UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

PREDICTION DU TAUX DE DEFAUT

Master 2 Modélisations Statistiques, Économiques et Financières

Matière : Scoring

Fancello Marie Clara, Germini Eva, Gutfreudn Eloise

TABLE DES MATIERES

0. INTRODUCTION

Le renforcement de la solidité du système financier est un enjeu mondial porté par le comité de Bâle. Ce dernier regroupe des superviseurs provenant de 27 pays afin de faire appliquer une réglementation bâloise commune. L'un des piliers fondamentaux du comité de Bâle est la mesure du risque.

La mesure du risque consiste à mesurer la probabilité de défaut d'un emprunteur. Plus généralement, il s'agit de déterminer sa solvabilité. Dans ce rapport, nous proposons plusieurs modèles afin d'expliquer la variable "BAD" (prenant la modalité 1 si l'individu a fait défaut et 0 sinon) par d'autres variables explicatives. L'objectif principal est donc de trouver le bon compromis biais-variance qui permet un bon ajustement du modèle et une bonne capacité de prédiction sur un nouveau jeu de données.

La base de données que nous utilisons regroupe 5960 prêts sur valeurs domiciliaires et donne des informations sur la délinquance bancaire. Lorsqu'un emprunteur contracte un prêts sur valeur domiciliaire, la valeur de sa résidence est utilisée comme garantie.
Dans cette perspective, nous commencerons par étudier notre base de données par différentes analyses statistiques univariées et bivariées. Dans une deuxième partie, nous effecturons des transformations sur la matrice design afin d'optimiser nos modèles. Dans une troisième partie, nous proposerons trois modèles afin de comparer leur capacité d'ajustement et de performance prédictives.

1. PRESENTATION DES DONNEES

1.1. PRESENTATION DE LA BASE

La base de données est composée de 13 variables et 5960 observations représentant 5960 individus. Nous sommes donc en présence de données individuelles. Les informations disponibles décrivent leur situation. Elles peuvent être décrites en trois catégories. Une première décrivant le prêt : nous disposons du montant de leur prêt, le montant encore dû sur

le prêt, la valeur de sa propriété, et de la raison de la demande de prêt de l'individu. Une seconde catégorie décrit la situation financière de l'individu : son travail, le nombre d'années d'ancienneté dans ce poste, et son ratio dette/revenu. Une dernière catégorie informe sur le risque d'insolvabilité de l'individu tel que : son nombre de rapports de dérogations majeurs, de lignes de crédits en souffrance, nombre d'enquêtes de crédit récentes, et son nombre de lignes de crédit.

Finalement, la dernière variable dont nous disposons est celle nous disant si l'individu est en défaut de crédit ou non, et cette variable sera notre variable d'intérêt.

1.2. GESTION DE LA BASE

Maintenant que nous connaissons notre base de données, nous nous assurons premièrement qu'elle ne comporte aucun doublon. Ensuite, nous séparons la base en deux afin d'avoir une base d'entraînement et une base de test de notre modèle. Cette étape est cruciale car elle permet d'entraîner notre modèle pour qu'il soit le plus performant possible, sans impacter la base de test qui nous permettra d'avoir un modèle applicable à de nouvelles données entrantes, et ainsi inconnues du modèle.

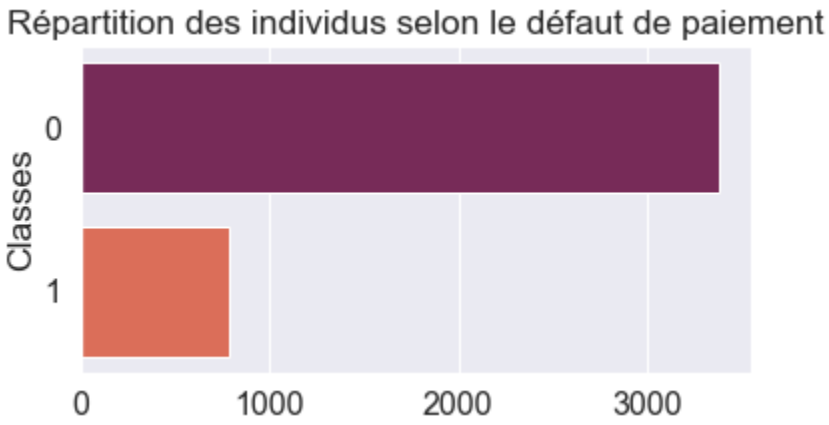
Nous séparons aléatoirement nos données en deux parties. Nous gardons 70% de notre base pour l'entraînement, et 30% pour la base de test.

1.3. STATISTIQUES DESCRIPTIVES

À partir de cette étape, et jusqu'au test du modèle, nous ne nous basons que sur la base d'entrainement. Toutes les modifications que nous faisons sur le dataset est par la suite reproduite à l'identique sur la base de test.

1.1.1. DESCRIPTION DE LA VARIABLE D'INTERET

Notre variable d'intérêt est une variable binaire qui prend ses valeurs dans {0,1}. Elle suit donc une loi de Bernoulli étant donné que la variable a uniquement deux issues possibles :l'une pouvant être assimilée à un succès et l'autre à un échec. Elle comporte 3382 individus ayant la valeur "0", c'est-à-dire ayant remboursé leur prêt, et 790 étant en défaut. La valeur 1 que nous allons prédire est donc représentée plus faiblement que la modalité 0.



1.1.2. DESCRIPTION DES VARIABLES NUMERIQUES

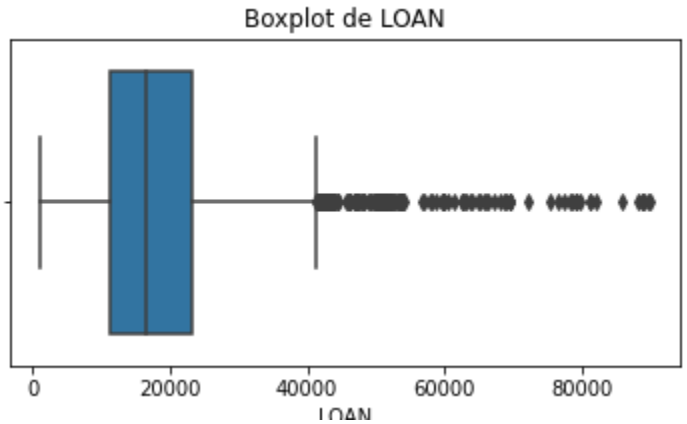
Nous regardons premièrement les distributions de ces variables.

	LOAN	MORTDUE	VALUE	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	YOJ
count	4172.0	4172.0	4172.0	4172.0	4172.0	4172.0	4172.0	4172.0	4172.0	4172.0
mean	18661.79	73219.42	101687.28	0.23	0.38	180.01	1.17	21.25	33.91	8.80
std	11056.28	43102.39	57079.50	0.81	1.01	84.33	1.68	9.96	7.14	7.35
min	1100.0	2063.0	8000.0	0.0	0.0	0.0	0.0	0.0	0.52	0.0
25 %	11300.0	48228.5	67000.0	0.0	0.0	118.06	0.0	15.0	30.69	3.0
50 %	16400.0	65106.5	89539.5	0.0	0.0	173.67	1.0	20.0	34.76	7.0
75 %	23300.0	88265.0	118884.5	0.0	0.0	228.29	2.0	26.0	37.94	12.0
max	89900.0	399550.0	855909.0	10.0	12.0	1168.23	17.0	71.0	133.53	41.0

Par exemple, la moyenne des prêts est de $18661.79, le minimum est de $1100 et le maximum de $89900. Pour le montant restant à rembourser, le minimum est de $2063 et le maximum de $399550. Au niveau de la variable décrivant le nombre de lignes de crédit (CLNO), la moyenne est de 21 lignes avec un écart type de 9.96, le maximum est lui de 71 lignes.

Nous observons aussi les boîtes à moustaches de chaque variable afin d'observer la distribution des valeurs continues, de détecter les valeurs aberrantes et de déterminer notre façon de gérer les valeurs manquantes.

Notre premier constat est qu'il existe des valeurs extrêmes dans toutes les variables. Nous préférons ainsi remplacer les valeurs manquantes par la médiane.

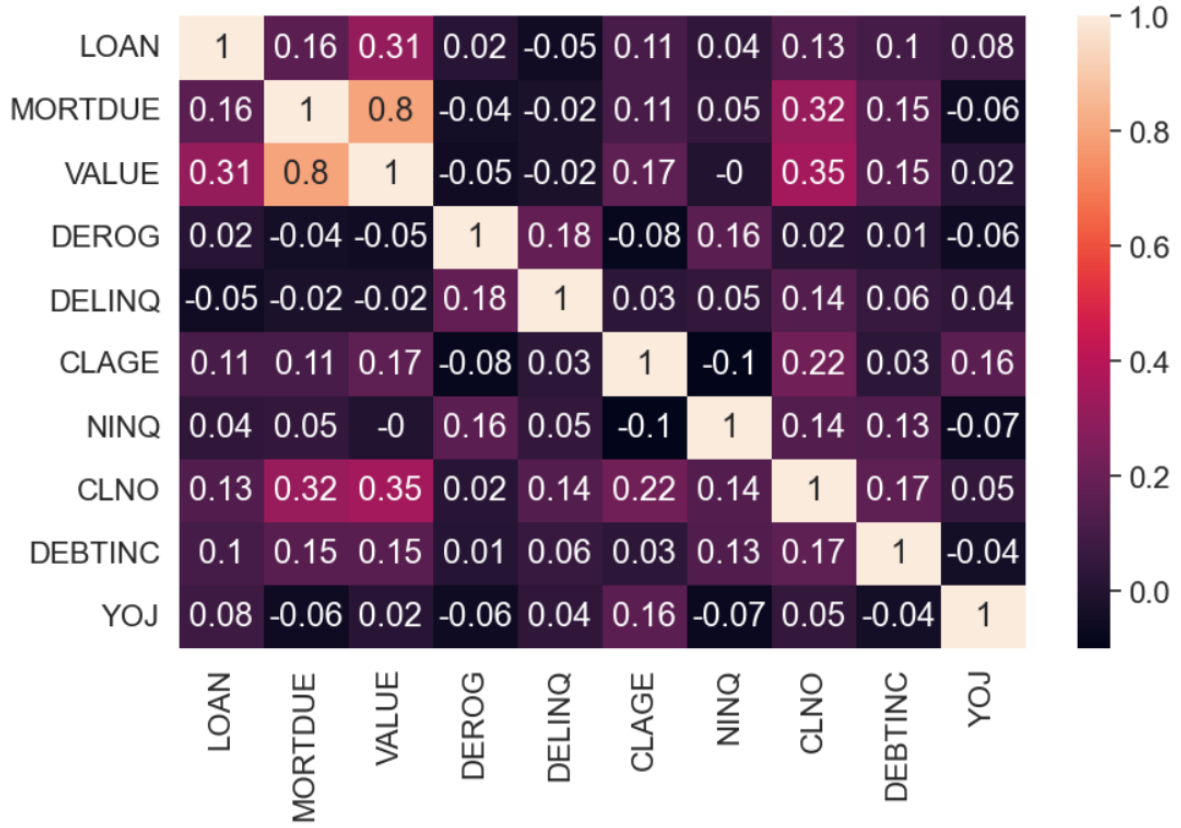


Par exemple, la distribution de la variable du prêt a une médiane légèrement inférieure à 20000. Elle comporte un maximum à 40000 et beaucoup d'outiliers allant de 40000 à plus de 80000. 50% de la population a un prêt comprit entre 11300 et 23300.

Nous décrirons plus amplement les distributions de nos variables par la suite, lorsque nous observerons leurs distributions en fonction de leur classement en défaut ou non.

A présent que nous avons déterminé la façon dont nous allons transformer les valeurs manquantes, nous les remplaçons par leur médiane du set d'entraînement, et appliquons cette transformation aux deux sets de données (entraînement et test).En effet, comme dit précédemment, le test dataset est présent pour tester notre modèle sur des données inconnues au modèle. Nous ne pouvons donc pas utiliser des transformations "inconnues" sur ces données.

Nous regardons enfin les corrélations entre nos variables numériques afin de déterminer si certaines variables sont trop corrélées. Nous utilisons la corrélation de rang Spearman. En effet, le jeu de données comporte des valeurs extrêmes.



Nous constatons que les données ne sont pas fortement corrélées, ayant des coefficients de corrélation entre -0.09 et 0.32, excepté pour les variables “MORTDUE” et “VALUE”. La première définit le montant encore dû par l’individu sur son prêt, tandis que la seconde définit la valeur de sa propriété actuelle. Notre problématique étant d’estimer la probabilité qu’un individu soit scoré 1 et soit alors en défaut de paiement, nous préférons garder la variable “MORTDUE” qui a un pouvoir informatif plus important quant à la créance détenue de l’individu, et supprimons donc “VALUE”.

1.1.3. Description des variables qualitatives

Nous regardons à présent nos variables qualitatives. Elles sont au nombre de deux : “REASON” et “JOB”. La première décrit la raison du prêt et prend deux catégories : la consolidation d’une dette et l’amélioration de l’habitat. Un peu plus de 1000 individus ont contracté leurs prêts pour améliorer leur lieu d’habitation tandis que plus de 2500 individus l’ont contracté pour consolider leur dette.

La seconde variable catégorielle en comporte 6 et décrit la catégorie du travail de l’individu. La plupart de ces derniers sont classés dans la classe “Other” décrivant un autre type de travail que les 5 présents dans les possibilités de réponse.

Ces deux variables comportent respectivement 3.93% et 4.7% de valeurs manquantes, nous décidons ainsi de les remplacer par une nouvelle catégorie que nous nommons “Missing”. Cela nous permet de ne plus avoir de valeurs manquantes, sans pour autant perdre l’information de la valeur manquante (pouvant elle-même être une information).

Nous calculons finalement la corrélation des variables qualitatives grâce au V de Cramer :

Statistique :	Résultat :
p-value	0.0000
Cramer’s V	0.2856

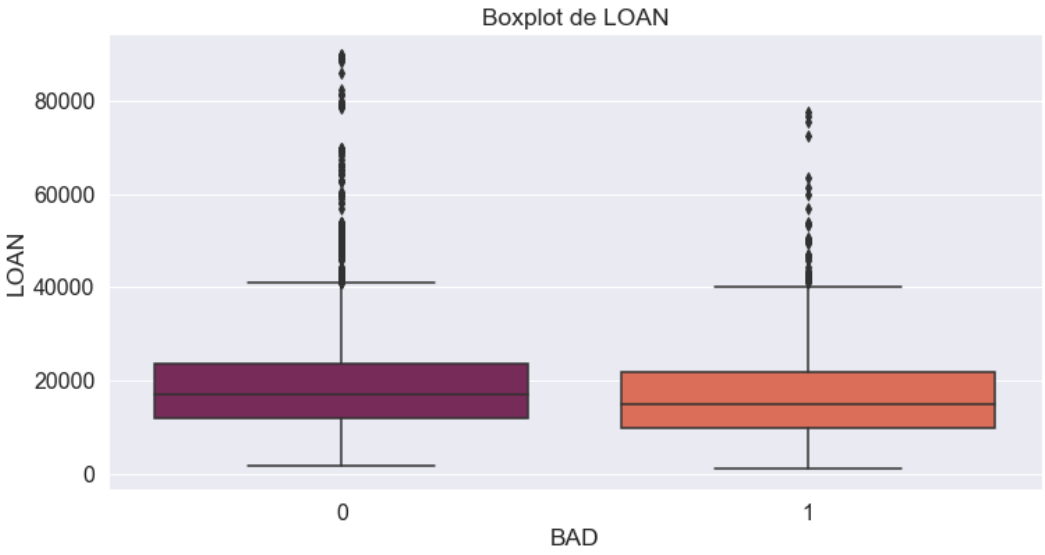
La p_value est inférieure à 0.05 qui est notre seuil critique, la valeur de la corrélation du V de Cramer est donc significative au seuil de 95%. De plus, la statistique du V de Cramer est inférieure à 0.3, ce qui nous indique que cette corrélation n’est pas trop importante. Nous ne supprimons donc aucune de ces deux variables.

1.1.4. ANALYSE CROISEE DES VARIABLES EXPLICATIVES ET DE LA VARIABLE EXPLIQUEE

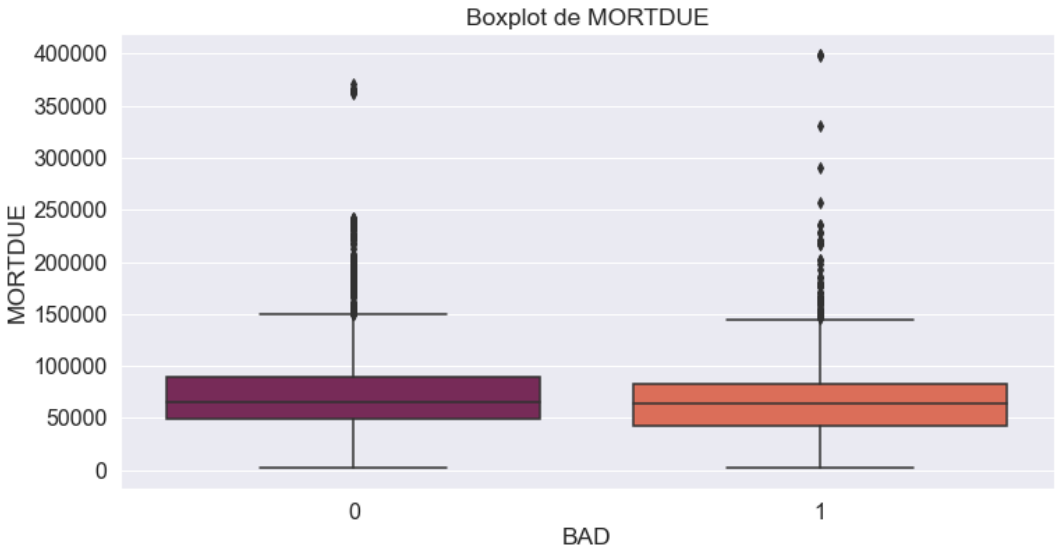
Dans cette dernière partie, nous étudions les variables en fonction de la variable d’intérêt. Cela nous permet d’observer si les distributions des observations diffèrent entre les individus ayant remboursé leur prêt, et ceux étant en défaut de paiement.

DISTRIBUTION DES VARIABLES NUMERIQUES SELON LA VARIABLE D’INTERET

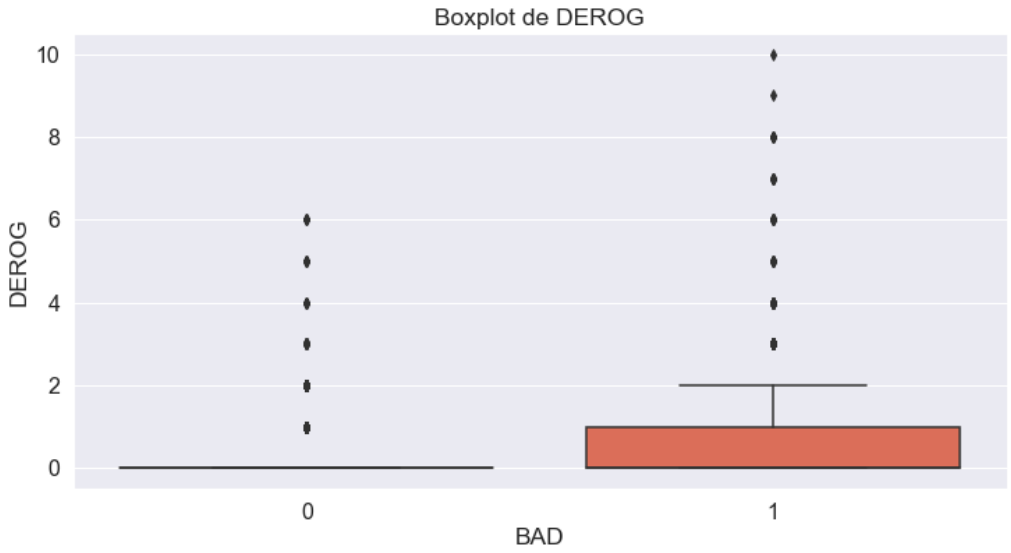
Nous nous intéressons premièrement aux variables numériques. La première de ces variables est le montant du prêt accordé :



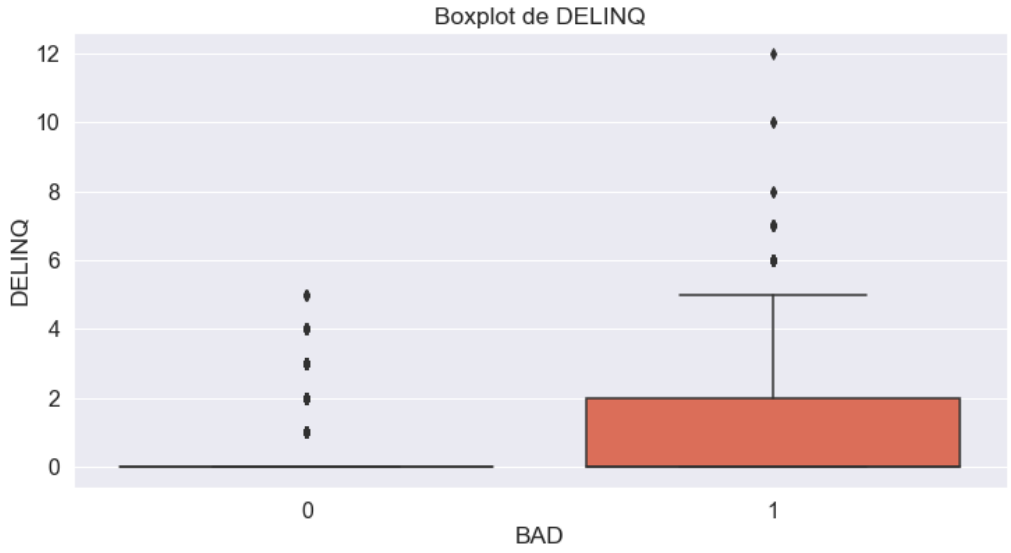
Nous remarquons que les deux distributions ne varient pas particulièrement. Pour les personnes n'ayant pas remboursé leurs prêt, le montant de ces prêts est légèrement moins élevé (car tous leurs quantiles sont inférieurs).



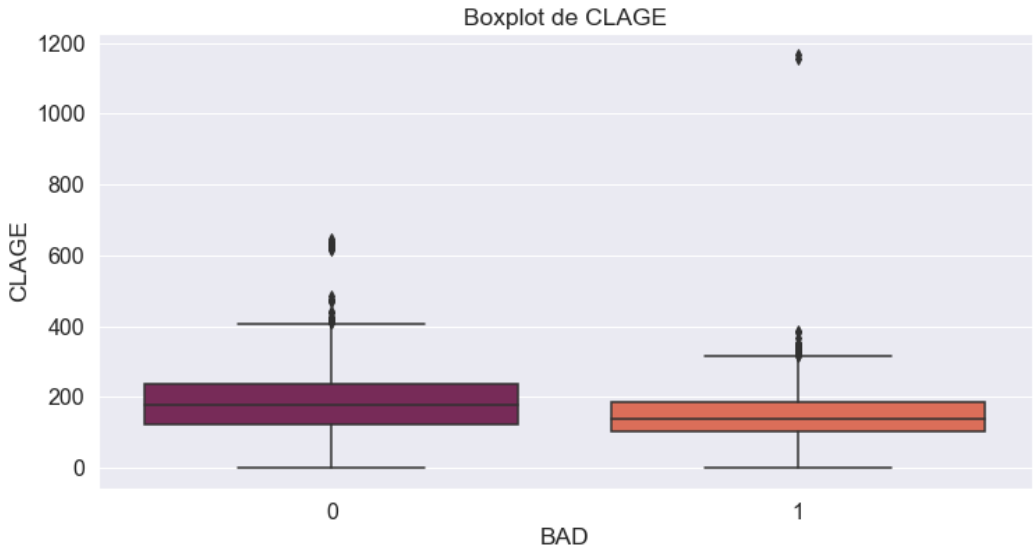
Pour la variable décrivant le montant encore dû par l'individu, les distributions sont semblables. En effet, les quantiles sont légèrement moins élevés, cependant, il y a un nombre plus important d'outliers élevés pour les personnes en défaut.



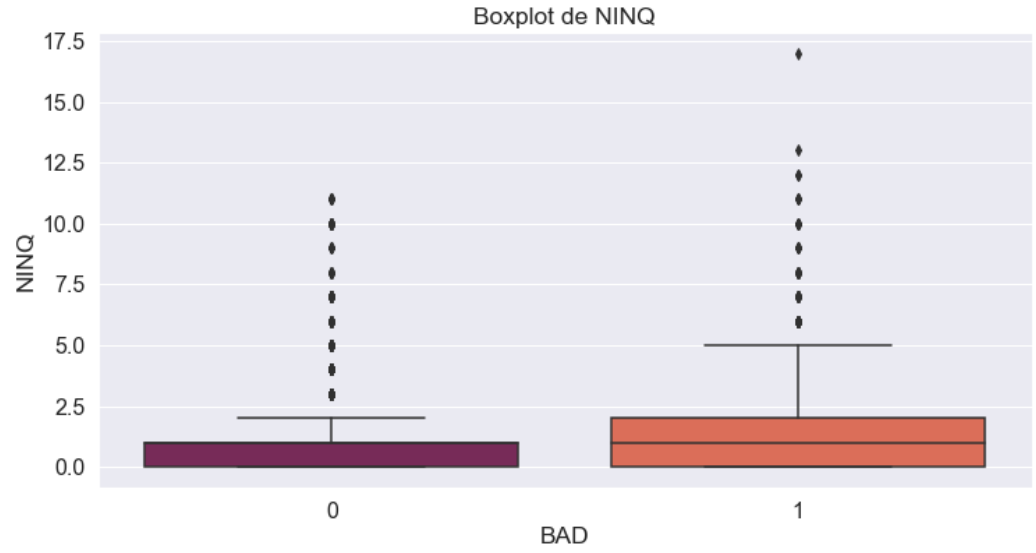
Les distributions pour la variable du nombre de rapports de dérogations majeurs sont-elles tout à fait différentes. Pour les personnes ayant remboursé leurs prêts, il n'y a que 6 outliers et tous les autres individus sont à 0. Pour les personnes en défaut, 50% des individus seulement ont 0 rapports.



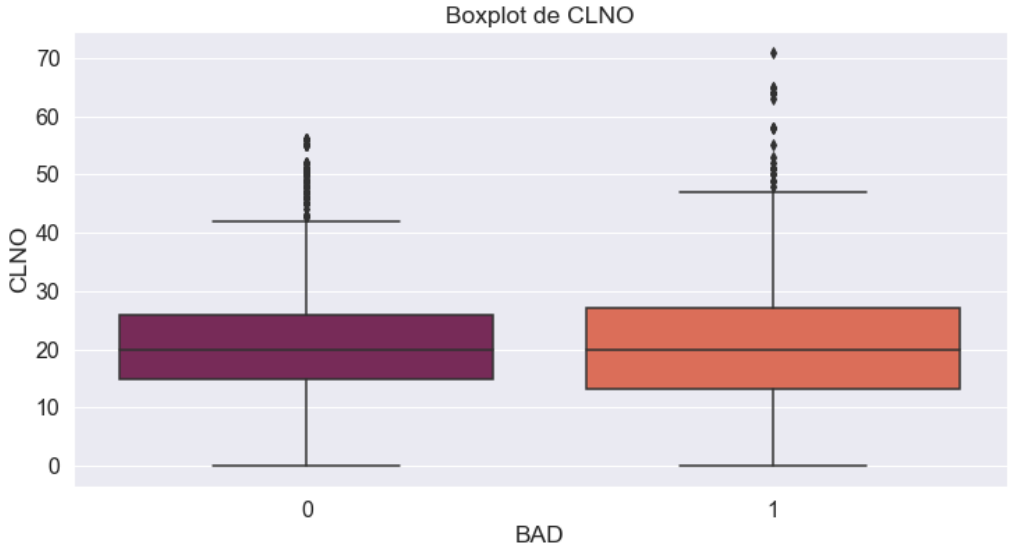
Pour la variable décrivant le nombre de lignes de crédits en souffrance, le résultat est semblable : il existe 5 outliers n'ayant pas un nombre égal à 0 pour les personnes ayant remboursé le prêt, tandis que 50% des individus en défaut en ont plus de 0.



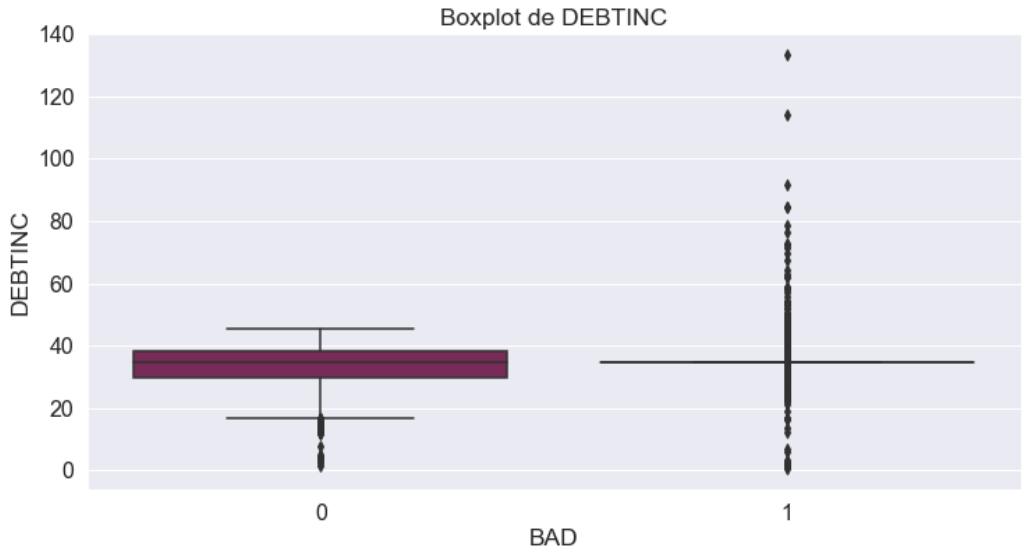
Pour l'âge de la plus vieille ligne de crédit, nous constatons que les individus en défaut ont une distribution plus regroupée autour de sa médiane, cette dernière est inférieure à celle des individus ayant remboursé leurs prêts tout comme le troisième quantile et le maximum de la distribution.



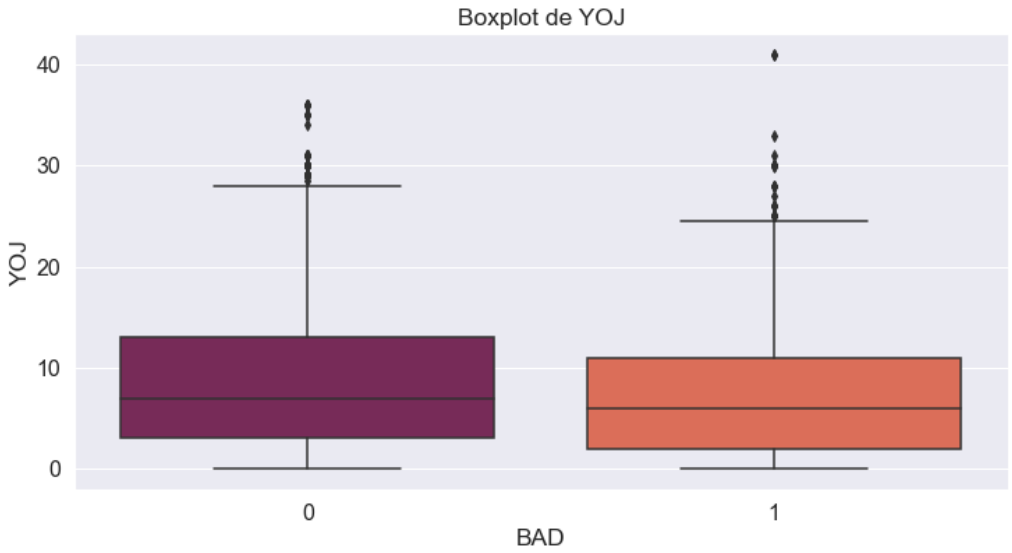
La variable décrivant le nombre d'enquêtes de crédit récentes est également très différente pour les personnes catégorisées en défaut ou non. Ces dernières ont beaucoup de valeurs aberrantes, un maximum de 5 et une médiane de 2.5, égale au troisième quantile des personnes ayant remboursé leurs prêts.



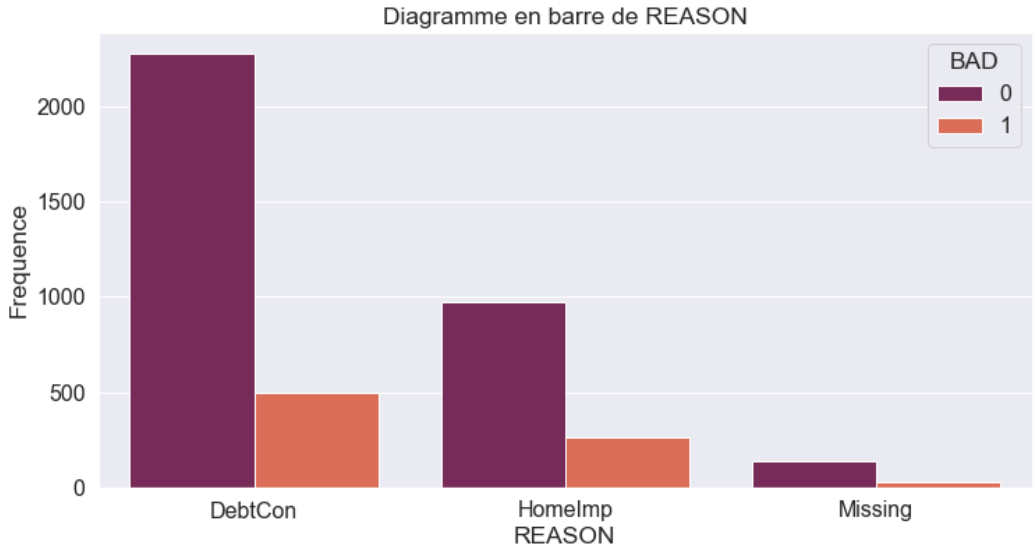
Pour le nombre de lignes de crédit, la distribution des personnes en défaut est moins centrée sur sa médiane. La médiane des deux populations est égale mais les maximums, premiers et troisièmes quantiles sont plus éloignés de cette dernière pour les personnes en défaut. Ils ont également plus de valeurs aberrantes vers le haut.



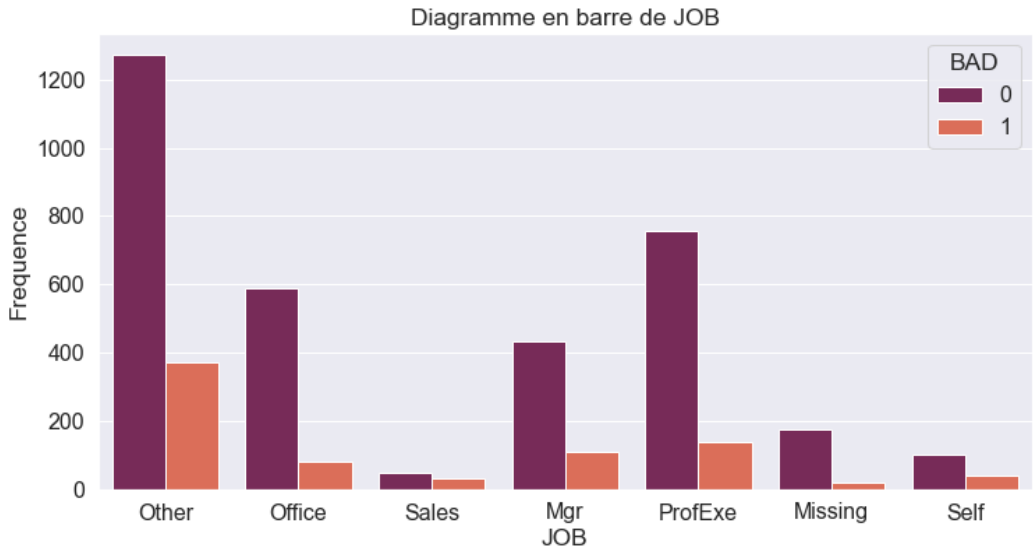
Au niveau des ratio dette-revenu, la population ayant remboursé son prêt a une distribution centrée sur sa médiane, comportant quelques valeurs aberrantes vers le bas. La population en défaut a elle une distribution égale à sa médiane avec des valeurs aberrantes allant de 0 à 140.



Finalement, les distributions de la variable du nombre d'année au métier actuel, la population en défaut est globalement en place depuis moins longtemps. Tous ses quantiles sont inférieurs à ceux de la population ayant remboursé son prêt. Elle comporte cependant quelques valeurs aberrantes, et la personne ayant le plus d'ancienneté dans son travail (environ 40 ans) est une personne en défaut.



Au niveau de la raison de l'emprunt, le nombre de personnes empruntant pour consolider sa dette est majoritaire quel que soit la valeur de la variable d'intérêt. On constate que la moitié des personnes de chaque catégorie (en défaut ou non) a fait un emprunt pour cette raison.



Au niveau de leur travail, les résultats sont similaires, bien que les personnes en défaut soient en minorité, nous constatons qu'ils suivent la même tendance que les personnes ayant remboursées leurs prêts. En effet, les modalités majoritaires pour les personnes ayant remboursées leurs prêts (Other, ProfExe, Office et Mgr) sont également majoritaires pour les personnes en défaut, et les modalités minoritaires (Self, Missing et Sales) sont aussi minoritaires pour les personnes en défaut.

2. TRANSFORMATIONS FINALES

2.1. DISCRETISATION

La technique MDLP (Minimum Description Length Principle) a été utilisée pour discrétiser les variables continues en catégorielles. Elle est basée sur l'entropie. Elle est utilisée dans les problèmes de classification, c'est pourquoi nous avons choisis de l'utiliser. C'est une méthode de discrétisation récursive avec un critère d'arrêt basé sur le principe de longueur minimale de description. Les points de coupure (cut-points) sont déterminés automatiquement.

Nous avons pu utiliser la méthode généralisée afin d'avoir plusieurs intervalles. Sur Python, elle est implémentée sur la librairie scikit-mine. Nous avons discrétisé les variables numériques (LOAN, MORTDUE, DEROG, DELINQ, CLAGE, NINQ, CLNO, DEBTINC et YOJ) présentent dans X_train. Nous avons fait de même pour X_test mais en utilisant les points de coupure trouvés pour X_train.

2.2. ANALYSE DES CORRELATIONS APRES DISCRETISATION

Après discrétisation des variables, nous calculons les corrélations de Cramer afin de mesurer les corrélations entre les variables explicatives et les corrélations entre la variable d'intérêts et les variables explicatives.

	LOAN	MORTDUE	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC	YOJ	REASON	JOB
LOAN	1.0	0.05	0.02	0.06	0.11	0.03	0.05	0.06	0.05	0.19	0.10
MORTDUE	0.05	1.0	0.04	0.02	0.01	0.12	0.05	0.04	0.07	0.05	0.16
DEROG	0.02	0.04	1.0	0.11	0.06	0.11	0.02	0.10	0.04	0.02	0.064
DELINQ	0.06	0.02	0.11	1.0	0.05	0.05	0.04	0.09	0.024	0.06	0.06
CLAGE	0.11	0.01	0.06	0.05	1.0	0.09	0.12	0.05	0.09	0.03	0.17
NINQ	0.03	0.12	0.11	0.05	0.09	1.0	0.08	0.14	0.04	0.07	0.09
CLNO	0.05	0.05	0.02	0.04	0.12	0.08	1.0	0.09	0.08	0.06	0.11
DEBTINC	0.06	0.04	0.10	0.09	0.05	0.14	0.09	1.0	0.07	0.10	0.09
YOJ	0.05	0.07	0.04	0.02	0.09	0.04	0.08	0.07	1.0	0.05	0.09
REASON	0.19	0.05	0.02	0.06	0.03	0.07	0.06	0.10	0.05	1.0	0.29
JOB	0.10	0.16	0.06	0.06	0.17	0.09	0.11	0.09	0.09	0.29	1.0

Un coefficient de Cramer supérieur à 0.3 signifie que la corrélation entre les variables est forte. L'analyse des corrélations confirme qu'il n'existe aucune corrélation entre les variables explicatives car leurs coefficients sont inférieurs à 0.3. Également, aucune variable explicative n'est corrélée à la variable d'output. Cela nous permet de valider le théorème d'Albert et Anderson. En effet, les variables explicatives et la variable expliquée sont bien complètement séparables, ce qui nous confirme l'existence du maximum de vraisemblance.

2.3. ONE HOT ENCODING ET MODALITES DE REFERENCES

Après avoir discrétisé les variables continues, nous transformons les variables catégorielles en variables numériques par la technique du One-Hot-Encoding. Cette technique consiste à représenter des variables catégorielles sous forme de vecteurs binaires. Nous avons effectué cela afin de ne pas avoir d'hypothèse de linéarité. En effet, elle ne nécessite pas d'hypothèses sur la forme de la relation entre les variables explicatives et la variable expliquée. Par ailleurs, cela permettra une interprétation et une comparaison plus simple des résultats. Pour effectuer cela, nous avons utilisé le One-Hot-Encoding. Nous obtenons alors une variable par modalité prenant la valeur 0 si l'individu n'a pas cette caractéristique et 1 s'il l'a.

Afin d'éviter toute multi colinéarité parfaite, nous retirons les catégories les plus discriminantes : ce sont celles contenant davantage d'individus avec la modalité 1 pour la variable d'intérêt. Les catégories de référence non intégrées dans les modèles sont donc :

- LOAN_3 représentants les personnes ayant un montant de prêt supérieur à 15100.
- MORTDUE_1 représentant les personnes ayant un montant de prêt hypothécaire supérieur à environ 39 000.
- DEROG_0 étant le nombre de dérogation (inférieur à 1)
- DELINQ_0 correspondant au nombre de crédit en souffrance (inférieur à 1)
- CLAGE_1 caractérisé par une ancienneté du plus vieux crédit supérieur à environ 174 mois
- NINQ_0 correspondant au nombre de demandes de crédits récents inférieures à 2
- CLNO_2 représentants le nombre de crédits inférieur à 3
- DEBTINC_3 représentants les personnes avec un ratio dettes/revenus compris entre 34.7 et 44,5.

- YOJ_1 donnant un nombre d'années dans l'emploi actuel supérieur à 5
- DEBTCON pour la catégorielle REASON qui indique des dettes consolidées
- OTHER pour la catégorie JOB qui indique une autre catégorie socio-professionnelle.

2.4. IDENTIFIABILITE DES MODELES

Lorsqu'un modèle présente de la multi colinéarité parfaite, il ne peut être correctement estimé car sa matrice de design n'est pas identifiable. La multi colinéarité se présente lorsqu'une variable indépendante est une combinaison parfaite d'autres variables explicatives.

	Variables	**VIF**
21	JOB_Missing	1.33
19	REASON_Missing	1.25
0	LOAN_0	1.15
10	NINQ_2	1.12
25	JOB_Self	1.10
3	DEROG_1	1.08
11	CLNO_0	1.07
12	DEBTINC_0	1.05
24	JOB_Sales	1.05
4	DEROG_2	1.05
6	DELINQ_2	1.05
15	DEBTINC_5	1.04
14	DEBTINC_4	1.03
7	DELINQ_3	1.03
9	NINQ_1	0.37
2	MORTDUE_0	0.31
22	JOB_Office	0.30
17	YOJ_2	0.29
18	REASON_HomeImp	0.27
16	YOJ_0	0.19
23	JOB_ProfExe	0.19
1	LOAN_1	0.13
8	CLAGE_0	0.12

13	DEBTINC_2	0.08
20	JOB_Mgr	0.07
5	DELINQ_1	0.00

Le VIF (Variance Inflation Factor) mesure la gravité de la multi colinéarité dans l'analyse de régression. Si le coefficient du VIF est supérieur à 10, la multi-colinéarité est élevée : la variable apparaît alors plus influente qu'elle ne l'est réellement. Si le VIF est proche de 1, alors le modèle est beaucoup plus robuste, car le poids des variables n'est pas influencé par la corrélation avec d'autres variables. Dans notre étude, le VIF est compris entre 0 et 1.33 pour chacune de nos variables. Nous pouvons donc conclure que notre matrice de design est identifiable.

3. MODELISATION

3.1. METHODOLOGIE

La variable que nous cherchons à estimer et à prédire est la variable binaire "BAD". Cette variable peut prendre deux valeurs 0 et 1. Après discrétisation des variables explicatives, la matrice de design contient 25 variables. L'entraînement du modèle se fait sur l'échantillon nommé train. Le train contient 70% des observations tirées aléatoirement (représentant 4172 observations). La prédiction du modèle est obtenue sur la partie test qui représente 30% de la base de données soit 1788 observations.

Dans cette partie, nous allons procéder à l'exécution de trois modèles: un modèle linéaire généralisé (GLM) sans régularisation des hyper paramètres avec une composante aléatoire qui suit une loi de Bernoulli et une fonction de lien que sera la fonction logistique via la librairies statsmodel de python, un modèle logit avec régulation des hyper paramètres via la librairie sklearn de python et une random forest, également avec régulation via la librairie sklearn. Le choix de ces modèles porte sur deux objectifs : l'estimation des coefficients associés aux variables explicatives et une capacité de prédiction de la variable d'intérêt. Il s'agit ainsi de trouver le bon compromis biais-variance qui permet un bon ajustement du modèle et une bonne capacité prédictive. Les critères d'information bayésiens permettent de mesurer la capacité

d'ajustement d'un modèle. La Receiver Operating Characteristic (ROC) curve et l'Area Under the Cruve (AUC) sont des indicateurs de capacité de prédiction d'un modèle.

3.1.1. DESCRIPTION DU COMPROMIS BIAIS-VARIANCE

Lorsque le modèle est bien entraîné, les coefficients estimés associés aux variables explicatives sont optimaux et interprétables. Cependant, sans régularisation, le modèle peut être confronté à un problème d'overfitting. L'overfitting, également appelé le sur apprentissage, peut être présent au sein d'un modèle lorsque ce dernier est sans biais, c'est-à-dire qu'il a un bon ajustement, cependant, sa variance importante génère de mauvaises prédictions. Dans ce cas, le modèle n'est pas généralisable. L'overfitting peut-être causé par un nombre de variables trop important en comparaison à un nombre d'observation trop faible ou par de la multi colinéarité parfaite.

La régularisation des hyper paramètres pallie aux problèmes d'overfitting. Il existe deux types de régularisation. Premièrement, la régularisation par pénalisation est appliquée dans cette étude sur le modèle logit de sklearn. Intuitivement, elle consiste à imposer une pénalité sur le vecteur de poids des coefficients. Ainsi, un biais est introduit au sein du modèle, cependant, la variance étant plus faible, le modèle est généralisable sur un nouveau jeu de données. Deuxièmement, la régularisation par vote majoritaire de plusieurs modèles est appliquée à la random forest. Effectivement, la random forest construit un grand nombre d'arbres de décision en utilisant des échantillons bootstrap. Les arbres de décision créent une hiérarchie de classement des observations selon différents critères. Pour chaque échantillon bootstrap, nous obtenons les estimateurs ainsi que leur espérance et leur variance. Lorsque nous agrégeons ces estimations, le biais n'est pas modifié cependant, la variance baisse car elle est divisée par le nombre d'échantillons bootstrapés. La variance étant plus faible, le modèle est donc généralisable sur un nouveau jeu de données.

Ainsi, le modèle linéaire généralisé est utilisé sans régularisation afin d'éviter une interprétation biaisée des coefficients, de leurs significations et des odd ratios. La régularisation des hyper paramètres est appliquée sur le modèle logit de sklearn et sur la random forest afin d'optimiser les prédictions de la variable d'output. Les hyper paramètres choisis seront expliqués ultérieurement.

3.1.2. AJUSTEMENT DES MODELES

ESTIMATION DES COEFFICIENTS DES MODELES LOGIT

Le premier modèle que nous avons effectué est un GLM sans régularisation afin d'estimer les coefficients, leurs significativités et leurs odds ratio, sans biais. Ce modèle a trois composantes : une composante aléatoire correspondant à la loi de Bernoulli, une composante déterministe correspondant à notre matrice design et une fonction de lien logit. Ayant supprimé les variables les plus discriminantes après discrétisation afin d'éviter toute multi colinéarité, nous ajoutons une constante dans ce modèle.

Lorsqu'une variable est significative (p-value < 0.05), seul le signe du coefficient peut être interprété : un coefficient positif signifie que la variable indépendante à un impact positif sur la variable dépendante. A contrario, un coefficient négatif signifie que la variable indépendante a un impact négatif sur la variable dépendante. La quantification de cet impact est mesurée par les odds ratio. Un odd ratio de 1 correspond à une absence d'effet. Un odd ratio inférieur à 1 signifie que l'événement est moins fréquent dans le groupe 1 que dans le groupe 0. A l'inverse, lorsque l'odd ratio est supérieur à 1, l'événement est plus fréquent dans le groupe 1 que dans le groupe 0.

MESURE DE L'AJUSTEMENT DU MODELE : LES CRITERES D'INFORMATION BAYESIENS

Les critères d'informations bayésiens permettent de comparer la capacité d'ajustement d'un modèle. Les critère d'information que nous avons choisi pour cette étude sont l'Akaike Information Criterion (AIC) et le Bayesian Information Criterion (BIC):

$$AIC = - 2\ LL + 2p$$

$$BIC = - 2\ LL + p*\log(n)$$

Avec n le nombre d'observations et p le nombre de variables présentent au sein de la matrice de design.

Le calcul de l'AIC et du BIC requiert le calcul de la Log Likelihood (LL). Nous pouvons donc utiliser ces critères de comparaison pour le modèle GLM et le modèle logit de sklearn.

Cependant, la Random Forest n'étant pas ajustée selon le maximum de vraisemblance, nous ne pouvons utiliser ces critères de comparaison.

3.1.3. CAPACITE PREDICTIVES DES MODELES

REGULARISATION DES HYPER PARAMETRES PAR UNE STRATIFIED K-FOLD CROSS VALIDATION

L'échantillon test est utilisé pour tester la performance de notre modèle après régularisation des hyper paramètres. Ainsi, la régularisation s'effectue sur un échantillon de validation. Cependant, notre échantillon d'entraînement étant composé de seulement 4172 observations, l'utilisation de 30% de cette base réserverait seulement 2920 observations pour l'entraînement représentant un risque de faible performance lors de la généralisation du modèle.

La cross validation en k fold est une technique de ré échantillonnage permettant de séparer de manière aléatoire le dataset train en k groupes d'une taille à peu près égale. L'entraînement du modèle et la validation du modèle sont donc répétés k fois. Ensuite, un score est évalué pour chaque groupe k. La moyennisation de ces scores donne un taux de prédiction moyen.

Dans cette étude, nous utilisons une cross validation en cinq k-fold (paramètre n_splits) stratifiés. Dans notre jeu de données train, la modalité 0 représente plus de 3000 observations tandis que la modalité 1 représente moins de 1000 observations. En conséquence, la stratification permet d'échantillonner des ensembles de données où chaque échantillon contient approximativement le même pourcentage de classes cibles. Les autres paramètres entrés au sein du ré échantillonnage sont "shuffle" qui permet un mélange des données avant leur division en cinq k-fold et "random_state" fixant le seed à 42 ce qui permet de maintenir le même tirage aléatoire lorsque l'algorithme est de nouveau généré.

DETERMINATION DES MEILLEURS HYPER PARAMETRES PAR UNE GRIDSEARCHCV

La procédure de la GridSearchCV effectue une sélection des meilleurs hyper-paramètres à utiliser au sein d'un modèle sur une grille de valeurs possibles. Elle est appliquée sur les échantillons train et valide de la cross validation abordée précédemment.

Dans ce paragraphe, nous allons analyser les paramètres pour le modèle logit de sklearn. Premièrement, le paramètre “class_weight” prenant la valeur “balanced” permet de spécifier que le pourcentage de modalité 0 et 1 au sein de la variable d’output est inéquitable. Deuxièmement, le paramètre “penality” spécifie trois pénalités. La pénalité L2 (appelée également pénalité Ridge) est fondée sur une norme euclidienne imposée sur le vecteur de poids afin de pondérer l’importance de certaines variables. La pénalité L1 (appelée également pénalité Lasso) permet de faire de la feature selection en gardant les variables les plus importantes au sein du modèle. La pénalité elastic net utilise la pénalité L1 pour supprimer les variables qui sont inutiles dans l’analyse et la pénalité L2 pour diminuer le poids des variables moins importantes. L'hyper paramètre “C” régule le paramètre de poids pour la régularisation s'appliquant à L1 et L2. L’hyper paramètre “l1_ratio” a cette même fonctionnalité pour la pénalisation elastic net. Plus les hyper paramètres “C” et “l1_ratio” sont élevés, plus la pénalisation est importante.

Dans ce paragraphe, nous allons analyser les paramètres pour le modèle de la forêt aléatoire. Premièrement, l’hyper paramètre “n_estimators” donne le nombre d'arbres composant la random forest. Plus le nombre d'arbres est important, moins la variance sera élevée. Deuxièmement, le paramètre “max_depth” détermine la profondeur d’un arbre. Plus un arbre est profond, plus la probabilité d’overfitting est importante.

Ainsi, pour chaque combinaison de paramètres qui lui sont entrés dans un dictionnaire nommé “parametres”, la GridSearch commence par ajuster le modèle sur l’échantillon d’entraînement X_train. Deuxièmement, elle évalue le modèle sur l’ensemble de validation et sélectionne le modèle avec le meilleur score. Le score que nous avons choisi de maximiser est l’accuracy. Cette métrique calcule la somme des observations bien classées (correspondant au taux de vrais positifs et au taux de vrais négatifs) sur le total des observations (correspondant aux taux de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs). Elle mesure donc le pourcentage de bon classement. Le paramètre n_job fixe le nombre de tests que l'algorithme tourne simultanément.

La dernière étape appliquée aux trois modèles est utilisée pour estimer la capacité de prédiction de notre modèle. La variable dépendante "BAD" est prédite sur le jeu de données de test. L'évaluation des performances se fait par la courbe ROC et la métrique AUC.
La courbe ROC est une représentation graphique de la relation entre le taux de vrais positifs (sensibilité) et le taux de faux positifs (spécificité) pour différents seuils. Cette courbe permet de savoir pour quel seuil nous minimisons le taux de faux positifs et maximisons le taux de vrais positifs. La métrique AUC est un score qui calcule la qualité de précision de notre modèle indépendamment de tous seuils. Plus l'AUC est proche de 1, plus le modèle est précis. Un AUC proche de 0.5 signifie que nos prédictions sont proches de l'aléatoire.

3.2. RESULTATS DES MODELES

3.2.1. INTERPRETATION DES COEFFICIENTS, DE LEURS SIGNIFICATIVITES ET DES ODD RATIOS DU MODELE GLM SANS REGULARISATION

Le premier modèle que nous avons effectué est un modèle GLM sans régularisation afin d'estimer les coefficients, leurs significativités et leurs odds ratio. Les variables ayant une p-value associée à leurs coefficients supérieure à 0.05 sont DELINQ_3, DEBTINC_4, DEBTINC_5, REASON_HomeImp, REASON_Missing, JOB_ProfExe et JOB_Self. Leurs coefficients et odds ratio ne peuvent donc être interprétés.

Les variables ayant un coefficient négatif sont DEBTINC_2, YOJ_0, JOB_Mrg, JOB_Missing et JOB_Office. Les autres variables ont un coefficient positif. A titre d'exemple, un individu ayant un travail dans la vente (JOB_Sales) a plus de probabilité d'être en défaut qu'un individu ayant un travail faisant partie de la catégorie socio-professionnelle classée comme autre (JOB_Other). A l'inverse, un individu ayant un travail dans un bureau (JOB_Office) a moins de probabilité d'être en défaut qu'un individu ayant un travail faisant partie de la catégorie socio-professionnelle classée comme autre (JOB_Other). Autre exemple, un individu avec un nombre de crédit récent supérieur à 2 (NINQ_1, NINQ_2) a plus de probabilité d'être en taux de défaut qu'un individu avec un nombre de crédit récent inférieur à 2 (NINQ_0).

La quantification de cet impact est mesurée par les odds ratio. Reprenons les exemples précédents, un individu ayant un travail dans la vente (JOB_Sales) a deux fois plus de chance d'être en taux de défaut qu'un individu ayant un travail faisant partie de la catégorie socio-professionnelle classée comme autre (JOB_Other). A l'inverse, un individu ayant un travail dans un bureau (JOB_Office) a environ 0.5 fois moins de chance d'être en taux de défaut qu'un individu ayant un travail faisant partie de la catégorie socio-professionnelle classée comme autre (JOB_Other). Un emprunteur avec un nombre de crédit récent compris entre 2 et 4 (NINQ_1) et un emprunteur avec un nombre de crédit récent supérieur à 4 (NINQ_2) ont respectivement 1.7 et 2 fois plus de chance d'être en taux de défaut qu'un individu avec un nombre de crédit récent inférieur à 2.

3.2.2. Comparaison de l'ajustement des modeles par la mesure des critere bayesiens (AIC, BIC)

Nous avons calculé les critères d'information bayésien pour le modèle GLM et pour un modèle logit avec une pénalité lasso et un hyperparamètre "C" fixé à 1 (résultat du meilleur modèle selon la GridSearchCV).

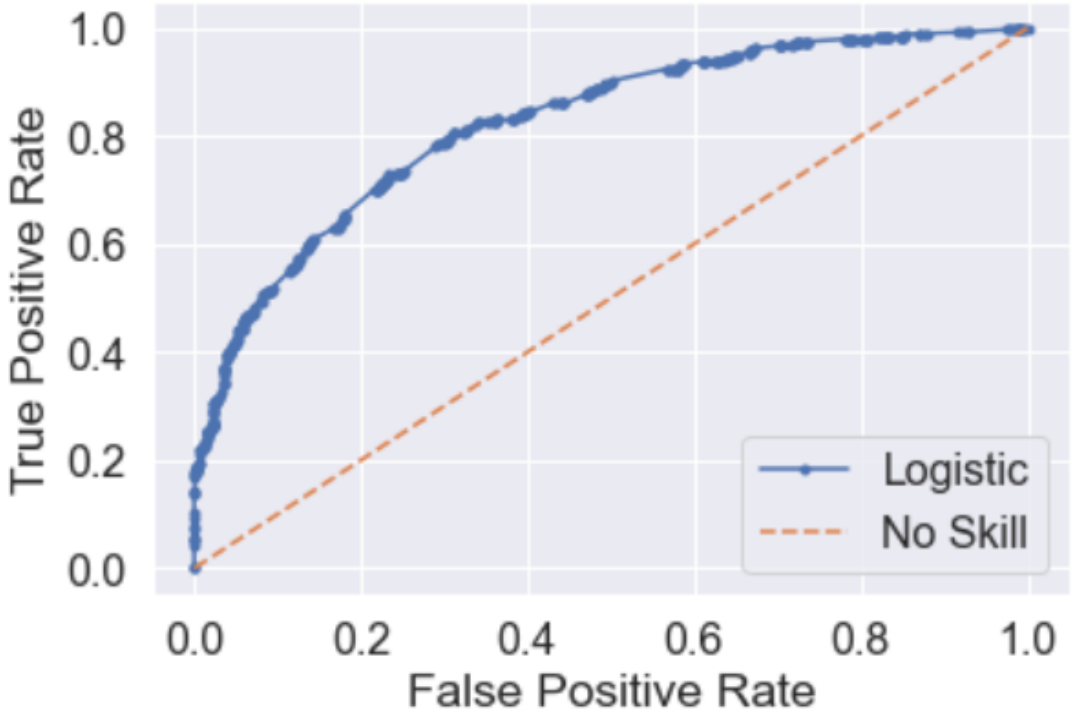
	AIC	BIC
GLM	2733	2911
Logit Sklearn	2688	2701

La minimisation des critères bayésiens indique que le modèle qui ajuste le mieux les données est celui du logit de sklearn. Ce résultat semble tout à fait cohérent car la pénalité L1 permet de réduire le nombre de variables p utilisées au sein d'un modèle.

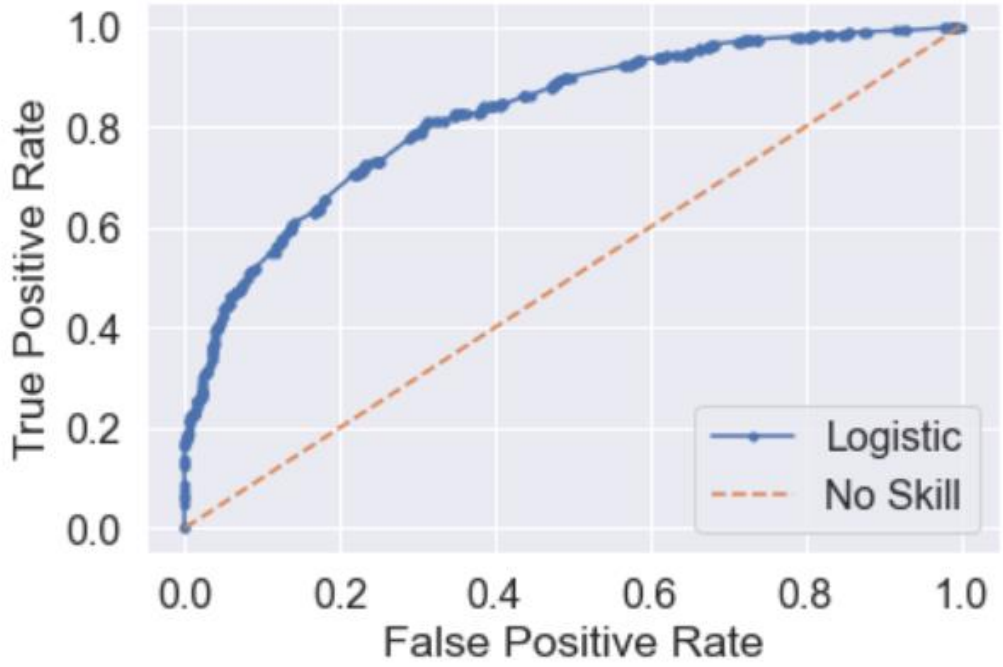
3.2.3. Comparaison des predictions des modeles

Dans cette partie, nous comparons les performances prédictives du modèle GLM sans régularisation, du modèle logit avec une pénalité lasso et un hyper paramètre C=1 et du modèle Random Forest avec un "n_estimators" de 800 et un max_depth de 20 (résultat du meilleur modèle selon la GridSearchCV).

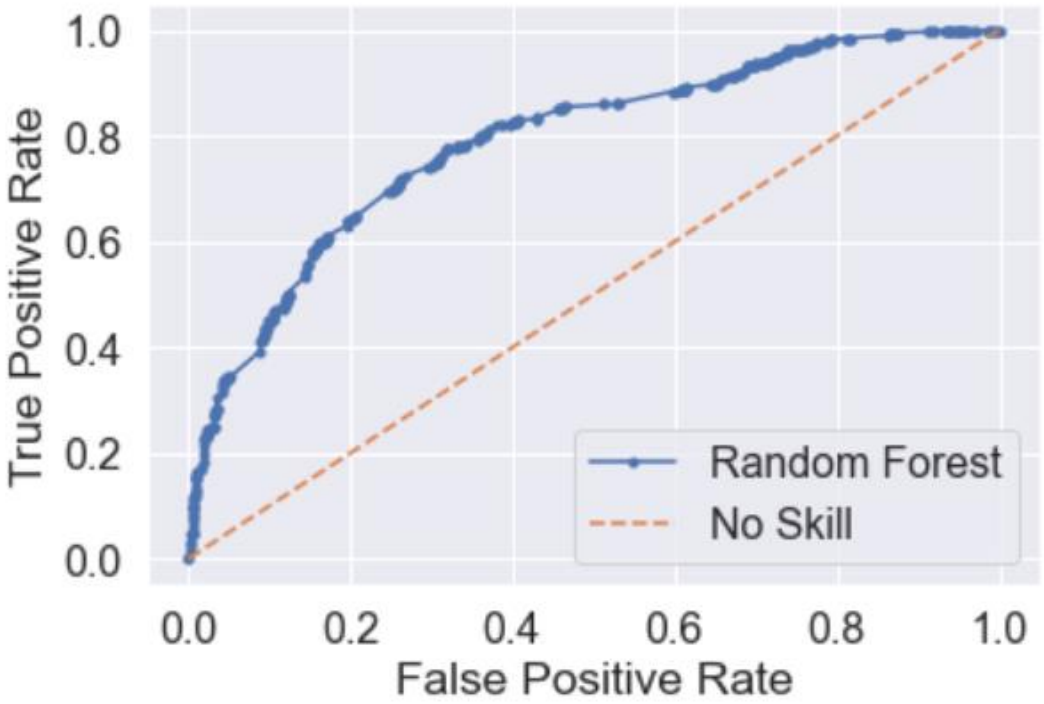
COURBE ROC DU MODELE GLM



COURBE ROC DU MODELE LOGIT DE SKLEARN



COURBE ROC DU MODELE RANDOM FOREST



Lorsque l'on observe la courbe ROC du modèle GLM et celle du modèle logit nous constatons qu'elles sont relativement similaires. Cependant, si l'on compare ces deux dernières avec celle de la random forest, nous observons un point de divergence. Lorsque le taux de faux positif est compris entre environ 0.5 et environ 0.6, le taux de vrai positif croît plus lentement pour la random forest. Ainsi, nous pourrions s'attendre à une performance plus faible de la random forest sur l'échantillon test. Comparons les AUC de chacun des modèles :

	GLM	Logit Sklearn	Random Forest
AUC	0.83	0.83	0.79

L'AUC de la random forest est plus faible que celle des modèles de régression logistique. Nous savons ainsi, qu'indépendamment du seuil choisi, le modèle GLM et le modèle logit de sklearn ont une performance prédictive plus importante.

4. CONCLUSION

Nous avons choisi d'effectuer trois modèles afin d'expliquer le taux de défaut. Tout d'abord, le modèle GLM sans régularisation avait pour objectif de quantifier les relations entre la variable d'intérêt "BAD" et les variables explicatives. Ensuite, nous avons choisi de développer un autre modèle logit avec régularisation des hyper paramètres par l'utilisation d'une GridSearchCV afin de proposer un modèle avec un meilleur ajustement et une meilleure capacité de prédiction. Enfin, un modèle Random Forest est abordé (également par l'utilisation d'une GridSearchCV) afin de comparer deux techniques de régulation des hyper paramètres : par pénalisation et par agrégation de plusieurs modèles. Au regard des critères d'information bayésien, le logit pénalisé est celui qui présente le meilleur ajustement du modèle. Selon le critère de l'AUC, les modèles GLM et logit de sklearn meilleurs que celui de la random forest, ont une capacité prédictive plus optimale sur un nouveau jeu de données. En conclusion, le modèle logit régularisé avec lasso est celui qui offre le meilleur compromis biais-variance comparativement à une régression logistique et une random forest.