PROJET SCORING : Risque de défaut

École d'Économie de la Sorbonne
Master 2 Modélisations Statistiques, Économiques et Financières
MoSEF - Data science

04 DECEMBRE 2022

GAOUSSOU DIAKITE, EUNICE KOFFI, CAMIL ZAHI
SUPERVISE PAR MOHAMED-SNEIBA HAMOUD
Paris, France

Table des matières

Introduction

Le Système Expert BEST Immo est un outil **d'aide à la décision à l'octroi d'un crédit immobilier**. L'outil permet de rendre un avis en risque pour un projet immobilier à partir des données saisies par le conseiller.

Cette note de risque résulte d'un croisement entre un **diagnostic de solvabilité** qui a été revu fin 2019 par la Direction des engagements et d'un **score statistique** construit et maintenu par la Direction Risques et Contrôles Permanents.

Le score statistique a été mis en production en 2015 sur la base des clients connus. La variable cible du score est le défaut actuel observé à 36 mois à la suite de l'octroi.

L'objectif du projet est de **modéliser la tombée en défaut à 36 mois de dossiers immobiliers** et de **construire une grille de score sur le périmètre clients connus**. Ce score a pour but d'aider à la décision au conseiller pour acceptation d'un dossier de crédit immobilier.

Cet objectif se traduit par :

- La prise en compte d'un profil emprunteur actualisé avec l'utilisation d'un historique de données plus récent ;
- L'exploration et l'étude de l'apport de nouvelles variables discriminantes ;
- Le test de l'apport de données externes (données socio-démographiques par exemple).
- La prise en compte de la Nouvelle Définition du Défaut Bâlois (NDB) comme variable cible (implémentée en août 2020 chez LCL).

Pour se faire, notre projet s'est découpé en plusieurs grandes parties :

- La première partie a consisté à **la construction de la base d'analyse**, à savoir la définition du critère à modéliser, la collecte des données, et les premières statistiques descriptives sur les variables explicatives.
- La deuxième partie a permis de s'occuper de **la sélection des variables**. Basé sur les statistiques descriptives, les tests de liaisons (Kruskall-Wallis, Khi-2, et Pearson), la transformation des variables via la discrétisation et la enfin l'étude de la stabilité temporelle.
- La troisième partie avec **l'estimation du modèle**. De la sélection du modèle, à la création de la grille de score en passant par l'estimation des coefficients.
- La quatrième partie a permis **l'analyse des performances**. Les tests de densités conditionnelles, la courbe de ROC et l'indice de Gini nous ont permis d'observer la qualité de nos modèles.
- La cinquième partie qui a consisté en **la création de classes de risque** avec l'estimation des seuils pour segmenter les individus, l'estimation du taux de défaut par classe et l'étude de la stabilité temporelle de ces derniers.
- Une sixième et dernière partie pour le **challenge de notre modèle** avec des algorithmes de machine learning.

1. Construction de la base d'analyse

La base de données des clients connus est une base de données composées des **163 614 dossiers de clients** ayant fait l'objet d'un octroi de crédit entre avril 2014 et avril 2018 et dont le défaut observé correspondant à au moins 36 mois sur la période. Ce taux de défaut à 36 mois, représentant 0,65% du volume total des dossiers.

a. Le critère à modéliser : le défaut Bâlois

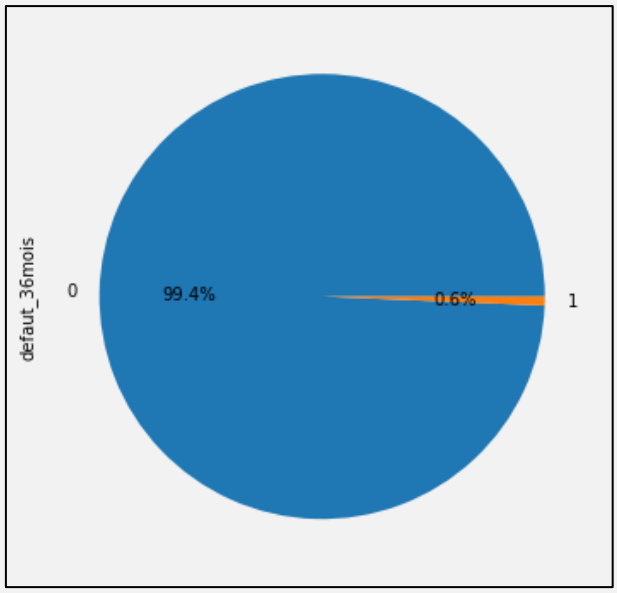
Les critères d'entrée en défaut Bâlois sont :

- Un seuil d'arriérés franchis de plus de 90 jours consécutifs ;
- Une faillite/ procédure judiciaire ;
- Les restructurations pour risque avec perte économique supérieure à 1% ;
- Des provisions sur le compte.

Les critères de sortie en défaut Bâlois :

- Une période de surveillance avant retour en statut non-défaut ;
- De 3 mois pour les clients sans contrat RR (cas général) et dès qu'il n'y a plus d'élément déclencheur défaut ;
- De 12 mois pour les clients avec contrat RR dès le début de la RR.

Notre target : « **defaut_36mois** » est une **variable binaire** composée de 0 et de 1 (0 représentant le non-défaut et 1 représentant le défaut). Ce que le graphique ci-dessous nous aide à constater, c'est que la **base est très déséquilibrée entre la présence de 0 et de 1**. En effet, le nombre d'observations constatant un défaut n'est présent que dans 1059 observations contre 162555 observations sans défaut, soit 0,6% de défaut.



Répartition du défaut dans la base de données connus

b. Les variables explicatives

Sur les 122 autres variables présentes dans la base :

- 72 d'entre elles ont le format : float64 ;
- 40 d'entre elles ont le format : int64 ;
- 10 d'entre elles ont le format : object.

Cependant ce que l'on remarque c'est que certaines variables sont dans le mauvais format. Il conviendra de les transformer par la suite.

Dans la base, on retrouve des **variables relatives à l'emprunteur** :

- L'épargne ;
- La classe de risque ;
- L'ancienneté professionnelle ;
- La situation socio-professionnelle ;
- Le contrat de travail ;
- Le logement actuel ;
- Le nombre d'enfants à charge ;
- La nationalité ;
- L'existence d'un prêt conso ;
- La nature de l'intervenant…

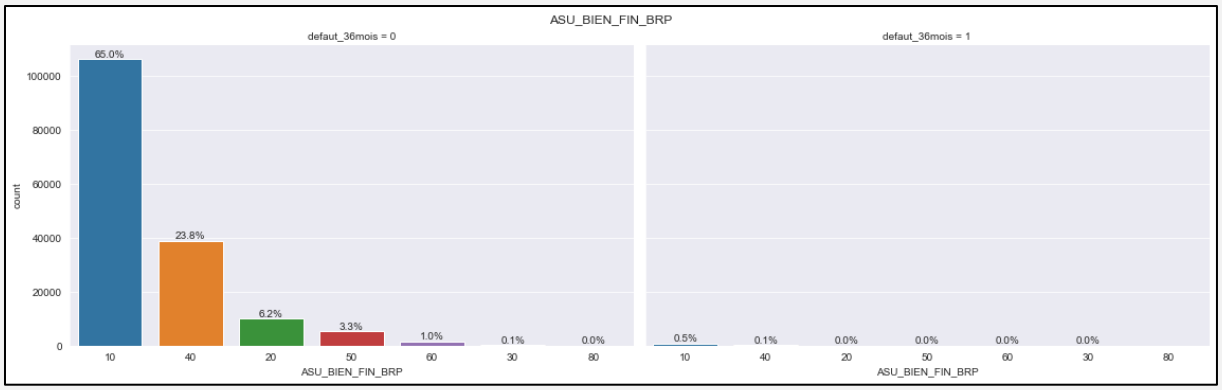
On trouve aussi dans la base des **variables relatives au projet** :

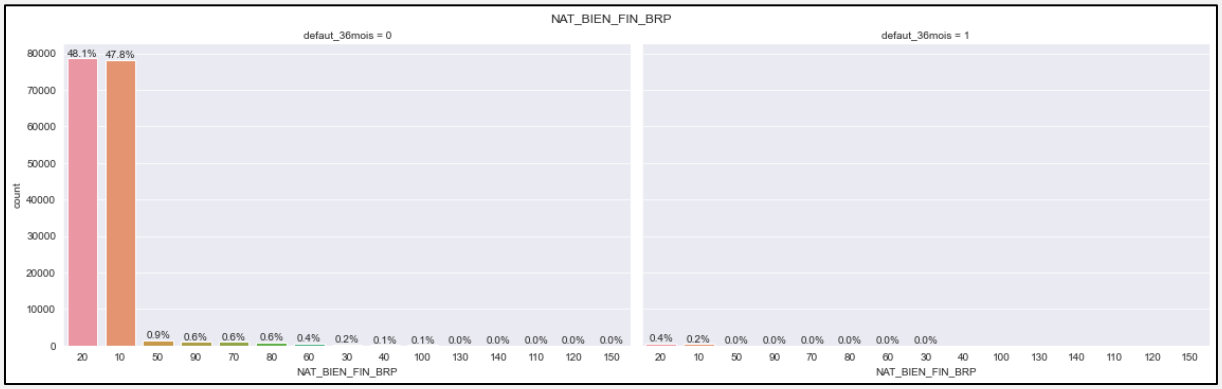
- Le taux d'intérêt ;
- Le ratio d'un prêt à la valeur d'un actif acheté ;
- Le taux d'apport ;
- Le coût des travaux ;
- La nature du projet ;
- L'état du bien ;
- L'usage du bien ;
- La durée du prêt ;
- Le prêt relais ;
- La localisation du bien…

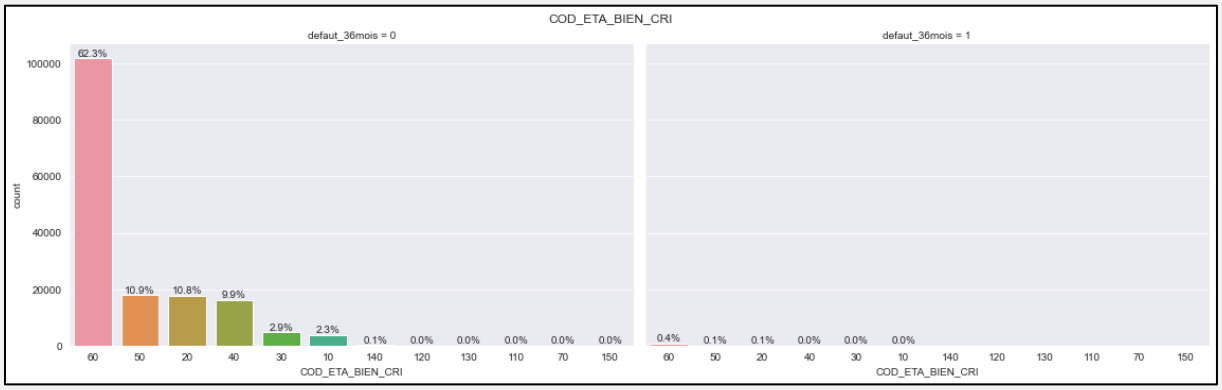
	PCT_ENDETTMT_CRI	TX_APPORT_SE_BRP	LTV_OCTROI_BRP	MNT_PRET_CRI	PCT_NEGOC_CRI	PCT_TEG_TAEG_CRI	MNT_TOT_ASSURANCE_CRI
count	163614.000000	163614.000000	133865.000000	1.636140e+05	163614.000000	163614.000000	160255.000000
mean	33.572435	0.192973	0.763841	1.735032e+05	1.827250	2.834497	9403.724226
std	17.105735	0.232756	0.267408	1.697095e+05	0.659781	0.807864	12181.231467
min	0.000000	0.000000	0.000452	1.000000e+02	0.000000	0.000000	0.000000
25%	25.750988	0.000000	0.570718	7.680625e+04	1.400000	2.287747	1276.490000
50%	32.550930	0.092200	0.835449	1.410170e+05	1.750000	2.763869	5665.650000
75%	39.645657	0.329700	1.000000	2.210000e+05	2.230000	3.305687	13127.895000
max	998.092480	1.025300	4.395022	1.280000e+07	4.970000	16.870178	512038.000000

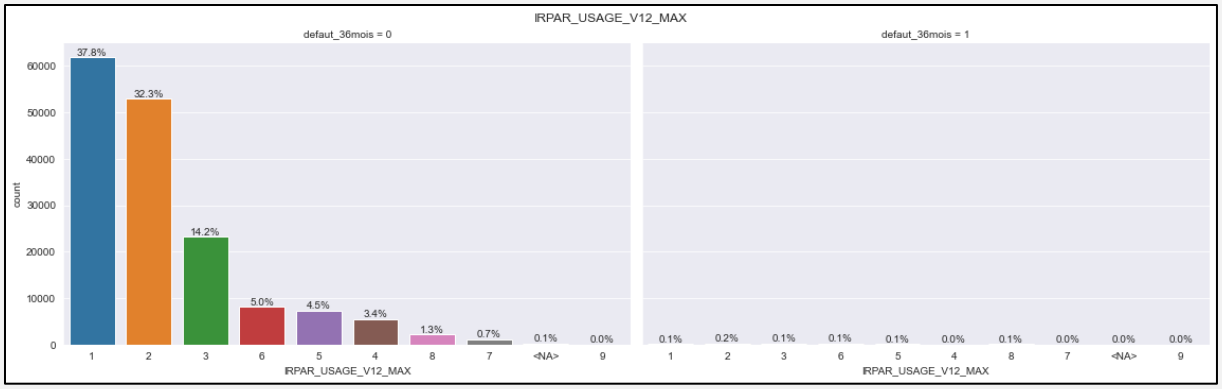
Exemples de statistiques descriptives des variables de la base

En analysant les statistiques descriptives, nous remarquons que la moyenne est souvent assez éloignée de la médiane. En regardant les distributions, on peut remarquer qu'elles sont très souvent asymétriques. Il faudra donc imputer les valeurs manquantes par la médiane.









Exemples de distributions des variables de la base selon la modalité de la variable défaut_36mois

Certaines variables contiennent trop de modalités avec de faibles effectifs. Il faudra donc faire un regroupement de modalités pour plus de clarté dans notre code.

2. La sélection des variables

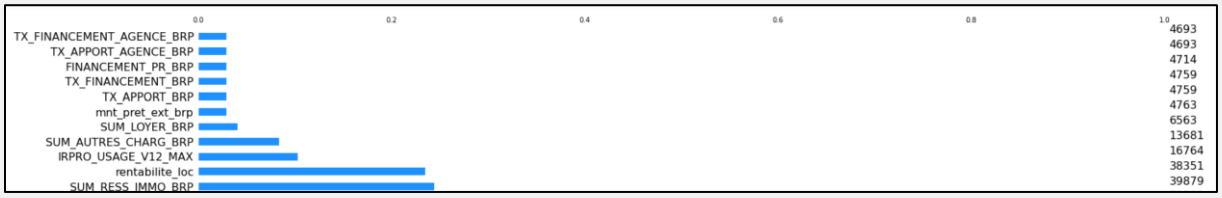
La base de données comprend 123 variables.

a. Traitement des valeurs manquantes

Il est important de comprendre pourquoi on a des données manquantes pour une colonne donnée dans un ensemble de données. Parmi les raisons possibles de l'absence de données, le champ peut ne pas être applicable à tous les clients, informations facultatives ou non disponibles…

On considère qu'au-delà de 70% de valeurs manquantes, une stratégie d'imputation risque d'introduire un biais dans l'analyse. Nous décidons dans un premier temps de **supprimer les variables qui ont au moins 70% de valeurs manquantes.**

Cela nous a permis d'exclure 11 variables. Nous nous retrouvons avec 112 variables.



Taux de valeurs présentes pour chacune des variables supprimées

A gauche, le nom des variables, accompagné de la barre horizontale représentant le taux de valeurs présentes. En haut, les taux de valeurs présentes et à droite le nombre de valeurs présentes dans la base.

Aide à la lecture : La variable « TX_FINANCEMENT_AGENCE_BRP » admet une valeur présente dans la base seulement dans moins de 5% des observations.

b. Présélection de variables niveau métier

Pour chaque critère nous retrouvons plusieurs variables qui y sont rattachées dans la base de données. Une sélection des variables adéquates à chaque critère s'est faite selon certaines règles :

- **la suppression des doublons** ;
- **la non-pertinence des valeurs pour une variable** ;
- **un nombre important de valeurs manquantes** ;
- **un grand nombre de modalités de la variable** (par exemple nationalité car il y a trop de pays).

Ces études nous ont permis d'exclure 63 variables. Nous nous retrouvons avec 49 variables restantes.

c. Le test de Kruskall-Walis

Ce test est usuellement utilisé pour les variables quantitatives. Il permet de **vérifier l'égalité des médianes de la variable conditionnellement aux modalités de la variable cible**.

	Feature	Statistique	p-value
0	defaut_36mois	130890.000000	0.000000e+00
1	PCT_TEG_TAEG_CRI	65.631159	0.000000e+00
2	MNT_TOT_ASSURANCE_CRI	81.405208	0.000000e+00
3	MNT_COUT_TOT_CREDIT_CRI	75.171593	0.000000e+00
4	top_exist_conso_revo_BRP	122.756453	0.000000e+00
5	Anc_banc_INT_jr_MAX_BRP	54.038911	0.000000e+00
6	SUM_EPARGNE_BRP	75.812306	0.000000e+00
7	NBR_DUREE_TOT_PRET_CRI	42.754829	1.000000e-10
8	TX_APPORT_SE_BRP	40.376342	2.000000e-10
9	PCT_NEGOC_CRI	41.028207	2.000000e-10
10	MNT_PRET_CRI	26.524401	2.602000e-07
11	LTV_OCTROI_BRP	26.030963	3.360000e-07
12	financement_tot	20.202154	6.967400e-06
13	NBR_ENF_CHAR_BRP	16.089890	6.040570e-05
14	top_locatif	14.632117	1.306686e-04
15	TOP_NAT_FR_CRI	14.451221	1.438366e-04
16	quotite	13.700505	2.143969e-04
17	PCT_ENDETTMT_CRI	12.245597	4.663574e-04
18	COUT_TRAVAUX_BRP	10.477871	1.208129e-03
19	MNT_RESSOURCES_CRI	8.043179	4.567531e-03
20	TOP_PRET_RELAIS_BRP	4.435407	3.520082e-02
21	NBR_INT_BRP	4.191601	4.062470e-02
22	SUM_PATR_IMMO_BRP	2.170529	1.406779e-01
23	cout_projet_tf_ht_BRP	1.722104	1.894224e-01
24	top_pret_int_ext	1.502833	2.202360e-01
25	TOP_SURFINANCEMENT_BRP	1.437020	2.306221e-01
26	MNT_PRET_ENC_LCL_CRI	0.837320	3.601643e-01
27	nb_pret	0.525130	4.686615e-01
28	COUT_BIEN_FINANCE_BRP	0.180522	6.709253e-01
29	ROL_INT_MAX_BRP	0.131742	7.166331e-01
30	top_pers_seule	0.097078	7.553648e-01
31	NBR_AGE_CLIENT_CRI	0.026022	8.718468e-01
32	MNT_PRET_ENC_HRSLCL_CRI	0.015270	9.016551e-01
33	NBR_AUT_CHAR_BRP	0.014441	9.043470e-01
34	TOP_BIEN_FR_CRI	0.001080	9.737788e-01

Résultats du test de Kruskall-Walis

Au seuil de 5%, ce test nous a permis d'exclure 13 variables quantitatives. Nous nous retrouvons avec 36 variables.

d. Le test du Khi-2

Ce test adapté pour les variables discrètes consiste à **vérifier la présence ou l'absence d'une dépendance entre les modalités de la variable et modalités de la cible.**

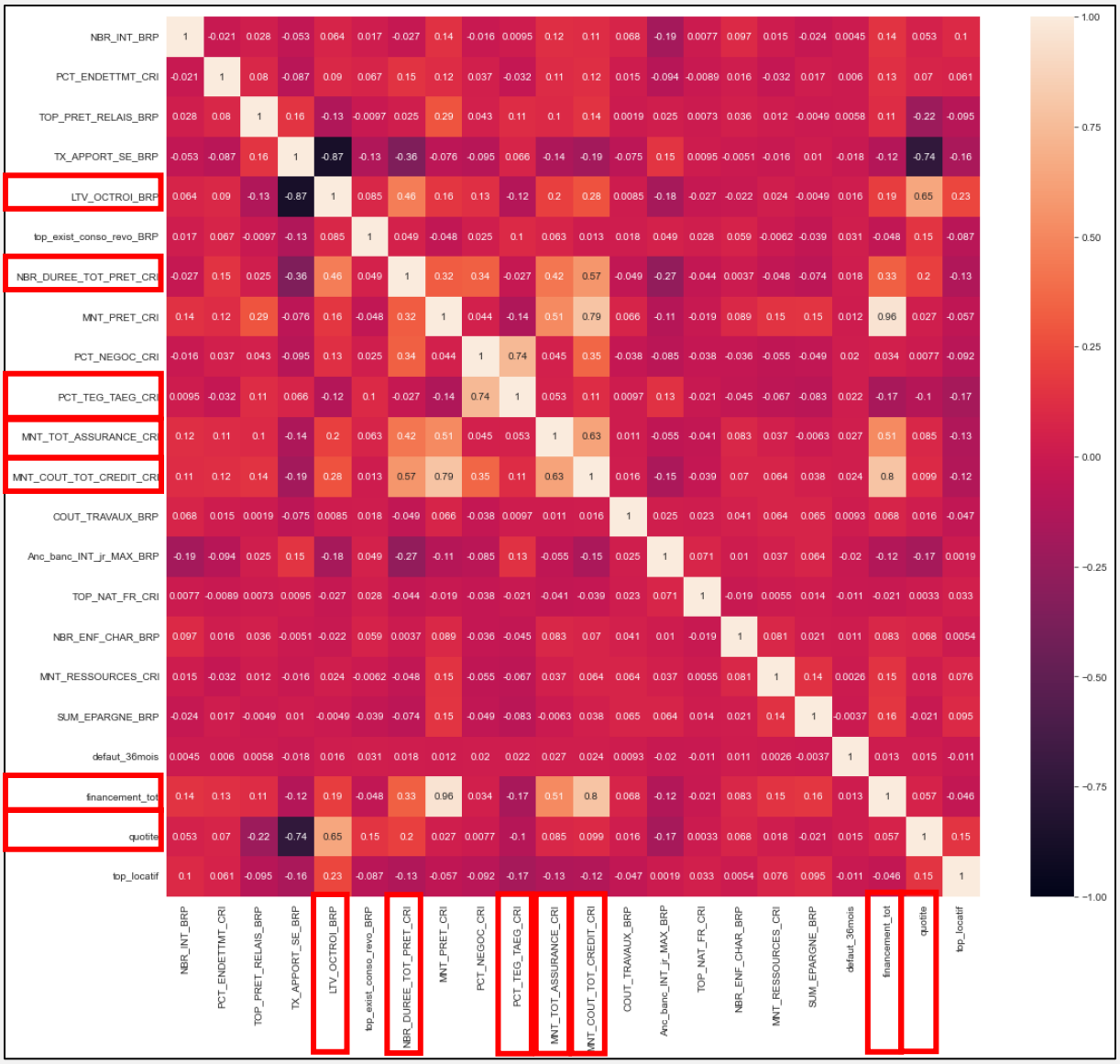
	Feature	Statistique	p-value
0	IRPAR_USAGE_V12_MAX	1511.536937	0.000000e+00
1	TYP_CNT_TRA_MAX_BRP	114.032337	0.000000e+00
2	COD_TYPE_MARCHE_CRI	74.510288	0.000000e+00
3	CSP_RGP_BRP	70.431594	0.000000e+00
4	NAT_BIEN_FIN_BRP	77.409246	1.000000e-10
5	COD_SITU_LOGT_CRI	32.028730	1.968643e-04
6	QUA_INT_MAX_BRP	14.646680	2.144887e-03
7	STA_CLP_BRP	16.253021	2.697778e-03
8	ASU_BIEN_FIN_BRP	19.245903	3.767881e-03
9	COD_ETA_BIEN_CRI	24.880831	9.489923e-03
10	IND_PRIMO_ACCEDT_CRI	8.777836	1.241415e-02
11	SIT_FAM_INT_BRP	6.796127	3.401136e-01
12	IND_INCIDENT_BDF_CRI	0.000000	1.000000e+00

Résultats du test de Khi-2

Au seuil de 5%, ce test nous a permis d'exclure 2 variables catégorielles. Nous nous retrouvons avec 34 variables.

e. L'analyse des corrélations

La matrice de corrélation nous permet **d'identifier les variables colinéaires**. Deux variables très corrélées ne doivent pas être ensemble dans un même modèle. Nous garderons donc les variables qui ont des corrélations inférieures à 30%.



Analyse des corrélations

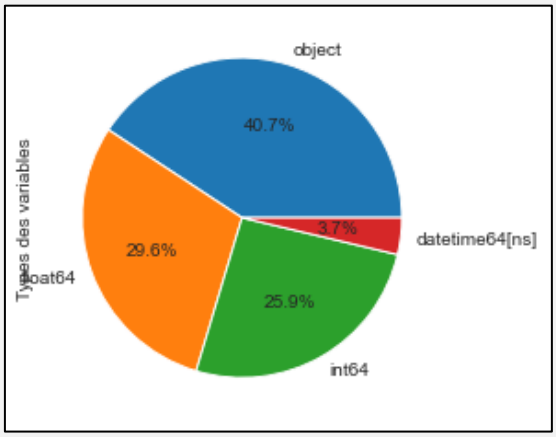
L'analyse des corrélations nous a permis d'exclure 7 variables. Nous nous retrouvons avec 27 variables.

	Nombre de variables supprimées	Nombre de variables restantes
Base connus	---	123
Traitement des valeurs manquantes	11	112
Présélection de variables niveau métier	63	49
Test de Kruskall-Walis	13	36
Test de Khi-2	2	34
Test de Pearson	7	27

Récapitulatif des étapes de la suppression des variables

Ainsi, notre base est composée de 25 variables expliquées, en addition de la date de déblocage et de la variable expliquée « défaut_36mois ».

A ce stade et après avoir modifier le format des variables dans leur bon format, nous obtenons la répartition suivante sur les formats des 27 variables :



Répartition du format des variables présélectionnées

Concernant l'imputation des variables quantitatives, on va se baser sur la distribution de la variable. Comme les distributions semblent souvent asymétriques, on va imputer par la médiane.

Concernant l'imputation des variables qualitatives, on décide d'imputer en créant une classe « NA ».

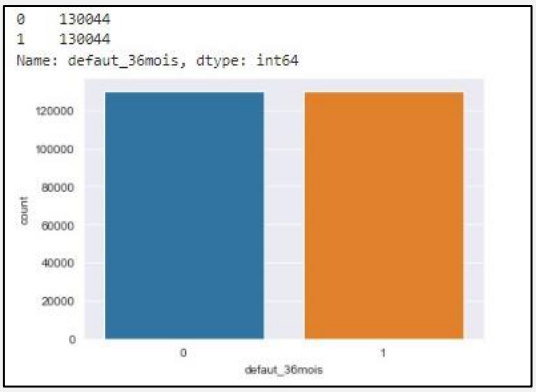
f. Répartition des données entre train et test

On va répartir maintenant nos données dans les ensembles suivants : entraînement (80%) et test (20%). Nous allons effectuer des manipulations sur la partie entraînement pour évaluer préliminairement notre modèle, tandis que l'ensemble de test nous servira pour l'évaluation finale du modèle. Notons que les transformations effectuées dans le train set seront aussi effectuées dans le test.

g. Rééchantillonnage

Notre **jeu de données est très déséquilibré**. En effet, nous avons une très faible proportion de défaut dans notre base de données. Ce déséquilibre de classe est un assez fréquent dans plusieurs cas réels tels que : la détection de fraude, la détection d'intrusion, la détection d'activité suspecte pour n'en nommer que quelques-uns. Cette structure de la base données nous est défavorable dans le sens où notre classification binaire donnera un poids plus important à la population majoritaire (la plus fréquente) au détriment de la classe minoritaire (la moins fréquente). Si nous travaillons avec un tel jeu de données, notre modèle identifiera mal la classe minoritaire mais prédira le plus souvent la classe majoritaire sur laquelle il a appris le plus. Dans notre contexte, mal prédire un défaut coûtera énormément comme pour toute activité. Pour avoir une meilleure performance du modèle, il faudra donc rééquilibrer notre base de données.

Pour se faire nous utiliserons la **méthode de sur-échantillonnage** avec la fonction resample de scikit-learn. Cette méthode rajoutera des observations en faisant un tirage aléatoire avec remise des individus de la classe minoritaire afin de rehausser leur proportion.



Distribution du défaut après rééchantillonnage

h. La discrétisation

Il est important **de discréditer nos variables pour permettre une meilleure performance de notre modèle de classification**. Aussi, dans la suite de notre analyse, une discrétisation facilitera la création de notre grille de score. Il faut cependant faire attention à certaines conditions :

- Chaque modalité doit avoir du sens d'un point de vue métier,
- Limiter le nombre de modalités à 6 maximum selon les variables,
- S'assurer de la stabilité en risque ainsi qu'en volume sur la période.

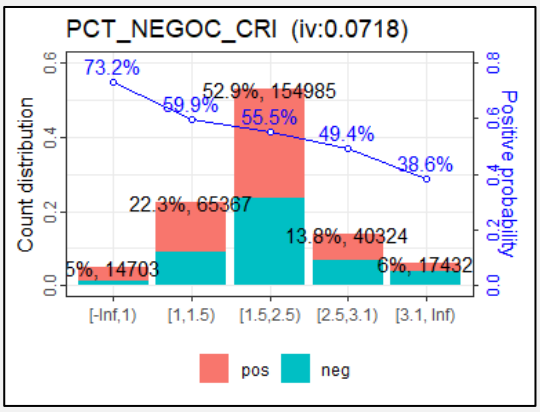
Le poids de la preuve (WOE) et la valeur de l'information (IV) sont des techniques simples mais puissantes pour effectuer une transformation et une sélection de variables.

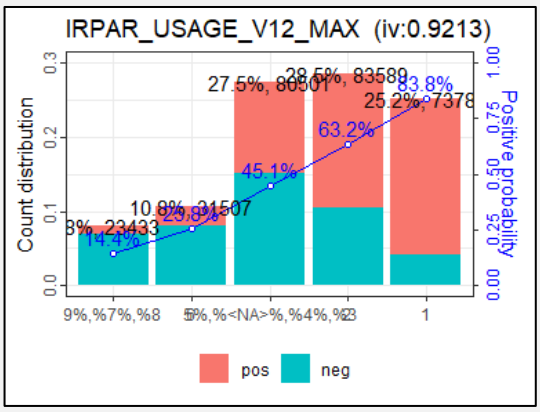
Cet algorithme permet de faire à la fois la discrétisation des variables continues et un regroupement des modalités des variables catégorielles de manière optimale. En effet, nous lui demandons de se baser sur le taux de défaut afin de s'assurer qu'on ait au moins 5% d'effectif dans chaque classe et que le taux de défaut dans chaque classe soit bien différencié. On parle de binning optimal.

L'algorithme utilise des méthodes de segmentation arborescente ou le Chi-2 pour fusionner des modalités. C'est une mesure du pouvoir prédictif d'une variable indépendante par rapport à la variable cible. L'analyse des résultats nous indique donc dans quelle mesure une variable peut différencier les clients sains des clients en défaut. L'un des avantages de l'utilisation de cet algorithme est la gestion des valeurs manquantes et des valeurs aberrantes.

La formule pour calculer le WOE est la suivante : $WoE = ln(\frac{\%\ de\ non\ defaut}{\%\ de\ defaut})$

Un WOE positif signifie que la proportion de bons clients est supérieure à celle des mauvais clients. Un WOE négatif signifie que la proportion de mauvais clients est supérieure à celle des bons clients.





Weight of Evidence (WoE) : Exemples de certaines discrétisation

On attend de l'algorithme de classement qu'il divise un ensemble de données d'entrée en classes de telle sorte que si vous passez d'une classe à l'autre dans la même direction, il y ait un changement monotone de l'indicateur de risque de crédit, c'est-à-dire qu'il n'y ait pas de sauts soudains dans le score de crédit si on passe d'une catégorie socio-professionnelle à une autre par exemple. Cette monotonicité de la proportion des clients en défaut n'est pas respectée pour certaines variables.

L'Information Value (IV) aide à classer les variables explicatives (caractéristiques des clients) en fonction de leur importance relative. Cette importance relative est calculée à partir de la force prédictive univariée de la variable.

Valeur de l'information	Pouvoir prédictif
Moins de 0.02	Non prédictif
0.02 de 0.1	Faible pouvoir prédictif
0.1 de 0.3	Pouvoir prédictif moyen
0.3 de 0.5	Fort pouvoir prédictif
Plus de 0.5	Trop grand pour être vrai

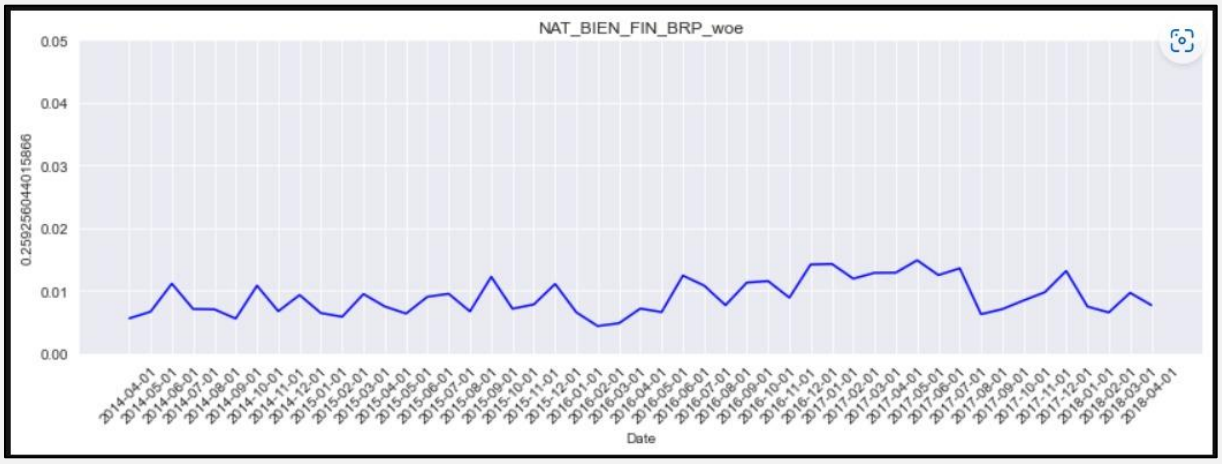
Valeurs possibles de l'IV

variable	total_iv
TOP_PRET_RELAIS_BRP	0.00000
STA_CLP_BRP	0.00000
NBR_INT_BRP	0.00586
IND_PRIMO_ACCEDT_CRI	0.00990
TOP_NAT_FR_CRI	0.01310
QUA_INT_MAX_BRP	0.01337
COD_ETA_BIEN_CRI	0.01591
NBR_ENF_CHAR_BRP	0.01895
top_locatif	0.02031
ASU_BIEN_FIN_BRP	0.02212
COUT_TRAVAUX_BRP	0.02796
PCT_ENDETTMT_CRI	0.03081
COD_SITU_LOGT_CRI	0.03263
MNT_RESSOURCES_CRI	0.04146
NAT_BIEN_FIN_BRP	0.04402
MNT_PRET_CRI	0.05244
CSP_RGP_BRP	0.05810
COD_TYPE_MARCHE_CRI	0.06865
PCT_NEGOC_CRI	0.07180
TX_APPORT_SE_BRP	0.07349
Anc_banc_INT_jr_MAX_BRP	0.09626
TYP_CNT_TRA_MAX_BRP	0.10402
top_exist_conso_revo_BRP	0.13158
SUM_EPARGNE_BRP	0.14706
IRPAR_USAGE_V12_MAX	0.92129

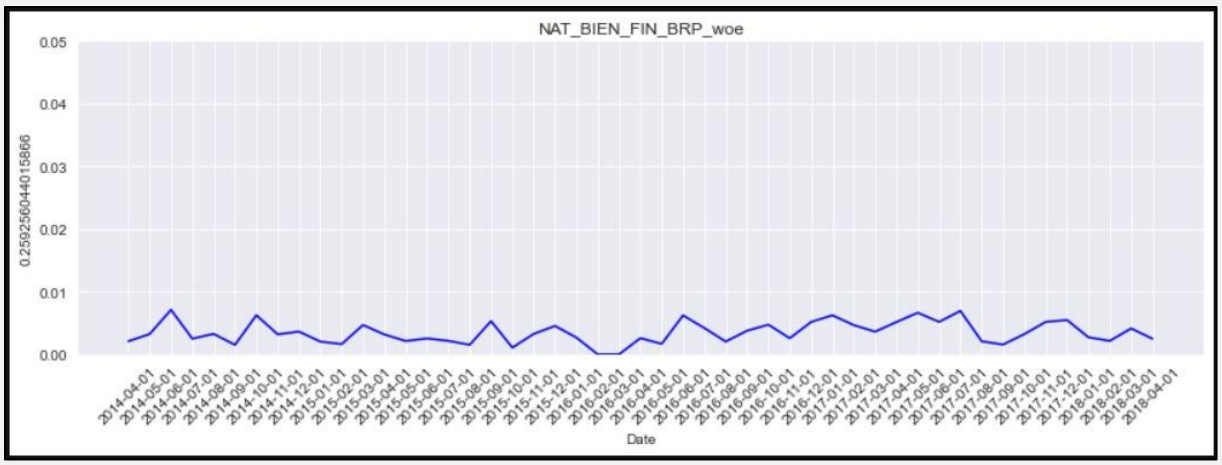
L'Information Value des variables

Les variables avec une IV inférieure à 0,02 ne doivent pas être dans la modélisation, en raison de leur faible pouvoir prédictif. On décide donc de supprimer ces 8 variables et on se retrouve avec 17 variables explicatives, en ajout de la date de déblocage et de la target.

i. Etude la stabilité temporelle



Stabilité mensuelle en volume sur une classe de la variable NAT_BIEN_FIN_BRP



Stabilité mensuelle en risque sur une classe de la variable NAT_BIEN_FIN_BRP

Comme observé sur les graphiques ci-dessus, l'algorithme WOE nous permet d'avoir une discrétisation stable dans le temps en volume et en risque.

3. L'estimation du modèle

a. Choix des variables explicatives

La régression logistique permet de déterminer la probabilité de défaut d'un client en fonction de ses caractéristiques. C'est un cas particulier de modèle linéaire généralisé où l'on utilise le logit comme fonction de lien et une loi de Bernoulli comme composante aléatoire.

On a sélectionné les variables à inclure dans le modèle avec 3 méthodes différentes :

- RFE
- ExtraTreeClassifier
- Corrélations à la variable dépendante

Avec les différentes variables sélectionnées, on estime notre modèle et on a choisit la meilleure combinaison de variables à inclure dans le modèle selon le critère de l'AUC.

Voici les variables que nous avons sélectionnées pour notre modèle :

Variables	**Libellés**
MNT_RESSOURCES_CRI	Ressources Emprunteurs
NAT_BIEN_FIN_BRP	Nature du bien financé
MNT_PRET_CRI	Montant du prêt
CSP_RGP_BRP	Catégorie Socio-Professionnelles
COD_TYPE_MARCHE_CRI	Marché de l'emprunteur (particulier, personne morale)
PCT_NEGOC_CRI	Taux négocié
TX_APPORT_SE_BRP	Taux d'apport
Anc_banc_INT_jr_MAX_BRP	Ancienneté du client ayant les ressources maximales
TYP_CNT_TRA_MAX_BRP	Type de contrat de travail de l'emprunteur
top_exist_conso_revo_BRP	Top existence d'un crédit conso ou crédit revolving
SUM_EPARGNE_BRP	Somme de l'épargne
IRPAR_USAGE_V12_MAX	Classe de risque de l'intervenant principal

Voici les résultats que nous obtenons avec notre régression logistique.

Train score	0,71
Test score	0,70
Cut-Off	0,4

Test	Précision	Recall	F1-score	support
0	1,00	0,56	0,72	32511
1	0,01	0,01	0,02	212
Accuracy			0,57	32923
Macro avg	0,50	0,69	0,37	32923
Weighted avg	0,99	0,57	0,72	32923

b. Grille de score

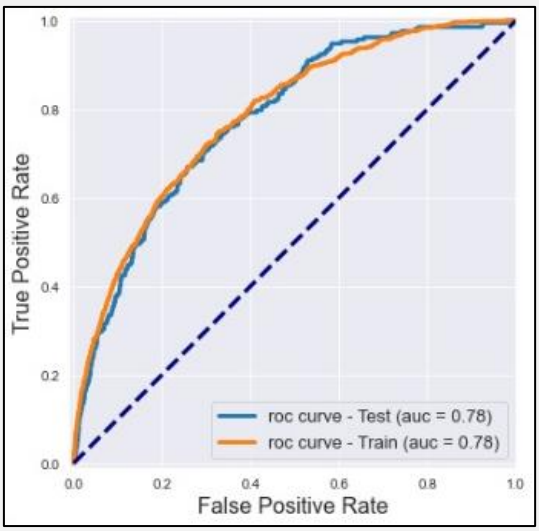
Une grille de score d'une échelle de score allant de 0 (très risqué) à 1000 (peu risqué) a été construit. Pour cela, nous avons associé un poids à chacune des modalités des variables du modèle.

Variables	Classes	Poids
MNT_RESSOURCES_CRI	<40000	67
	[40000;70000[	59
	[70000;85000[	82
	[85000;125000[	62
	[125000,220000[	48
	>220000	36
NAT_BIEN_FIN_BRP	Local mixte, Terrain constructible, Garage, Box, Parking, Annexes, Piscines, Maison individuelle	55
	Appartement, Terrain non constructible, Local Professionnel, Parts de SCPI d'habitation, Parts de SCI,Bien non destiné au logement,Péniche	68
MNT_PRET_CRI	<30000	76
	[30000;50000[	57
	[50000;70000[	80
	[70000;90000[	71
	[90000;190000[	63
	>190000	53
CSP_RGP_BRP	Agriculteurs, Artisans commercants	60
	Ouvriers, Autres	61
	Cadres, professions intermédiaires, employés	61
COD_TYPE_MARCHE_CRI	Particuliers des Professionnels	57
	Marché des particuliers, marché des professionnels	62
PCT_NEGOC_CRI	<1	89
	[1;1,5[	67
	[1,5;2,5[	61
	[2,5;3,1[	52
	>3,1	37
TX_APPORT_SE_BRP	<0,06	58
	[0,06;0,14[	63
	[0,14;0,2[	59
	[0,2;0,56[	65
	>0,56	71
Anc_banc_INT_jr_MAX_BRP	<2000	53
	[2000;3000[	60
	[3000;4500[	66
	[4500;7000[	56
	>7000	69
TYP_CNT_TRA_MAX_BRP	Fonctionnaire exercant pour la ville ou la région, autres (libérale, artisan…)	48
	Privé - CDI, Privé - CDD, RMI, RMA, emploi jeune, Retraité, Inactif (Etudiant, ...)	63
	Privé - CDI période d'essai ou CNE, Privé - CDD, RMI, RMA, emploi jeune, Public - Titulaire, Public - Titulaire période d'essai, Public - Non titulaire, Intérimaire ou intermittent, Chomeurs	68
top_exist_conso_revo_BRP	1	49
	0	68
SUM_EPARGNE_BRP	<5000	53
	[5000;20000[	62
	[20000;30000[	67
	[30000;35000[	55
	>35000	67
IRPAR_USAGE_V12_MAX	Classes 7,8,9	-13
	Classe 6	14
	Classes 5,4,3	46
	Classe 2	73
	Classe 1	113

4. L'analyse des performances

a. Indicateurs de performances

La courbe de ROC permet de représenter l'arbitrage entre les vrai et les faux positifs.



Courbe de ROC

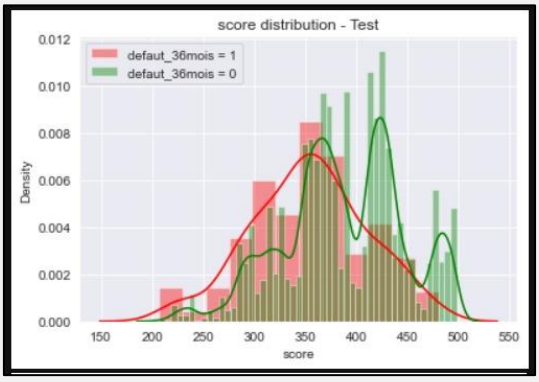
Notre modèle est plutôt satisfaisant puisqu'il performe par rapport à l'aléa.

L'indice de Gini est un indicateur compris entre 0 et 1 qui rend compte de la répartition d'une variable dans une population. Dans un score, on s'attend à ce qu'il soit proche de 1, signe que les défauts sont concentrés dans les classes les plus risquées. Il est égal à : $2 * AUC - 1$.

AUC	0,78
Gini test	0,56
Gini train	0,57

b. Densités conditionnelles

La distribution des scores des individus conditionnellement au défaut. Plus les distributions sont éloignées, plus le score est discriminant.



Densités conditionnelles

On peut voir une différenciation entre les courbes de densités conditionnelles. Le score est discriminant.

c. Stabilité des performances

Le score construit doit faire l'objet d'une étude afin de vérifier qu'il est robuste dans le temps : un score est robuste (stabilité temporelle) s'il est indépendant du temps.

Il est ainsi nécessaire de s'assurer de la stabilité des performances dans le temps.

Nous avons testé la stabilité des deux distributions de scores dans le train et le test set, on obtient un PSI = 0.0002 (PSI : Population Stability Index).

Le résultat ainsi obtenu atteste de la stabilité de nos classes dans le temps.

5. La création de classes de risque

Afin d'orienter la décision du conseiller, on regroupe les individus dans différentes classes en fonction de la probabilité de défaut estimée par le score (exemple : risque faible, risque modéré, risque élevé, risque très élevé).

La segmentation doit répondre aux critères suivants :

- Avoir des classes homogènes vis-à-vis du critère d'intérêt (i.e. défaut);
- Avoir une différenciation appropriée entre les classes ;
- Avoir un nombre minimal de clients par classe.

Pour déterminer les classes, il est possible d'utiliser la méthode des clustering par K-means.

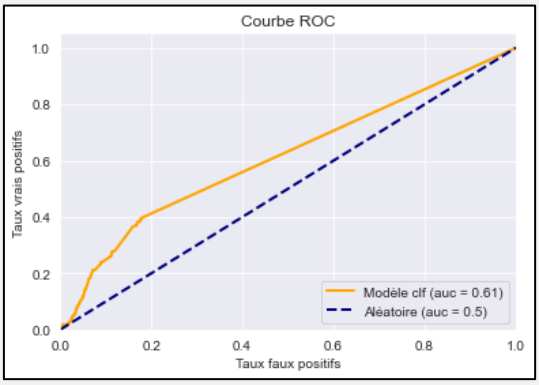
Les taux de défaut par classe de risque doivent être stables sur la période d'étude.

Voici les classes de risque obtenus :

Numéro de classe	Classes	Effectif		Probabilité de défaut	
		test	train	test	train
0	[200,300[	304	6329	0,006579	0,205088
1	[300,350[	416	8651	0,007212	0,198821
2	[350,400[	2037	40978	0,009327	0,205476
3	[400,450[	1196	26139	0,006689	0,206244
4	[450,600[	2595	48794	0,004810	0,206501

6. Machine Learning

Nous avons tenté de challenger les résultats du modèle de scoring avec la régression logistique avec un RandomForest. Cependant, les performances de ce modèle ne surpassent pas les résultats de notre régression logistique.



AUC du Random Forest

---------- Les métriques du modèle ----------

	precision	recall	f1-score	support
0	0.99	0.98	0.99	32511
1	0.01	0.02	0.01	212
accuracy			0.98	32723
macro avg	0.50	0.50	0.50	32723
weighted avg	0.99	0.98	0.98	32723

Métriques du modèle

Conclusion

Nous arrivons à la fin de notre étude qui nous a permis de faire un modèle de scoring en partant de la collecte de données jusqu'à la création de classe de risque. Les différentes étapes parcourues au cours de ce projet nous ont permis d'acquérir de réelles compétences avec des données réelles.

Nous avons pu monter un outil d'aide à la décision avec de bonnes performances. Toutefois, la nature de cet outil ne serait se substituer aux compétences métiers et aux analyses faites par le métier quant à la décision d'octroi d'un crédit à un client.