods in the three years prior to the last payment date		-0.2339***	0.791	0.787 0.796
Number of hurricanes in the three years prior to the last payment date		-0.1839***	0.832	0.829 0.835
Intercept		-6.0344***		

*** : pvalue <.0001

4.1.3 Interpretations

From tables 4.1 and 4.2 presented in the previous section, we can see that all the variables used are significant at the 5% threshold both in the model without the climate variables and in the one with the climate variables. Concerning the credit related variables, the effects remain globally the same in the two models constructed except for the **average revenue per capita** where the effect has almost tripled with the addition of the climate variables. However, we note from the odds ratios of the modalities of this variable (greater than 1) that individuals with low incomes have a much greater risk of defaulting than those with high incomes. Also, through the variable **term**, we note that long term loans present slightly more risk in terms of default than short term loans (odds ratio= 1.199). The variable **grade** confirms the importance of the rating agencies in the loan granting process. Indeed, compared to an A rated individual, a B rated individual is twice as likely to default while an E rated individual is about 8 times more likely to default.

Regarding the climate variables, the variables **insurance** and **Ecart_temp2** show the expected

effects. In states where individuals pay higher insurance premiums, the risk of default is higher. Individuals who pay insurance premiums above \$1700 are twice as likely to default as those who pay premiums below \$1000 (odds ratio=2.167). Similarly, as the temperature differential increases, the risk of default increases. By contrast, the number of occurrences has the opposite effect to what one might expect. Indeed, we see from the odds ratios (less than 1) that when the number of occurrences increases, the risk of default decreases. These results, at first sight counter-intuitive, could be explained by the existence of grant policies in United States for natural disasters. Victims of disasters could benefit from subsidies, which would prevent them from defaulting if they had loans to repay.

4.1.4 Model performances

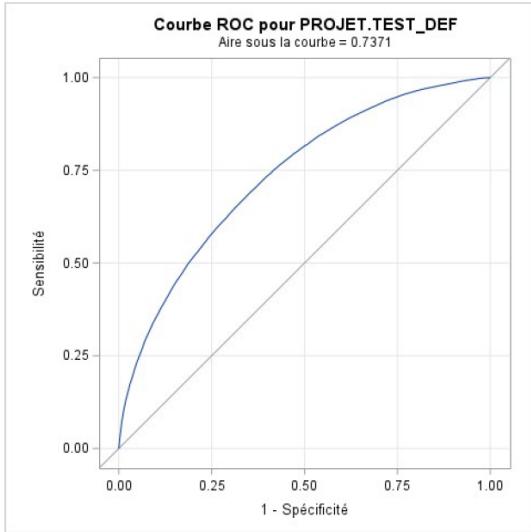


Figure 4.1: ROC curves obtained with credit variables only

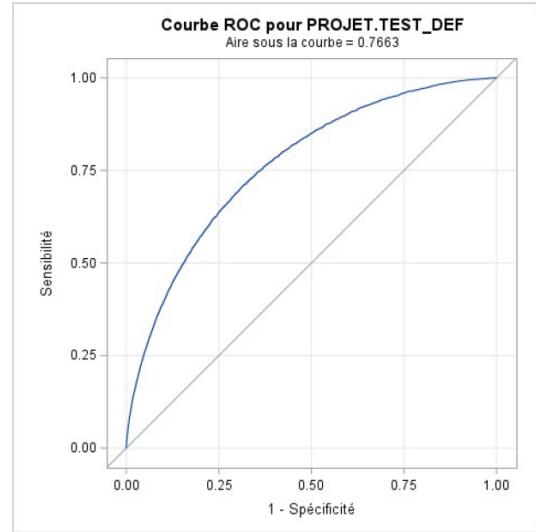


Figure 4.2: ROC curves obtained with model including climate variables

Model with credit related variables only

	Train	Test
AUC	0.738	0.737
GINI	0.476	0.474
Error rate	11.8%	12.1%

Model including climate variables

	Train	Test
AUC	0.768	0.767
GINI	0.535	0.534
Error rate	11.7%	12.06%

Figure 4.3: Performance of logistic regression models

The AUC on the test base obtained with the first model (with only the credit variables) is 73.7% while this value is 76.7% in the second model (including climate variables). Therefore, the model gains in performance by taking into account the climatic variables. Moreover, we also notice that in both models the AUC values on the learning base and the test base are quite close, which indicates that there is no overfitting.

For the first model (credit variables only), at the optimal threshold ², we obtain a specificity of 75.9% and a sensitivity of 57.1%; while for the second model (including climate variables) we obtain a specificity of 75.6% and a sensitivity of 62.9%. Another important observation is that there is a large increase in sensitivity (True Positive Rate) by including climate variables. Overall, we see that both models tend to predict non-default better than default. This can be explained by the imbalance of the database, which contains a majority of non-defective customers

4.2 Projection of the default rate under 3 scenarios of temperature rise

The figure below 4.4 describes how the default probability would behave under scenarios of temperature rise. From the left to right, different increments were applied. From the scenario (a) to scenario (d), default probability would increase by **9.2% on average**. We can notice that the default rate would increase more in the **Mid-South** of the United States. States in that area such as Arkansas, would register very important rise in default rate. In that state the default probability would go from 24% to 37%, meaning an increment of 13% in default probability.

Evidences from GIEC lastest report³ showed that rise in temperature is mainly caused by human activities. Therefore, policies for transition should be made and banks and their counterparties should be ready for that change.

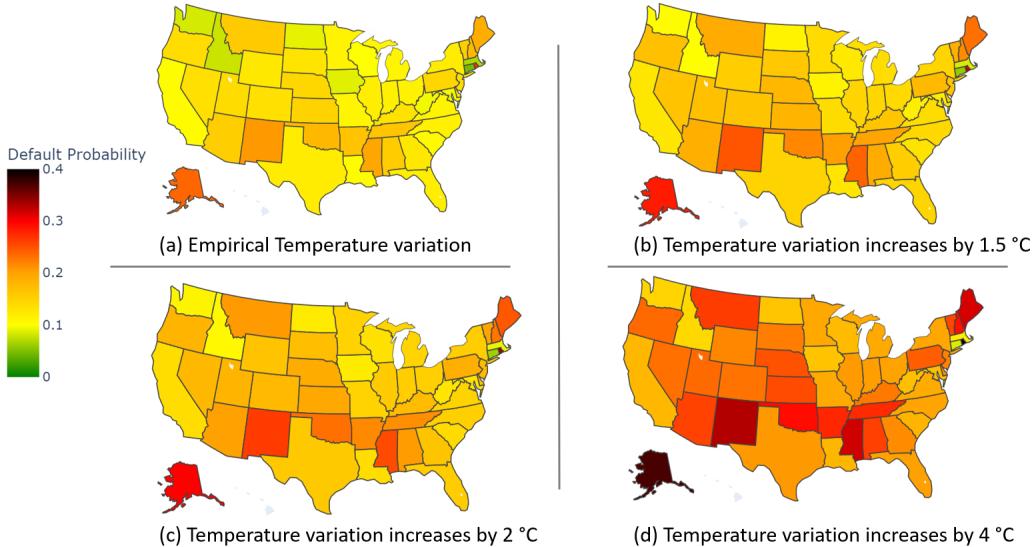


Figure 4.4: Scenarios of default when temperature rises

After laying the theoretical foundations of logistic regression in the previous section, we proceeded in this section to estimate the model. Then, we evaluated the performance of our logistic regression using indicators such as the Gini index or the ROC curve.

²We recall that the optimal threshold in terms of balance of sensitivity and specificity is defined as the threshold corresponding to the point (1-specificity,sensitivity) closest to the point (0,1)

³<https://www.ipcc.ch/report/sixth-assessment-report-cycle/>

In the next section, we will apply machine learning models to our data in order to compare their predictive quality to that of the logistic regression. We will focus on **Random Forests**, **XGBoost** and **Neural Network** which are commonly used and generally give good performances.

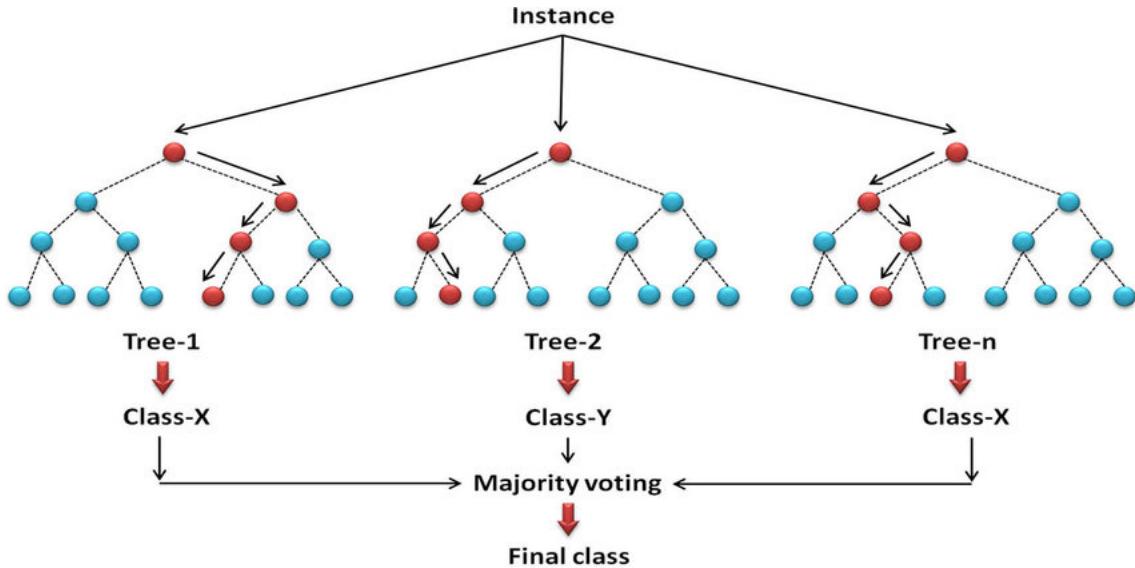
4.3 Challenger models

4.3.1 Random Forest: "wisdom of crowds" (Chloé-Agathe Azencott, 2018)

The Random Forest belongs to the class of methods known as ensemble methods. It was proposed by Brieman in 2001. The term "ensemble method" refers to the fact that these classes of algorithms combine many weak learners (weak trees, figure below) to obtain a powerful model. This algorithm is based on Bagging (Bootstrap Aggregating) with a particular way of features selection for each weak learner. The Bagging allows one to obtain diversity for each weak learner. In addition, in the process of weak learners training, random forests propose to search for the best split not among all features, but on a random subset of the features.

As presented below, the final class predicted by the model is obtained by majority voting.

Figure 4.5: Illustration of Random Forest



Source:<https://larevueia.fr/random-forest/>

4.3.2 XGBoost: (Tianqi Chen et al., 2016)

XGBoost (Extreme Gradient Boosting) is an ensemble algorithm that combines gradient descent and boosting algorithm. It is a sequential algorithm which tries to improve the model at iteration t based on the obtained one at $t - 1$. In the optimisation process, it minimises a cost function with a penalisation that accounts for model complexity to avoid overfitting. Several hyperparameters can be tuned through a **grid search**. These parameters can be divided into three groups:

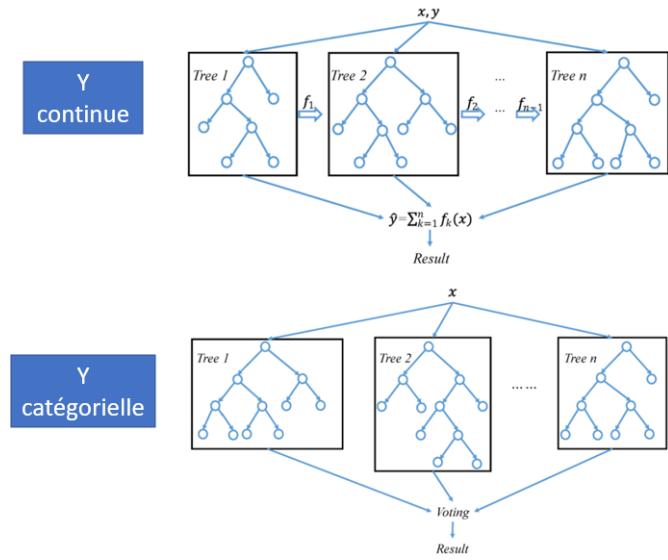
- the ones related to the numeric calculus

- the hyperparameters of the trees (the weak learners)
- and the ones specific to the optimisation.

This algorithm has many advantages. First, the optimisation process account for overfitting which is a real problem in machine learning. Second, it often gives good performance when we have a large amount of data (which is the case here).

However, it is very power-consuming and may require time to train.

Figure 4.6: Illustration of XGBoost



Source:

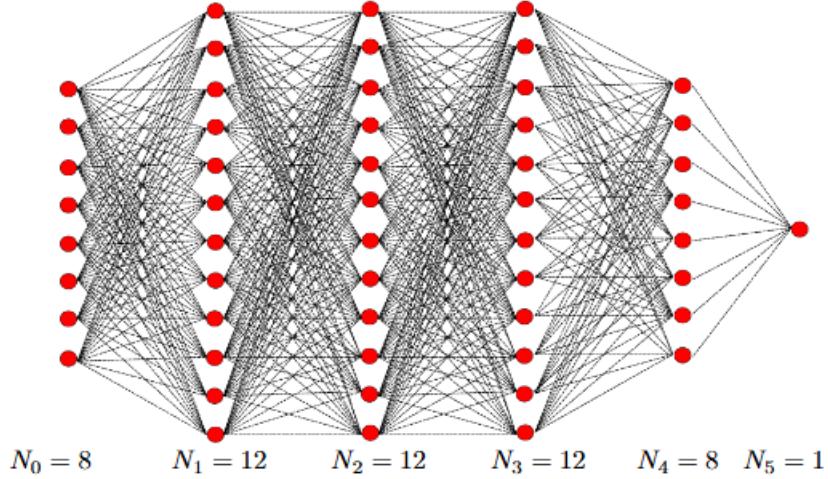
https://www.researchgate.net/figure/A-general-architecture-of-XGBoost_fig3_335483097

4.3.3 Neural network

The artificial neural network is an information processing system that has some performance characteristics similar to biological neural networks. Artificial neural networks are generalizations of mathematical models of human cognition or neural biology, based on several assumptions, including the fact that information processing occurs at the level of many simple elements called neurons or that signals are transmitted between neurons by connecting links.

A neural network is characterized by its pattern of connections between the neurons (called its architecture), its method of determining the weights on the connections (called its training, or learning, algorithm), and its activation function.

Figure 4.7: Illustration of a multi-layer perception with 5 layers. The red dots correspond to the neurons



Source: http://pc-petersen.eu/Neural_Network_Theory.pdf

4.3.4 Results of the machine learning models

With machine learning models we obtained better results with both climate variables and only credit variables. In fact, the random forest model gives an AUC of 0.84 (gini index of 0.68), the XGBoost model gives an AUC of 0.826 (gini index of 0.652) and the neural network model an AUC of 0.81 (gini index of 0.61) as shown in the graphs below with the climate variables (see appendix 4.24 for the model without climate variables).

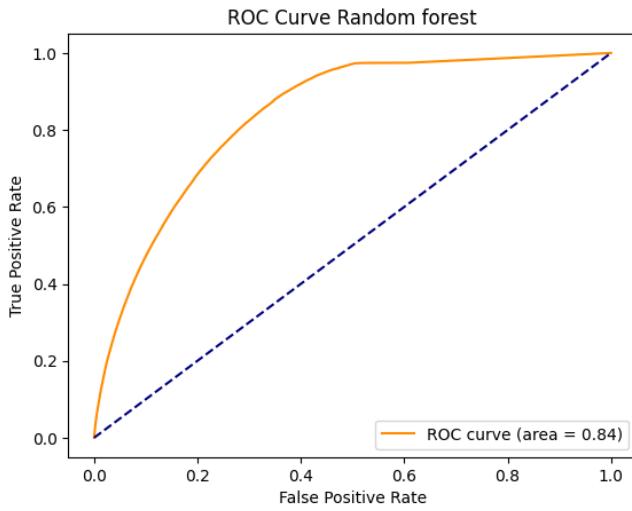


Figure 4.8: ROC curve with Random Forest model

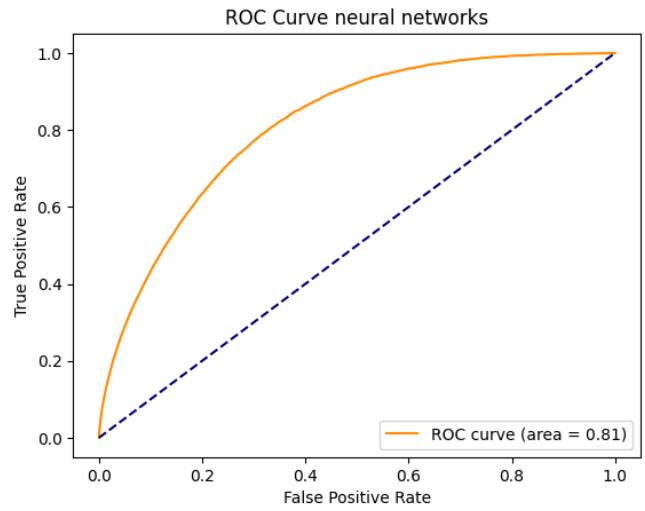


Figure 4.9: ROC curve with Neural Network model

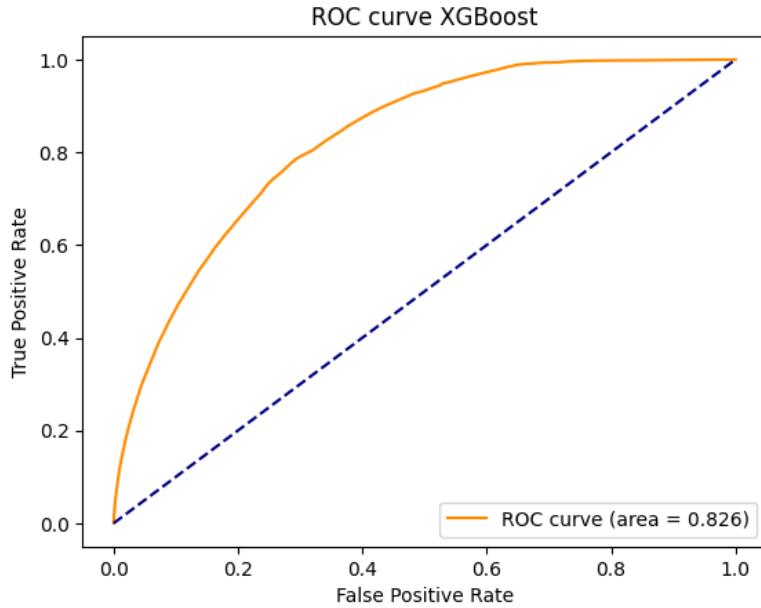


Figure 4.10: ROC curve with XGBoost model

The table below summarizes the performance of the machine learning models. We see that the **Random Forest** model performs best compared to the other two models. It is worth noting that all machine learning models outperform logistic regression.

	Random Forest	XGBoost	Neural Network
AUC	0.84	0.826	0.81
Gini index	0.68	0.652	0.61
Accuracy	0.880	0.883	0.881
Balanced accuracy	0.576	0.557	0.545

Table 4.3: Results from Machine Learning models

Now that we have seen that machine learning models outperform logistic regression in this section, in the next section we will see how to interpret these black box models using Shapley values.

4.3.5 Explanation with Shapley values

The purpose of this section is to explain how the covariables impact the probability of default. To do so, a very used method is the Shapley values. Inspired from game theory, the Shapley values allows to estimate the contribution of each variables by supposing an additivity between each contribution to the estimated function. Let's denote by f the model prediction function, in the framework of the Shap values, the prediction function can be additively decomposed as follow, where p is number of variables and $\phi(i)$ the contribution of each variable and ϵ is the random part :

$$f(x) = \epsilon + \sum_{i=1}^p \phi(i) \quad (4.2)$$

The contributions are estimated as follow:

$$\phi(i) = \sum_{S \subseteq \{1, \dots, p\} / \{i\}} \frac{|S|!(p-1-|S|)!}{p!} (E_S \cup \{i\} - E_S)$$
 (4.3)

This formula is equivalent to:

$$\phi(i) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} E_{before(\pi, i) \cup \{i\}} - E_{before(\pi, i)}$$
 (4.4)

Where:

- $E_S = E[f(x)|x_S = x_{S^*}]$: model prediction with all S variables;
- p : number of covariates in the model;
- Π : the set of arrangements of the p explanatory variables;
- $E_{before(\pi, i)}$: set of variables selected before variable i in order π .

As we can see in equation 4.4, for each arrangement of the p explanatory variables, one estimates the mean contribution of the variable i . That contribution of the variable, the shapley value, will give the type of influence (positive or negative) and the importance of the variable.

However, one drawback of the Shapley values is that it computationally expensive. Also, the Shapley values are symmetrical, meaning that if two variables have the same effect on the model's output, they will receive equal attributions. In practice, the impact of one variable can be decomposed as an indirect effect and direct effect, then the symmetrical axiom will not take it into account.

In figure bellow, we present the estimated Shapley values of variables used in the Random Forest, best model among all tested. We highlighted in green variables with bigger impact on the model's prediction. Results from this model corroborate with those from the logistic regression. At the bottom of the graphic, we note that the **average revenue per capita** is the most important variable. The higher this variable is, the lower the default risk will be. This variable shows that the state's richness has significant impact on client default. The second most important variable is the **notation grade**, as in the logistic regression. The lower the note is (from A to E), the higher the risk is. We can notice that good notation variable can help to grasp of client default, but is not sufficient.

The rise in temperature is also a very important variable in the model. The more the state's temperature rises, the higher the default will be. Increase in temperature encompasses different types of events and may have significant effect on the economy. In our study, by including it directly in the model, we both represent the direct and indirect effect of the temperature. In addition to the economic impact of rise in temperature, it may also be responsible of many deceases and thus will put the counter parties (mainly the particular) in position of high expense and low financial health.

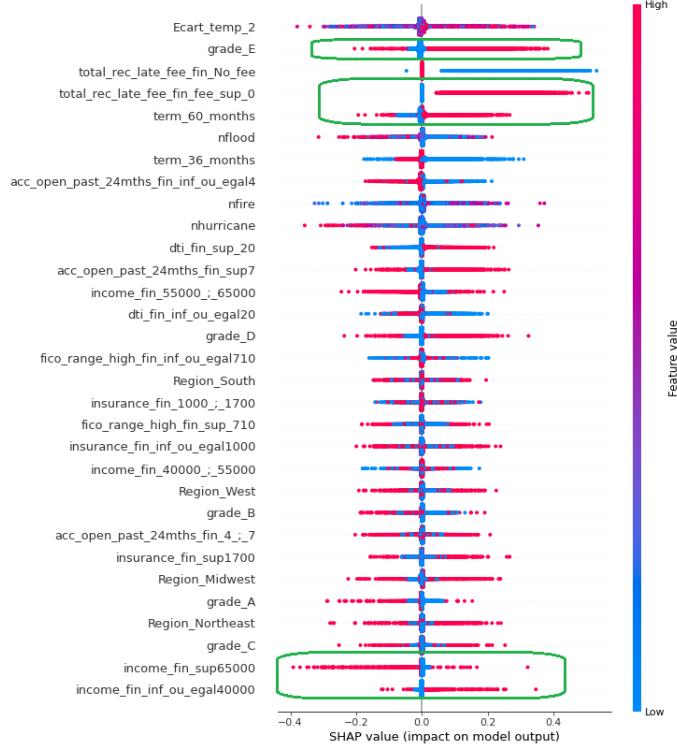


Figure 4.11: Shapley Values from Random Forest Model

In this chapter, we reviewed the theoretical framework of the different models we applied to model the probability of default (PD). We found that machine learning models perform better than logistic regression. Finally, we used Shapley values to better interpret the machine learning models which give same interpretation as the logistic regression. However, it is useful to recall that these machine learning models are difficult to put into practice because even if their interpretation is possible, notably thanks to Shapley values, it is very costly in time and algorithmic complexity. Our results show that climate variable can add more precision in default modelling by increasing the Gini index (4.3).

Our work will now end with a conclusion that will resume the essential elements of this study.

CONCLUSION

Climate risk has become a major issue not only for companies but also for ordinary citizen around the world, as natural disasters occur at an unprecedented rate. This climate risk can have a significant impact on their financial performance and their ability to repay their debt. Taking this risk into account in the calculation of the probability of default is therefore essential for an accurate assessment of financial risks.

First, the literature review allowed us to gather and analyze existing knowledge on the subject, to identify sources of climate risk, to assess the potential impacts on businesses, communities and ecosystems, and to identify the tools and methodologies used to assess climate risk. This literature review therefore allowed us to better understand the state of the art in climate risk research.

Then, after having performed all the necessary treatments and transformations of the variables, we did a modeling through logistic regression. We thus obtained quite satisfactory performances with an AUC of 0.738 (gini index of 0.476) for the model without climate variables and 0.768 (0.534 for the gini index) with climate variables. We thus found that the introduction of climate variables improves the model.

In addition, we estimated the probability of default using machine learning methods. In fact, machine learning models can help to better incorporate climate risk into the calculation of the probability of default (PD) by allowing for a more complete and accurate analysis of climate and financial data while overcoming the very strong assumptions of logistic regression. Thanks to these models, we obtained better performances compared to the logistic regression with an AUC of 0.84 (0.68 for Gini index) for the Random Forest model for example.

One of the main conclusions of this study is the existence of an impact of climate risk on a borrower's ability to repay. Some variables, notably the occurrence of claims, show effects contrary to what one might expect. However, these results should be treated with caution because the study has certain methodological limitations. First, the choice of the state level for the granularity of the climate data does not necessarily allow the effect of climate variables to be correctly captured in the models. Also, the choice of temporality in the constitution of certain climate variables could have been different and provided different effects. In the data processing step, the discretization of the variables could be done with other thresholds that could make the model more discriminating.

Finally, it is important to note that the use of logistic regression or machine learning models for climate risk assessment also has challenges, including data quality and reliability, biases and uncertainties associated with the models, and interpretability for the machine learning models. Therefore, companies should use these models with caution and supplement them with other risk assessment tools to ensure a complete and accurate assessment of climate risk.

BIBLIOGRAPHY

- [1] Ulrich and al. L'impact du risque climatique sur la modélisation de la pd, 2022. [Athea conseil].
- [2] Carolyn Kousky and al. Flood damage and mortgage credit risk: A case study of hurricane harvey, 2020. [Journal of Housing Research].
- [3] Pierre Monnin. Integrating climate risks into credit risk assessment: Current methodologies and the case of central banks corporate bond purchases, 2018. [Research Gate].
- [4] R.M. Walle. macro-econometric model for climate-related credit risk, 2020. [Erasmus University Rotterdam].
- [5] Claire Zhang. How to integrate climate risk into credit risk assessments. <https://towardsdatascience.com/how-to-integrate-climate-risk-into-credit-risk-assessments-f7eac5c01850>, 2021.
- [6] Banque de France. Élaborer des scénarios de transition climatique pour gérer les risques financiers, 2021.
- [7] Novela and al. The impact of climate change on credit risk, 2022. [LUISS University].
- [8] Andrea Cruz and al. Integrating climate risk into credit risk modeling, 2019.
- [9] Basel Committee on Banking Supervision. Climate-related financial risks – measurement methodologies, 2021.

APPENDIX

4.4 Correlation coefficient

Pearson's correlation coefficient is an indicator of the existence of a linear relationship between two quantitative variables. It varies between -1 and 1. A negative correlation means that when one of the variables increases, the other decreases; while a positive correlation indicates that both variables move in the same direction. It is calculated by the following formula:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

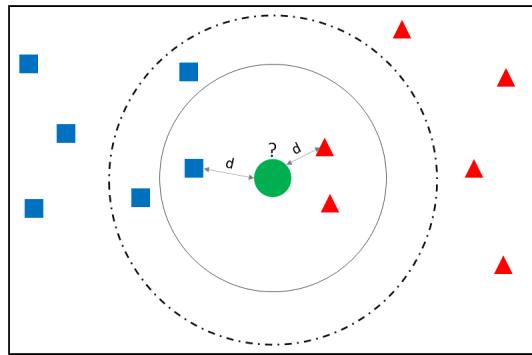
The Spearman correlation coefficient plays the same role as the Pearson correlation coefficient with the only difference being that it detects the existence of any relationship. For a sample of size n , the rank variables (rgX_i , rgY_i) are calculated from the data (X_i, Y_i) . The Spearman correlation is defined by:

$$r_s = \frac{cov(rgX, rgY)}{\sigma_{rgX} \sigma_{rgY}}$$

where $cov(rgX, rgY)$ is the covariance of rank variables, and $\sigma_{rgX}, \sigma_{rgY}$ are the standard deviations of rank variables

4.5 K-nearest neighbor algorithm (KNN)

The k-nearest neighbors algorithm (KNN) is a supervised learning method used for classification and regression. The basic principle of the KNN algorithm is to find the k nearest observations of a new sample to be classified, and then to assign the class (or the value for a regression) from these observations. To calculate the proximity between samples, the KNN algorithm uses a distance measure such as the Euclidean distance. The value of k is chosen based on the data and the desired accuracy. For a regression problem, the prediction is obtained from the average of the nearest neighbors.



Source: KNN imputer algorithm

4.6 Correlation to the target variable

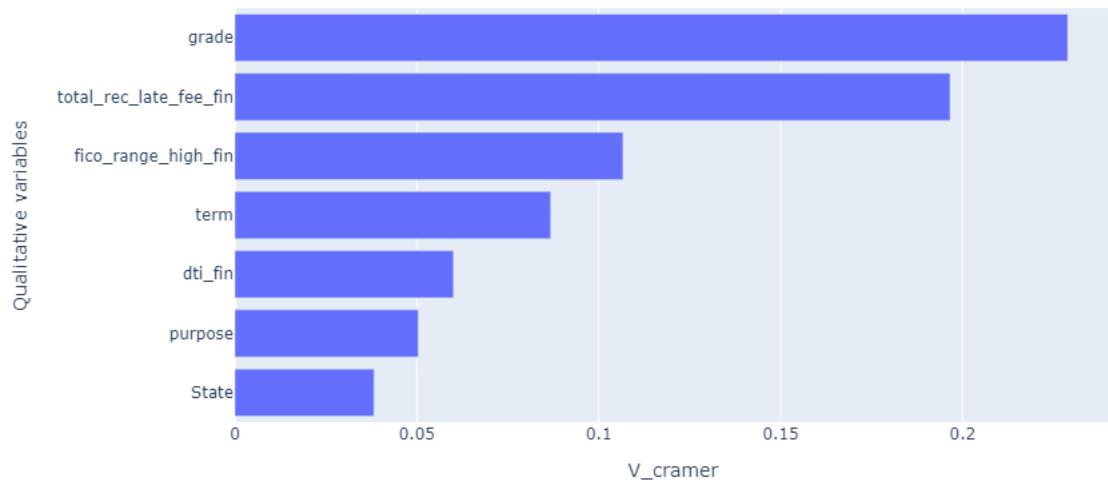


Figure 4.12: Correlation between categorical variables and the default indicator

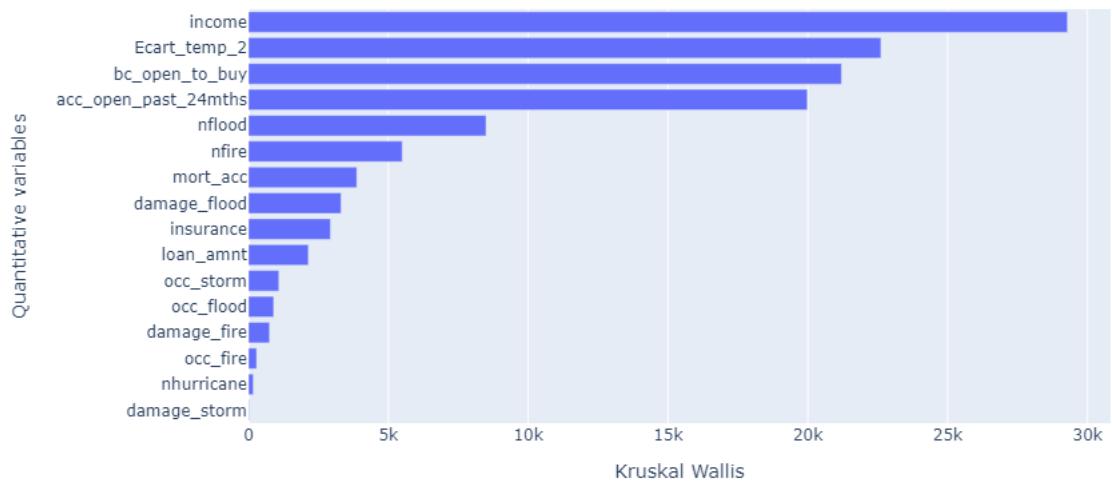


Figure 4.13: Correlation between quantitative variables and the default indicator

4.7 Descriptive charts

4.7.1 Grade: univariate distribution

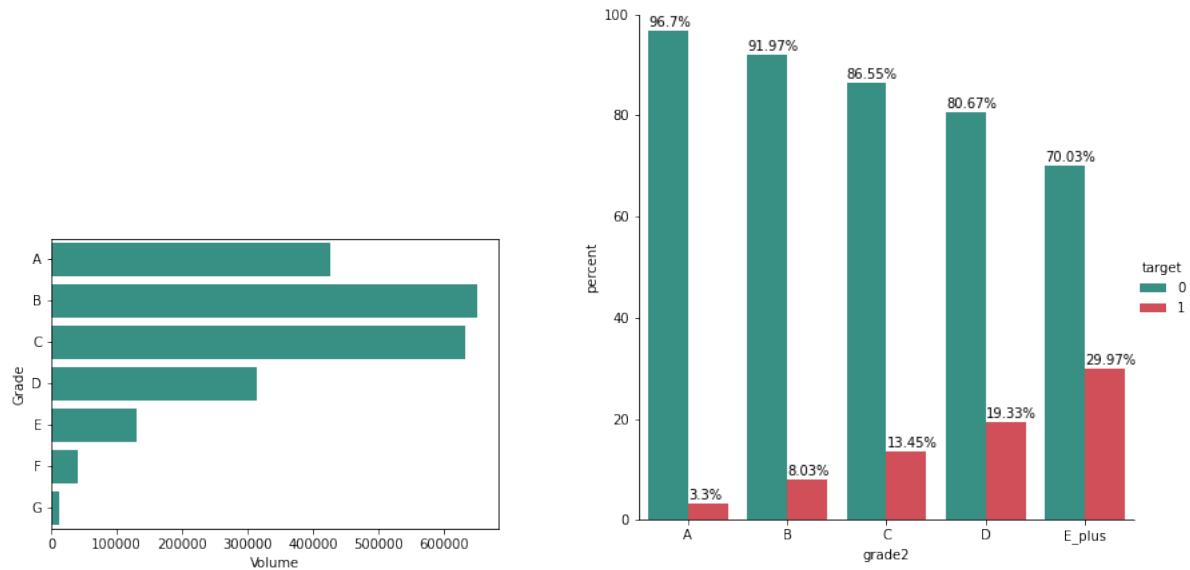


Figure 4.14: Distribution according the grade

4.7.2 Late fees received to date

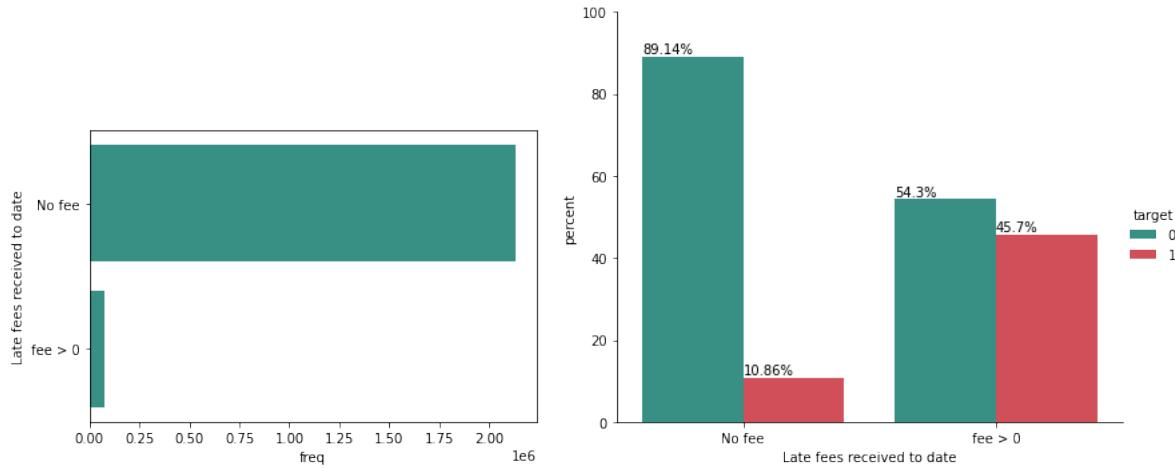


Figure 4.15: Distribution of the late fees received to date

4.7.3 Fico score

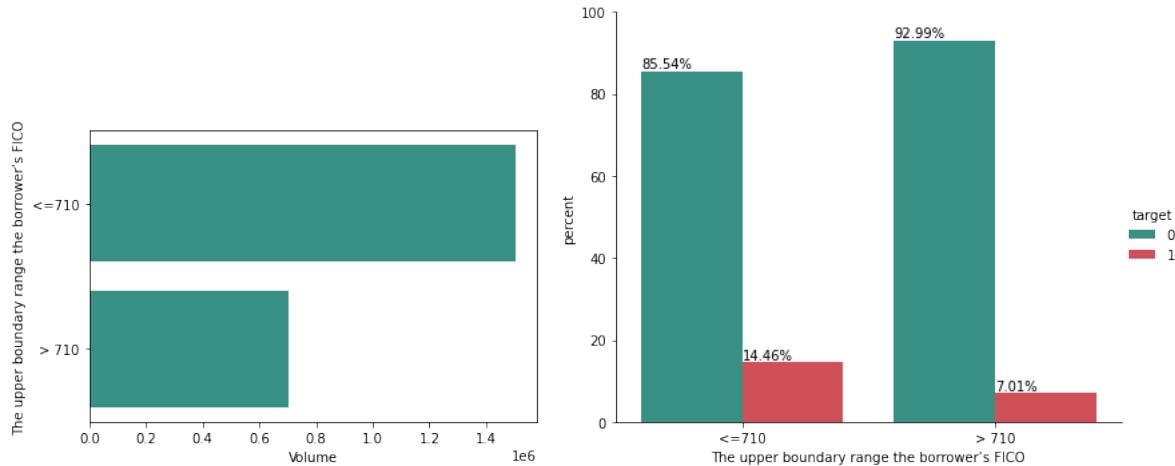


Figure 4.16: Distribution of the upper boundary of the Fico score

4.7.4 Loan term: number of payments of the loan

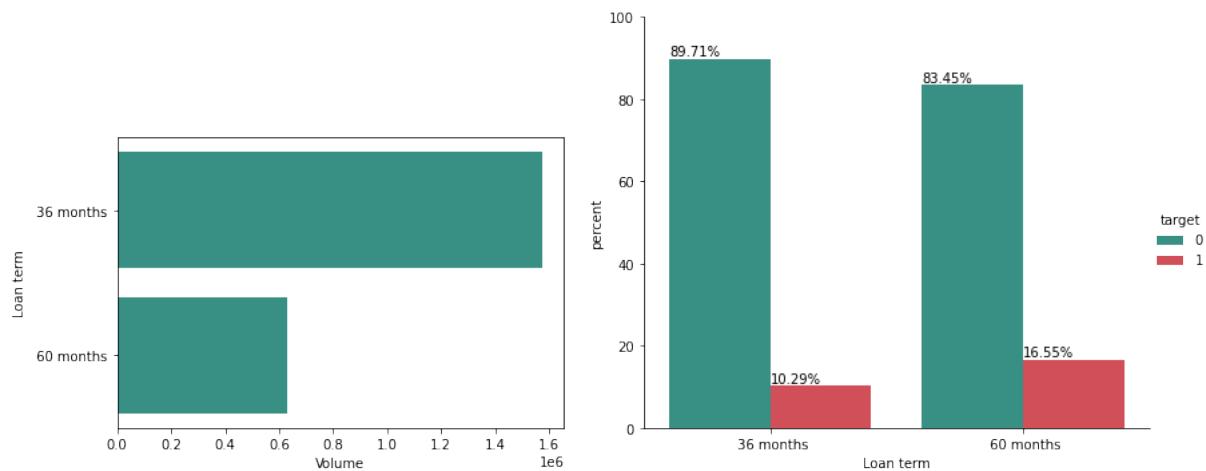


Figure 4.17: Loan term distribution

4.7.5 Monthly debt ratio

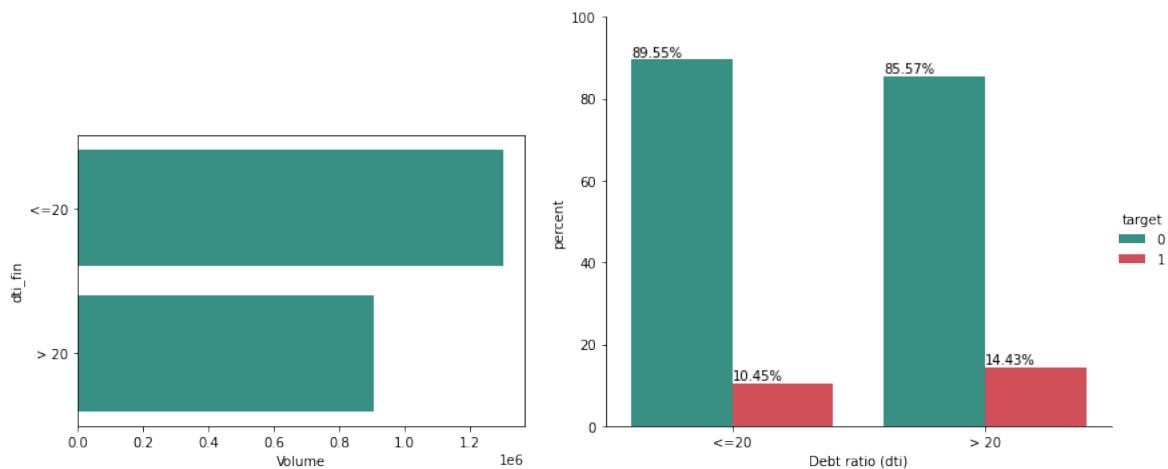


Figure 4.18: Distribution of the monthly debt ratio

4.7.6 Average revenue per capita

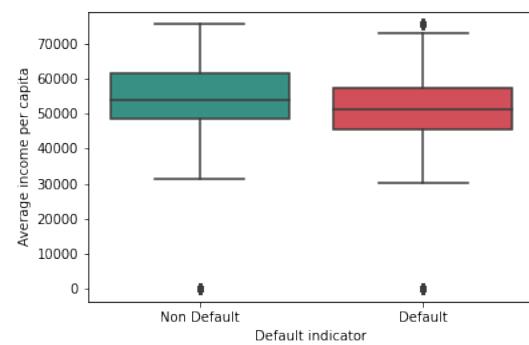


Figure 4.19: Distribution of revenue per capita

4.7.7 Number of trades opened in past 24 months

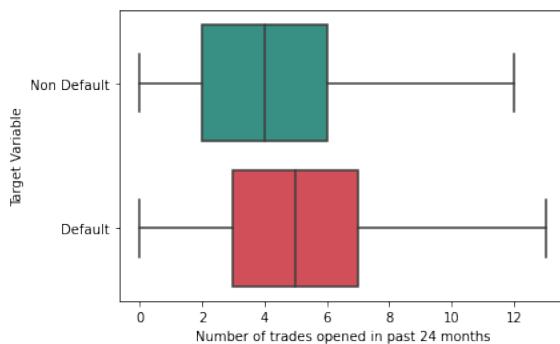


Figure 4.20: Number of trades opened in past 24 months

4.7.8 Region

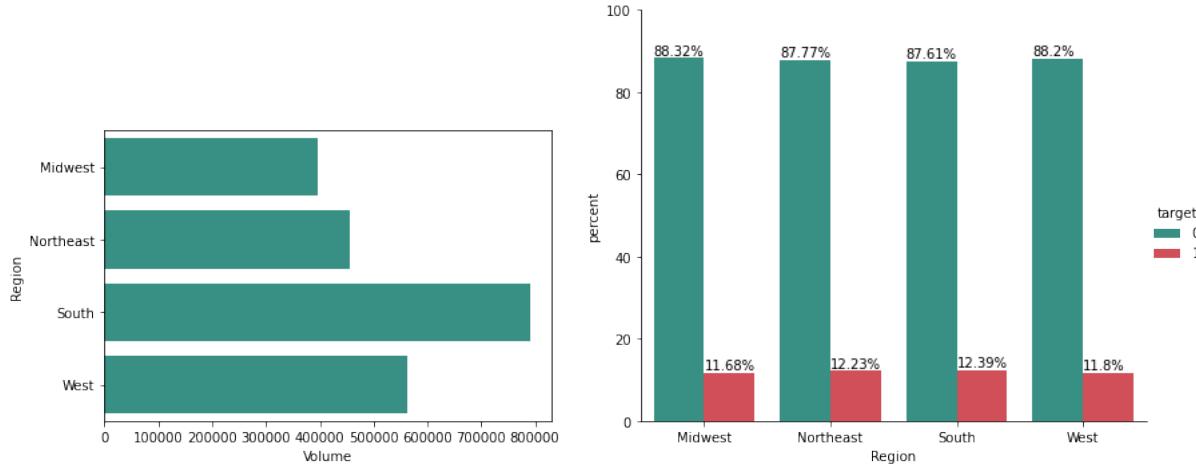


Figure 4.21: Distribution of the Region variable

4.8 Performances on machine learning models without climate related variables

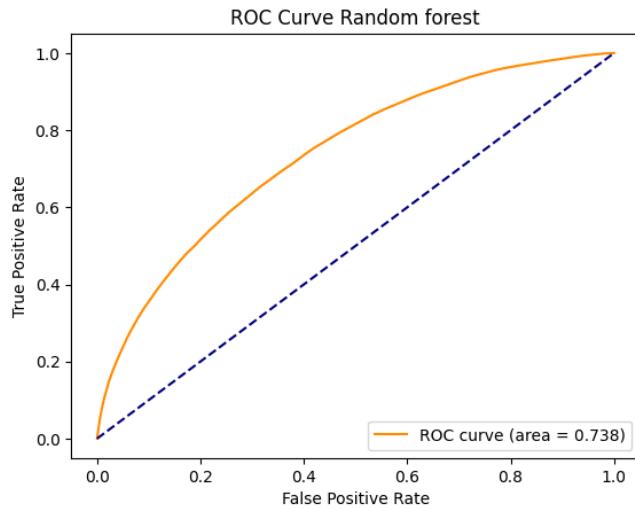


Figure 4.22: ROC curve of Random Forest model without climate variables

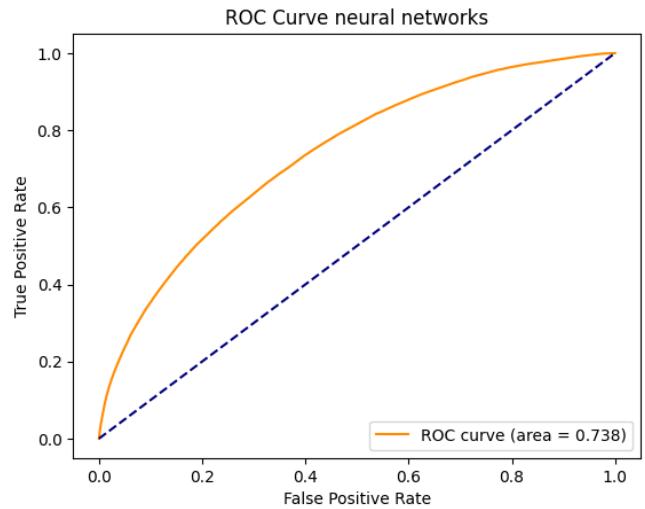


Figure 4.23: ROC curve of Neural Network model without climate variables

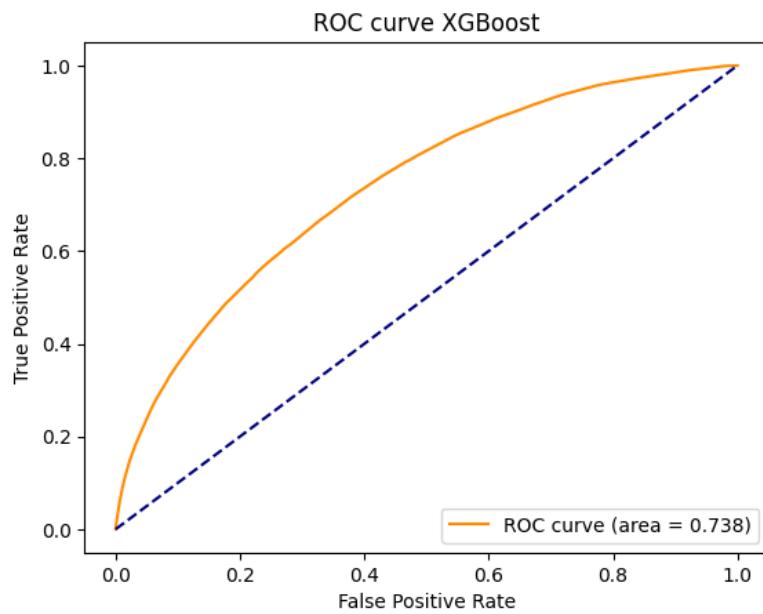


Figure 4.24: ROC curve of XGBoost model without climate variables