Scoring in the Bank Industry

Alexandra AMANI Colombe BECQUART Marie PHILIPPE Claire SERRAZ

M2 D3S & SE, TSE

March - April 2022

Contents

List of Figures

List of Tables

1 Introduction

Credit is associated with a risk which can cause losses. These are, for instance, customers not refunding their loans. The losses are covered by the provisions and the equity capital of the bank, depending on their nature. Not only banks need to anticipate their losses to manage their resources, but risk management is also a legal obligation. Indeed, banks must assess their exposure at default, which is the predicted amount of loss they can be exposed in case of default for a customer. Thus, banks use statistical models to measure the risk and anticipate nonpayment, using information about their clients.

The aim of this report is to implement a model to predict the probability of default and give a score to the customers. The model is trained on current and old customers to then be applied on potential customers. A score is a function that aims to measure the risk of customers defaulting (i.e. not refunding its loan). This score is explained by several features that may increase or decrease the risk of default for a given customer. It can be used in decisions regarding granting of credits or pricing of the rates for loans. To achieve this objective, cross-sectional data are available, containing information about clients with a credit in the bank. The database is at the client level. There are 99 variables: one ID client, *top_def_12m_90* is the variable to explain, it is a dummy variable equal to 1 if the situation within the 12 months after the observation date is doubtfull and 0 otherwise. There are 28 qualitative explanatory variables about the situation of the clients and the type of accounts they hold; and there are 69 quantitative explanatory variables such as the age of the client, the length of the customer relationship as well as other indicators about their accounts.

First, the available data is described and explored. The methodology of variables selection/creation is then detailed. After the presentation of the dataset, one selects the variables. For the model, one has to discretize the quantitative variables and group the modalities of the qualitative ones in order to have balanced classes. The variables that are similar must be grouped. Then, a model is built to explain the default status. One wants to have significant variables and significant modalities to explain the default status. Finally, scoring techniques are used to segment the clients, their score reflecting their associated risk.

2 Data cleaning and summary statistics

2.1 Presentation of the dataset

This first subsection aims to present the dataset without doing any analysis yet. The dataset includes 91 105 observations, one observation being a client, and 99 variables. The variable *top def 12m 90j* being the variable of interest and the variable *matricule* being the matricule of the client, it remains 97 explanatory variables. This first part aims to leave behind some variables that can't be exploited and select around 10 variables that will help us later to build scores. Among these 97 variables, there are 69 quantitative variables and 28 qualitative variables. The rest of this first section is divided between the analysis and selection of qualitative variables and then the analysis and selection of quantitative variables. Also, it is important to point out that the dataset is in french.

The repartition of the variable of interest is as follows:

top_def_12m_90j				
top_def_12m_90j	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	89745	98.51	89745	98.51
1	1360	1.49	91105	100.00

Figure 1: Frequency table of the variable of interest

2.2 Selection of the qualitative variables

This part aims to drop the unexploitable qualitative variables to keep only the most relevant ones. First, one will look at the univariate statistics and then the bivariate statistics will be exploited on the remained candidates. As a reminder, there are 28 qualitative variables.

2.2.1 Using the univariate analysis

Univariate analysis aims to describe and summarize the variables to ease the understanding of the dataset. Here, to analyze the qualitative variables, one will look at the frequencies of each modality.

The decision to leave behind a variable lays on the proportion of missing values, and on the fact that a modality of the variable has a frequency higher than 95%.

One can notice on the table below that the variables *statut_juridique* and *statut_eco* could be dropped using that criterion, as the first one has 97% of the observations in the modality "Personne physique" and as the second one has almost 95% of the observations in the modality "Particulier". The analysis done here is mostly interesting for the physical person rather than the moral person,

i.e. the institute, association and for individuals ("particulier"). Thus, a subdataset is created, only including these two modalities: "Particuliers" for the variable *statut_eco* and "Personne physique" for the variable *statut_juridique*. After that, the two variables are dropped.

statut_juridique				
statut_juridique	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Non renseigné	12	0.01	12	0.01
Personne morale	2488	2.73	2500	2.74
Personne physique	88605	97.26	91105	100.00

Figure 2: Frequency table for the variables statut_juridique

statut_eco				
statut_eco	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Adm prive	1	0.00	1	0.00
Association	7	0.01	8	0.01
Entrepreneur individuel	7	0.01	15	0.02
Institut financier	1	0.00	16	0.02
Non renseigné	12	0.01	28	0.03
Particulier	85737	94.11	85765	94.14
Personnel et assimile	1977	2.17	87742	96.31
Profession libérale	884	0.97	88626	97.28
Société	40	0.04	88666	97.32
Société civile	2436	2.67	91102	100.00
Sociétés offshores	1	0.00	91103	100.00
Trust et fiducie	2	0.00	91105	100.00

Figure 3: Frequency table for the variables statut_eco

The next variable excluded from the analysis is the *situation_matrimonial*. Indeed, there are 39.38% of missing values ('Non renseigné'), as seen in the following table.

sit_matrimonial				
sit_matrimonial	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Anc. communaute acquets	874	1.02	874	1.02
Communauté universelle	5158	6.02	6032	7.04
Indivision de biens	6	0.01	6038	7.04
Non renseigné	33761	39.38	39799	46.42
Participation aux acquets	1247	1.45	41046	47.87
Regime legal	28410	33.14	69456	81.01
Séparation de biens	16281	18.99	85737	100.00

Figure 4: Frequency table for the variable situation_matrimonial

Then, one drops the 13 following variables for the following reasons:

- **top_pret_infine**: There are 99.73% of the observations in the modality 0.
- **top_pret_relais**: There are 99.99% of the observations in the modality 0.
- **top_Gar_autre**: There are around 99.2% of the observations in the modality 0.
- **top_Gar_Hyp**: There are 95.32% of the observations in the modality 0.
- **top_Gar_Crelog**: There are 100% of the observations in the modality 0.
- **top_DAV**: There are around 98.88% of the observations in the modality 1.
- **top_fac_caisse**: There are 99.98% of the observations in the modality 0.
- **topEpargne**: There are 99.42% of the observations in the modality 1.
- **topCred**: There are 95.49% of the observations in the modality 0.
- **topAutre**: There are 99.41% of the observations in the modality 1.
- **topService**: There are 96.79% of the observations in the modality 0.
- **topDecNonAuto**: There are 95.51% of the observations in the modality 0.
- **topDecAuto**: There are 99.86% of the observations in the modality 0.

After this first cleaning, it remains 85737 observations and 12 qualitative candidates.

2.2.2 Using the bivariate analysis

The bivariate analysis aims to explore the relationship of the explanatory variables and the variable of interest, to measure the strength of this relationship, if any.

In the case of qualitative variables, one looks at the Cramer's V test. The formula is as follows:

$Cramer'V = \sqrt{\frac{(\frac{X^2}{n})}{min(c-1,r-1)}}$

With:

- X^2: The Chi-square statistic
- n: Total sample size
- r: Number of rows
- c: Number of columns

The interpretation is as follows:

- If $\|V\|$ tends to 0, then the two variables are not dependent.
- if $\|V\|$ tends to 1, the two variables have a strong dependence.

The table below shows the Cramer's V statistic for the 12 remaining candidates. It measures their dependence with the variable of interest.

Obs.	Value	ABS_V_CRAMER	Variable
1	-0.0471	0.04713	topFacilite
2	0.0438	0.04382	CSP
3	-0.0263	0.02629	topAssuIARD
4	0.0181	0.01806	type_compte
5	0.0166	0.01660	sit_familiale
6	-0.0143	0.01434	top_Gar_Cnp
7	-0.0134	0.01335	top_compte_joint
8	-0.0129	0.01294	topAssurVIE
9	-0.0125	0.01254	topGestionPTF
10	-0.0073	0.00730	topCreditImmo
11	-0.0038	0.00382	top_credit
12	-0.0004	0.00040	topTitrePEA

Figure 5: Table of the Cramer's V-test

The first line concerns the link of the variable of interest with itself and shall not be looked at. One decides to only keep the eight variables with the highest statistic. Those eight candidates are

then: *topFacilite*, *CSP*, *topAssuIARD*, *type_compte*, *sit_familiale*, *top_gar_Cnp*, *top_compte_joint* and *topAssurVIE*.

One can notice that the Cramer's V statistic between the variable of interest and the variables *topFacilite* and *CSP* is the highest with a value around 0.04. All the other statistics are bounded between 0.01 and 0.027.

Nonetheless, it is also important to check that these variables don't give the same information, i.e have a Cramer's V test statistic higher than 0.8 between them. It is the case for the *type_compte* and *top_compte_joint* with a statistic around 0.98. This result isn't surprising since, as one can see on the contingency table, they both have two very similar modalities:

- 0 and "Compte individuel uniquement"
- 1 and "Compte joint uniquement"

Table de type_compte par top_compte_joint

type_compte(type_compte)	top_compte_joint(top_compte_joint) 0	1	Total
Compte individuel uniquement	46047 53.71 98.28 100.00	805 0.94 1.72 2.03	46852 54.65
Compte joint et individuel	0 0.00 0.00 0.00	4950 5.77 100.00 12.47	4950 5.77
Compte joint uniquement	0 0.00 0.00 0.00	33935 39.58 100.00 85.50	33935 39.58
Total	46047 53.71	39690 46.29	85737 100.00

Figure 6: Contingency table between type_compte and top_compte_joint

Since these variables give the same information one of them has to be delete. Only *type_compte* is kept since it is the variable with the highest Cramer's V test statistics, i.e has the highest dependence with the variable of interest. Moreover, it is the most precise variable since it has three modalities.

Finally, after an univariate and a bivariate analysis on the qualitative variables, these seven variables are kept:

- **topFacilite**: It is a binary variable equal to 1 if the customer has a credit card and 0 otherwise.
- **CSP**: It is the socio-professional category.

- **topAssuIARD**: It is a binary variable that shows if the customer owns an IARD insurance (=1) or not (=0). In France, IARD stands for "Incendie, Accidents et Risques Divers". It is an insurance related to the goods protection.
- **type_compte**: It is a binary variable that shows the type of the bank account owned by the consumer (only individual account, joint and shared account or only joint account).
- **sit_familiale**: It is the family situation (single, married, divorced...)
- **top_Gar_Cnp**: It is a binary variable that shows if the consumer owns a CNP guarantee (=1) or not (=0). A CNP guarantee is linked to a CNP insurance that concerns the home loan and the risk of deaths.
- **topAssurVIE**: It is a binary variable equal to 1 if a life insurance is hold and 0 otherwise.

The univariate and bivariate analysis enable to keep 85 737 observations and 7 qualitative variables.

2.2.3 Modalities creation

One might need to group some modalities of the seven qualitative variables that are kept. Indeed, one wants at least 5% of the clients in each modality and one wants all modalities to be significant, so one has to group modalities with a similar effect on the variable of interest *top_def_12m_90j*.

The variables *topFacilite*, *topAssuIARD*, *top_Gar_Cnp* and *top_Assur_VIE* are binary. As one can see on Figures 7 to 10, each modality of these binary variables represent at least 5% of the clients.

Table de topFacilite par top_def_12m_90j

topFacilite(topFacilite)	top_def_12m_90j(top_def_12m_90j) 0	1	Total
0	62346 73.91 98.15 75.05	1173 1.39 1.85 91.93	63519 75.30
1	20731 24.58 99.51 24.95	103 0.12 0.49 8.07	20834 24.70
Total	83077 98.49	1276 1.51	84353 100.00

Figure 7: Contingency table between topFacilite and top_def_12m_90j

Table de topAssuIARD par top_def_12m_90j

topAssuIARD(topAssuIARD)	top_def_12m_90j(top_def_12m_90j) 0	1	Total
0	56702 67.22 98.27 68.25	1000 1.19 1.73 78.37	57702 68.41
1	26375 31.27 98.96 31.75	276 0.33 1.04 21.63	26651 31.59
Total	83077 98.49	1276 1.51	84353 100.00

Figure 8: Contingency table between topAssuIARD and top_def_12m_90j

Table de top_Gar_Cnp par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
top_Gar_Cnp(top_Gar_Cnp)	0	1	Total
0	78226 92.74 98.44 94.16	1237 1.47 1.56 96.94	79463 94.20
1	4851 5.75 99.20 5.84	39 0.05 0.80 3.06	4890 5.80
Total	83077 98.49	1276 1.51	84353 100.00

Figure 9: Contingency table between top_Gar_Cnp and top_def_12m_90j

Table de topAssurVIE par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
topAssurVIE(topAssurVIE)	0	1	Total
0	78188 92.69 98.45 94.12	1233 1.46 1.55 96.63	79421 94.15
1	4889 5.80 99.13 5.88	43 0.05 0.87 3.37	4932 5.85
Total	83077 98.49	1276 1.51	84353 100.00

Figure 10: Contingency table between top_Assur_VIE and top_def_12m_90j

The variable *type_compte* has 3 modalities. Figure 11 shows that each modality contains at least 5% of the clients and the impacts on *top_def_12m_90j* seem different enough so there is no need to group the modalities here. Even if there is no modality to group for these five variables, one creates new variables to rename the modalities.

For the binary variables, *topAssurVIE2*, *topAssuIARD2*, *top_Gar_Cnp2* and *topFacilite2* are created. They are equal to "1. avec" if the corresponding dummy is equal to 1 and they are equal to "2. sans" otherwise. Similarly, *type_compte2* is created with the modalities "1. Compte individuel uniquement", "2. Compte joint et individuel" and "3. Compte joint uniquement".

Table de type_compte2 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
type_compte2	0	1	Total
1. Compte individuel uniquement	44740 53.04 98.31 53.85	768 0.91 1.69 60.19	45508 53.95
2. Compte joint et individuel	4862 5.76 98.22 5.85	88 0.10 1.78 6.90	4950 5.87
3. Compte joint uniquement	33475 39.68 98.76 40.29	420 0.50 1.24 32.92	33895 40.18
Total	83077 98.49	1276 1.51	84353 100.00

Figure 11: Contingency table between type_compte and top_def_12m_90

Figure 14 shows that the variable *CSP* has 10 modalities and they need to be grouped because some of them do not contain 5% of clients. Thus, one creates a new variable named *CSP2* whose modalities represent at least 5% of clients. To group the modalities, one looks for similar percentages of clients in default in each modality. The groups in *CSP2* are as follows:

- "1. Ouvriers agriculteurs": if *CSP* is "Agriculteurs","Artisans, commercants, chef entreprise","Autre" or "Ouvriers",
- "2. Cadres": if *CSP* is "Cadre, profession superieures" or "Professions intermediaires",
- "3. Autres": if *CSP* is "Sans activite","Employes" or "Non renseigne"
- "4. Retraites": if *CSP* is "Retraites".

One repeats this method for *sit_familiale* whose modalities are displayed in Figure 13. One creates a new variable named *sit_familiale2*. The groups are the following:

- "1. Celibataire seul divorce": if *sit_familiale* is "Célibataire","Divorcé(e)" or "Séparé(e)"
- "2. Autres": if *sit_familiale* is Non renseigné","Veuf(ve)","Pacs"
- "3. Marie": if *sit_familiale* is "Marié(e)"

Table de CSP par top_def_12m_90j

CSP(CSP)	top_def_12m_90j(top_def_12m_90j) 0	1	Total
Agriculteurs	179 0.21 96.24 0.22	7 0.01 3.76 0.55	186 0.22
Artisans, commercants, chef entreprise	4245 5.03 97.36 5.11	115 0.14 2.64 9.01	4360 5.17
Autre	1580 1.87 96.93 1.90	50 0.06 3.07 3.92	1630 1.93
Cadre, profession superieures	26305 31.18 98.29 31.66	459 0.54 1.71 35.97	26764 31.73
Employes	4484 5.32 98.31 5.40	77 0.09 1.69 6.03	4561 5.41
Non renseigné	1 0.00 100.00 0.00	0 0.00 0.00 0.00	1 0.00
Ouvriers	249 0.30 95.77 0.30	11 0.01 4.23 0.86	260 0.31
Professions intermediaires	10678 12.66 98.28 12.85	187 0.22 1.72 14.66	10865 12.88
Retraites	23142 27.43 99.17 27.86	194 0.23 0.83 15.20	23336 27.66
Sans activite	12214 14.48 98.58 14.70	176 0.21 1.42 13.79	12390 14.69
Total	83077 98.49	1276 1.51	84353 100.00

Figure 12: Contingency table between CSP and top_def_12m_90j

Table de sit_familiale par top_def_12m_90j

sit_familiale(sit_familiale)	top_def_12m_90j(top_def_12m_90j) 0	1	Total
Célibataire	21229 25.17 98.38 25.55	350 0.41 1.62 27.43	21579 25.58
Divorcé(e)	5921 7.02 98.37 7.13	98 0.12 1.63 7.68	6019 7.14
Marié(e)	51166 60.66 98.50 61.59	780 0.92 1.50 61.13	51946 61.58
Non renseign	100 0.12 100.00 0.12	0 0.00 0.00 0.00	100 0.12
Pacs	36 0.04 100.00 0.04	0 0.00 0.00 0.00	36 0.04
Séparé(e)	288 0.34 96.64 0.35	10 0.01 3.36 0.78	298 0.35
Veuf(ve)	4337 5.14 99.13 5.22	38 0.05 0.87 2.98	4375 5.19
Total	83077 98.49	1276 1.51	84353 100.00

Figure 13: Contingency table between sit_familiale and top_def_12m_90j

The contingency tables between the new variables and the variable of interest are displayed in Figures 14 and 15.

Table de CSP2 par top_def_12m_90j			
CSP2	top_def_12m_90j(top_def_12m_90j)		
	0	1	Total
1. Ouvriers agriculteurs	6253 7.41 97.16 7.53	183 0.22 2.84 14.34	6436 7.63
2. Cadres	36983 43.84 98.28 44.52	646 0.77 1.72 50.63	37629 44.61
3. Autres	16698 19.80 98.51 20.10	253 0.30 1.49 19.83	16951 20.10
4. Retraites	23143 27.44 99.17 27.86	194 0.23 0.83 15.20	23337 27.67
Total	83077 98.49	1276 1.51	84353 100.00

Figure 14: Contingency table between CSP2 and top_def_12m_90j

Table de sit_familiale2 par top_def_12m_90j			
sit_familiale2	top_def_12m_90j(top_def_12m_90j)		
	0	1	Total
1. Celibataire seul divorce	27438 32.53 98.36 33.03	458 0.54 1.64 35.89	27896 33.07
2. Autres	4373 5.18 99.14 5.26	38 0.05 0.86 2.98	4411 5.23
3. Marie	51266 60.78 98.50 61.71	780 0.92 1.50 61.13	52046 61.70
Total	83077 98.49	1276 1.51	84353 100.00

Figure 15: Contingency table between sit_familiale2 and top_def_12m_90j

2.3 Selection of the quantitative variables

As done previously for the qualitative variables, a selection of quantitative variables is done. To do so, once again an univariate analysis is performed, and then a bivariate one. As said previously, there are 69 quantitative variables.

2.3.1 Using the univariate analysis

The univariate analysis consists most of all in looking at the summary statistics (minimum, maximum, mean) as well as the distribution of the variables.

When looking at the minimum and maximum of the age, one can notice that babies seem to be customers as well as very old people.

Variable d'analyse : age age				
N	Moyenne	Ec-type	Minimum	Maximum
84404	61.5484100	16.7372172	0	137.0000000

Figure 16: Summary statistics of the variable age

Since people being 137 years old are rare, one has a closer look at the frequency of this variable.

age				
age	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	2	0.00	2	0.00
2	1	0.00	3	0.00
3	13	0.02	16	0.02
4	25	0.03	41	0.05
5	34	0.04	75	0.09
109	9	0.01	84380	99.97
120	5	0.01	84385	99.98
121	5	0.01	84390	99.98
122	1	0.00	84391	99.98
123	1	0.00	84392	99.99
124	1	0.00	84393	99.99
125	1	0.00	84394	99.99
126	3	0.00	84397	99.99
127	1	0.00	84398	99.99
128	3	0.00	84401	100.00
130	1	0.00	84402	100.00
133	1	0.00	84403	100.00
137	1	0.00	84404	100.00

Figure 17: Frequency table of the age (from 0 to 5 and 109 to 137 only)

The frequency table in Figure 17 shows there is a gap of people being 109 and 120 years old. Thus, all the individuals being more than 109 are deleted as they are likely to be mistakes. People of the age of 0 and 2 are deleted as well because it doesn't make sense to be a bank's customer at this age. The individuals with missing ages are deleted too.

Not seeing more incoherence for the moment, one decides to look at the probability distribution of the variables. To see if some variables are very asymmetric, one looks at the quantiles. One considers three quantiles: Q1-25%, Q2-50%, and Q3-75%. For some variables, it happens that until Q3, the other quantiles are equal to 0 or only Q1 is different from 0. In such case the variable is deleted. It means that the variable is very right-skewed or left-skewed. Then the mean is often much smaller or larger than the variance.

Among the 69 variables, quite a lot have asymmetric distributions. Indeed, 22 are highly right-skewed and 5 are highly left-skewed. One can see on the table below the quantiles for the variables *montant_garantie*, that has a right-skewed distribution, and *capital_restant_du*, that has a left-skewed distribution.

Quantiles (Définition 5)	
Niveau	Quantile
100Max 100%	0.00
99%	0.00
95%	0.00
90%	0.00
75% Q3	0.00
50% Médiane	0.00
25% Q1	0.00
10%	-9783.16
5%	-65769.03
1%	-183329.30
0% Min	-1987853.93

Figure 18: Quantiles of the variable capital_restant_du

One can clearly see that most of the quantiles are equal to 0, especially Q2 and Q3. In the case of the variable *capital_restant_du*, the mean is -9200.37 whereas the median is 0 which proves once again that this variable is left-skewed. Only a very few individuals have a lot of capital they still owe.

Quantiles (Définition 5)	
Niveau	Quantile
100Max 100%	1.11111E+11
99%	3.37709E+05
95%	4.70465E+04
90%	0.00000E+00
75% Q3	0.00000E+00
50% Médiane	0.00000E+00
25% Q1	0.00000E+00
10%	0.00000E+00
5%	0.00000E+00
1%	0.00000E+00
0% Min	0.00000E+00

Figure 19: Quantiles of the variable montant_garantie

For the variable *montant_garantie* we have the same issue but this time Q1, Q2, and Q3 are equal to 0. It means a very small part of individuals is concerned by a rather high guaranty. Actually the mean is 1416839. This proves this variable is right-skewed.

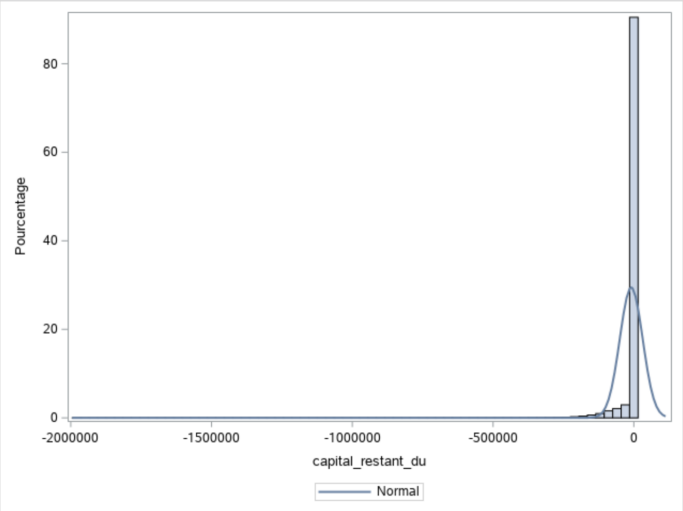

Figure 20: Histogram and density of the capital_restant_du

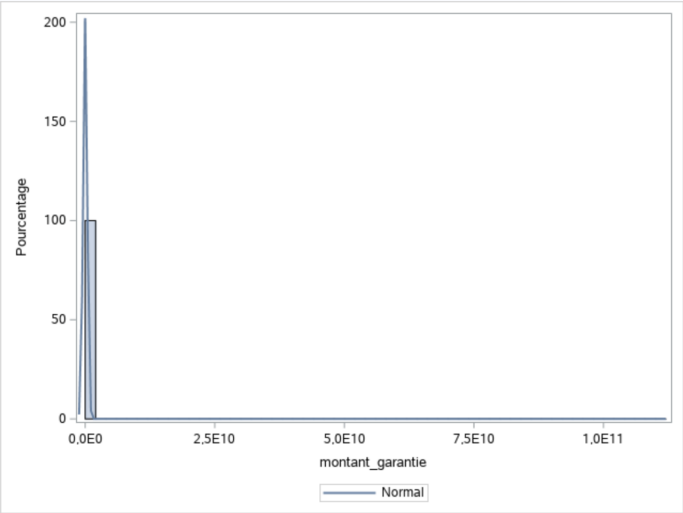

Figure 21: Histogram and density of the montant_garantie

The density and histogram of these variables confirm our previous conclusion. The distribution of the other right or left-skewed variables look roughly the same.

At this step one has 41 quantitative variables left.

2.3.2 Using the bivariate analysis

To decide which variables are kept among the 41 remaining variables, one uses a bivariate analysis.

To do so, one especially looks at the correlation between two variables. The Pearson correlation coefficient is used here. The coefficient goes from -1 to 1. The closer to 1 or -1, the higher the correlation and the closer to 0 the smaller the correlation is. If the coefficient is negative it means when one variable increases the other decreases and the other way around. On the contrary, when the coefficient is positive, then the variables evolve in the same way, i.e if one increases or decreases then

the other does the same.

First, one will look at the correlation between the variables and the variable of interest to keep the variables the most correlated to it. Then, one makes sure the variable chosen aren't correlated between them or if they are, one only keeps the variable the most correlated to the variable of interest and deletes the others.

The 15 variables the most correlated with the variable of interest are:

- **nb_credit**: -0.03916
- **nb_debit**: -0.03880
- **nb_debit_6m_min**: -0.03849
- **nb_credit_6m_min**: -0.03823
- **nb_credit_6m_mean**: -0.03786
- **nb_credit_6m_max**: -0.03728
- **nb_debit_6m_mean**: -0.03685
- **nb_debit_6m_max**: -0.03530
- **age**: -0.03291
- **sld_liqu_6m_mean**: -0.03036
- **sld_avoirs_6m_mean**: -0.02989
- **sld_avoirs_6m_min**: -0.02898
- **mtn_cptecourant**: -0.02823
- **sld_courant_6m_mean**: -0.02774
- **mtn_cptecourant_min**: -0.02770

When seeing the list of variables, one can directly see that some variables seem to give a similar information and might thus be very correlated, i.e have a correlation coefficient higher than 0.8. It could for example be the case for: *nb_credit*, *nb_credit_6m_min*, *nb_credit_6m_max*, *nb_credit_6m_mean*.

Coefficients de corrélation de Pearson, N = 84377 Proba > |r| sous H0: Rho=0				
	nb_credit	nb_credit_6m_min	nb_credit_6m_max	nb_credit_6m_mean
nb_credit nb_credit	1.00000	0.94332 <.0001	0.93099 <.0001	0.95748 <.0001
nb_credit_6m_min nb_credit_6m_min	0.94332 <.0001	1.00000	0.91526 <.0001	0.96657 <.0001
nb_credit_6m_max nb_credit_6m_max	0.93099 <.0001	0.91526 <.0001	1.00000	0.98069 <.0001
nb_credit_6m_mean nb_credit_6m_mean	0.95748 <.0001	0.96657 <.0001	0.98069 <.0001	1.00000

Figure 22: Correlation coefficients of the variables nb_credit, nb_credit_6m_min, nb_credit_6m_max and nb_credit_6m_mean

As expected the four variables are very correlated between them, almost exactly since their coefficient are all above 0.91. Thus, only *nb_credit* is kept because it is the one with the highest correlation with the variable of interest.

The same reasoning is used for: *nb_debit*, *nb_debit_6m_min*, *nb_debit_6m_max*, *nb_debit_6m_mean*

Coefficients de corrélation de Pearson, N = 84377 Proba > |r| sous H0: Rho=0				
	nb_debit	nb_debit_6m_min	nb_debit_6m_max	nb_debit_6m_mean
nb_debit nb_debit	1.00000	0.94545 <.0001	0.92666 <.0001	0.95584 <.0001
nb_debit_6m_min nb_debit_6m_min	0.94545 <.0001	1.00000	0.91195 <.0001	0.96631 <.0001
nb_debit_6m_max nb_debit_6m_max	0.92666 <.0001	0.91195 <.0001	1.00000	0.97936 <.0001
nb_debit_6m_mean nb_debit_6m_mean	0.95584 <.0001	0.96631 <.0001	0.97936 <.0001	1.00000

Figure 23: Correlation coefficients of the variables nb_debit, nb_debit_6m_min, nb_debit_6m_max and nb_debit_6m_mean

Since here *nb_debit* is the one that is the most correlated to the variable of interest, it is the one that is kept.

The variables *sld_avoirs_6m_mean* and *sld_avoirs_6m_min* are highly correlated as well since their coefficient is around 0.87. Thus, only *sld_avoirs_6m_mean* is kept.

Finally, one looks at *mtn_cptecourant*, *sld_courant_6m_mean* and *mtn_cptecourant_min*.

Coefficients de corrélation de Pearson, N = 84377 Proba > |r| sous H0: Rho=0			
	mtn_cptecourant	sld_courant_6m_mean	mtn_cptecourant_min
mtn_cptecourant mtn_cptecourant	1.00000	0.84410 <.0001	0.85954 <.0001
sld_courant_6m_mean sld_courant_6m_mean	0.84410 <.0001	1.00000	0.83337 <.0001
mtn_cptecourant_min mtn_cptecourant_min	0.85954 <.0001	0.83337 <.0001	1.00000

Figure 24: Correlation coefficients of the variables mtn_cptecourant, sld_courant_6m_mean, and mtn_cptecourant_min

Once again these variables are very correlated between them. Therefore, one only keeps *mtn_cptcourant* which has the highest correlation coefficient with the variable of interest.

Among the 15 most correlated variables, *age* and *sld_liqu_6m_mean* aren't very correlated to any other and are therefore kept.

To have the same number of quantitative and qualitative variables, one looks at the next most correlated variables with the variable of interest. The three next ones are variables very correlated with other variables as well that would thus be deleted anyway. However, the fourth one is *anc_cli_bqe* which coefficient is -0.02580 with the variable of interest. This variable only has a rather high positive correlation of 0.47 with the age that one will look into in the next subsection.

The table below shows that there is no strong correlation between the remaining variables, since no coefficient is higher than 0.8. Therefore, they can be kept.

Coefficients de corrélation de Pearson Nombre d'observations							
	age	anc_cli_bqe	nb_debit	nb_credit	sld_liqu_6m_mean	sld_avoirs_6m_mean	mtn_cptecourant
age age	1.00000 84377	0.47931 84353	-0.14431 84377	-0.08455 84377	0.02564 84377	0.11583 84377	0.04249 84377
anc_cli_bqe anc_cli_bqe	0.47931 84353	1.00000 84353	-0.08042 84353	-0.02662 84353	0.00440 84353	0.07930 84353	0.03220 84353
nb_debit nb_debit	-0.14431 84377	-0.08042 84353	1.00000 84377	0.41278 84377	0.09550 84377	0.02349 84377	0.10031 84377
nb_credit nb_credit	-0.08455 84377	-0.02662 84353	0.41278 84377	1.00000 84377	0.09103 84377	0.03895 84377	0.09390 84377
sld_liqu_6m_mean sld_liqu_6m_mean	0.02564 84377	0.00440 84353	0.09550 84377	0.09103 84377	1.00000 84377	0.44083 84377	0.67953 84377
sld_avoirs_6m_mean sld_avoirs_6m_mean	0.11583 84377	0.07930 84353	0.02349 84377	0.03895 84377	0.44083 84377	1.00000 84377	0.31357 84377
mtn_cptecourant mtn_cptecourant	0.04249 84377	0.03220 84353	0.10031 84377	0.09390 84377	0.67953 84377	0.31357 84377	1.00000 84377

Figure 25: Correlation coefficients of the quantitative variables kept

Hence, the quantitative variables kept after the bivariate analysis are the following:

- **nb_credit**: Number of credit flows.
- **nb_debit**: Number of debit flows.
- **age**: Age of the customers.
- **sld_liqu_6m_mean**: Average amount of the liquid assets during the last 6 months.
- **sld_avoirs_6m_mean**: Average amount of assets during the last 6 months.
- **mtn_cptcourant**: Amount on the current account.
- **anc_cli_bqe**: Length of the time the individual is a customer.

After deleting some rows with missing values, coming from the variable *anc_cli_bqe*, there are 84353 individuals left and 7 quantitatives variables.

2.3.3 Discretization

The discretization consists in creating groups of values from a quantitative variable in order to transform the variable in a qualitative one. To discretize the 7 quantitative variables, the SAS code provided in class is used: it maximizes the normalized correlation according to the number of classes. For instance, for the variable *age*, Figure 26 below displays the correlation indicator as a function of the number of classes. One can see that the number of classes maximizing this measure is 3.

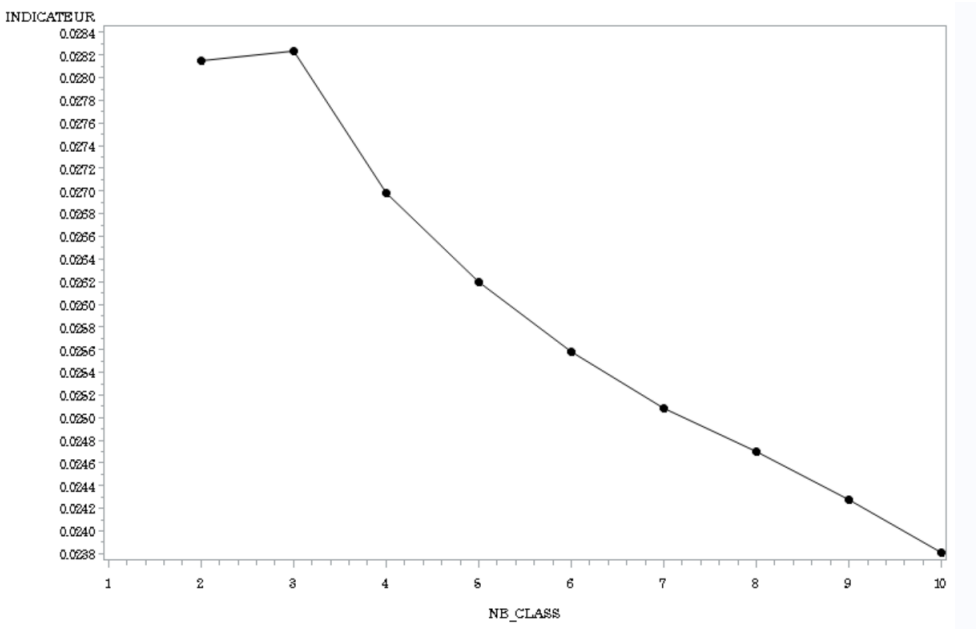

Figure 26: Correlation indicator as a function of the number of classes of the variable age

With this code, the following results are obtained:

- **age2**: it is equal to "1. inf 63" if $age <= 63$, to "2. btw 63 and 83" if $62 < age <= 83$ and to "3. sup 83" otherwise.

Table de age par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
age(age)	0	1	Total
age<=63	44190 52.39 98.11 53.19	851 1.01 1.89 66.69	45041 53.40
63<age<=83	30854 36.58 98.76 37.14	386 0.46 1.24 30.25	31240 37.03
age>83	8033 9.52 99.52 9.67	39 0.05 0.48 3.06	8072 9.57
Total	83077 98.49	1276 1.51	84353 100.00

Figure 27: Contingency table between age2 and top_def_12m_90j

- **anc_cli_bqe2**: it is equal to "1. inf 22" if *anc_cli_bqe* <= 22 and to "2. sup 22" otherwise.

Table de anc_cli_bqe par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
anc_cli_bqe(anc_cli_bqe)	0	1	Total
anc_cli_bqe<=22	62543 74.14 98.29 75.28	1090 1.29 1.71 85.42	63633 75.44
anc_cli_bqe>22	20534 24.34 99.10 24.72	186 0.22 0.90 14.58	20720 24.56
Total	83077 98.49	1276 1.51	84353 100.00

Figure 28: Contingency table between anc_cli_bqe2 and top_def_12m_90j

- **nb_debit2**: it is equal to "1. inf 0" if *nb_debit* <= 0 and to "2. sup 0" otherwise.

Table de nb_debit par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
nb_debit(nb_debit)	0	1	Total
nb_debit<=0	34103 40.43 97.49 41.05	877 1.04 2.51 68.73	34980 41.47
nb_debit>0	48974 58.06 99.19 58.95	399 0.47 0.81 31.27	49373 58.53
Total	83077 98.49	1276 1.51	84353 100.00

Figure 29: Contingency table between nb_debit2 and top_def_12m_90j

- **nb_credit2**: it is equal to "1. inf 0" if *nb_credit* <= 0 and to "2. sup 0" otherwise.

Table de nb_credit par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
nb_credit(nb_credit)	0	1	Total
nb_credit<=0	35831 42.48 97.46 43.13	933 1.11 2.54 73.12	36764 43.58
nb_credit>0	47246 56.01 99.28 56.87	343 0.41 0.72 26.88	47589 56.42
Total	83077 98.49	1276 1.51	84353 100.00

Figure 30: Contingency table between nb_credit2 and top_def_12m_90j

- **sld_liqu_6m_mean2**: it is equal to "1. inf 72.3" if $sld\_liqu\_6m\_mean <= 72.389$ and to "2. sup 72.3" otherwise.

Table de sld_liqu_6m_mean par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
sld_liqu_6m_mean(sld_liqu_6m_mean)	0	1	Total
sld_liqu_6m_mean<=72.389	8512 10.09 91.73 10.25	767 0.91 8.27 60.11	9279 11.00
sld_liqu_6m_mean>72.389	74565 88.40 99.32 89.75	509 0.60 0.68 39.89	75074 89.00
Total	83077 98.49	1276 1.51	84353 100.00

Figure 31: Contingency table between sld_liqu_6m_mean2 and top_def_12m_90j

- **sld_avoirs_6m_mean2**: it is equal to "1. inf 107.6" if $sld\_avoirs\_6m\_mean <= 107.673$ and to "2. sup 107.6" otherwise.

Table de sld_avoirs_6m_mean par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
sld_avoirs_6m_mean(sld_avoirs_6m_mean)	0	1	Total
sld_avoirs_6m_mean<=107.673	3757 4.45 89.07 4.52	461 0.55 10.93 36.13	4218 5.00
sld_avoirs_6m_mean>107.673	79320 94.03 98.98 95.48	815 0.97 1.02 63.87	80135 95.00
Total	83077 98.49	1276 1.51	84353 100.00

Figure 32: Contingency table between sld_avoirs_6m_mean2 and top_def_12m_90j

- **mtn_cptcourant2**: it is equal to "1. inf -0.09" if $mtn\_cptcourant <= -0.09$ and to "2. sup -0.09" otherwise.

Table de mtn_cptecourant par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
mtn_cptecourant(mtn_cptecourant)	0	1	Total
mtn_cptecourant<=-0.09	4403 5.22 86.98 5.30	659 0.78 13.02 51.65	5062 6.00
mtn_cptecourant>-0.09	78674 93.27 99.22 94.70	617 0.73 0.78 48.35	79291 94.00
Total	83077 98.49	1276 1.51	84353 100.00

Figure 33: Contingency table between mtn_cptcourant2 and top_def_12m_90j

2.4 Variables grouping

2.4.1 Qualitative variables

To be able to have a stable model, one wants to have modalities with enough individuals (at least more than 5% of the individuals). To do so, modalities are grouped if they have the similar rate of default. If one take the example of the variable CSP which corresponds to the socio-professional category of the individual (figure 12), one can see that the modalities Agriculteurs, Artisans, commercants, chef entreprise, Autre and Ouvriers have similar default rate and represent also a small proportion of the population. Thus, these modalities are grouped together. In the same logic, the modalities Cadre, profession superieures and Professions intermediaires are grouped together in the modality "2. Cadres" and the modalities Sans activite, Employes and Non renseigne are grouped togetherin the modality "3. Autres CSP".

Below, in the Table 1, the modalities grouped are listed and summed up:

Variable	Old Modalities	New modalities
CSP	Agriculteurs Artisans, commercants, chef entreprise Autre Ouvriers	1. Ouvriers agriculteurs
	Cadre profession superieures Professions intermediaires	2. Cadres
	Sans activite Employes Non renseigne	3. Autres CSP
Family	Celibataire, Divorce(e) Separe(e)	1. Celibataire seul divorce
	Non renseigne Veuf(ve Pacs	2. Autres
	Marie	3. Marie

Table 1: Modalities grouping

2.4.2 Creating interaction variables

Variables should be grouped when their common Cramer's V statistic is high in absolute value, especially when it is not high enough to remove one of the variable (a threshold is fixed at 0.8) but is too high to keep both variables separately. One chooses a threshold for which if the absolute value of the Cramer's V statistic is higher than this threshold then the variables are crossed.

Variable 1	Variable 2	Cramer's V statistic
nb_debit2	nb_credit2	0.6880
sld_liqu_6m_mean2	sld_avoirs_6m_mean2	0.5556
age2	CSP2	0.4750
nb_credit2	topFacilite2	-0.4099
nb_debit2	topFacilite2	-0.3899
anc_cli_bqe2	age2	0.3814
topAssuIARD2	topAssurVIE2	0.3667
type_compte2	sit_familiale2	0.3570
mtn_cptecourant2	sld_liqu_6m_mean2	0.3175
sld_liqu_6m_mean2	nb_debit2	0.2761
mtn_cptecourant2	sld_avoirs_6m_mean2	0.2711
anc_cli_bqe2	CSP2	0.2710
sld_liqu_6m_mean2	nb_credit2	0.2404
sit_familiale2	CSP2	0.2039
age2	sit_familiale2	0.2862

Table 2: Cramer's V statistic greater than 0.2 in absolute value

One decides to fix the threshold at 0.5 in absolute value and cross related variables one wants to include in the model. Thus, one ends up with the following groupings:

- **deb_cred**: is a grouping of *nb_debit2* and *nb_credit2*

Table de cred_deb par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
cred_deb	0	1	Total
1. inf 01. inf 0	28619 33.93 97.26 34.45	805 0.95 2.74 63.09	29424 34.88
1. inf 02. sup 0	7212 8.55 98.26 8.68	128 0.15 1.74 10.03	7340 8.70
2. sup 01. inf 0	5484 6.50 98.70 6.60	72 0.09 1.30 5.64	5556 6.59
2. sup 02. sup 0	41762 49.51 99.36 50.27	271 0.32 0.64 21.24	42033 49.83
Total	83077 98.49	1276 1.51	84353 100.00

Figure 34: Contingency table between cred_deb and top_def_12m_90j

- **avoir_liq**: is a grouping of *sld_liqu_6m_mean2* and *sld_avoirs_6m_mean2*

Table de avoir_liq par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
avoir_liq	0	1	Total
1. inf 107.61. inf 72.3	3218 3.81 87.92 3.87	442 0.52 12.08 34.64	3660 4.34
1. inf 107.62. sup 72.3	539 0.64 96.59 0.65	19 0.02 3.41 1.49	558 0.66
2. sup 107.61. inf 72.3	5294 6.28 94.22 6.37	325 0.39 5.78 25.47	5619 6.66
2. sup 107.62. sup 72.3	74026 87.76 99.34 89.11	490 0.58 0.66 38.40	74516 88.34
Total	83077 98.49	1276 1.51	84353 100.00

Figure 35: Contingency table between avoir_liq and top_def_12m_90j

One still wants at least 5% of the clients in each modality. It is the case for the new variable *cred_deb* but not *avoir_liq*. This is going to be modified.

Table de avoir_liq2 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
avoir_liq2	0	1	Total
1. avoir sup 107 liq sup 72.3	74026 87.76 99.34 89.11	490 0.58 0.66 38.40	74516 88.34
2. autres avoir liq	9051 10.73 92.01 10.89	786 0.93 7.99 61.60	9837 11.66
Total	83077 98.49	1276 1.51	84353 100.00

Figure 36: Contingency table between avoir_liq2 and top_def_12m_90j

As it can be seen on Figure 36, *avoir_liq* had 4 modalities and ends up with 2. When the amount of the liquid assets is above 107 and the amount of assets is above 72.3 then the modality is "1. avoir sup 107 liq qup 72.3". The other modalities are in "2. autres avoir liq" because they have a higher rate of customers in default. This grouping enables to have at least 5% customers in each modality and to avoid collinearity issues.

2.4.3 Variables dependence

As a recall, there are still:

- 7 qualitative variables: topFacilite2, CSP2, topAssuIARD2, type_compte2, sit_familiale2,

top_Gar_Cnp2 and topAssurVIE2.

- 2 crossed variables: deb_cred and avoir_liq2.
- 3 former quantitative variables which are now qualitative : age2, anc_cli_bqe2 and mtn_cptecourant2.

It is then once again checked that the variables are not too dependent between them. It is indeed the case since the highest Cramer's V statistic is only around 0.4 maximum as it can be seen in the following table.

Variable 1	Variable 2	Cramer's V statistic
CSP2	age2	0.4749
cred_deb	topFacilite2	0.4496
age2	anc_cli_bqe2	0.3814
topAssuIARD2	topAssurVIE2	0.3666

Table 3: Cramer's V statistics greater than 0.35 in absolute value

Finally, Figure 37 shows that all the variables kept are still linked to the dependent variable even though the Cramer's V statistics aren't very high.

Obs.	Value	ABS_V_CRAMER	Variable
1	-0.2382	0.23818	mtn_cptecourant2
2	0.1928	0.19282	avoir_liq2
3	0.0779	0.07794	cred_deb
4	0.0478	0.04778	topFacilite2
5	0.0435	0.04352	CSP2
6	0.0372	0.03715	age2
7	-0.0288	0.02875	anc_cli_bqe2
8	0.0266	0.02656	topAssuIARD2
9	0.0184	0.01845	type_compte2
10	0.0145	0.01453	top_Gar_Cnp2
11	0.0137	0.01366	sit_familiale2
12	0.0131	0.01308	topAssurVIE2

Figure 37: Cramer's V statistics between the dependent variable and the variables kept for the model

The dependent variable is the most dependent to *mtn_cptecourant2*, *avoir_liq2* and *cred_deb*.

All the 12 variables are kept for the modeling part. It could seem too many variables are kept, but actually when building our models some will very likely be deleted.

- **topFacilite2**
- **CSP2**
- **TopAssuIARD2**

- **type_compte2**
- **sit_familiale2**
- **top_Gar_Cnp2**
- **topAssurVIE2**
- **deb_cred**
- **avoir_liq2**
- **age2**
- **anc_cli_bqe2**
- **mtn_cptecourant2**

The table still contains 84 353 of the clients.

3 Train - Test split

3.1 Utility of a train - test split

Before fitting a model on the variables selected in the previous section, one has to split the data into a train set and a test set. The train set is used to fit the parameters of the model and the test set is used to evaluate the performance of the model. It is important to evaluate the performance on a different data set than the one used to train the model as the model can overfit or underfit the data used to train the model. Here, the train dataset contains 70% of the customers (59 048) and the test the remaining 30% (25 305 customers).

3.2 Dependent variable balance in the train and test datasets

Here, the exploratory variable is the default (*top_def_12m_90j*) and one knows that only very few customers are in default. Hence, the split is done is a way that the rate of customers in default is approximately the same in the train and test datasets. A stratification is used.

top_def_12m_90j				
top_def_12m_90j	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	58154	98.49	58154	98.49
1	894	1.51	59048	100.00

Figure 38: Frequency of *top_def_12m_90j* in the train dataset

top_def_12m_90j				
top_def_12m_90j	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	24923	98.49	24923	98.49
1	382	1.51	25305	100.00

Figure 39: Frequency of *top_def_12m_90j* in the test dataset

Figures 38 and 39 show that the train and test sets both have 1.51% of customers in default. Thus, the stratification was well done when splitting the data.

3.3 Checking the explanatory variables balance in the train and test datasets

Not just the dependent variable must be balanced between the train and test sets for the models to work on both, but also the explanatory variables. Thus, one checks that the explanatory variables are well balanced between both datasets. The result is shown only for one variable for a question of redundancy.

Table de anc_cli_bqe2 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
anc_cli_bqe2	0	1	Total
1. inf 22	43842 74.25 98.28 75.39	768 1.30 1.72 85.91	44610 75.55
2. sup 22	14312 24.24 99.13 24.61	126 0.21 0.87 14.09	14438 24.45
Total	58154 98.49	894 1.51	59048 100.00

Figure 40: Contingency table between anc_cli_bqe2 and top_def_12m_90j in the train dataset

Table de anc_cli_bqe2 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
anc_cli_bqe2	0	1	Total
1. inf 22	18701 73.90 98.31 75.04	322 1.27 1.69 84.29	19023 75.17
2. sup 22	6222 24.59 99.04 24.96	60 0.24 0.96 15.71	6282 24.83
Total	24923 98.49	382 1.51	25305 100.00

Figure 41: Contingency table between anc_cli_bqe2 and top_def_12m_90j in the test dataset

As one may notice, the percentage of customers in "1. inf 22" is 75.55% in the train dataset and 75.17% in the test dataset. It means the number of customers in each modality for the train and test datasets is very similar, which is what one was looking for. In terms of customers in default in each modality for each dataset, the balance is good as well since if one takes again as example the first modality, one may see that 1.72% of the customers in the first modality in the train dataset are in default against 1.69% in the test dataset, which is very close.

The conclusion is the same for all the variables:

- The number of customers in each modality is very similar in the train and test datasets.
- The number of customers in default in each modality is very similar as well for both datasets.

It confirms that the split was well done and that one can start the modeling part.

4 Model estimation

4.1 Logistic regression theory

The aim of a logistic regression is to model the probability of an event, here being the variable *top_def_12m_90j*, characterized by the fact that whether or not a customer will be in default within the next 12 months.

As it is a qualitative variable with two modalities (1 or 0), the logistic model estimates the odds of a client not being in default for the upcoming 12 months. The formula for the odd is:

$$Odds_A = \frac{P(A)}{1 - P(A)}$$

One has to assume that, using $X = x$ (the covariates) and *top_def_12m_90j* following a random bernouilli distribution with parameter $\pi(x) = P(Y = 1|x) = \mathbb{E}(Y|X = x)$,

$$logit(\pi(x)) = ln(Odds_{\pi(x)}) = ln\frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

The score of the individual x is:

$$s(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

where the β are the model's coefficients to be estimated.

4.2 Finding the reference modalities

In order to find the reference modalities for the models, one choose the modality with the highest risk in the train dataset. This enables the models to be stable and to perform well.

For instance, when looking at *anc_cli_bqe2* again (on Figure 42), one can see that among the customers in the modality "1. inf 22", 1.72% are in default, whereas only 0.87% are in default among the customers in "2. sup 22". Thus, here the reference modality chosen is "1. inf 22" because it is the riskier.

Table de anc_cli_bqe2 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
anc_cli_bqe2	0	1	Total
1. inf 22	43842 74.25 98.28 75.39	768 1.30 1.72 85.91	44610 75.55
2. sup 22	14312 24.24 99.13 24.61	126 0.21 0.87 14.09	14438 24.45
Total	58154 98.49	894 1.51	59048 100.00

Figure 42: Contingency table between anc_cli_bqe2 and top_def_12m_90j in the train dataset

The following list enumerates the reference modality for each variable.

- **topFacilite2**: "2. sans" (see Figure 7),
- **TopAssuIARD2**: "2. sans" (see Figure 8),
- **top_Gar_Cnp2**: "2. sans" (see Figure 9),
- **topAssurVIE2**: "2. sans" (see Figure 10),
- **type_compte2**: "2. Compte joint et individuel" (see Figure 11),
- **CSP2**: "1. Ouvriers agriculteurs" (see Figure 14),
- **sit_familiale2**: "1. Celibataire seul divorce" (see Figure 15),
- **age2**: "1. inf 63" (see Figure 27),
- **anc_cli_bqe2**: "1. inf 22" (see Figure 28),
- **mtn_cptecourant2**: "1. inf -0.09" (see Figure 33),
- **deb_cred**: "1. inf 01. inf 0" (see Figure 34),
- **avoir_liq2**: ="2. autres avoir liq" (see Figure 36).

4.3 First model: using all the variables

A first model is done, it is the most "naive" one since it includes all of the selected and created variables. It gives a first insight of the variables that still need to be deleted or modified.

To establish how well a model is performing, one can look at the Somers'D. It is computed as following $Somers'D = (NC–ND)/(NC + ND + NT)$, where NC is the number of concordant pairs,

ND is the number of discordant pairs and NT the number of tied pairs. Being concordant means that the prediction corresponds to the actual value. It ranges between -1 and 1, with -1 indicating that all pairs are discordant and 1 meaning that all pairs are concordant. Hence, the goal is to get the Somers'D statistics the closest to 1. Here, it is actually of 0.724 which is quite good and means that the model performed well. The concordant percentage is actually of 86.1% and the discordant one of 13.7 %. Since one wants a concordant percentage close to 100% and a discordant one close to 0%, it confirms that the model is quite good.

Nonetheless, even if the model, in theory, seems to predict well the customers in default, it has some issues.

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	4.1919	0.1798	543.8389	<.0001
cred_deb	1. inf 02. sup 0	1	-0.3100	0.0981	9.9862	0.0016
cred_deb	2. sup 01. inf 0	1	0.000376	0.1173	0.0000	0.9974
cred_deb	2. sup 02. sup 0	1	0.5545	0.0786	49.7257	<.0001
avoir_liq2	1. avoir sup 107 liq sup 72.3	1	0.6871	0.0440	243.8517	<.0001
age2	2. btw 63	1	-0.2721	0.0857	10.0920	0.0015
age2	3. sup 83	1	0.6287	0.1574	15.9613	<.0001
anc_cli_bqe2	2. sup 22	1	0.1994	0.0535	13.9025	0.0002
mtn_cptecourant2	2. sup -0.09	1	1.1755	0.0421	781.0356	<.0001
type_compte2	1. Compte individuel uniquement	1	0.2461	0.0658	14.0106	0.0002
type_compte2	3. Compte joint uniquement	1	0.3916	0.0668	34.3974	<.0001
topAssurVIE2	1. avec	1	0.1870	0.1105	2.8622	0.0907
topAssulARD2	1. avec	1	0.1203	0.0460	6.8454	0.0089
top_Gar_Cnp2	1. avec	1	0.3092	0.1042	8.8081	0.0030
topFacilite2	1. avec	1	0.5126	0.0692	54.8430	<.0001
sit_familiale2	2. Autres	1	-0.1845	0.1415	1.7021	0.1920
sit_familiale2	3. Marie	1	0.00642	0.0812	0.0063	0.9370
CSP2	2. Cadres	1	-0.1574	0.0598	6.9334	0.0085
CSP2	3. Autres	1	0.1412	0.0742	3.6177	0.0572
CSP2	4. Retraites	1	0.3755	0.0957	15.4026	<.0001

Figure 43: Coefficient estimators - Model 1

Figure 43 displays the coefficient estimators for each modality of each variable. When looking at them, one wants to make sure that all the modalities are significant at a 5% level, i.e that the p-value (last column on the right) is smaller than 0.05 and that all the estimations are positive. Indeed, a negative estimation means there is a collinearity issue, i.e that there exists a dependency between the variables used in the model or that there is an issue in the construction of the modalities.

Thus, one may notice several issues that are inside the red boxes. Some modality modifications are done to see if it helps dealing with the collinearity and significance.

4.4 Modalities modification

First, let us group the modalities of *cred_deb*. As one can see it in Figure 34, the 3 first modalities have similar rates of customers in default. They are therefore grouped together in a new modality called "2. autres" and when both the number of credit and debit is higher than 0 the modality is "1. cred deb sup 0". The reference becomes here "2. autres".

Table de cred_deb2 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
cred_deb2	0	1	Total
1. cred deb sup 0	41762 49.51 99.36 50.27	271 0.32 0.64 21.24	42033 49.83
2. autres	41315 48.98 97.63 49.73	1005 1.19 2.37 78.76	42320 50.17
Total	83077 98.49	1276 1.51	84353 100.00

Figure 44: Contingency table between cred_deb2 and top_def_12m_90j

The two next variables for which the modalities are going to be grouped are *age2* and *CSP2* in order to avoid collinearity issues.

Table de age3 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
age3	0	1	Total
1. sup 83	8033 9.52 99.52 9.67	39 0.05 0.48 3.06	8072 9.57
2. inf 83	75044 88.96 98.38 90.33	1237 1.47 1.62 96.94	76281 90.43
Total	83077 98.49	1276 1.51	84353 100.00

Figure 45: Contingency table between age3 and top_def_12m_90j

Table de CSP3 par top_def_12m_90j			
	top_def_12m_90j(top_def_12m_90j)		
CSP3	0	1	Total
1. Retraites	23143 27.44 99.17 27.86	194 0.23 0.83 15.20	23337 27.67
2. Autres	59934 71.05 98.23 72.14	1082 1.28 1.77 84.80	61016 72.33
Total	83077 98.49	1276 1.51	84353 100.00

Figure 46: Contingency table between CSP3 and top_def_12m_90j

The variable *age3* has 2 modalities depending if a customer is less or more than 83 years old and *CSP3* makes the distinction between retired customers and not retired ones. The reference for *age3* is "2. inf 83" and "1. Autres" for *CSP3*.

The new variables aren't very highly correlated to any other variable since the Cramer's V statistics are still smaller than 0.5

Variable 1	Variable 2	Cramer's V statistic
topFacilite2	cred_deb2	0.4466
CSP3	age3	0.4041
topAssuIARD2	topAssurVIE2	0.3666
sit_familiale2	type_compte2	0.3569

Table 4: Cramer's V statistics greater than 0.35 in absolute value

Figure 47 shows that the new variables are linked to the dependent variable.

Obs.	Value	ABS_V_CRAMER	Variable
1	-0.2382	0.23818	mtn_cptecourant2
2	0.1928	0.19282	avoir_liq2
3	0.0709	0.07087	cred_deb2
4	0.0478	0.04778	topFacilite2
5	0.0345	0.03452	CSP3
6	-0.0288	0.02875	anc_cli_bqe2
7	0.0274	0.02744	age3
8	0.0266	0.02656	topAssuIARD2
9	0.0184	0.01845	type_compte2
10	0.0145	0.01453	top_Gar_Cnp2
11	0.0137	0.01366	sit_familiale2
12	0.0131	0.01308	topAssurVIE2

Figure 47: Cramer's V statistics between the dependent variable and the variables kept for the model

The dependent variable is now the most dependent to *mtn_cptecourant2*, *avoir_liq2* and *cred_deb2*.

4.5 Second model: model with the new modalities

The second model, as the first one, includes all the variables, but this time modalities have been changed as seen just previously.

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	4.6656	0.1877	617.6853	<.0001
cred_deb2	1. cred deb sup 0	1	0.3927	0.0478	67.5184	<.0001
avoir_liq2	1. avoir sup 107 liq sup 72.3	1	0.6854	0.0420	265.7849	<.0001
age3	1. sup 83	1	0.4641	0.1162	15.9516	<.0001
anc_cli_bqe2	2. sup 22	1	0.2071	0.0518	16.0064	<.0001
mtn_cptecourant2	2. sup -0.09	1	1.1837	0.0417	804.8510	<.0001
type_compte2	1. Compte individuel uniquement	1	0.2410	0.0657	13.4706	0.0002
type_compte2	3. Compte joint uniquement	1	0.4048	0.0667	36.8606	<.0001
topAssurVIE2	1. avec	1	0.1801	0.1104	2.6646	0.1026
topAssulARD2	1. avec	1	0.1272	0.0459	7.6884	0.0056
top_Gar_Cnp2	1. avec	1	0.3255	0.1023	10.1259	0.0015
topFacilite2	1. avec	1	0.5087	0.0689	54.4767	<.0001
sit_familiale2	2. Autres	1	-0.1674	0.1404	1.4211	0.2332
sit_familiale2	3. Marie	1	-0.0180	0.0808	0.0497	0.8237
CSP3	1. Retraites	1	0.2694	0.0550	23.9632	<.0001

Figure 48: Coefficient estimators - Model 2

One may notice two issues in this model. The first one is that the modality "1. avec" of the variable *topAssurVIE2* isn't significant since the p-value is of 0.1026 which is higher than 0.05. The second problem is with the variable *sit_familiale2*. Indeed, both modalities coefficient estimators are negative. This indicates that these variables should probably be deleted to have a model that is stable and that will perform well on the test dataset. Before doing so, one tries other models.

Nonetheless, thanks to the new modality grouping of *age3*, *CSP3* and *cred_deb2* there are less collinearity issues and most of the modalities are significant. Moreover, the performance of the model is still very much the same since the Somers'D is 0.719. It is a little smaller than before but still good. This decrease means the predictions are a little less good. Indeed, the concordant percentage is 85.6% (against 86.1%) and the discordant one is 13.6% (against 13.7%). Even though the prediction seems less good, one expects the model to be more stable and robust.

Another measure one can look at is the area under the ROC curve, which is called AUC (Area Under Curve).

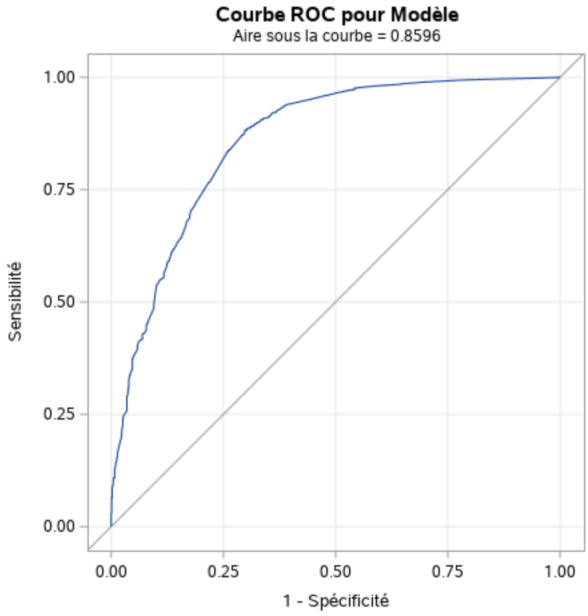

Figure 49: ROC curve - Model 1

The ROC curve is built from the sensitivity and the specificity:

- The sensitivity, also called True Positive Rate (TPR) is here the probability to be predicted to be in default, conditioned on being actually in default.
- The specificity, also called True Negative Rate (TNR) is here the probability to be predicted to not be in default, conditioned on actually not being in default.

They are both close to 1, which means once again that the predictions are well. Moreover, the area under the curve (AUC) is of 0.8596. When the AUC is equal to 1 it means the model is able to predict perfectly if a customer is in default or not and if it is equal to 0 it means all customers not in default would be predicted to be in default and all the customers in default would be predicted not in default. Thus, having an AUC close to 1 means once again that the model works well.

Before deciding what to do with the two variables causing problems, one uses another model, which is a stepwise model.

4.6 Third model: stepwise model

The third model is a stepwise model. This is a method that SAS allows to do. As in a forward selection, variables are added one at a time to the model. However, at each step, if an effect is not significant, then the corresponding modality is deleted. This implies that no effect/modality that isn't significant is added to the model. The process ends when no additional significant effect can be added.

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	4.5171	0.1629	769.2114	<.0001
cred_deb2	1. cred deb sup 0	1	0.3910	0.0478	66.9411	<.0001
avoir_liq2	1. avoir sup 107 liq sup 72.3	1	0.6850	0.0420	265.3931	<.0001
age3	1. sup 83	1	0.4649	0.1162	16.0008	<.0001
anc_cli_bqe2	2. sup 22	1	0.2082	0.0518	16.1827	<.0001
mtn_cptecourant2	2. sup -0.09	1	1.1858	0.0417	807.8781	<.0001
type_compte2	1. Compte individuel uniquement	1	0.2385	0.0657	13.1907	0.0003
type_compte2	3. Compte joint uniquement	1	0.4027	0.0667	36.4839	<.0001
topAssulARD2	1. avec	1	0.1542	0.0435	12.5574	0.0004
top_Gar_Cnp2	1. avec	1	0.3261	0.1023	10.1607	0.0014
topFacilite2	1. avec	1	0.5070	0.0689	54.1374	<.0001
sit_familiale2	2. Autres	1	-0.1632	0.1404	1.3517	0.2450
sit_familiale2	3. Marie	1	-0.0186	0.0808	0.0532	0.8176
CSP3	1. Retraites	1	0.2718	0.0550	24.4064	<.0001

Figure 50: Coefficient estimators - Model 3

The coefficient estimators are in Figure 50. The variable *topAssurVIE2* hasn't been included. It isn't surprising since one has noticed before that its only modality wasn't significant. However, the *sit_familiale2* variable hasn't been deleted. The best model gotten with the stepwise selection has a concordant percentage of 85.5% which is actually smaller than the one in Model 2. The Somers'D doesn't change and is still of 0.719.

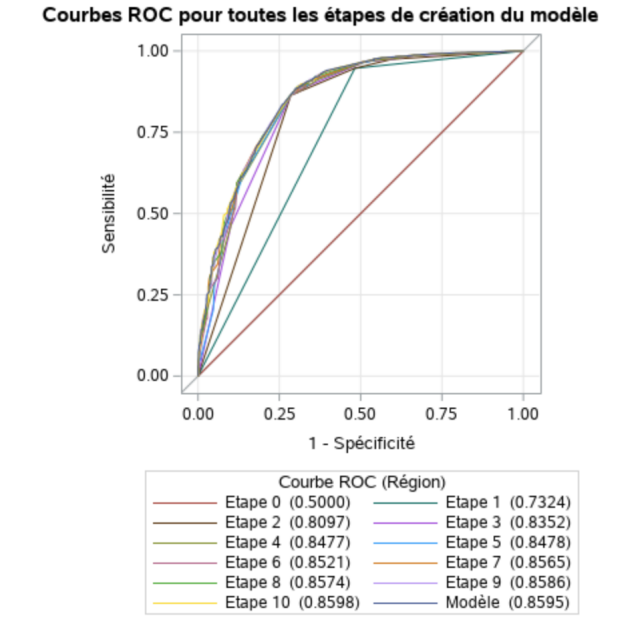

Figure 51: ROC curves - Model 3

Each ROC curve represents the ROC curve at one step. The best model gives an AUC of 0.8595.

The model still being very good, it suggests that the variable *topAssurVIE2* should definitely be deleted. To avoid the collinearity issues and be sure to have a stable model, *sit_familiale2* will also be deleted.

4.7 Fourth model: the best model

After trying three models, changing modalities and deleting variables, the fourth model is expected to be the best model in terms of prediction and stability.

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	4.5101	0.1575	820.3356	<.0001
cred_deb2	1. cred deb sup 0	1	0.3942	0.0477	68.1696	<.0001
avoir_liq2	1. avoir sup 107 liq sup 72.3	1	0.6849	0.0420	266.3822	<.0001
age3	1. sup 83	1	0.4381	0.1138	14.8291	0.0001
anc_cli_bqe2	2. sup 22	1	0.2022	0.0517	15.2960	<.0001
mtn_cptecourant2	2. sup -0.09	1	1.1859	0.0416	811.1656	<.0001
type_compte2	1. Compte individuel uniquement	1	0.2873	0.0613	21.9715	<.0001
type_compte2	3. Compte joint uniquement	1	0.3725	0.0652	32.6928	<.0001
topAssulARD2	1. avec	1	0.1520	0.0435	12.2188	0.0005
top_Gar_Cnp2	1. avec	1	0.3316	0.1022	10.5214	0.0012
topFacilite2	1. avec	1	0.5062	0.0689	54.0394	<.0001
CSP3	1. Retraites	1	0.2497	0.0541	21.2702	<.0001

Figure 52: Coefficient estimators - Model 4

As it can be seen on the coefficient estimators figure (Figure 52), there is no issue anymore: no collinearity and all the modalities are significant at a 99% level of significance. One might think that this would affect the performance of the model but actually the model is one of the best. Indeed, the Somers'D is 0.720 which is the highest of all the models. Moreover, the concordant percentage is 85.3% and the discordant percentage is 13.3%.

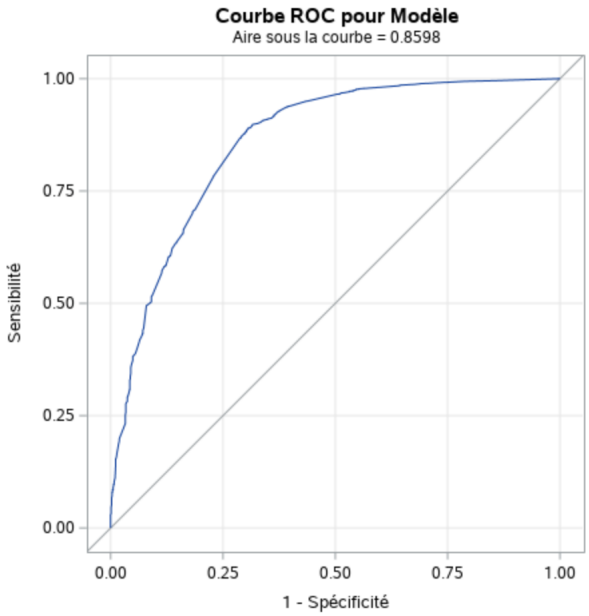

Figure 53: ROC curves - Model 3

The AUC is 0.8598 which is the highest among all the models. It is due to the fact that both the sensitivity and specificity are close to 1. It confirms that this model is the best.

Almost all performance measures are the best for this fourth model and there is no insignificance issue or collinearity issues. Thus, one can conclude that this model is the best one. It is the one that will be used to evaluate the model on the test dataset and for the scoring.

Estimation du rapport de cotes			
Effet	**Estimation du point**	**Intervalle de confiance de Wald à95%**	
cred_deb2 1. cred deb sup 0 vs 2. autres	2.200	1.824	2.653
avoir_liq2 1. avoir sup 107 liq sup 72.3 vs 2. autres avoir liq	3.935	3.338	4.638
age3 1. sup 83 vs 2. inf 83	2.402	1.538	3.751
anc_cli_bqe2 2. sup 22 vs 1. inf 22	1.498	1.224	1.835
mtn_cptecourant2 2. sup -0.09 vs 1. inf -0.09	10.716	9.102	12.616
type_compte2 1. Compte individuel uniquement vs 2. Compte joint et individuel	2.578	1.918	3.467
type_compte2 3. Compte joint uniquement vs 2. Compte joint et individuel	2.808	2.069	3.810
topAssulARD2 1. avec vs 2. sans	1.355	1.143	1.607
top_Gar_Cnp2 1. avec vs 2. sans	1.941	1.300	2.897
topFacilite2 1. avec vs 2. sans	2.752	2.101	3.605
CSP3 1. Retraites vs 2. Autres	1.648	1.333	2.037

Figure 54: Odds Ratio - Model 4

The model enables to know how each modality affects the chances of default:

- Individuals with credit and debit values higher than 0 have a lower probability to default than the ones with credit and debit values lower than 0. Actually, the odds to not default when the

credit and debit values are higher than 0 are 2.2 times higher than when the credit and debit values are lower than 0.

- Individuals with a bank liquidity higher than 72.3 euros and a credit higher than 107 euros have lower chances of default than the others. Their odds to pay their loan is multiplied by 3.935 with respect to the others.
- Individuals older than 83 years old have a higher probability to not default (odds to pay the loan multiplied by 2.402) than younger individuals.
- Retired persons have a lower probability of default compared to active persons. Their odds to pay the loan are 1.648 times higher than the odds of active persons.
- A person with a seniority higher than 22 years has a lower probability of default than another with a seniority lower than 22 years. Their odds to pay the loan are 1.498 times higher than the odds of active persons.
- Individuals with unique individual bank accounts or unique joint accounts have lower chances of default than individuals with others types of accounts. The odds to pay the loans are respectively multiplied by 2.578 and 2.808 compared to individuals not in these categories.
- Individuals with an IARD insurance have a lower probability of default (odds to pay the loan multiplied by 1.355) than the ones who do not.
- Individuals with a facility have a lower probability of default (odds to pay the loan multiplied by 2.572) than the ones who do not.
- Individuals with a CNP warranty have a lower probability of default (odds to pay the loan multiplied by 1.941) than the ones who do not.

4.8 Model evaluation

Once the best model has been trained, it is necessary to evaluate it with predictions on new individuals to see how well the model can generalize. Thus, one tries to predict the default status of the individuals in the test set with the model fitted with the train set.

To visualize the performance of the model, the first tool is the confusion matrix. The true value of the default status are displayed in rows and the predictions are displayed in columns in Figure 55:

- When the true value is 1 and the predicted value is also 1, it is called a True Positive (TP).
- When the true value is 1 and the predicted value is 0, it is called a False Negative (FN).

- When the predicted value is 0 and the true value is 0, it is called a True Negative (TN)
- When the true value is 1 and the predicted value is 1, it is called a False Positive (FP).

Table de top_def_12m_90j par l_top_def_12m_90j			
	l_top_def_12m_90j(Dans : top_def_12m_90j)		
top_def_12m_90j(top_def_12m_90j)	0	1	Total
0	24922 98.49 100.00 98.49	1 0.00 0.00 50.00	24923 98.49
1	381 1.51 99.74 1.51	1 0.00 0.26 50.00	382 1.51
Total	25303 99.99	2 0.01	25305 100.00

Figure 55: Confusion matrix on the test set

Then, one can compute some measures to evaluate the performance of the model. The most famous one is the accuracy: it is equal to the sum of the true positive and true negative observations, divided by the total number of observations. In this case, the accuracy is 98.49%. It is a very good one for a logistic regression. One can observe that the clients who are not in default are well classified: the row percentage is called the specificity and it is equal to 100%. Nevertheless, the clients in default are rarely detected, only 1 out of 382, this ratio is called the sensitivity. The data is very unbalanced and that explains this low result. In this case, the accuracy is not an appropriate measure, it is better to look at the sensitivity and the specificity. The latter measures depend on the threshold used to make the predictions: in this case, if the probability to be in default is greater than 0.5 then the client is predicted as in default. Changing the threshold impacts the predictions and the measures as well. The ROC curve below summarises this information.

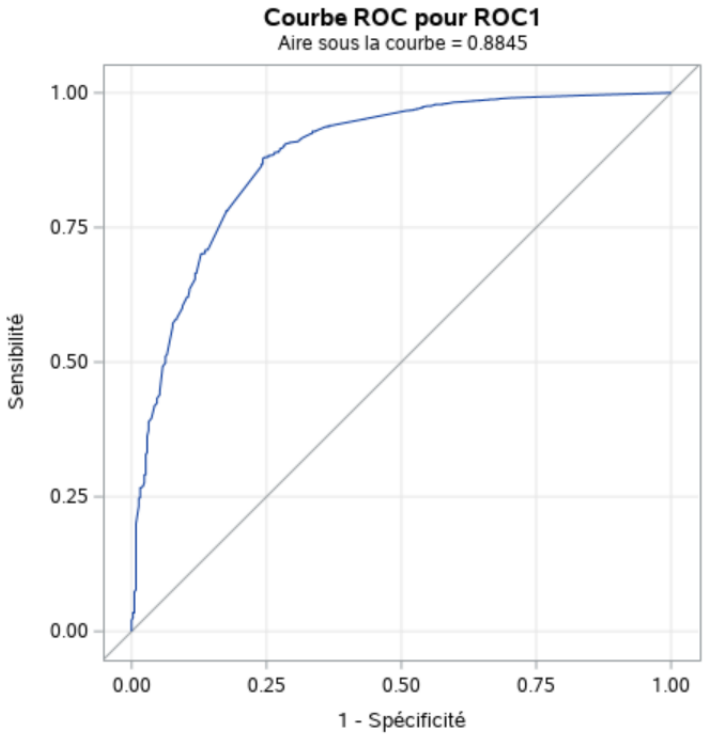

Figure 56: Roc curve - test set

The AUC is another measure to evaluate the performance of a model. The AUC of the model computed on the test set is 0.8845. It is higher than the AUC of the model computed on the train set which was 0.8598. Thus, the model seems very robust.

5 Scoring

5.1 Principle

In the previous part, the fourth model corresponds to the best model of this report. Looking at the model evaluation, this model appears to be robust and then applicable on the whole portfolio. Then, one takes the estimates and applies them to the test sample to get a score from 0 to 1000. The score is set as follows:

- The closer to 0 the score is, then the more likely to default the individual is.
- The closer to 1 the score is, then less likely to default the individual is.

It is called the risky score but is actually more like a safe score.

5.2 Weights

Using the results from the fourth model, one gets the weights.

The formula to get the weights is:

$$Weights(mod1, var1) = \frac{odd_{mod1,var1}}{\sum .(\text{all the odds for all the modalities of all the variables})}$$

On the left column are the features, which are actually dummies. On the middle column, there are the corresponding weights, and as the score is wanted on a 1000-base, the right column shows the weights multiplied by 1000.

Features	Weights	Weights on base 1000
Having credit and debit numbers superior to 0	0.0820485625	82.05
Having avoirs superior to 107 € & liquidity superior to 72.3€	0.1425561853	142.56
Being more than 83 years old	0.0911739745	91.17
Having a seniority as client higher than 22 years	0.0420898486	42.09
Having a amount on the main bank account higher than -0.09€	0.2468173218	246.82
Having an individual bank account only	0.0598045025	59.8
Having a joint bank account only	0.0775361207	77.54
Having a IARD insurance	0.0316415921	31.64
Having a CNP garantee	0.0690096082	69.01
Having a facility paiement	0.1053469106	105.35
Being retired	0.051975374	51.98

Table 5: Weights for each features

Without any surprise, the features the most important are whether there is a positive amount or not on the main bank account, followed by the high capital and liquidity, and by the facility payment accessibility. The least important feature is if a customer has an IARD insurance or not.

5.3 Score distribution and risk threshold

To get a threshold dividing the dataset between the safe customers and the risky customers (the ones more likely to default), one needs to check the quantiles and the distributions of the score.

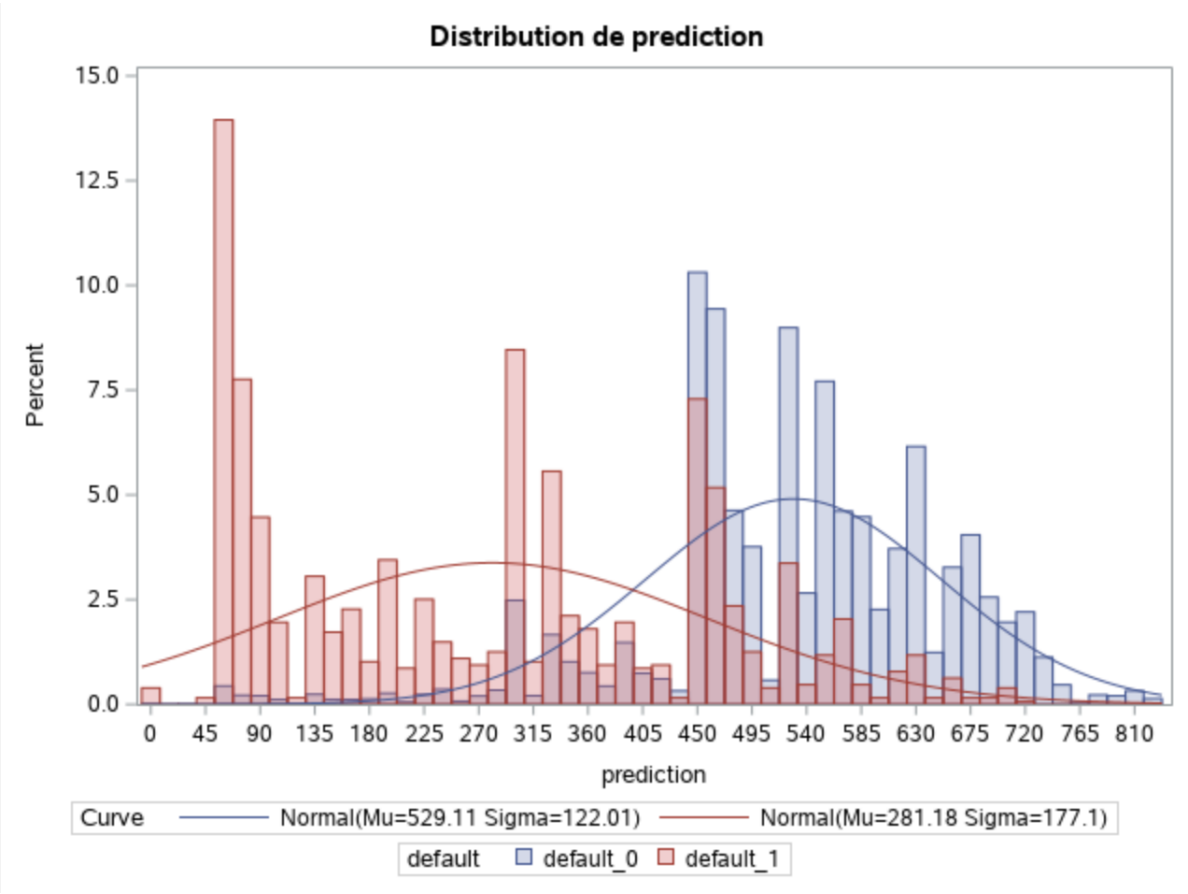

Figure 57: Distribution plot of the score

On Figure 57 above, the red part corresponds to the risky customers while the blue part corresponds to the safe customers. One can see a clear separation at around 465, even if above 465, some risky clients still overlap the safe clients but at a smaller proportions. Actually, looking at the quantiles tables, one can remark that 50% of the safe customers have a score higher than 531.23 and the other safe customers have a score lower than 531.23. On the contrary, 50% of the risky customers have a score lower than 306.62 and the other risky customers have a score higher than 306.62. These two values are confirmed on the plot by the normal density lines that show the center of the distributions.

Quantiles (Définition 5)	
Niveau	Quantile
100Max 100%	829.096
99%	754.956
95%	706.280
90%	674.376
75% Q3	614.844
50% Médiane	531.227
25% Q1	466.910
10%	389.374
5%	306.622
1%	111.780
0% Min	0.000

Figure 58: Quantiles for the score when default = 0

Quantiles (Définition 5)	
Niveau	Quantile
100Max 100%	720.1904
99%	655.1761
95%	576.7690
90%	518.8850
75% Q3	449.1780
50% Médiane	306.6218
25% Q1	91.4461
10%	59.8045
5%	59.8045
1%	59.8045
0% Min	0.0000

Figure 59: Quantiles for the score when default = 1

If one takes 306 as risk threshold, one remarks that 50% of the risky customers have a score higher and then are accepted as "safe customers". However, one remarks that with this threshold, only 5% of the safe customers are refused because considered as risky. It is a balance to do between not considering safe customers as risky ones and not considering risky customers as safe ones.

On Figure 57, one can clearly see a separation between safe and risky customers at a score equaling around 450. Actually, taking 449 as the risky threshold, one can see with the quantiles tables that now no more than 25% of the risky customers will be considered as safe, while only 23% (see Figure 60) of the safe customers will be considered as risky. It seems here as a good compromise.

The Figure 60 shows the contingency table between the score and the default value that sums up what said earlier. For the risky customers, there are exactly 20.45% of them that will have a score higher than 449.179 and then that will be considered safe. Almost 80% of the risky customers will be well considered as risky. On the contrary, there are about 23% of the safe customers that will have a score lower than 449.179 and then will be considered as risky, while 77% of the safe of customers are well considered as safe. It seems as a cautious and reasonable threshold.

Table de prediction par top_def_12m_90j			
prediction	top_def_12m_90j(top_def_12m_90j)		
	0	1	Total
prediction<=449.179	19101 22.64 94.95 22.99	1015 1.20 5.05 79.55	20116 23.85
prediction>449.179	63976 75.84 99.59 77.01	261 0.31 0.41 20.45	64237 76.15
Total	83077 98.49	1276 1.51	84353 100.00

Figure 60: Contingency table between the score and the default value

6 To go further: some tries

6.1 Apply the model to moral customers

One might want to see if the fourth model (the best one) works well on other types of customers too. Therefore, one applies the model to moral customers and professionals. One expects the model not to work well, as the public is not the same.

The variables and modalities used are the same but the observations have changed, thus so did the reference modalities. It is the case for the variable *top_Gar_Cnp2* for which the reference is here "1. avec" and *CSP3* for which the reference is "1. Retraites". Moreover, some variables are deleted because they have only one modality: *type_compte2* and *topFacilite2*.

Surprisingly, one gets a relatively good AUC (0.8366), concordant and discordant percentages (respectively 78% and 10.7%), as one can see on the following outputs.

Association des probabilités prédites et des réponses observées			
Pourcentage concordant	78.0	D de Somers	0.673
Pourcentage discordant	10.7	Gamma	0.760
Pourcentage lié	11.4	Tau-a	0.017
Paires	37840	c	0.837

Figure 61: Statistics values

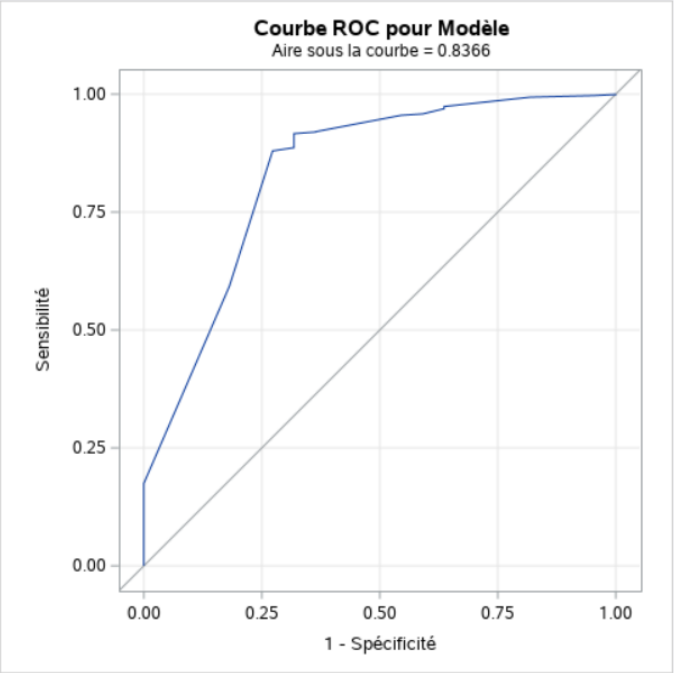

Figure 62: ROC curves

However, one can see that, without any surprise, there are coefficients that are not significant and correlated ones. That makes sense because this model, on moral persons, should not use the same variables.

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > khi-2
Intercept		1	21.8507	568.3	0.0015	0.9693
cred_deb2	1. cred deb sup 0	1	-0.1070	0.2457	0.1899	0.6630
avoir_liq2	1. avoir sup 107 liq sup 72.3	1	1.1054	0.2855	14.9877	0.0001
age3	1. sup 83	1	5.9903	327.5	0.0003	0.9854
anc_cli_bqe2	2. sup 22	1	0.6125	0.5213	1.3804	0.2400
mtn_cptecourant2	2. sup -0.09	1	0.6994	0.2789	6.2865	0.0122
topAssulARD2	1. avec	1	6.0618	271.1	0.0005	0.9822
top_Gar_Cnp2	2. sans	1	0.1590	0.4056	0.1537	0.6951
CSP3	2. Autres	1	5.9862	377.2	0.0003	0.9873

Figure 63: Estimation tables

6.2 Tarification score

To go further, if one wants to do tarification, using the same model, using the quantiles could be an idea. Thus, among the clients that are accepted as safe (i.e. with a score higher than 449.179), a tarification could be tried as follows:

- If the score is between 449.179 and 531.227 (which is the median for the safe customers), then the client would get a high tarification.

- If the score is between 531.227 and 614.844 (which is the third quantile for the safe customers), then the client would get a medium tarification.
- Finally, customers with a score higher than 615 would get a low tarification.

That is theoretical, and one could assume that it is a little risky to do so. Indeed, putting a high prices for the riskier (among the safest), could increase their risk of default.

Furthermore, for tarification, using sensible information (such as sex, origins....) can be forbidden. Here, the age, the family situation... are sensitive variables that can't be used for tarification. However, other variables can be added: the inflation, the score computed to determine if a customer is risky or safe....

Moreover, the prices must be determined before scoring the clients and allocating the tarification to the customers.

Finally, the tarification model must be built in the way that the losses is minimized in case customers (for which, let's recall, the credit has been accepted, i.e. they initially haven't been detected as risky) finally default.

Tarification is one of the hardest job for a data scientist. Here, an acceptation model was built, to decide whether or not a customer will be accepted. A tarification model is way more complex.

7 Conclusion

The best model found so far allows to predict well the non default customers, a little less the default customers. However, its AUC and other evaluations statistics were good, showing its performance. The model evaluation on the test set confirmed this outcome.

It is then adequate to separate the safe and risky customers for the next 12 months. Moreover, the score found seems reasonable and cautious.

Other models could have been used. The main limitation was the use of SAS Studio. Using other softwares such as Python or R, one could have tried other models.

The model is for physical persons and particulars. Applied to moral and professional customers didn't give satisfactory results. A new model should be created.

In the future, a more complex tarification model could also be tried. But this is a very complex model that varies with the current market, with the inflation, with the policies of the bank...etc.