



MASTER
ESA

Université d'Orléans

MOBILIZE
FINANCIAL SERVICES

MODÉLISATION D'UN SCORE D'OCTROI DE CRÉDIT

BARRAUD Lorenzo

MIRZA Simon

VIEIRA DE BARROS Mathias

Résumé non technique

Le **score d'octroi**, aussi appelé "score de crédit" ou "score de risque", est crucial dans la finance, la banque et la gestion du crédit, jouant un rôle central dans l'**évaluation des risques**. Son rôle principal est de :

- Évaluer le **risque de défaut** en estimant la probabilité qu'un emprunteur ne puisse pas rembourser sa dette.
- **Aider** à la prise de décision en déterminant si un prêt doit être accordé, sous quelles conditions.
- Optimiser les taux d'intérêt, offrant des taux avantageux aux emprunteurs à faible risque.
- **Justifier**, envers les organismes de régulation, l'octroi d'un crédit à un individu.

Résumé non technique

Dans le cadre de ce projet, nous développerons un **score d'octroi** pour les demandes financées par le groupe **Mobilize**. Après avoir effectué un nettoyage de nos données et sélectionné les prédicteurs pertinents, nous mettrons en œuvre plusieurs méthodes de modélisation, à savoir une **régression logistique** et deux techniques d'**apprentissage automatique**. Nous utiliserons la régression logistique dans le logiciel SAS et les deux méthodes d'apprentissage automatique avec Python. Ensuite, nous **comparerons** leurs performances afin de conclure sur les avantages et les inconvénients de chacune d'entre elles.

01

INTRODUCTION

02

MODÉLISATION

03

CONCLUSION



1. INTRODUCTION

1.1 Présentation de la base

1.2 Exploration de la base

1.3 Modifications effectuées

1.1 Présentation de la base



Base : 464 190 clients de chez Mobilize Financial Services qui ont bénéficié d'un financement sur le périmètre des véhicules, entre 2017 à 2020.

Variabes :

- 57 variables
- 37 variables avec description (22 quantitatives, 10 qualitatives, 2 de type date et une cible)
- Variable cible (WE12c) binaire prenant la valeur 1 si le client est tombé en défaut, et 0 dans le cas contraire

1.2 Exploration de la base

- **Variable d'intérêt**

WE12c	Fréquence	Pourcentage
0	455845	98,20
1	8334	1,80

- **Variables quantitatives** : histogramme, moyenne, médiane, écart-type et étendue
- **Variables qualitatives** : répartition de leurs modalités et boxplots
- En analysant la proportion d'**outliers** pour chaque prédicteur, nous avons conclu qu'ils ne sont **pas significatifs** et ne nécessiteront pas de traitement.
- **Peu de données manquantes** (moins de 0.0001%) pour chaque variables, à l'exception des variables "nb_imp_tot" et "nb_imp_an_0", qui présentent près de 70 % de valeurs manquantes. Cette situation s'explique par la nature de ces variables, car les valeurs manquantes correspondent aux nouveaux clients

1.3 Modifications effectuées

Les valeurs manquantes

Comme mentionné précédemment, notre base de données comporte très peu de valeurs manquantes, et sont uniquement associées aux variables qualitatives. Par conséquent, nous opterons pour une **imputation par le mode**.

Regroupement des modalités de variables qualitatives

Suite aux statistiques descriptives de nos variables, nous remarquons que certaines modalités peuvent et doivent être regroupées. En effet, certaines classes contiennent moins de **5% des observations** de la variable et/ou ne **respectent pas un taux d'ascendance**.

Obs.	eta_civ_prtc	EFF_TOT	EFF_DEF	TD_EFF	PART_EFF
1	M	237718	2307	0,97%	,51213
2	V	24156	422	1,75%	,05204
3	U	45425	953	2,10%	,09786
4	D	32391	751	2,32%	,06978
5	C	112479	3504	3,12%	,24232
6	S	12010	397	3,31%	,02587

Exemple de tableau qui permet de faire des choix sur les regroupements de modalités.

1.3 Modifications effectuées

Création d'une variable

Nous avons créé une variable "**part_finance_rev**", définie comme le résultat de **rev_tot** divisé par **mt_finance**, multiplié par 100. Cette variable exprime la proportion du revenu mensuel par rapport au montant financé. Par exemple, si "part_finance_rev" équivaut à 10, cela signifie que le revenu mensuel représente 10% du montant financé. Cependant, nous avons constaté ultérieurement que cette variable n'était **pas pertinente**.

Discrétisation des variables quantitatives

Technique : Khi-deux normé

- Permet de discréteriser des variables quantitatives
- Il doit avoir un taux de défaut strictement ascendant ou strictement descendant
- Une modalité doit au moins avoir 5% de l'effectif de la variable

Pourquoi ?

- Facilite l'interprétation des caractéristiques du client
- Explique la décision de la banque

1.3 Modifications effectuées

Discrétisation des variables quantitatives

Nb_Mod	Chi2	Chi2Norme	Mod	minimum	maximum	Bornes	Variable	mod2	NBR_DEFAUT	NBR_EFFECTIF	TX_DEFAUT	DIFF_TDF
2	20,2	0,0065985	Modalite 1	300000	2509976	[300000@2509976]	mt_ttc_veh	1	7070	385269	1,835081463	
2	20,2	0,0065985	Modalite 2	2509976	25597041]2509976@25597041]	mt_ttc_veh	2	1264	78910	1,601824864	-0,23326
3	49,8	0,0087087	Modalite 1	300000	1420000	[300000@1420000]	mt_ttc_veh	1	2157	125681	1,716249871	
3	49,8	0,0087087	Modalite 2	1420000	1890000]1420000@1890000]	mt_ttc_veh	2	2538	125569	2,0211995	0,30495
3	49,8	0,0087087	Modalite 3	1890000	25597041]1890000@25597041]	mt_ttc_veh	3	3639	212929	1,709020378	-0,31218
4	58	0,0084904	Modalite 1	300000	1420000	[300000@1420000]	mt_ttc_veh	1	2157	125681	1,716249871	
4	58	0,0084904	Modalite 2	1420000	1890000]1420000@1890000]	mt_ttc_veh	2	2538	125569	2,0211995	0,30495
4	58	0,0084904	Modalite 3	1890000	2509976]1890000@2509976]	mt_ttc_veh	3	2375	134019	1,772136787	-0,24906
4	58	0,0084904	Modalite 4	2509976	25597041]2509976@25597041]	mt_ttc_veh	4	1264	78910	1,601824864	-0,17031

Exemple permettant de faire la discrétisation de la variable mt_ttc_ech (variable quantitative)

Interprétation : la colonne “TX_DEFAUT” ne respecte pas un taux d’ascendance/descendance pour une discrétisation en 3 classes (lignes 3 à 5). La discrétisation se fera donc en 2 classes

2. MODÉLISATION

2.1 Sélection des variables

2.2 Méthode 1 : Régression logistique

2.3 Méthode 2 : Régression logistique avec SMOTE

2.4 Méthode 3 : Random Forest

2.5 Méthode 4 : XGBoost

2.1 Sélection des variables

Les différentes techniques

Analyse de la significativité statistique: évalue si les variables ont un pouvoir prédictif significatif pour expliquer la variation de la variable cible, en utilisant le test de Wald notamment.

Critères d'information: évalue la pertinence des variables en pénalisant les modèles qui incluent un grand nombre de variables. Les critères d'Akaike (AIC) et bayésien (BIC) sont couramment utilisés.

Validation croisée: entraîne notre modèle à l'aide de multiples échantillons d'apprentissage pour s'assurer de sa robustesse.

Corrélations: Facilite l'évaluation des relations entre les variables explicatives, ainsi que leur relation avec la variable cible.

2.1 Sélection des variables

Notre cas

Table	Cramers_V	Chisq	p_value	Cramers_V_abs
CSP_classe * rev_men_.autr2	0,6006657	334950,911	<.0001	,60067
appo_cptt_cnt2 * no_nat_prod2	0,6324753	185683,182	<.0001	,63248
REV_TOT2 * rev_men_autr2	0,7071141	464188,591	<.0001	,70711
diag_cli_rnva * nb_imp_tot2	0,7152119	474881,138	<.0001	,71521
diag_cli_rnva * nb_imp_an_0_2	0,7173528	477728,356	<.0001	,71735
mt_charges2 * tx_end_syex2	0,7440912	257002,755	<.0001	,74409
age_indv2 * anc_emp_indv2	0,7887155	288752,86	<.0001	,78872
anc_emp_indv2 * rev_men_autr2	0,8237262	314957,002	<.0001	,82373
nb_imp_an_0_2 * nb_imp_tot2	0,8286799	637513,054	<.0001	,82868
cpt_pai2_2 * nb_imp_an_0_2	0,8928043	739993,718	<.0001	,89280
CSP_classe * anc_emp_indv2	0,9693104	436125,221	<.0001	,96931

Variable	Cramers_V	Chisq	p_value
mod_habi_indv	0,0836	3244,3721	<.0001
cpt_pai2_2	0,0765	2716,9011	<.0001
nb_imp_an_0_2	0,0734	2504,0203	<.0001
age_indv2	0,0717	2388,9723	<.0001
eta_civ_prtc2	0,0690	2209,8516	<.0001
appo_cptt_cnt2	0,0644	1927,3936	<.0001
nb_imp_tot2	0,0623	1804,2429	<.0001
tx_end_syex2	0,0601	1677,7199	<.0001
part_ech2	0,0542	1361,3982	<.0001
CSP_classe	0,0483	1081,9221	<.0001
anc_adr_indv2	-0,0453	954,5919	<.0001
diag_cli_rnva	0,0429	853,1807	<.0001
anc_emp_indv2	0,0355	586,0988	<.0001
rev_men_autr2	0,0344	549,1923	<.0001
mt_ttc_veh2	0,0283	387,2584	<.0001
no_nat_prod2	0,0263	321,1919	<.0001
mt_charges2	-0,0263	320,4587	<.0001
secteur_2	0,0223	230,2806	<.0001
cd_natl_indv2	0,0169	133,2155	<.0001
region2	0,0135	84,7729	<.0001
MT_ECH2	-0,0124	71,3327	<.0001
REV_TOT2	0,0109	55,0079	<.0001
mt_finance2	0,0002	0,0183	0,8924

La première étape consiste à examiner les corrélations entre les variables explicatives. Le tableau ci-contre montre les interactions entre les variables fortement corrélées (V de Cramer > 0.6). En cas de corrélation entre deux variables explicatives, il est recommandé de supprimer celle moins corrélée à la variable cible pour respecter le principe de parcimonie.

Lors de l'évaluation de la corrélation entre les variables explicatives et la variable cible, aucun lien significatif n'a été trouvé (V de Cramer faible). Par conséquent, nous avons sélectionné les variables en fonction de critères tels que l'AUC, le gini, le 10/X, et la significativité des paramètres de la régression logistique associés aux modalités des variables.

Certaines variables, comme CSP_perphy, n'ont pas réussi à démontrer leur significativité au sein de toutes leurs modalités. Nous avons donc pris la décision de les supprimer afin d'optimiser la parcimonie de notre modèle.

2.1 Sélection des variables

Notre cas

Les variables explicatives retenues dans notre modèle finale sont :

- REV_TOT : **montant du revenu mensuel**
- part_ech : **part de l'échéance en pourcentage**
- mt_ttc_vech : **prix du véhicule**
- mt_finance : **montant financé**
- my_charges : **montant des charges**
- anc_adr_indv : **ancienneté de l'adresse**
- age_indv : **âge**
- nb_imp_an_0 : **nombre d'impayés**
- cd_natl_indv : **nationalité CEE ou hors CEE**
- etat_civ_prtc : **état civil**
- appo_cptt_cnt : **pourcentage d'apport**



2.2 Méthode 1 : Régression logistique

Comprendre la régression logistique

Régression logistique : modèle probabiliste pour réponse binaire

Son principe :

- estime des **coefficients** pour les variables explicatives via la méthode du maximum de vraisemblance
- Ces coefficients, combinés avec les valeurs des variables, génèrent un **score**
- Ce score est ensuite utilisé pour calculer une **probabilité** en utilisant la fonction de répartition logistique

Il est essentiel de noter que cette probabilité ne définit pas une règle de décision; la banque doit choisir le seuil pour juger si un client est susceptible de faire défaut ou non.

2.2 Méthode 1 : Régression logistique

Comprendre les indices de performances

Matrice de confusion :

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <i>Type II Error</i>	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <i>Type I Error</i>	True Negative (TN)	Specificity $\frac{TP}{(TN + FP)}$
	Precision	$\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TN + FP + FN)}$

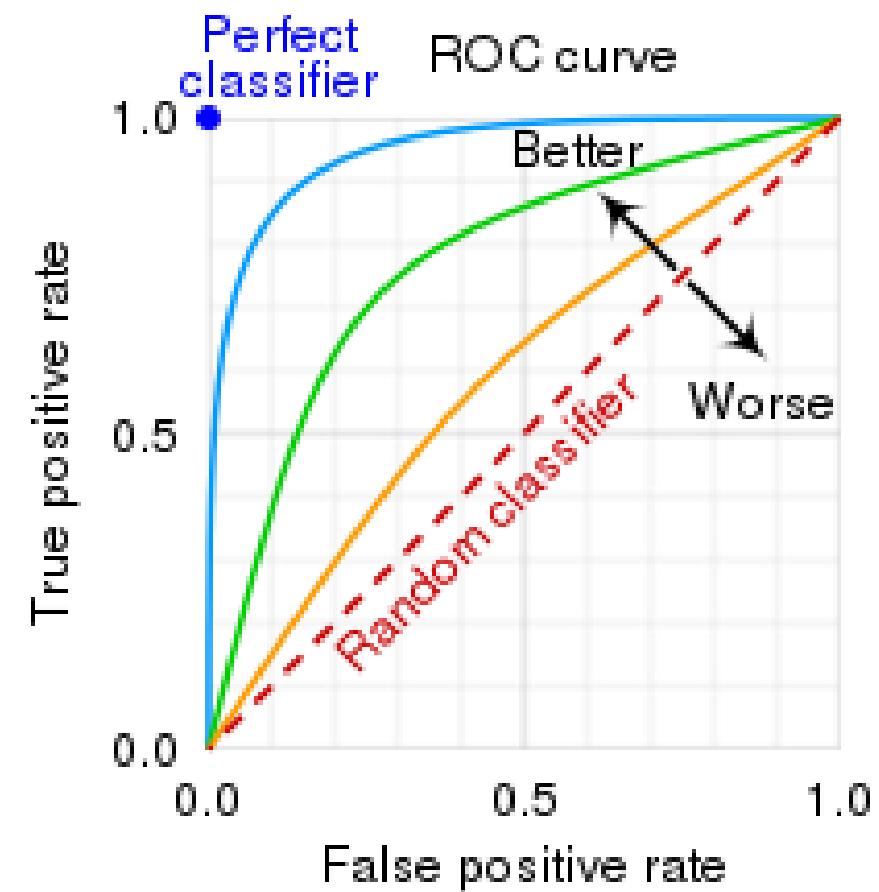
l'objectif est de déterminer le **pourcentage de défauts** parmi les **10% d'individus ayant les scores les plus bas**.

Indice 10/X : exemple : lorsque x est égal à 10% et que la courbe atteint un niveau de y égal à 65%, cela signifie que les 10% d'individus avec les scores les plus bas expliquent 65% des défauts, ce qui se traduit par un indice de 10/65.

2.2 Méthode 1 : Régression logistique

Comprendre les indices de performances

Courbe ROC :



AUC : Aire sous la courbe ROC

Indice de Gini : compris entre 0 et 1, il permet de mesurer le niveau d'inégalité dans la distribution d'une variable.

Caractérisé par : $\text{Gini} = (2 \times \text{AUC}) - 1$

2.2 Méthode 1 : Régression logistique

Modélisation

Nous avons divisé nos données en deux groupes : **apprentissage** et **test**. Ils sont stratifiés en raison de la sous-représentation des cas de défaut, essentielle pour l'estimation et l'évaluation du modèle. Ces groupes resteront **constants** pour la régression logistique et l'approche de machine learning, pour une comparaison **équitable**.

Echantillon
d'apprentissage

WE12c	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	319092	98,20	319092	98,20
1	5834	1,80	324926	100,00

Echantillon test

WE12c	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	136753	98,20	136753	98,20
1	2500	1,80	139253	100,00

2.2 Méthode 1 : Régression logistique

Modélisation

Analyse des valeurs estimées du maximum de vraisemblance						
Paramètre		DDL	Estimation	Erreur type	Khi-2 de Wald	Pr>khi-2
Intercept		1	-4,0374	0,1578	654,2131	<.0001
REV_TOT2	[0,04;130000]	1	1,2868	0,0641	403,1392	<.0001
REV_TOT2]130000;169266]	1	0,9610	0,0570	284,0699	<.0001
REV_TOT2]169266;199200]	1	0,8044	0,0589	186,4931	<.0001
REV_TOT2]199200;218000]	1	0,7163	0,0666	115,6824	<.0001
REV_TOT2]218000;250000]	1	0,6782	0,0583	135,1852	<.0001
REV_TOT2]250000;295000]	1	0,6480	0,0558	134,8768	<.0001
REV_TOT2]295000;335000]	1	0,4777	0,0590	65,6523	<.0001
REV_TOT2]335000;400900]	1	0,2476	0,0586	17,8520	<.0001
part_ech2	[0,01;1,08]	1	-0,4746	0,0694	46,7394	<.0001
part_ech2]1,08;1,23]	1	-0,5850	0,0641	83,2668	<.0001
part_ech2]1,23;1,36]	1	-0,5240	0,0494	112,6000	<.0001
part_ech2]1,36;1,52]	1	-0,4202	0,0372	127,3792	<.0001
part_ech2]1,52;1,66]	1	-0,2756	0,0372	55,0353	<.0001
mt_ttc_veh2	[199898;2509976]	1	-0,2721	0,0401	45,9413	<.0001
mt_finance2	[199898;799951.8]	1	-0,4118	0,1162	12,5582	0,0004
mt_finance2]1120069;1312850]	1	0,1757	0,0573	9,3971	0,0022
mt_finance2]1312850;23597912]	1	0,4811	0,0466	106,3673	<.0001
mt_charges2	[0;25000]	1	0,2722	0,0462	34,6720	<.0001

Extrait le table permettant d'observer la significativité des modalités associées à leur variables respectives

L'ensemble des coefficients estimés sont **significatifs**. Leur p-value étant pour tous inférieure au seuil de 5%.

2.2 Méthode 1 : Régression logistique

Performances

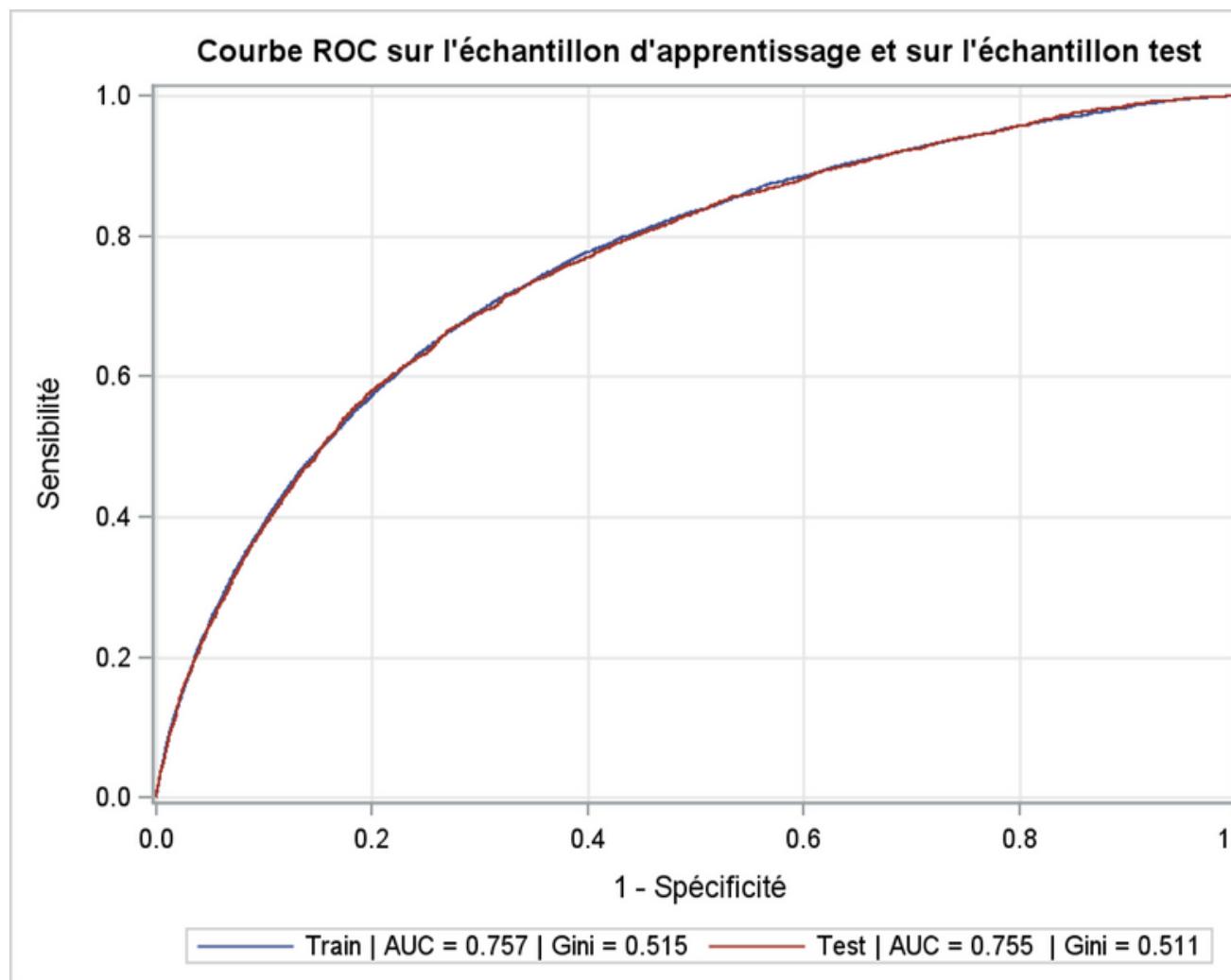
Matrice de confusion associée à WE12C : cut-off 0.0188			
Valeurs observées	Valeurs prédites		
Fréquence Pourcentage Pct de ligne Pct de col.	0	1	Total
0	94996 68.22 69.47 99.20	41757 29.99 30.53 96.01	136753 98.20
	763 0.55 30.52 0.80	1737 1.25 69.48 3.99	2500 1.80
	95759 68.77	43494 31.23	139253 100.00

Le cut-off sélectionné pour la matrice de confusion est celui qui égalise la **sensibilité** et la **spécificité**, on le considérera comme étant « optimal » : 0.0188

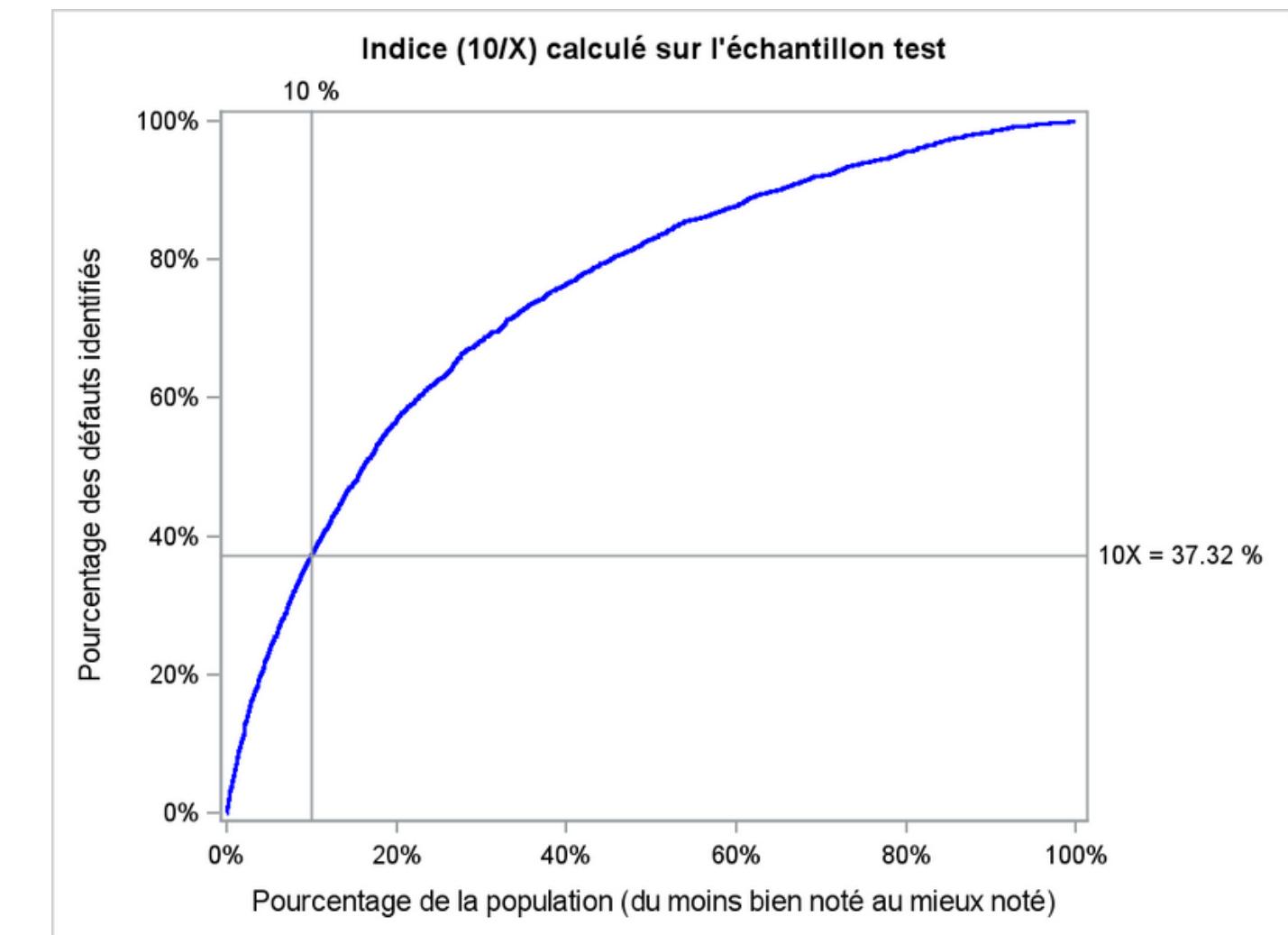
On observe qu'il y environ **69,50 %** de bonne prédictions pour les clients qui n'ont pas fait défaut et pour ceux qui ont fait défaut.

2.2 Méthode 1 : Régression logistique

Performances



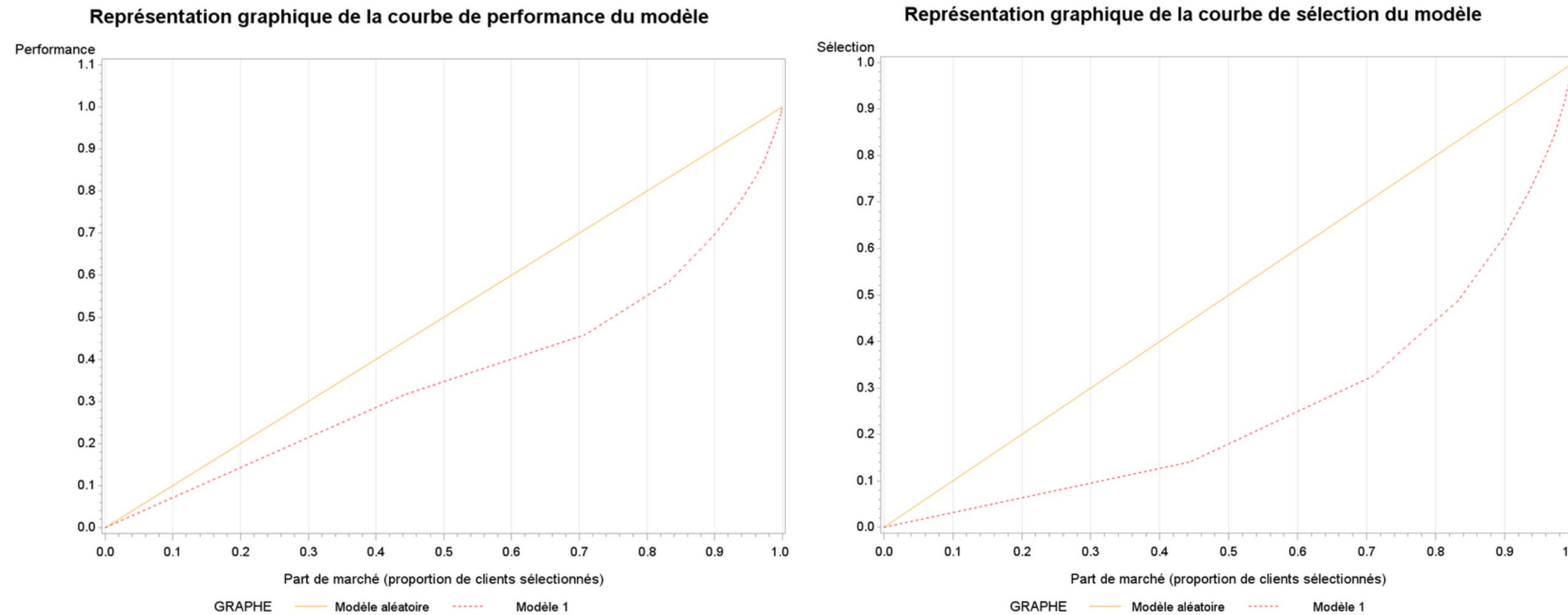
Le graphique ci-dessus compare la **ROC**, l'**AUC** et le **Gini** de notre échantillon d'apprentissage et du test.



Le graphique ci-dessus représente l'indice **10/X** de l'échantillon test.

2.2 Méthode 1 : Régression logistique

Performances



La courbe de performance et de sélection sont également des courbes qui sont **utilisés** pour tester la performance des modèles. Sur l'axe des ordonnées de la courbe de **performance**, on retrouve le % d'individus défaillants de la population sélectionnée par le score sur le % de défaillants de la population initiale. Pour la courbe de **sélection**, il correspond à la proportion de défaillants sélectionnée à tort par le modèle sur le nombre de défaillants total.

2.2 Méthode 1 : Régression logistique

Interprétation

La grille de score occupe une importance cruciale dans l'**interprétation** précise de la modélisation du score d'octroi. Grâce à elle, nous sommes en mesure de déterminer la **contribution** de chaque variable ainsi que le **taux de défaut** des individus associés à une modalité spécifique.

On voit clairement que les variables qui ont le plus d'importance dans la notation du client sont dans l'ordre décroissant: le **montant financé**, la **part de l'échéance**, le **revenu mensuel**, le **nombre d'impayés** et l'**âge**.

Label	Variable	Modalité	Poids	Taux de défaut	Répartition	Contribution
Age	AGE_INDV2	[17;26]	0	4,45%	5,41%	11,83%
]26;34]	15	3,48%	9,19%	
]34;38]	17	2,68%	5,61%	
]38;44]	23	2,10%	9,80%	
]44;48]	49	1,74%	8,41%	
]48;55]	76	1,52%	15,62%	
]55;66]	93	1,16%	22,83%	
]66;98]	107	1,00%	23,13%	
Ancienneté de l'adresse	ANC_ADR_INDV2	[0;107]	0	2,35%	53,12%	5,14%
Nationalité	CD_NATL_INDV2]107;110]	47	1,17%	46,88%	8,23%
		Hors CEE	0	5,11%	0,41%	
		CEE	92	1,78%	99,59%	
Etat civil	ETA_CIV_PRTC2	Séparé/Divorcé/Veuf	0	2,27%	14,88%	7,37%
		Celibataire	16	3,12%	24,05%	
		Union libre	48	2,19%	9,83%	
		Marié	72	0,96%	51,24%	
Montant des charges	MT_CHARGES2	[0;25000]	0	3,12%	5,58%	3,01%
]25000;1186200]	33	1,72%	94,42%	
Montant financé	MT_FINANCE2]1312850;23597912]	0	2,16%	61,96%	21,13%
]1120069;1312850]	45	1,92%	12,18%	
]799951.8;1120069]	74	1,27%	14,93%	
		[199898;799951.8]	168	0,32%	10,93%	
Prix du véhicule	MT_TTC_VEH2]2509976;25597041]	0	1,59%	16,95%	2,55%
		[199898;2509976]	26	1,84%	83,05%	
Nombre d'impayés	NB_IMP_AN_0_2	>1	0	9,77%	1,51%	12,47%
		nouveau client	162	1,88%	69,97%	
		0	195	1,17%	28,52%	
Part de l'échéance en pourcentage	PART_ECH2]1.66;75.75]	0	2,90%	20,01%	14,19%
]1.52;1.66]	32	2,31%	17,97%	
]1.36;1.52]	54	1,77%	22,17%	
]1.23;1.36]	79	1,50%	13,05%	
]1.08;1.23]	103	0,85%	9,85%	
		[0.01;1.08]	105	0,76%	16,96%	
Revenu mensuel	REV_TOT2	[0.04;130000]	0	3,21%	5,02%	14,09%
]130000;169266]	40	3,00%	9,97%	
]169266;199200]	60	2,42%	8,96%	
]199200;218000]	70	2,28%	6,06%	
]218000;250000]	75	1,89%	10,08%	
]250000;295000]	78	1,80%	11,93%	
]295000;335000]	98	1,62%	10,81%	
]335000;400900]	125	1,32%	13,96%	
]400900;76061300]	155	0,93%	23,21%	

La grille de score



2.3 Méthode 2 : Régression logistique avec SMOTE

Le SMOTE

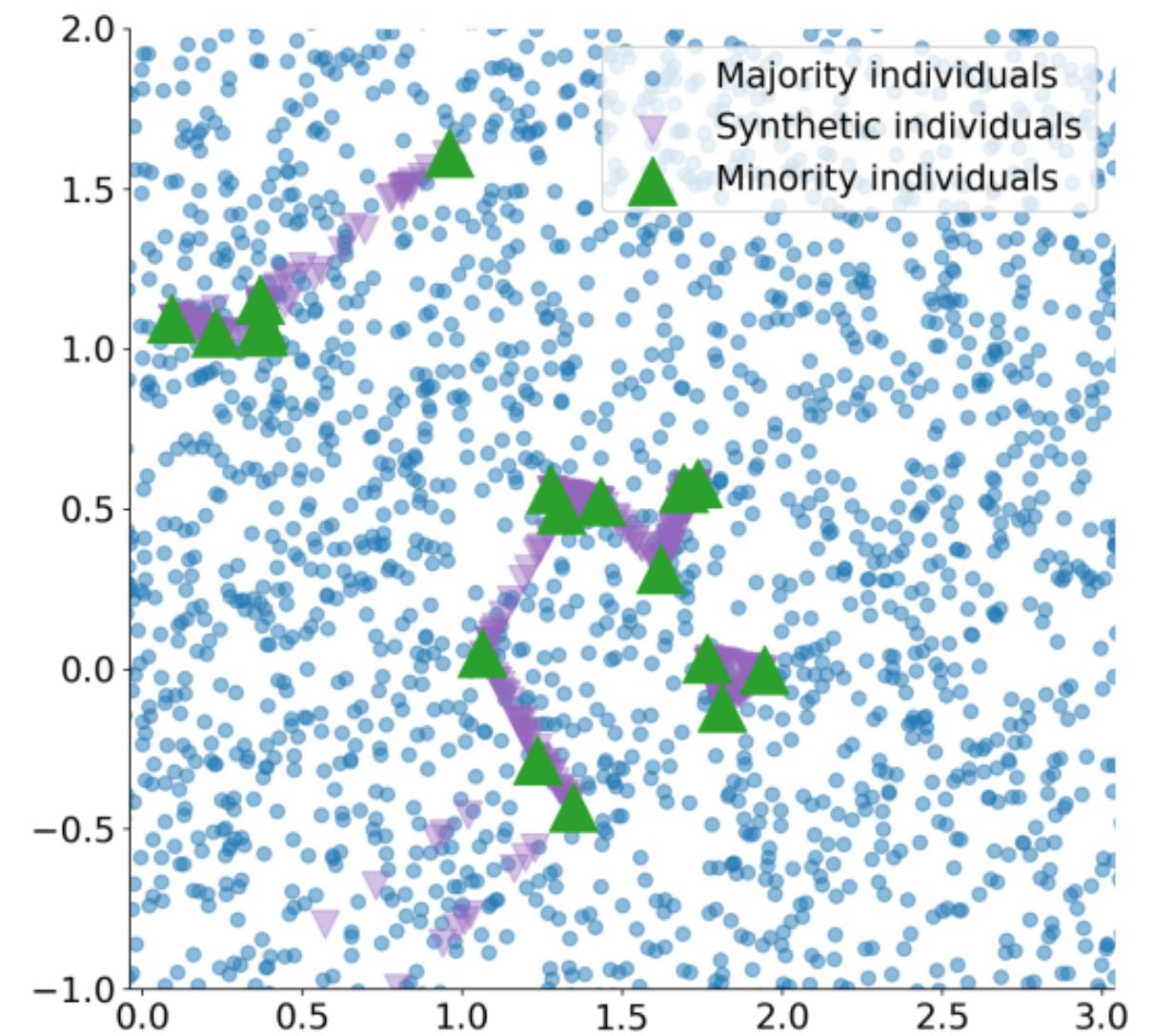
Méthode permettant de **générer** des individus afin **d'augmenter** l'effectif de la classe **minoritaire** (ici, les individus défaillants : 1.8% de l'effectif).

Objectif : fournir davantage d'informations à l'ensemble d'apprentissage pour permettre au modèle d'éviter la sous-estimation des cas au sein de la classe minoritaire.

2.3 Méthode 2 : Régression logistique avec SMOTE

Comprendre le SMOTE

1. Sélection aléatoire d'une observation **minoritaire initiale**.
2. Identification des **k voisins les plus proches** parmi les observations minoritaires.
3. Choix **aléatoire** d'un des k voisins les plus proches.
4. Génération aléatoire d'un **coefficent alpha** entre 0 et 1.
5. **Création** d'un nouvel individu positionné entre l'observation de départ et le voisin le plus proche, en fonction du coefficient alpha (par exemple, alpha=0,5 pour une position à mi-chemin).



Application du SMOTE à un couple de variables numériques

2.3 Méthode 2 : Régression logistique avec SMOTE

Modélisation

L'objectif d'un SMOTE à 5% (resp. 10%) est de faire **atteindre la classe minoritaire** à 5% (resp. 10%) de l'effectif total de la classe majoritaire.

Pourquoi 5% et 10% ?

- Calibrer le modèle sur un dataset reflétant la **réalité** : la proportion de défaut ne varie pas significativement de la réalité
- L'équilibrage 50/50 pourrait **biaiser** les données car les taux de défaut bancaire sont généralement bas

Base initiale

455 845 non défaillants
8 334 défaillants
total : 464 179 individus

SMOTE 5%

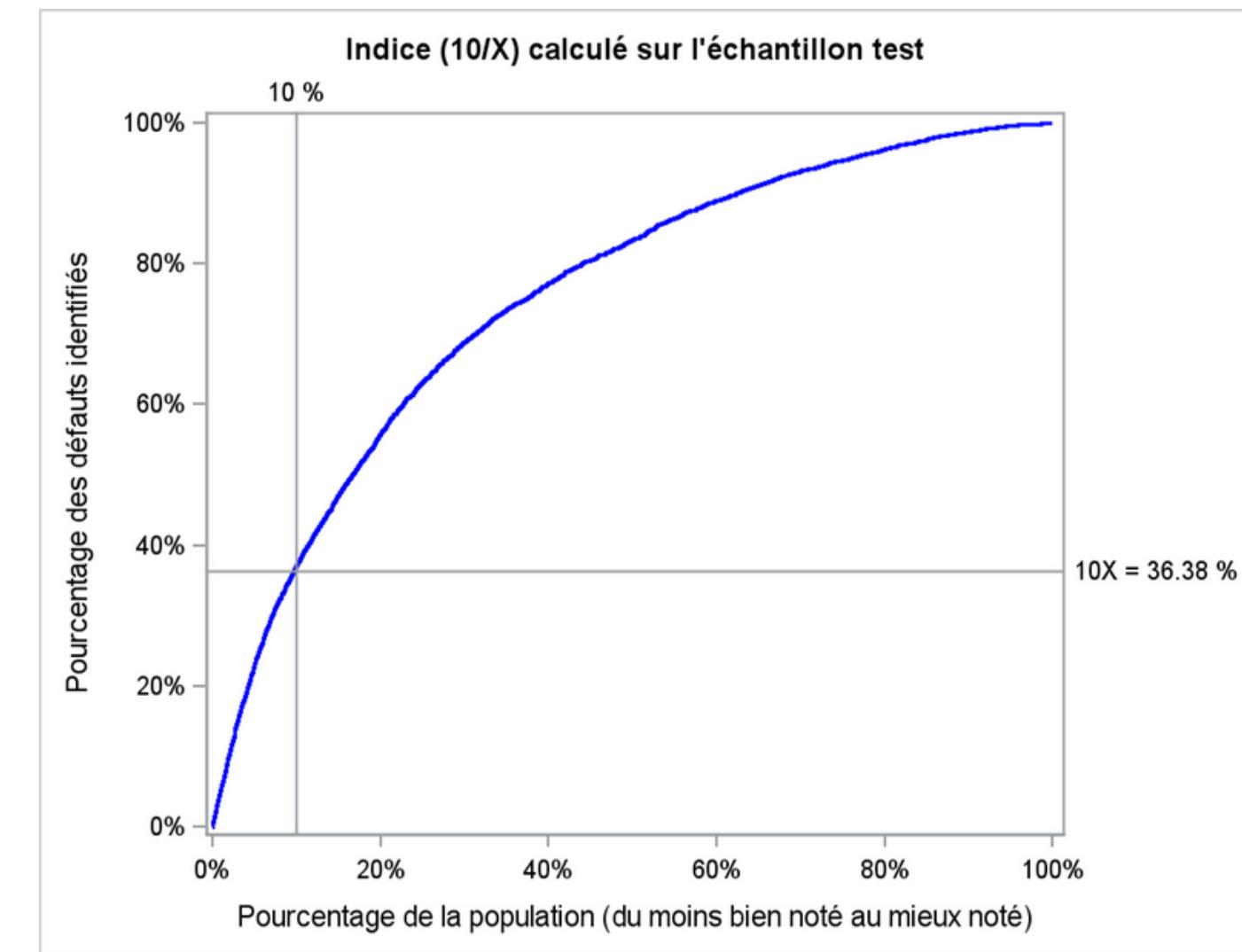
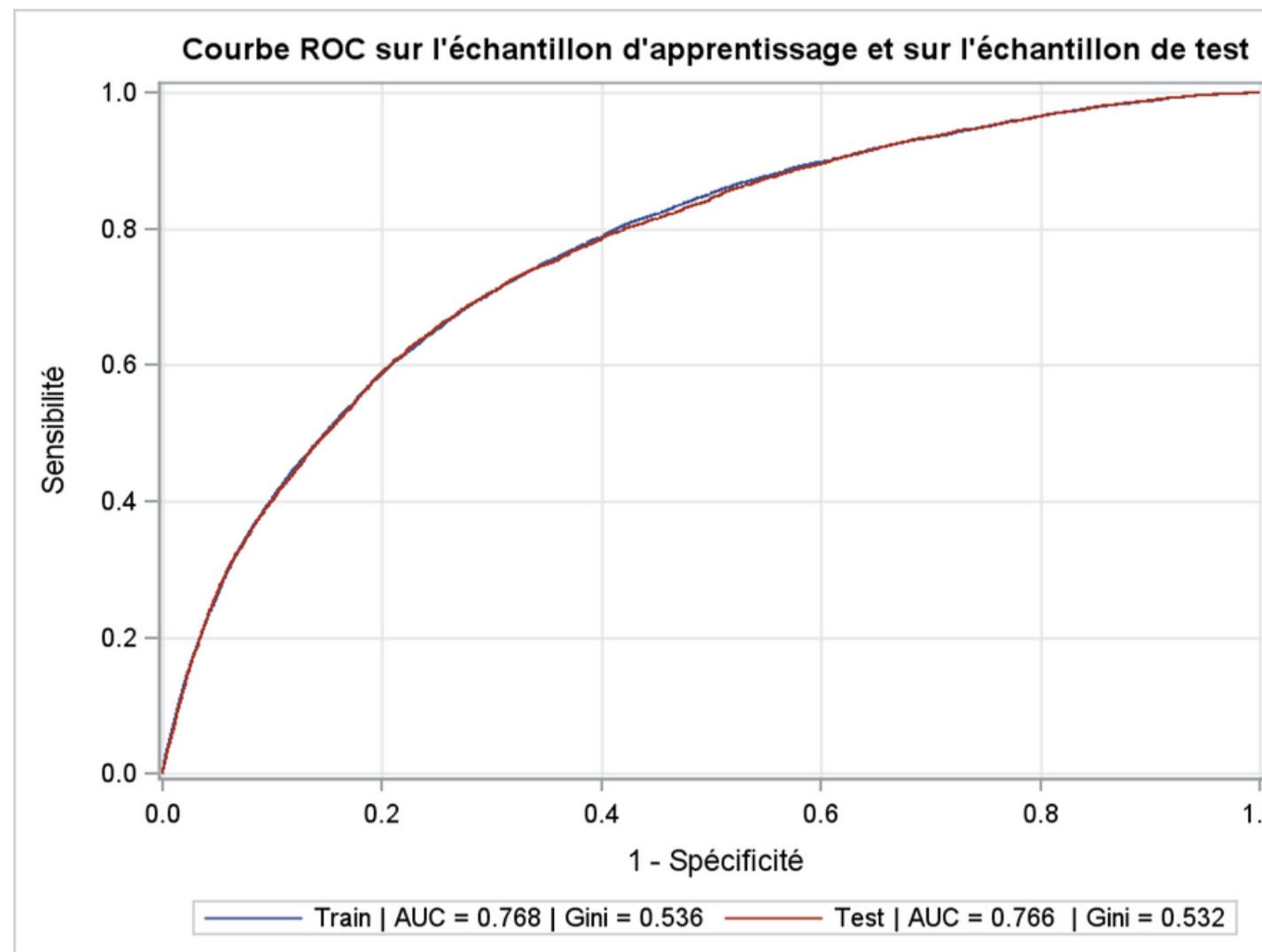
455 845 non défaillants
22 792 défaillants
total : 478 637 individus

SMOTE 10%

455 845 non défaillants
45 584 défaillants
total : 501 429 individus

2.3 Méthode 2 : Régression logistique avec SMOTE

Performances du SMOTE à 5%

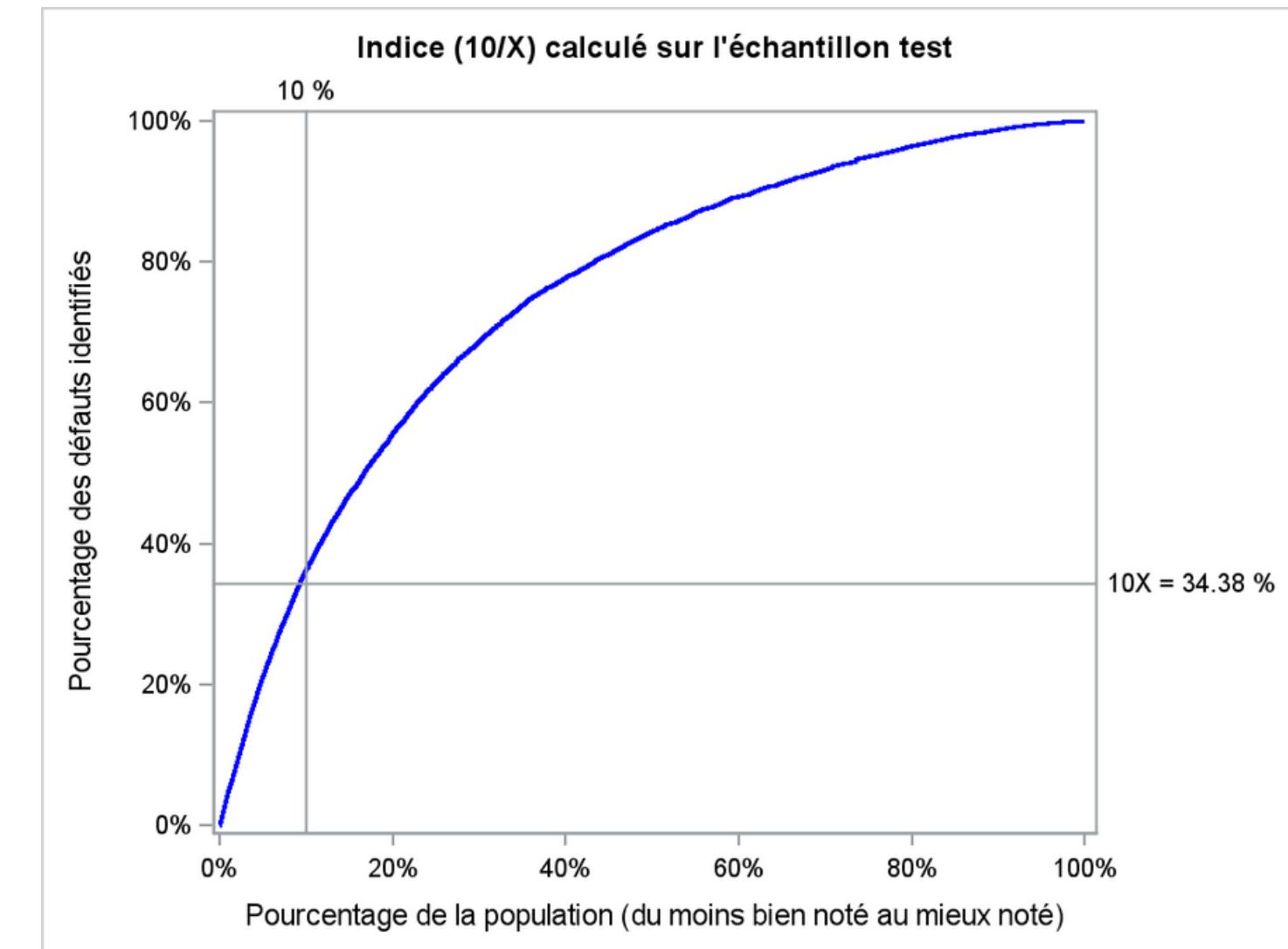
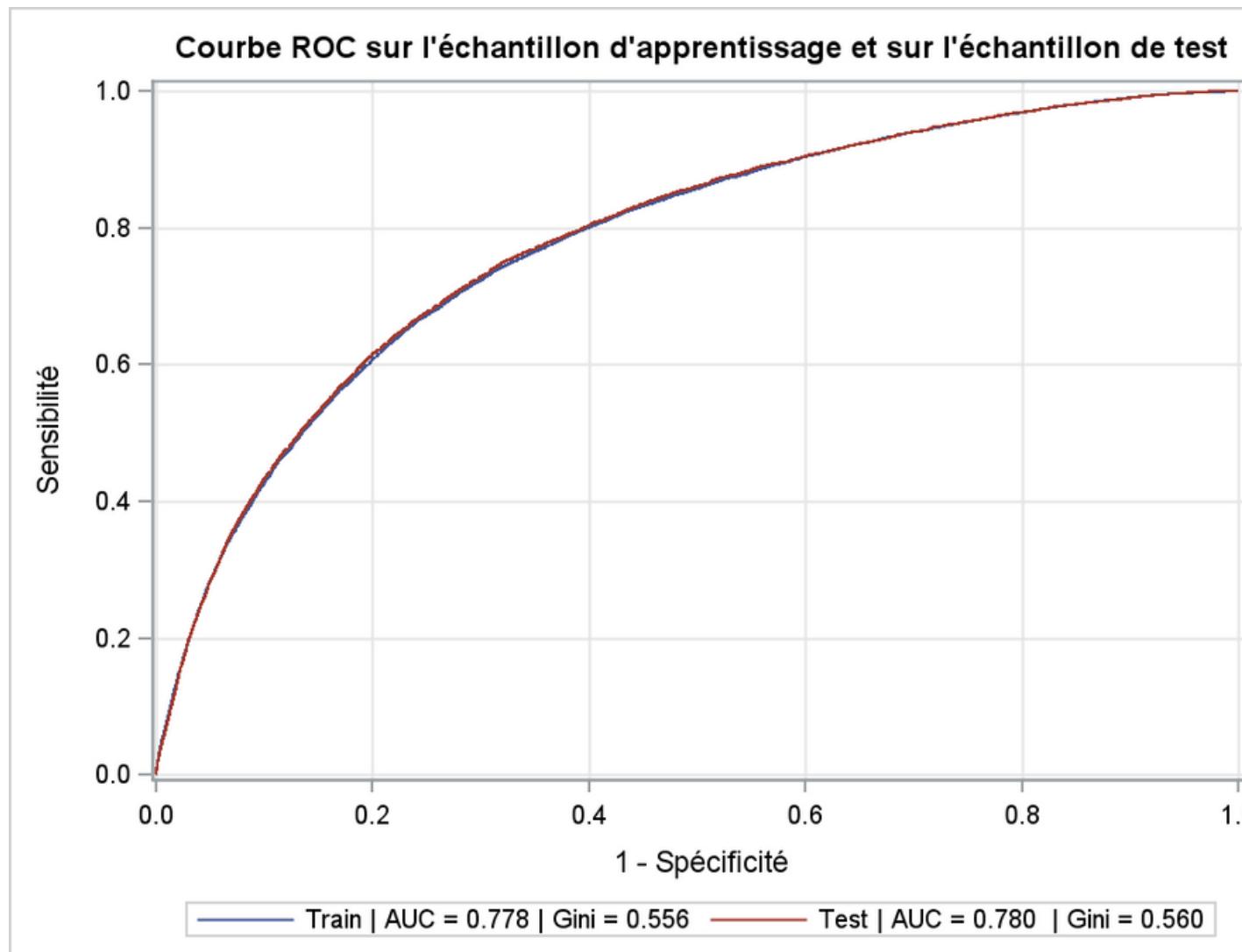


AUC sans smote (0.755) < AUC smote 5% (0.766)

10X sans smote (37.32%) > 10X smote 5% (36.38%)

2.3 Méthode 2 : Régression logistique avec SMOTE

Performances du SMOTE à 10%



AUC sans smote (0.755) < AUC smote 5% (0.766) <
AUC smote 10% (0.78)

10X sans smote (37.32%) > 10X smote 5% (36.38%) >
10X smote 10% (34.38%)

2.3 Méthode 2 : Régression logistique avec SMOTE

Conclusion

- Fourni **davantage** d'informations sur la classe minoritaire à l'apprentissage du modèle
- Les performances sont très **similaires**
- Impact **positif** sur l'AUC et le Gini
- Impact **négatif** sur l'indice 10/X



2.4 Méthode 3 : Random Forest

Comprendre le Random Forest

Son principe :

- Le Random Forest est l'**aggrégation** d'arbres de décision
- Chaque arbre individuel est construit en utilisant un **sous ensemble aléatoire** des données d'apprentissage et en utilisant **une partie** des caractéristiques
- Chaque arbre émet une prédiction : le résultat final est la variable qualitative **la plus représentée** ou la **moyenne** dans le cas d'une régression

Données utilisées :

- **Même** échantillon d'apprentissage et test que pour la méthode de régression logistique
- Transformation des variables catégorielles en variable **binaire**

Sélection des hyperparamètres :

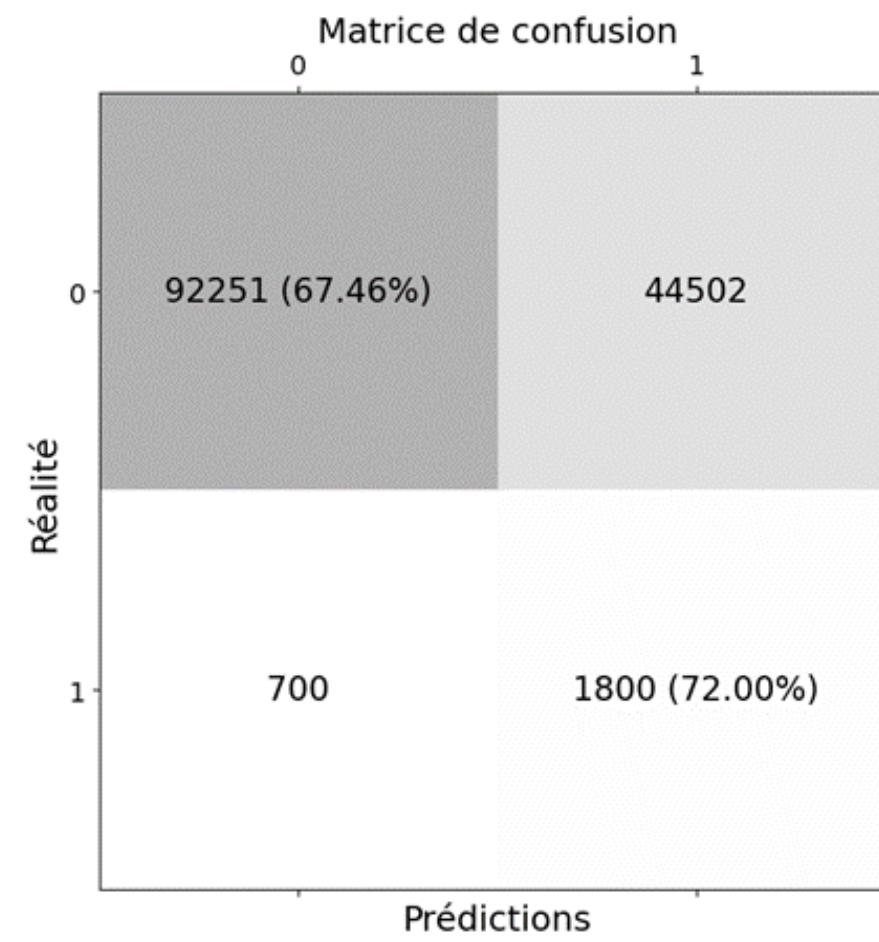
- Utilisation de la **validation croisée**
- **Optimisation** des performances du modèle en évitant l'overfitting

Modélisation :

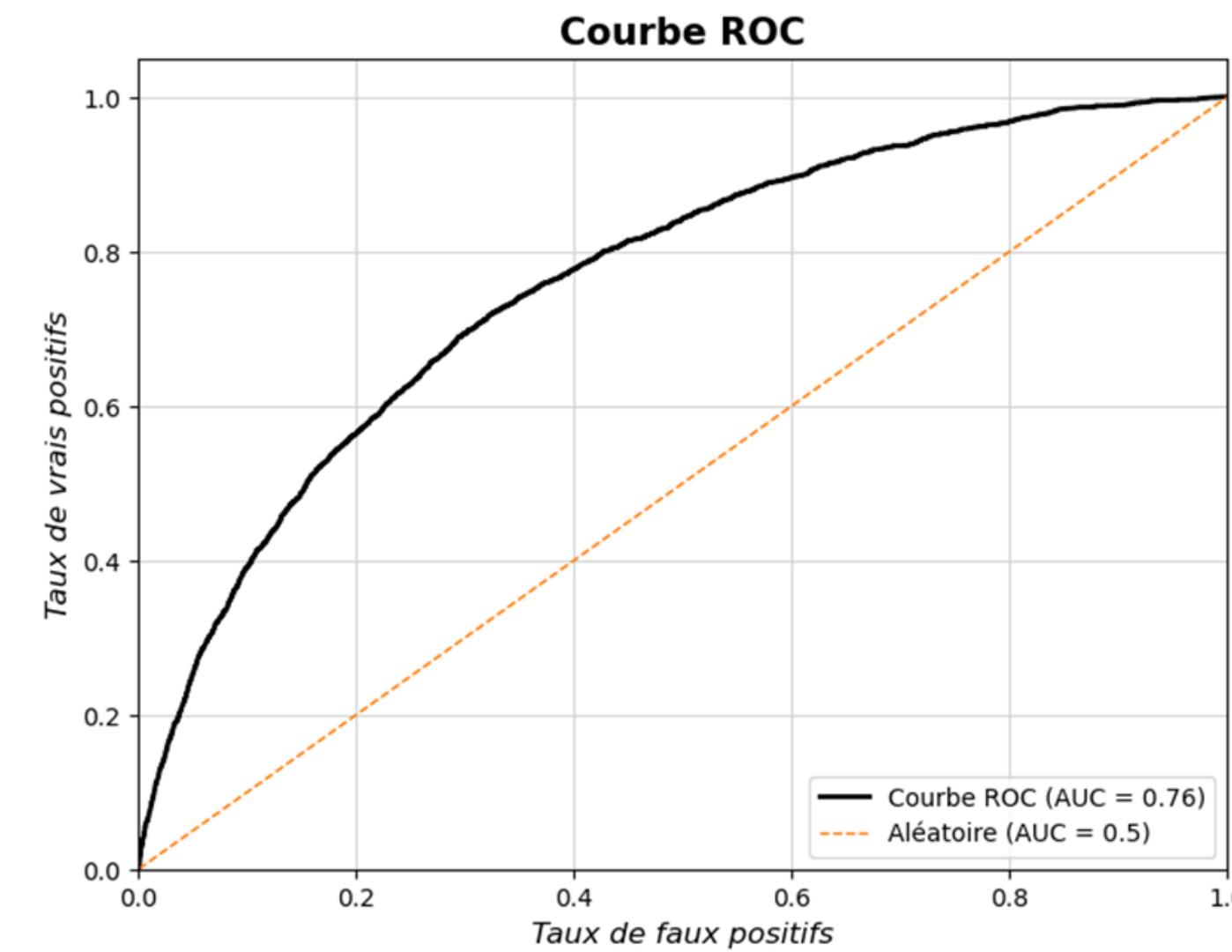
- Sélections des **meilleurs** paramètres
- Prédiction sur l'échantillon test

2.4 Méthode 3 : Random Forest

Performances



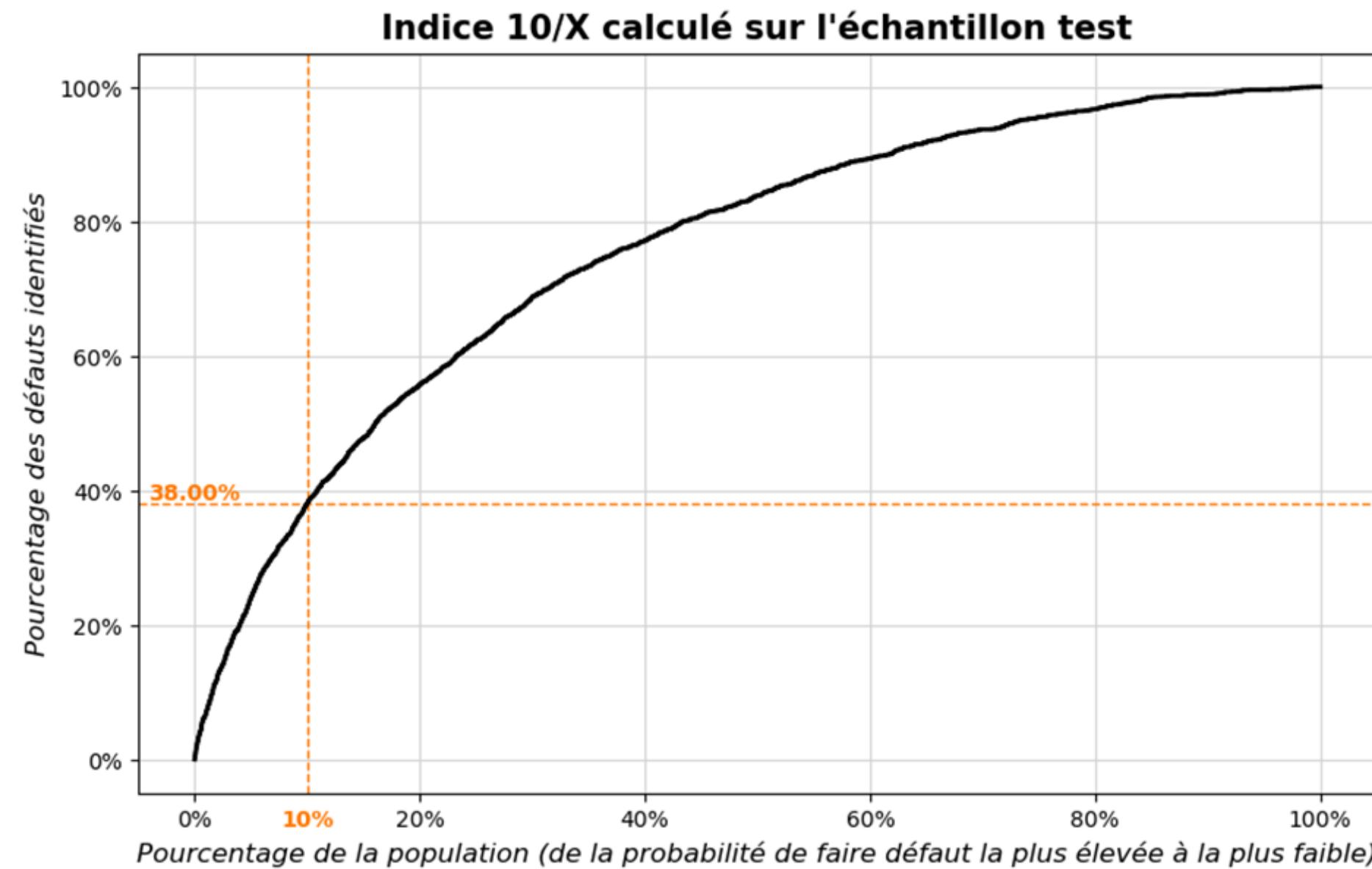
cut-off à 0.188



Gini = 0.52

2.4 Méthode 3 : Random Forest

Performances





SPRING

2.5 Méthode 4 : XGBoost

Comprendre le XGBoost

Son principe :

- Méthode ensembliste qui repose sur le **Boosting**. Chaque nouvel arbre est conçu pour **corriger** les erreurs de prédictions
- La technique du “gradient boosting” est utilisée pour **minimiser** une fonction de coût
- Le résultat final est la **combinaison** des arbres faibles afin de faire un arbre plus robuste

Données utilisées :

- **Même** échantillon d'apprentissage et test que pour les autres méthodes
- Transformation des variables catégorielles en variable **binaire**

Sélection des hyperparamètres :

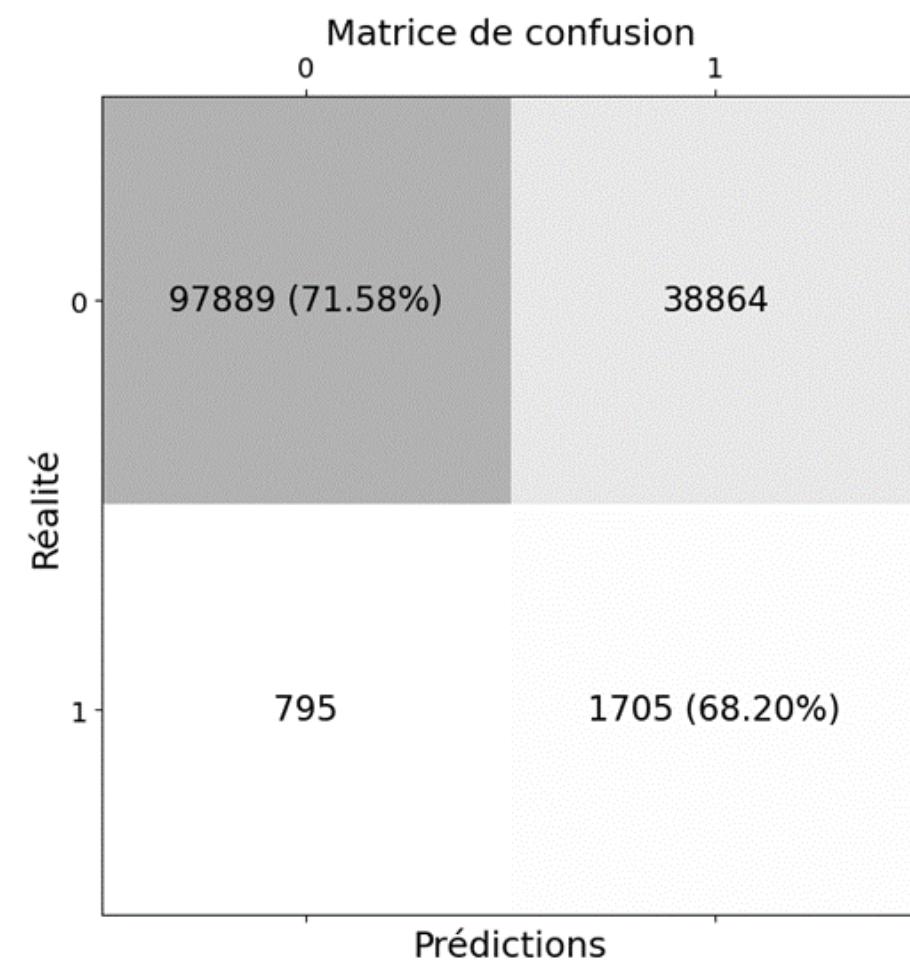
- Utilisation de la **validation croisée**
- **Optimisation** des performances du modèle en évitant l'overfitting

Modélisation :

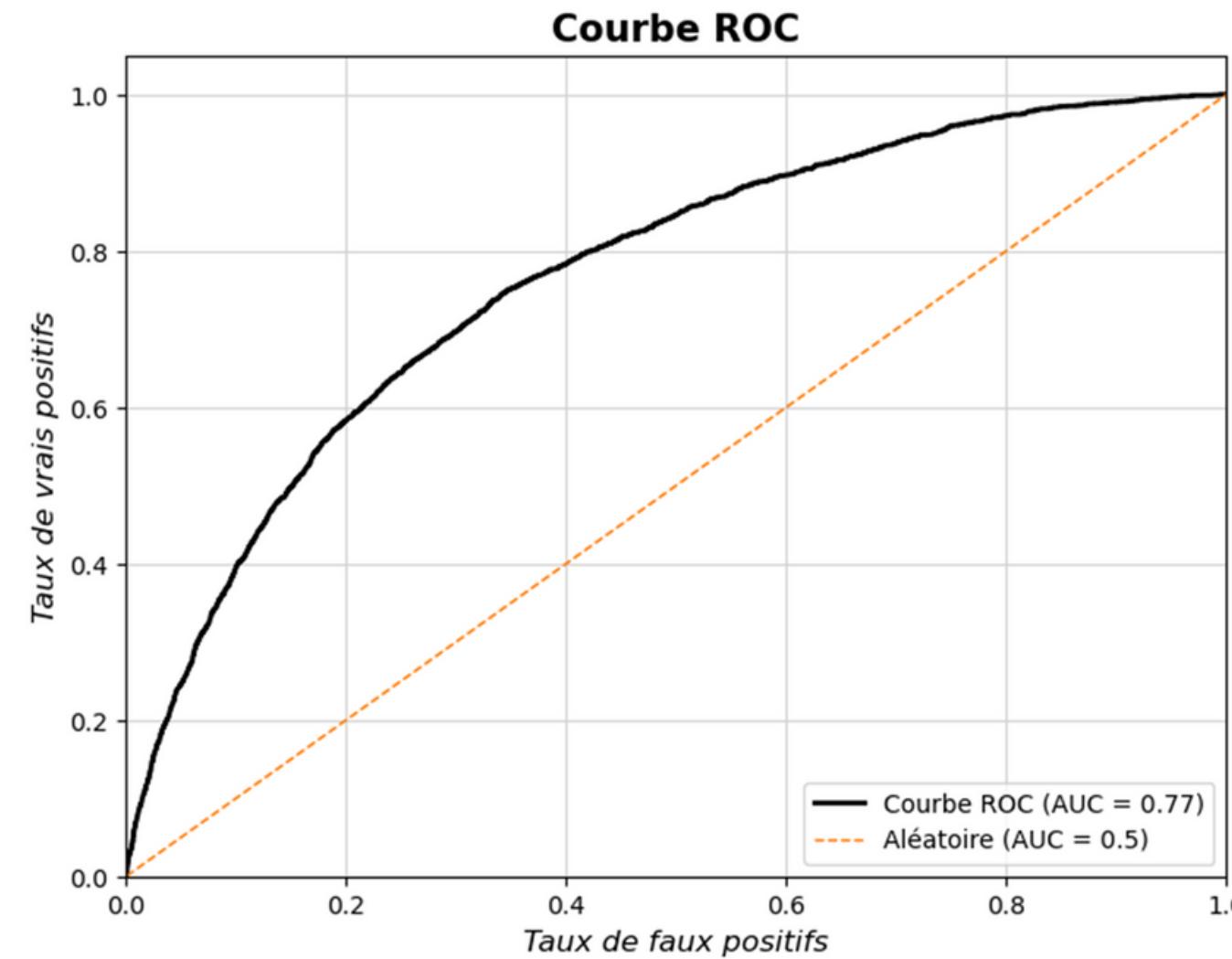
- Sélections des **meilleurs** paramètres
- Prédiction sur l'échantillon test

2.5 Méthode 4 : XGBoost

Performances



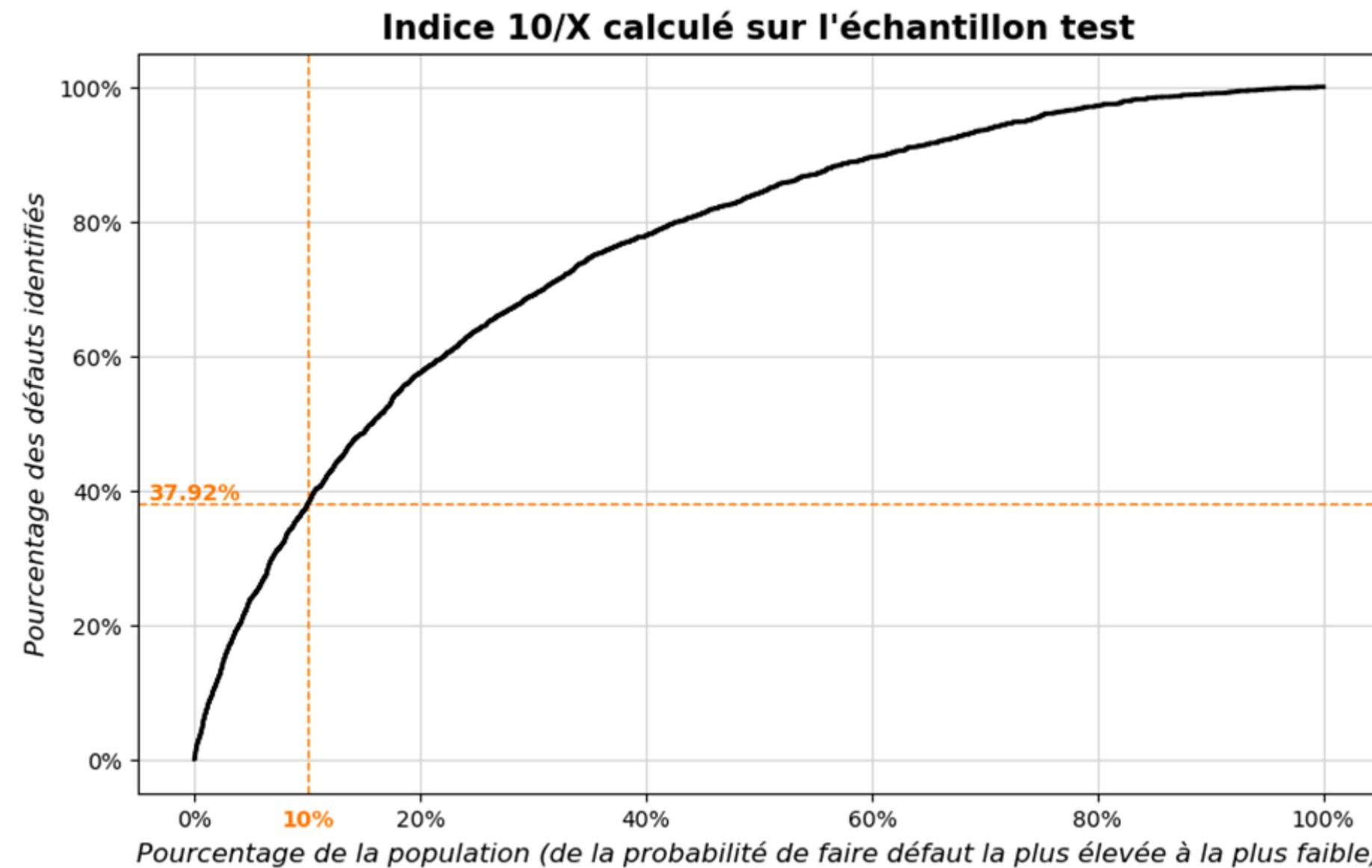
cut-off à 0.188



Gini = 0.54

2.5 Méthode 4 : XGBoost

Performances



3. CONCLUSION

3.1 Comparaison entre les modèles

3.2 Conclusion générale

3.3 Ouverture

3.1 Comparaison entre les modèles

	Régression logistique	Machine Learning
Avantages	<p>Interprétabilité: fournit des coefficients pour chaque variable, éclairant ainsi leur impact sur la décision de crédit, notamment en vue de la conformité réglementaire.</p> <p>Efficace avec de petits ensembles de données par rapport à des techniques de machine learning.</p> <p>Temps de calcul plus courts que les méthodes de machine learning</p>	<p>Capacité à modéliser des relations complexes: saisissent les relations non linéaires complexes courantes en crédit scoring.</p> <p>Haute précision: meilleure précision que la régression logistique, surtout avec un modèle bien paramétré.</p> <p>Capacité à gérer de grandes quantités de données</p>
Inconvénients	<p>Linéarité: suppose une relation linéaire entre les variables indépendantes et la variable dépendante, ce qui peut limiter son efficacité lorsque la relation est complexe.</p> <p>Moins adaptée à de grandes quantités de données</p>	<p>Moins interprétables: que la régression logistique. Ils ne fournissent pas de coefficients de variable directement interprétables.</p> <p>Plus de temps de calcul et de ressources que la régression logistique.</p> <p>Risque de surajustement: compromettant leur généralisation sur de nouvelles données.</p>

3.2 Conclusion générale

01

La construction d'un modèle de scoring est en effet un processus **complexe et exigeant**, nécessitant une approche **rigoureuse et critique**.

02

La construction d'un modèle de scoring nécessite que les données soient **cohérentes** avec la réalité et conformes aux aspects réglementaires.

3.3 Ouverture

Il est essentiel de reconnaître l'existence de nombreuses variables **inobservables** qui pourraient améliorer la précision du modèle. Ces données sont souvent **difficiles** à quantifier car elles ne sont pas systématiquement enregistrées, comme la fréquence des impayés pour le loyer, les paiements réguliers et la stabilité de l'emploi.

Le **comportement de consommation** peut influencer le risque de crédit. Certaines entreprises FinTech examinent même l'utilisation de données des médias sociaux pour évaluer le comportement financier. Cependant, cela suscite des questions de **confidentialité, de discrimination** et nécessite la conformité au **RGPD** en Europe.

Intégrer des **variables ESG** dans le scoring des prêts automobiles est essentiel. La consommation de carburant **influence** la capacité de remboursement, et le choix du véhicule peut entraîner des problèmes si de nouvelles réglementations l'interdisent, augmentant ainsi le risque de défaut de paiement en cas de perte de revenu.

01

02

03



CONTACTS



Lorenzo BARRAUD

barraudlorenzopro@gmail.com



Simon MIRZA

simon.mrza@gmail.com



Mathias VIEIRA DE BARROS

mathias.vieiradebarros@gmail.com

