MODÉLISATION D'UN SCORE D'OCTROI

NOVEMBRE 2024

MOBILIZE FINANCIAL SERVICES

MASTER ESA Université d'Orléans

I. INTRODUCTION

I.INTRODUCTION

> LE MARCHÉ DU VÉHICULE D'OCCASION NE CONNAÎT PAS LA CRISE. [...] SI BIEN QUE, DEPUIS LE DÉBUT 2024, IL AFFICHE UNE CROISSANCE DE 3,6 %
>
> **L'EXPRESS - OCT. 2024**

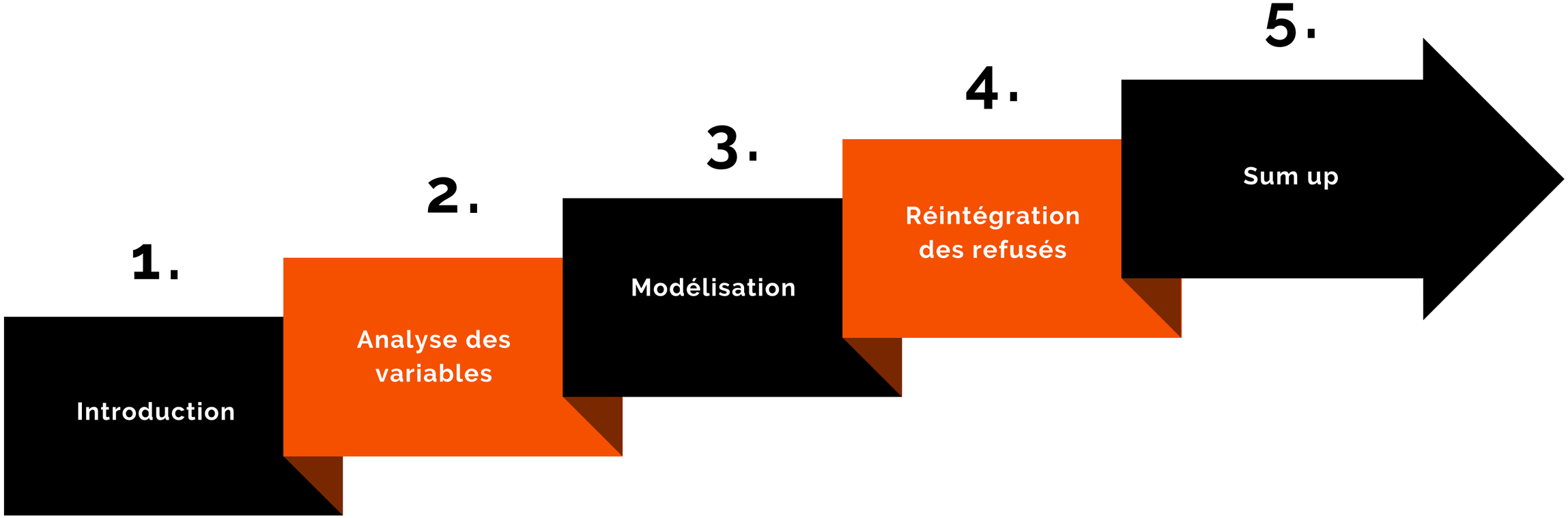
I.INTRODUCTION

Le marché européen des voitures d'occasion connaît une forte croissance, notamment grâce à la hausse des prix des véhicules neufs et aux progrès de la digitalisation.

Face à cette dynamique, **Mobilize Financial Services**, filiale de financement du groupe **Renault**, a sollicité les étudiants du **master ESA** pour développer un **score d'octroi** spécifique aux prêts automobiles dans ce secteur.

Le score sera fabriqué à partir des données des prêts **financés** et **refusés** accordés aux clients de **Mobilize Financial Services**

OVERVIEW


1.
Introduction
2.
Analyse des variables
3.
Modélisation
4.
Réintégration des refusés
5.
Sum up

II. ANALYSE DES VARIABLES

II. ANALYSE DES VARIABLES

Nous avons 2 bases de données, l'une avec les 173 205 financements acceptés, l'autre avec 18 824 financements refusés.

Variables :

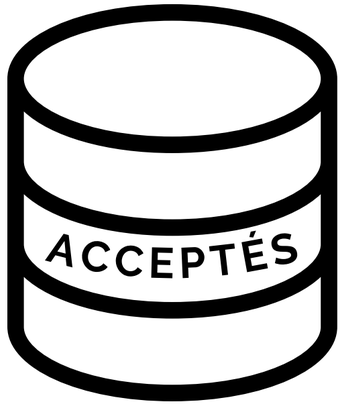
ID : Un identifiant unique attribué à chaque demande de prêt.

def12_nvd : notre variable cible

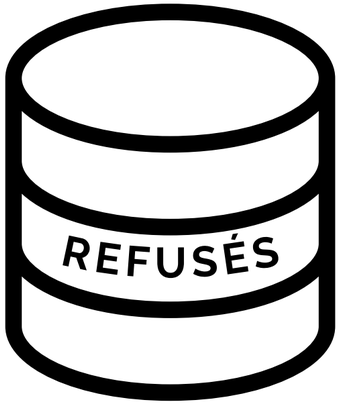
Mois de la demande et **mois d'entrée en gestion**

23 variables candidates continue : l'âge du client, l'âge du véhicule, le montant initial du financement etc...

9 variables candidates catégorielles : la marque du véhicule, la situation matrimoniale etc...



173 205
observations

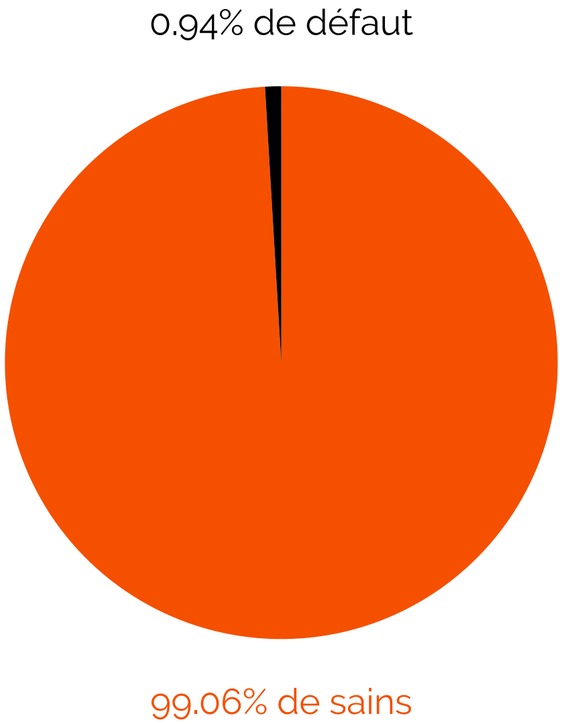


18 824
observations

II. ANALYSE DES VARIABLES

def12_nvd est notre variable cible indiquant si le client a fait défaut dans les 12 mois suivant l'octroi du prêt (1 en cas de défaut, 0 sinon)

def12_nvd	Fréquence	%
0	171571	99.06
1	1634	0.94



II. ANALYSE DES VARIABLES

Etudes des corrélations avec notre variable cible - variables continues

Nous utilisons la statistique de **Kruskal-Wallis** qui nous permet de retirer d'embler la variable **"Montant de la reprise du contrat"**

Tableau des corrélations de Kruskal-Wallis avec la variable cible

Variable	Statistique de Kruskal-Wallis	p-value
Montant de la reprise du contrat	3.41	0.065
Nombre de personnes à charge	6.21	0.012703
Montant de la cotation automatique	7.43	0.006419
Montant des remises	10.17	0.001424
...	...	...
Ancienneté relation avec la Schufa	305.13	< 0.0001
Ancienneté emploi	364.43	< 0.0001
Montant de l'apport	397.46	< 0.0001
Pourcentage d'apport	422.56	< 0.0001

II. ANALYSE DES VARIABLES

Etudes des corrélations avec notre variable cible - variables catégorielles

Nous faisons de même pour les variables catégorielles en utilisant cette fois-ci le **V de Cramer**.

Nous retirons les 4 variables **en rouge.**

Tableau des corrélations du V de Cramer avec la variable cible

Variable	X^2	V de Cramer	p-value
Indicateur client RCI	0.158	0.001142	0.69089
Code de la catégorie d'exposition	0.937	0.002779	0.33318
Co-contractant	2.915	0.004903	0.08777
Montant de la reprise du contrat qualitative	3.463	0.005344	0.06277
Type de bien	8.536	0.008391	0.00348
Remise	10.362	0.009245	0.00129
Montant de la cotation manuelle qualitative	17.573	0.012039	0.00003
Code marque du véhicule RCI	57.246	0.021729	< 0.0001
Type de profession	89.003	0.027094	< 0.0001
Mauvais paiement dans les 6 derniers mois	169.802	0.037423	< 0.0001
État civil	252.481	0.045634	< 0.0001
Mode d'habitation	278.925	0.047964	< 0.0001

II. ANALYSE DES VARIABLES

Discrétisations des variables quantitatives

Pourquoi ?

- Prendre en compte des effets non linéaires
- Réduire l'effet des valeurs extrêmes et manquantes
- Améliorer l'interprétabilité

Comment ?

Nous utilisons le critère de Belson pour avoir nos modalités.

Le choix du nombre de modalité est important:

- Trop peu amènerais à un faible pourvoir de prédiction
- Trop nombreuses peut amener à problème de généralisation

Ici nous nous limitons à 2,3 ou 4 modalités qui sont choisis sur base de la **monotonie** i.e évolution unidirectionnel du taux de défaut

Tableau du choix des modalités pour ancienneté de la relation avec rci

Nb_Mod	Chi2	Bornes	TX_DEFAUT	DIFF_TDF
2	108,63	[0@1]	1,14459	.
2	108,63	]1@38	0,52701	-0,61758
3	109,972	[0@1]	1,14459	.
3	109,972	]1@38	0,48873	-0,65586
3	109,972	[13@38]	0,61	0,12127
4	110,343	[0@1]	1,14459	.
4	110,343	]1@5	0,5194	-0,62519
4	110,343	]5@13	0,44682	-0,07258
4	110,343	]13@38	0,61	0,16318

II. ANALYSE DES VARIABLES

Prise en compte de la multicolinéarité

Variables quantitatives

Pour les variables quantitatives, une manière simple de détecter la corrélation entre deux variables consiste à calculer leur coefficient de corrélation de Pearson ou de Spearman

Les variables avec un coefficient de Spearman supérieure à 0.7 sont en rouge

Afin de **réduire le risque de multicolinéarité**, nous supprimons les **variables** :
montant de l'apport, prix du véhicule, prix catalogue du véhicule, montant automatique et montant de la valeur résiduelle

10 combinaisons des variables quantitatives les plus corrélées

1ère variable	2ème variable	p_value	rs_value
Prix Catalogue	**Prix Véhicule**	0,97898	0,99303
Montant Ballon	**Montant Valeur Résiduelle**	0,97513	0,97857
Montant Apport	Pourcentage Apport	0,84104	0,95973
Variable Binaire Ballon	**Montant Valeur Résiduelle**	0,79393	0,94483
Montant Automatique	**Prix Véhicule**	0,85042	0,86726
Montant Automatique	**Prix Catalogue**	0,83607	0,86435
Montant initial du financement	**Prix Véhicule**	0,84739	0,82294
Montant initial du financement	**Prix Catalogue**	0,83054	0,81771
Montant Automatique	Montant initial du financement	0,72206	0,71647
Age du véhicule	Prix Véhicule	0.59120	0.68067

II. ANALYSE DES VARIABLES

Prise en compte de la multicolinéarité

Variables qualitatives

Pour les variables qualitatives, il nous suffit de calculer le V de Cramer entre chaque variable

Les variables avec un coefficient de V de Cramer supérieure à 0.3 sont en rouge

Afin de **réduire le risque de multicolinéarité**, nous supprimons les **variables** :
âge du client et montant initial du financement

7 combinaisons des variables qualitatives les plus corrélées

1ere_variable	2eme_variable	V de Cramer
Montant initial du financement discrétisé	Montant du versement mensuel discrétisé	0,50686
Age client discrétisé	Code état civil regroupé	0,38833
Montant de l'engagement de reprise discrétisé	**Montant initial du financement**	0,33237
Age client discrétisé	Code mode d'habitation	0,30462
Ancienneté relation RCI discrétisé	Ancienneté relation Schufa discrétisé	0,30327
Présence de leasing ou non	Durée prévisionnelle du financement discrétisé	0,29829
Présence de leasing ou non	Ancienneté emploi discrétisé	0,29716

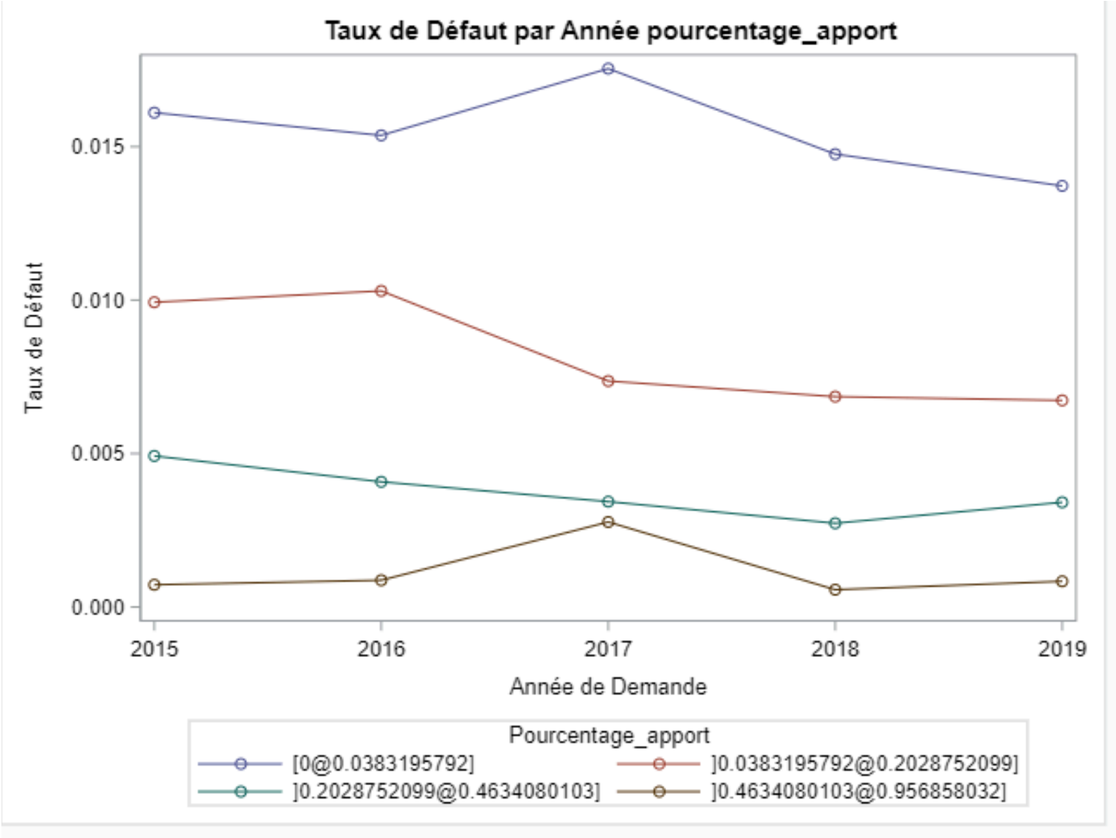
II. ANALYSE DES VARIABLES

Analyse de la stabilité temporelle

Il est important qu'une variable **fortement discriminante** à une période le soit toujours à la période suivante. À cet égard, une **analyse historique** des variables doit être menée.

Nous devons vérifier que chaque modalité entretiennent **une relation stable et bien hiérarchisée** avec l'indicateur de défaut.

La plupart de nos variables ont des modalités
stables dans le temps



II. ANALYSE DES VARIABLES

Analyse de la stabilité temporelle

Non Stable

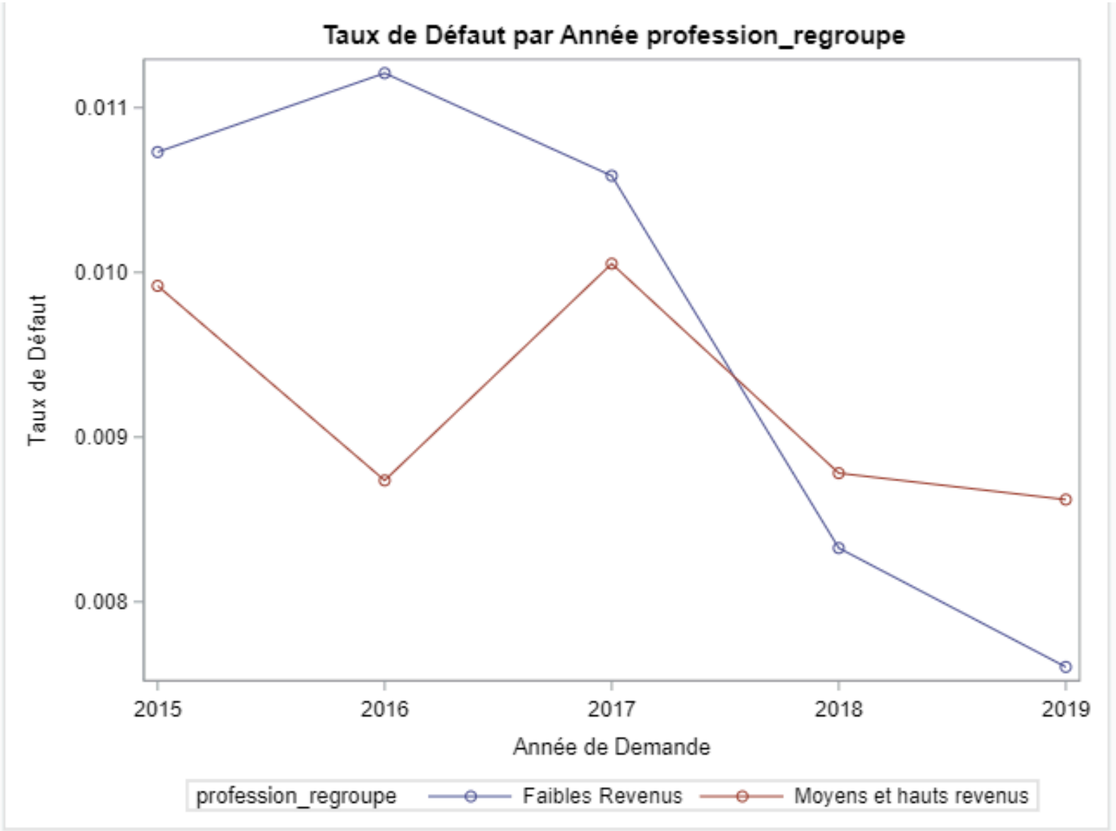
Certaines variables ne le sont pas

Pourquoi ?

- Problèmes de discrétisations
- Manque de données
- Changements structurels (Covid, guerre, etc ...)

Conséquences

- Détriment de la performance
- Interprétation erronée au cours du temps
- Incompatible avec des prévisions à long terme



III. MODÉLISATION

III. MODELISATION

Le Logit

Après avoir séparé notre échantillon en 2, un échantillon **d'apprentissage** de 70% de notre échantillon originelle et un échantillon **test** de 30% de notre échantillon originelle

Nous avons sélectionné **17 variables explicatives** pour la modélisation

Nous estimons une **régression logistique** (logit) sur notre ensemble d'apprentissage

Pourquoi le Logit ?

- Le logit permet de prédire une probabilité comprise entre 0 et 1, ce qui est intuitif pour évaluer le risque de défaut d'un client.
- Ses coefficients estimés par maximum de vraisemblance permettent de générer un **score** aisément interprétable

$$Y_i = \begin{cases} 1 & \text{si defaut} \\ 0 & \text{sinon} \end{cases}$$

$$\mathbb{P}(Y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i \boldsymbol{\beta}}}$$

III. MODELISATION

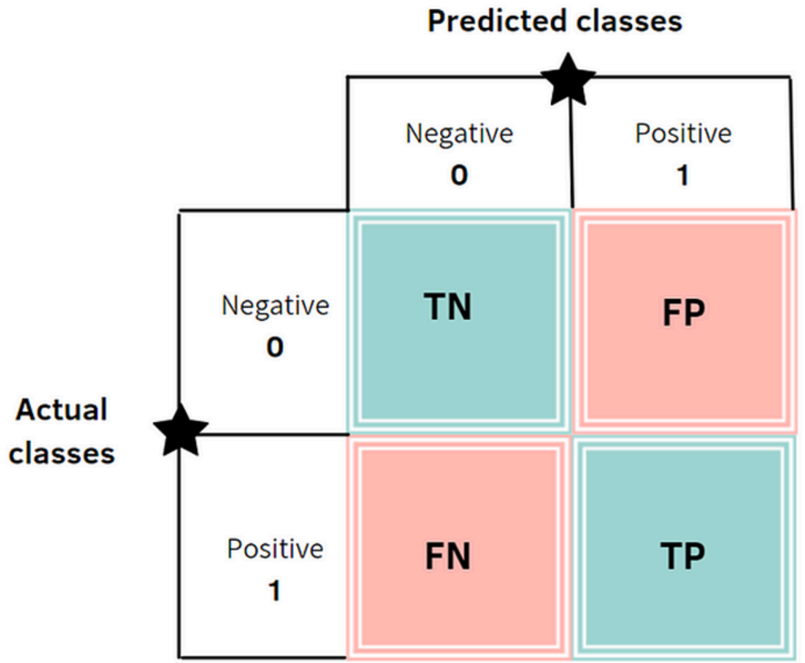
Matrice de confusion

- **Vrais positifs (TP)** : individus correctement classés comme positifs (défaut correctement prédit). On appelle le taux de vrais positifs la **sensibilité**.

- **Faux positifs (FP)** : individus incorrectement classés comme positifs (prédit comme défaut alors qu'il n'y en a pas).

- **Vrais négatifs (TN)** : individus correctement classés comme négatifs (absence de défaut correctement prédit). On appelle le taux de vrais positifs la **spécificité**.

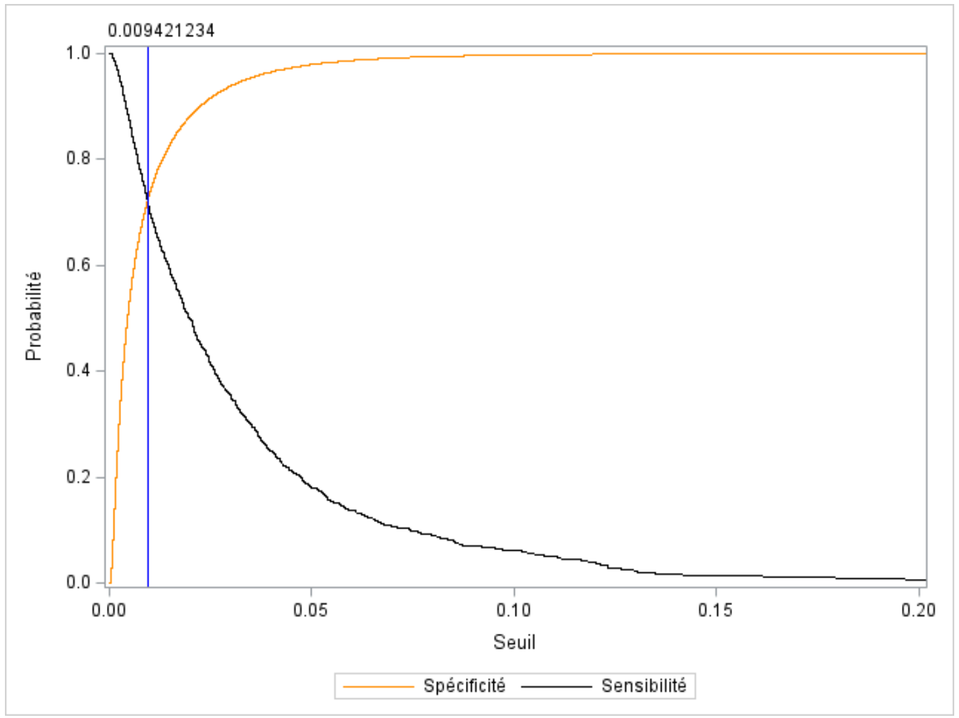
- **Faux négatifs (FN)** : individus incorrectement classés comme négatifs (défaut non détecté).



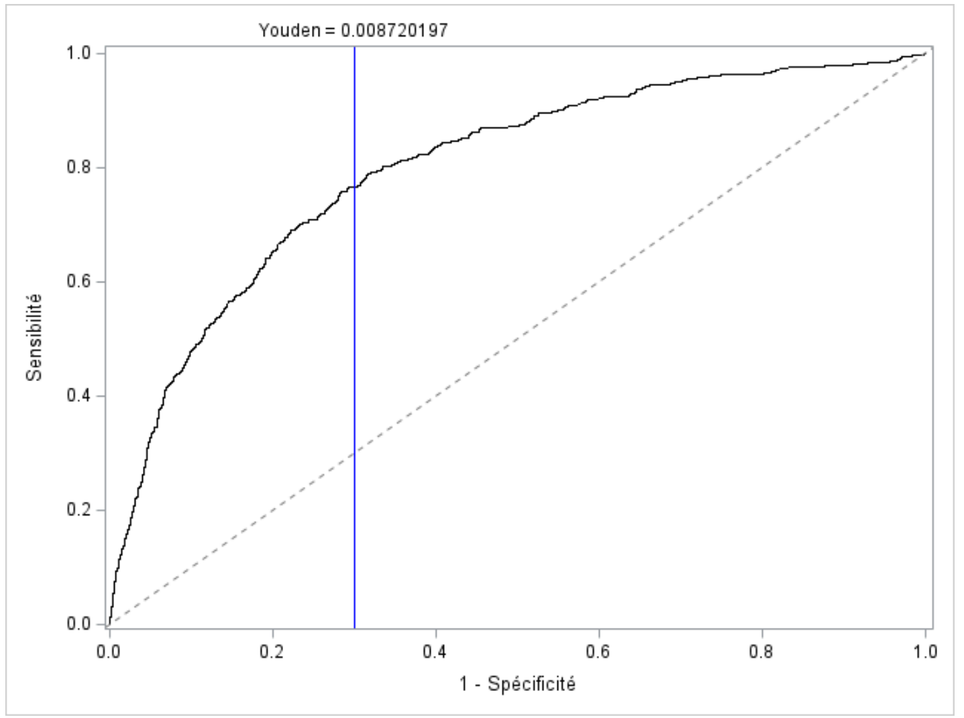
III. MODELISATION

Matrice de confusion - Cut-off optimal

Selon le critère **sensibilité = spécificité**



Selon **l'Indice de Youden** : le point le plus loin sur la l'aléatoire sur la courbe ROC.



III. MODELISATION

Matrice de confusion

Selon le critère **sensibilité = spécificité**

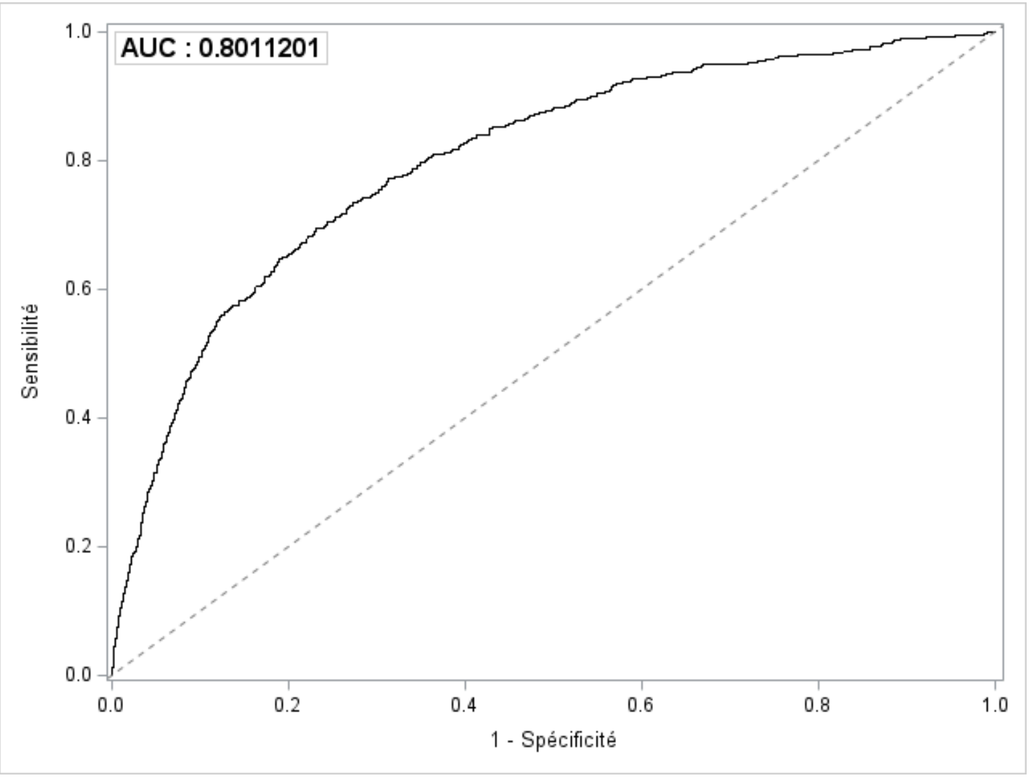
def12_nvd		Y Prédit	
		0	1
Y	0	**37 040** (71.96%)	**14 431** (28.04%)
	1	**123** (25.10%)	**0.94** (74.90%)

Selon **l'Indice de Youden**

def12_nvd		Y Prédit	
		0	1
Y	0	**36 002** (69.95%)	**15 469** (30.05%)
	1	**114** (23.27%)	**376** (76.73%)

III. MODELISATION

Performances prédictives sur l'échantillon test - Courbe ROC & AUC

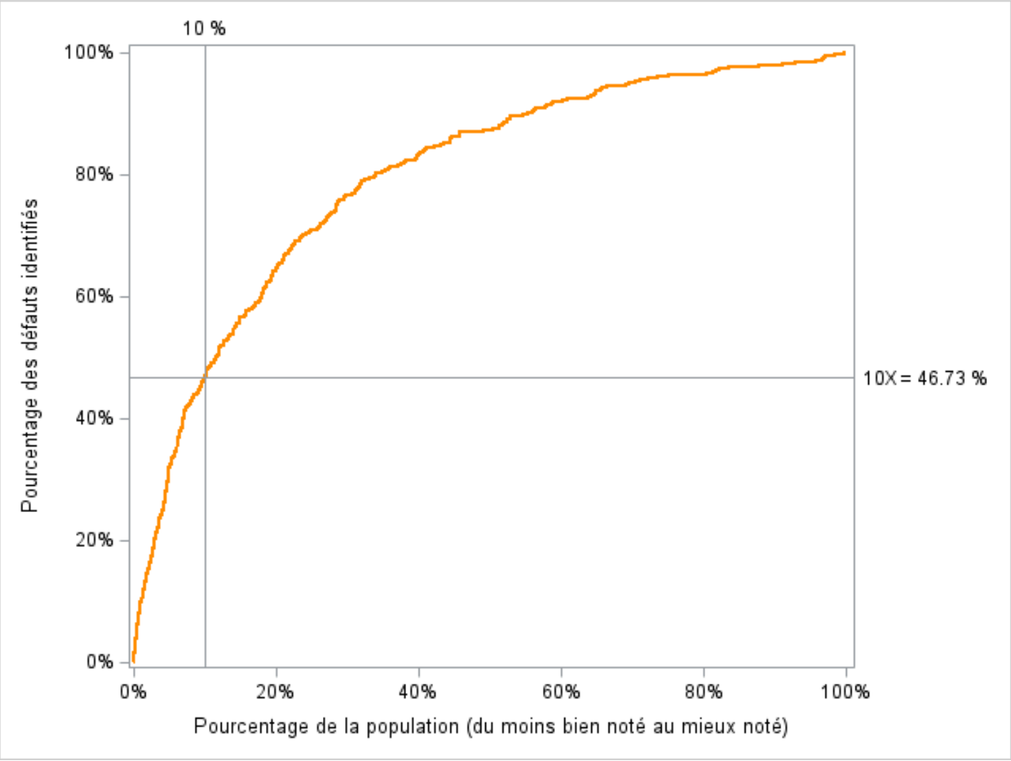


La **courbe ROC** représente le taux de vrais positifs en fonction du taux de faux positifs . Elle balaie donc la capacité discriminante pour chaque cut-off.
Idéalement la courbe roc doit s'éloigner de la diagonale (classification aléatoire) pour se rapprocher du coin supérieur droit.

L'**Area Under the Curve (AUC)** est la mesure chiffré de la courbe roc. Elle mesure l'aire sous la courbe ROC.
Une AUC supérieur à 0.8 signifie un très bon pouvoir discriminant.

III. MODELISATION

Performances prédictives sur l'échantillon test - Courbe LIFT & Indice 10X



La **courbe LIFT** représente le taux de vrais positifs en fonction des parts d'individus les moins bien scorés.

Cela permet en pilotage stratégique de fixer des objectifs par part de marché.

L'indice **10/X** indique pour les 10% ayant les scores les plus bas les X% de vrais positifs.

Ici pour les 10% des individus les plus risqués on détecte **46.73%** de défaut correctement identifiés.

III. MODELISATION

Gestion du déséquilibre entre les classes - Oversampling

On est confronté à un problème de **sous-représentation des défauts** (0,94 %), car les individus analysés ont déjà été sélectionnés sur leur capacité de remboursement. Ce déséquilibre pousse les modèles de classification à privilégier la classe majoritaire, entraînant une précision globale élevée mais une détection faible des défauts. En conséquence, le modèle devient **moins fiable pour identifier les défauts** et moins capable de généraliser sur de nouvelles données.
C'est pourquoi nous allons explorer les méthodes de **rééquilibrage des classes**.

def12_nvd	Fréquence	%
0	12 0100	99.06
1	1 144	0.94

def12_nvd	Fréquence	%
0	12 0100	90
1	13 344	10

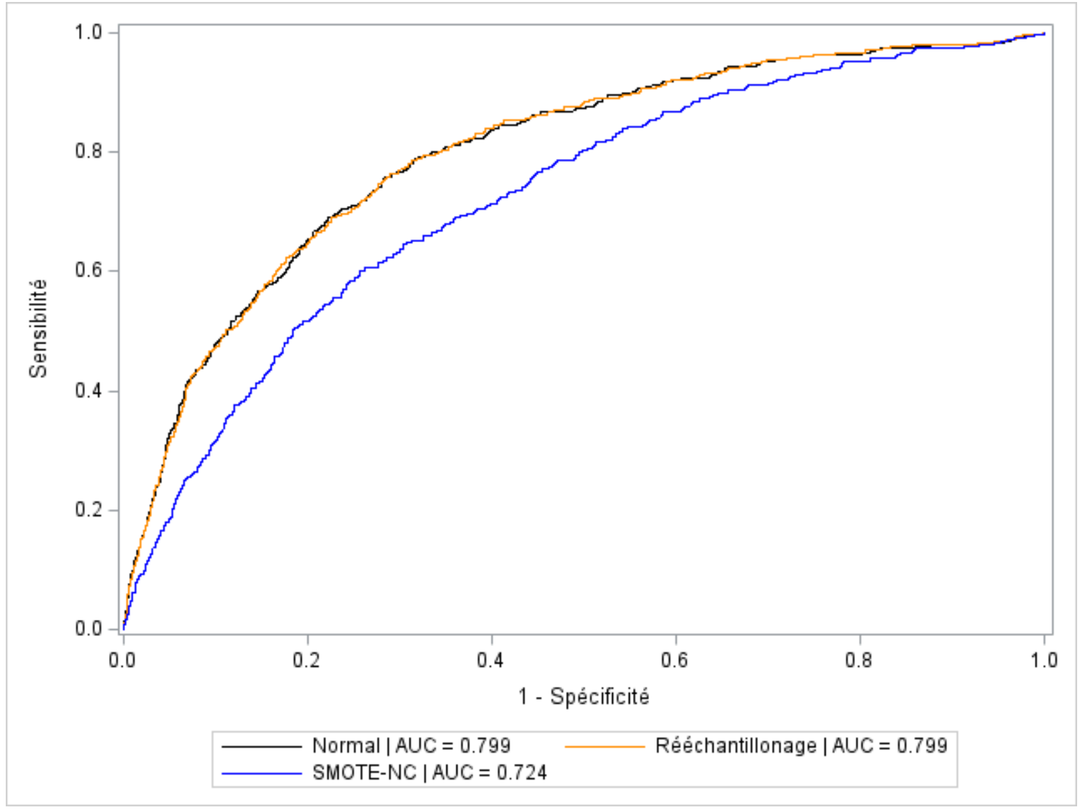
III. MODELISATION

Gestion du déséquilibre entre les classes - Oversampling

Rééchantillonnage par tirage avec remise

On fait un tirage aléatoire avec remise dans les individus en défaut et on les duplique tel quel.

Ici on voit quasiment la même courbe et ROC et la même AUC



SMOTE -NC

On fait un tirage comme pour la méthode précédente mais au lieu de dupliquer,
on va créer un individus synthétiques basé sur la méthodes des **K-nearest neighbors**. (Ici K=5)

On a des résultats détériorés à cause du sur-apprentissage.

Nous allons nous en tenir à notre modèle normal.

III. MODELISATION

Grille de score 1/2

Âge du véhicule

Les véhicules plus vieux nécessitent généralement plus d'entretien et sont moins fiables. Cela peut engendrer des coûts imprévus pour l'emprunteur.

Ces dépenses supplémentaires réduisent la capacité de remboursement des crédits.

Libellé	Modalité	Coef	Distribution	Tx de défaut	Contribution	Coef Recalibré (Score)
Marque du véhicule	Infiniti-Autres	0.3097	18.78%	1,38%	3,12%	25
	Renault	0.2174	52.37%	0,91%		18
	Dacia-Nissan-Infiniti	0,000	28.85%	0,72%		0
Type d'emploi	Moyens et hauts revenus	-0.3211	74.29%	0,92%	3,81%	0
	Faibles revenus	0,000	25.71%	0,99%		26
Type de véhicule	Véhicule Utilitaire	0.4776	2.30%	1,51%	1,95%	39
	Véhicule Particulier	0,000	97.70%	0,93%		0
Mauvais paiement dans les 6 derniers mois	Oui	1.7575	1,45%	3,79%	5,73%	144
	Non	0,000	98,55%	0,90%		0
Âge du véhicule (mois)	[0 ; 2]	0,000	6.41%	0,52%	9,41%	0
	] 2 ; 63]	0.7353	76.10%	0,87%		60
	> 63	1.4266	17.49%	1,41%		117
Ancienneté de l'emploi (années)	[0 ; 2]	0,000	25,30%	1,77%	5,60%	40
	] 2 ; 5]	-0.2651	56,20%	1,11%		18
	] 5 ; 67]	-0.4894	18,50%	0,52%		0
Ancienneté de l'habitation (années)	[0 ; 1]	0,000	13.61%	1,80%	3,17%	30
	] 1 ; 5]	-0.1724	24.44%	1,19%		16
	] 5 ; 18]	-0.2676	35.96%	0,73%		8
	] 18; 87]	-0.3697	25.99%	0,57%		0
Anciennet relation RCI (années)	[0 ; 21]	0,000	67.45%	1,15%	5,36%	35
	] 21; 38]	-0.4212	32.55%	0,52%		0

III. MODELISATION

Grille de score 2/2

Pourcentage d'apport

Un apport élevé diminue le risque de défaut, indiquant une capacité d'épargne et une gestion financière saine.

Montant du versement mensuel

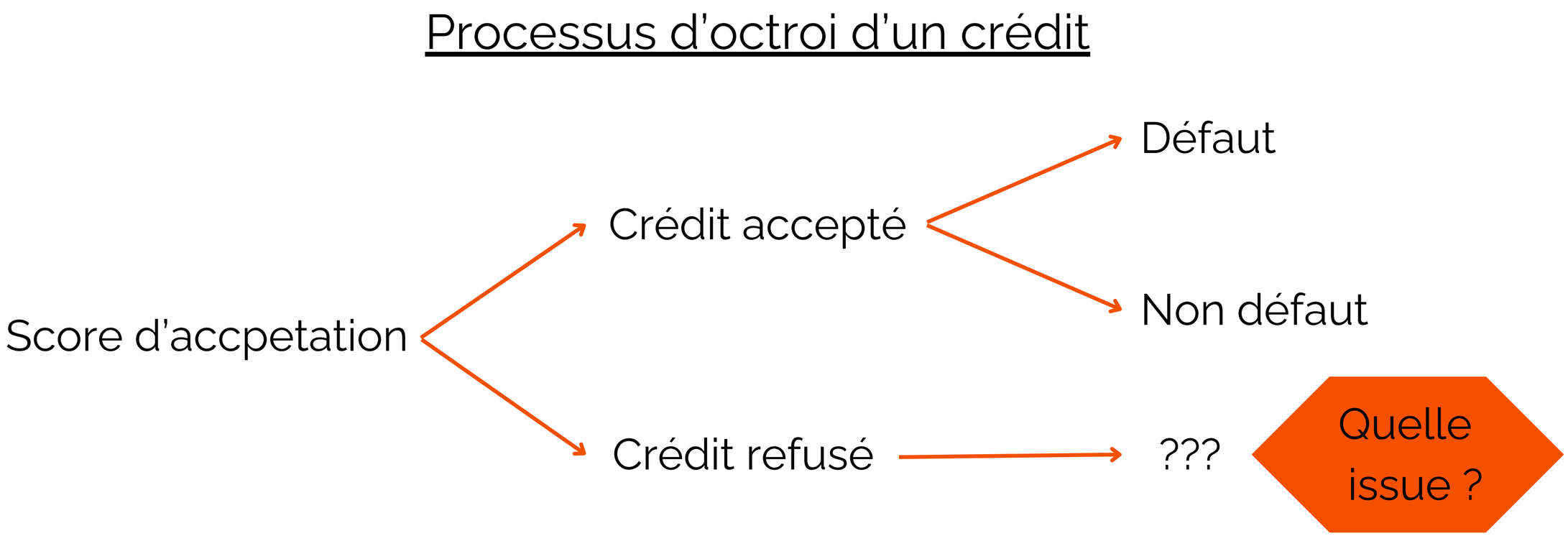
Les clients avec des paiements mensuels bas présentent des scores plus faibles, signifiant un risque de défaut plus faible. Ces clients tendent à contracter des crédits moins importants, reflétant une gestion plus prudente de leurs finances.

Libellé	Modalité	Coef	Distribution	Tx de défaut	Contribution	Coef Recalibré (Score)
Ancienneté relation Schufa (années)	[0 ; 2]	0,000	31,15%	1,65%	7,69%	57
	] 2 ; 6]	-0.5244	40.82%	0,74%		14
	> 6	-0.6919	28.03%	0,45%		0
Crédit Ballon	oui	0.3128	8.02	1,16%	4,08%	26
	non	0	91.98	0,82%		0
Situation Matrimoniale	Divorcé(e) ou Veuf/Veuve	0.2587	12.62%	0,86%	4,85%	21
	Marié(e)	0	57.16%	0,62%		0
	Célibataire ou Séparé(e)	0.3896	57.16%	1,59%		32
Mode d'habitation	Propriétaire	0,000	43,11%	0,46%	5,84%	0
	Locataire	0.4421	49,23%	1,26%		36
	Autres	0.3616	7,66%	1,66%		30
Durée prévue du financement (mois)	[4 ; 48]	0,000	46,07%	0,55%	9,17%	0
	] 48 ; 60]	0.4393	29.59%	1,08%		36
	] 60 ; 84]	0.6545	18.22%	1,40%		54
	> 84	1.1439	6.13%	1,90%		94
Montant de l'engagement repris	[0;7500]	0	93.95%	0,89%	2,95%	0
	]7500;41870.33]	0.4540	6.05%	1,80%		37
Montant du versement mensuel (€)	[17.13;117.31]	0,000	18.00%	0,67%	9,50%	0
	]117.31;253.76]	0.3167	65.00%	0,84%		26
	]253.76;310.51]	0.9059	10,00%	1,40%		74
	> 310.51	1.3530	7,00%	1,92%		111
Montant Revenue	[100;1894]	0	50.21%	1,19%	2,58%	16
	]1894;30000]	-0.1894	49.79%	0,70%		0
Pourcentage d'apport	[0 ; 3,83%]	0,000	42.00%	1,55%	15,47%	135
	] 3,83% ; 20,28%]	-0.6106	22.00%	0,84%		85
	] 20,28% ; 46,34%]	-1.0241	24.00%	0,40%		51
	> 46,34%	-1.6403	12.00%	0,14%		0

IV. RÉINTÉGRATION DES REFUSÉS

IV. REINTÉGRATION DES REFUSÉS

L'approche classique du scoring d'octroi se base exclusivement sur la **population de clients financés.** Ce même modèle est appliqué à l'ensemble de **nouveaux clients**.

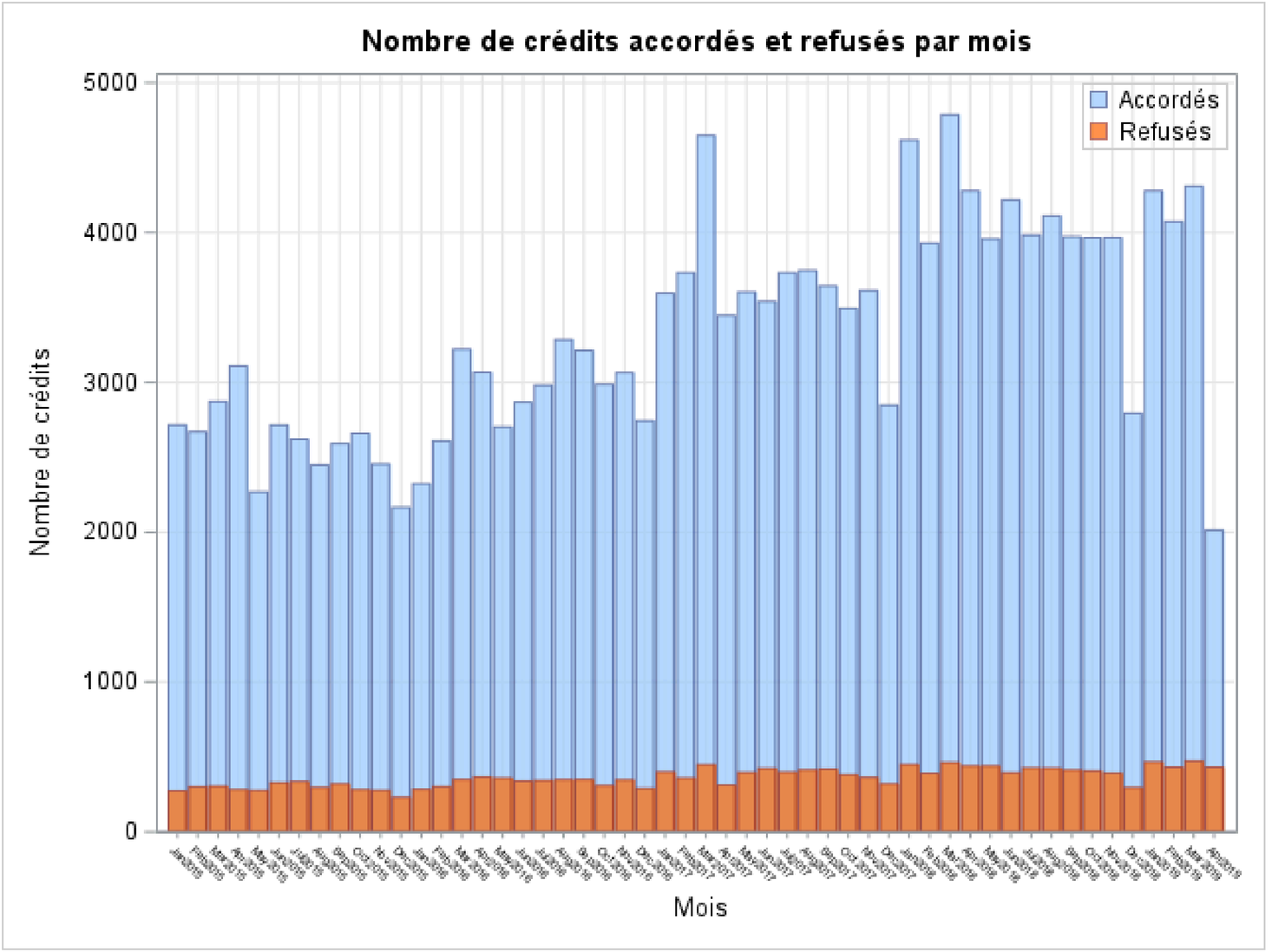


IV. REINTÉGRATION DES REFUSÉS

Entre janvier 2015 et avril 2019, **10,86 % des ménages** se sont vu refuser un crédit.

On peut imaginer qu' 1 futur client sur 10 ne sera pas considéré par le modèle de scoring.

Un test du X^2 va nous indiquer que **l'échantillon des clients financés a un distribution différente de celle des réfusés.**



IV. REINTÉGRATION DES REFUSÉS

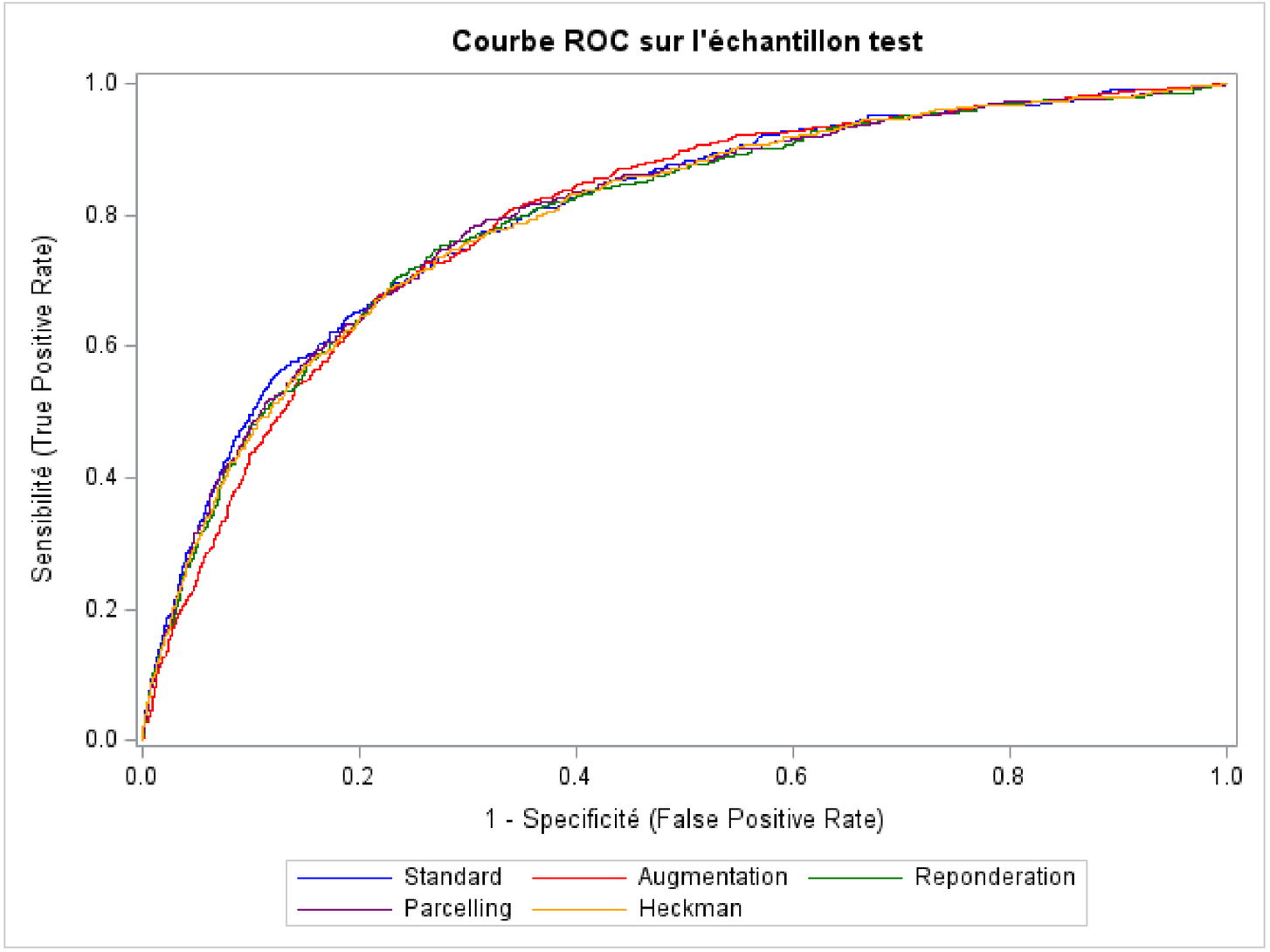
Le **biais de sélection** survient lorsque l'échantillon utilisé pour l'analyse n'est pas représentatif de la population cible, ce qui peut fausser les résultats et conduire à des conclusions erronées.

Augmentation simple	Repondération	Parcelling	Heckman
• Attribution d'étiquette "bon-mauvais" payeur des refusés sur la base du modèle initial. • Réestimation du scoring en tenant compte des "nouveaux" refusés.	• Estimation de la probabilité d'acceptation pour chaque client (acceptés et refusés). • Repondération des clients acceptés selon l'inverse de leur probabilité d'acceptation, puis recalcul du modèle de score sur les données pondérées.	• Assignation des refusés aux CHR correspondantes • Affectation aléatoire des refusés en "défaut" ou "non-défaut" selon la proportion observée dans chaque classe.	• Modèle d'acceptation de crédit sur l'échantillon d'apprentissage et des refusés • Calcul du ratio de l'inverse Mills avec les paramètres du modèle d'acceptation • Estimation du modèle de défaut en intégrant l'inverse de ratio de Mills

IV. REINTÉGRATION DES REFUSÉS

Evaluation des performances à l'aide de la **courbe ROC** sur l'échantillon test avec les différents modèles.

Difficulté de démarquer une méthode au détriment d'une autre.



IV. REINTÉGRATION DES REFUSÉS

Convergence des métriques d'AUC avec un **léger avantage** pour l'approche classique.

Surperformance plus notable concernant **l'indice 10X** pour l'approche classique

Modèle	AUC (test)	Indice 10X
Approche classique	0.801	48.37%
Augmentation simple	0.794	41.84%
Repondération	0.794	46.33%
Parcelling	0.797	46.53%
Heckman	0.795	45.51%

V. SUM UP

V. SUM UP

- Modélisation standard satisfaisante avec un **AUC de 0.801** et un **indice 10X de 48.37%.**

- Les méthode de réintégration des refusés ne semblent **pas bénéficier de résultats significativement meilleurs**.
  - Réel biais de sélection ?
  - Echantillon de test biaisé

- Test des données sur échantillon out-of-time.

- Changement structurel du marché de l'automobile d'occasion à partir de 2035.

MERCI POUR VOTRE ATTENTION

BLECON Gaetan
gaetanblecon@gmail.com

LANGER Marin
marin.langer1@gmail.com

AYIVI-TOGBASSA Joris
joris.ayivi12@outlook.fr

MOBILIZE FINANCIAL SERVICES

MASTER ESA Université d'Orléans