

Modélisation du risque de crédit



BARRAUD Lorenzo
MIRZA Simon
VIEIRA DE BARROS Mathias



Contexte.



Rentabilité et concurrence

Lorsqu'une personne demande un prêt à une banque, celle-ci examine la capacité de remboursement du client. La fixation du taux d'intérêt du prêt devient alors importante pour maximiser les bénéfices et rester compétitive par rapport aux autres banques.



Information client

Une analyse statistique du profil du client est établie afin d'obtenir un score similaire à une Probabilité de Défaut. La régression logistique est la méthode de modélisation la plus fréquemment employée par les institutions pour ce faire.



Amélioration des performances

Ces dernières années, de nombreuses banques explorent l'utilisation des techniques de Machine Learning plus avancées pour optimiser le processus de prise de décision lié à l'octroi. Toutefois, il est nécessaire de démontrer concrètement les bénéfices potentiels en termes de discernement.



Non transparence

Les algorithmes opaques posent des défis d'interprétation. Pour garantir l'acceptation des experts métier et prévenir les biais discriminatoires indésirables (qui pourraient nuire à la réputation de l'établissement), des méthodes sont disponibles pour mieux comprendre ces modèles avancés.

Exploration des données.

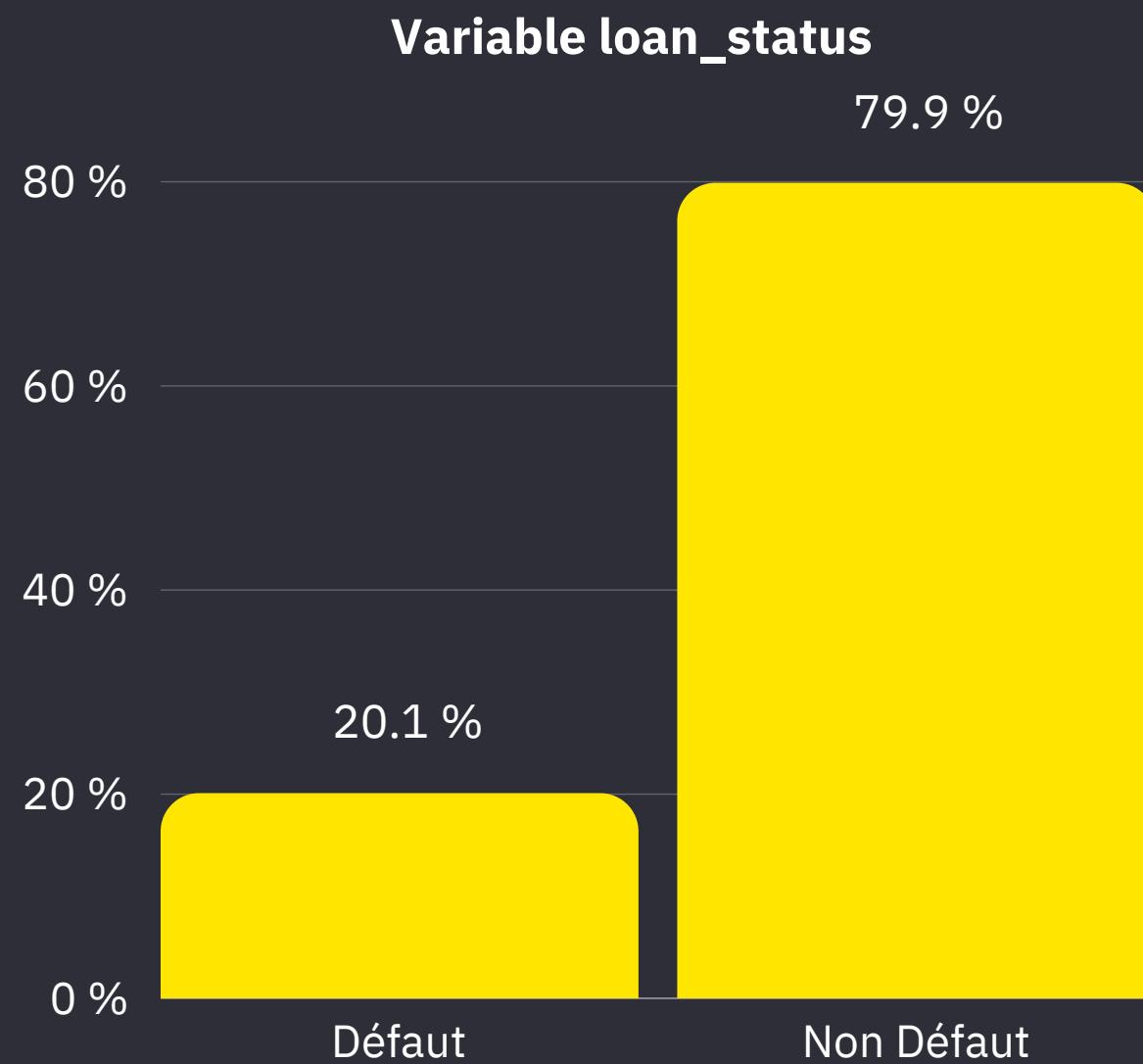
Présentation de la base.

Base : [Lending Club 2007-2020Q3 | Kaggle](#) (pré traitée)

↪ fait référence à 159 611 emprunts de 2007 à 2020 de chez LendingClub, la plus grande plateforme de prêt peer-to-peer au monde

Variables :

- 127 variables donnant des informations sur l'individu, son score comportemental et sur son crédit
- 1 variable cible (loan_status) donnant une information sur l'état du prêt :
 - intégralement remboursé
 - considéré comme une perte pour le prêteur



Sélection des variables.



La base de données Lending Club contient des informations sur leur **clients actuels**. Or, le **but** ici est de développer un modèle de **score d'octroi** (étude de la solvabilité d'un nouveau client)

Sélection des variables.



La base de données Lending Club contient des informations sur leur **clients actuels**. Or, le **but** ici est de développer un modèle de **score d'octroi** (étude de la solvabilité d'un nouveau client)

1

Suppression des variables **impossibles** à obtenir en tant que **nouveau client** (exemple : le grade du client pour l'octroi)
(suppression de 21 variables)

Sélection des variables.



La base de données Lending Club contient des informations sur leur **clients actuels**. Or, le **but** ici est de développer un modèle de **score d'octroi** (étude de la solvabilité d'un nouveau client)

1

Suppression des variables **impossibles** à obtenir en tant que **nouveau client** (exemple : le grade du client pour l'octroi)
(suppression de 21 variables)

2

Suppression des **valeurs manquantes** au seuil de **30%**
(suppression de 23 variables)

Sélection des variables.



La base de données Lending Club contient des informations sur leur **clients actuels**. Or, le **but** ici est de développer un modèle de **score d'octroi** (étude de la solvabilité d'un nouveau client)

1

Suppression des variables **impossibles** à obtenir en tant que **nouveau client** (exemple : le grade du client pour l'octroi)
(suppression de 21 variables)

2

Suppression des **valeurs manquantes** au seuil de **30%**
(suppression de 23 variables)

3

Suppression des variables **inutiles** (1 modalité, 99% de la même modalité, variable ID...)
(suppression de 19 variables)

Sélection des variables.



La base de données Lending Club contient des informations sur leur **clients actuels**. Or, le **but** ici est de développer un modèle de **score d'octroi** (étude de la solvabilité d'un nouveau client)

- 1 Suppression des variables **impossibles** à obtenir en tant que **nouveau client** (exemple : le grade du client pour l'octroi)
(suppression de 21 variables)
 - 2 Suppression des **valeurs manquantes** au seuil de **30%**
(suppression de 23 variables)
 - 3 Suppression des variables **inutiles** (1 modalité, 99% de la même modalité, variable ID...)
(suppression de 19 variables)
 - 4 Suppression des variables **corrélées** entre elles (tests de Spearman, V de Cramer, KW)
(suppression de 34 variables)
- Il reste dès à présent **31 variables**.

Modifications effectuées.

Variable	% VM
il_util	14,75
emp_length	8,88
mths_since_recent_inq	8,71
mo_sin_old_il_acct	2,69
bc_util	1,6
dti	0.24

- Pour les variables ayant **moins de 2%** de valeurs manquantes:
 - ↪ imputation par la **médiane**
- Pour les variables avec **plus de 2%** de valeurs manquantes :
 - ↪ imputation par **KNN** des autres variables avec utilisation de la validation croisée afin de déterminer le nombre optimal de voisins

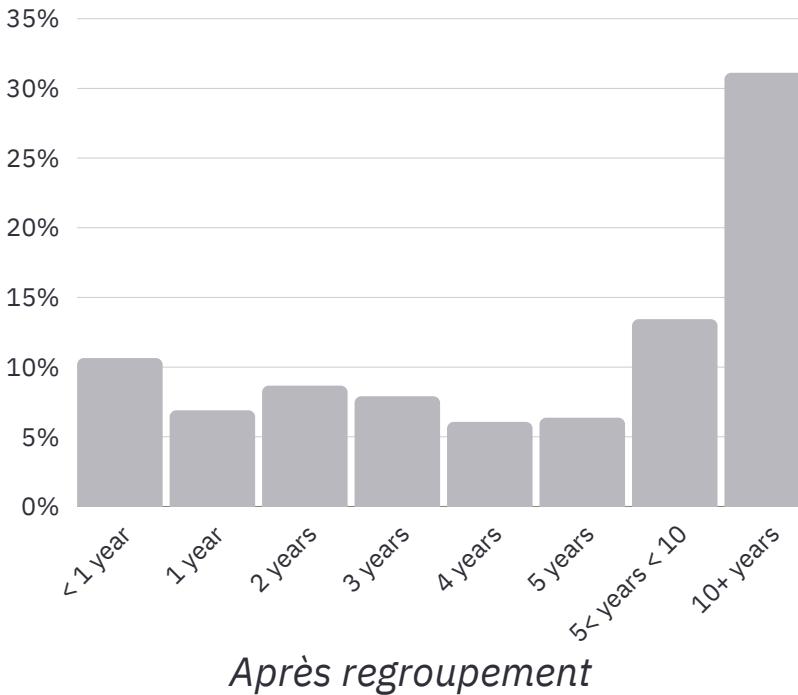
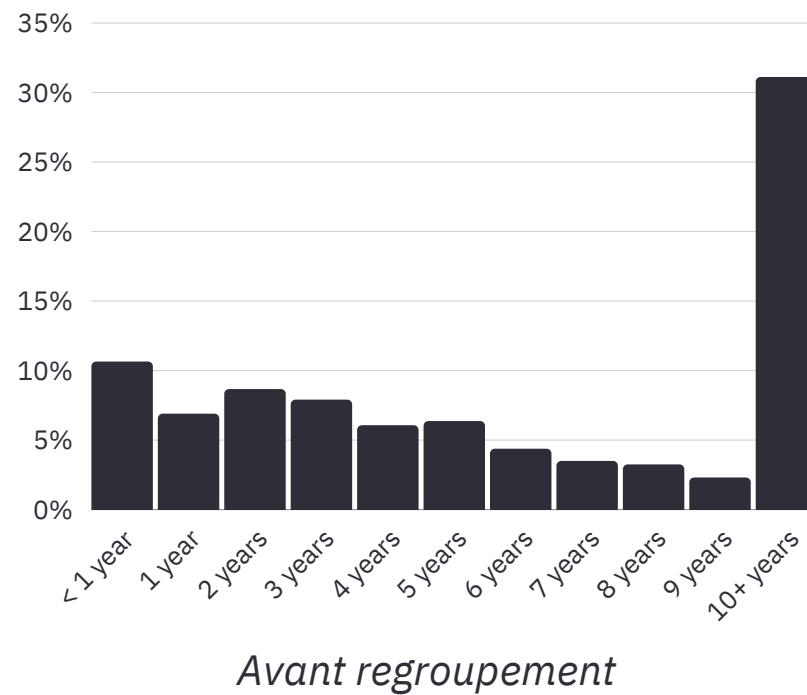
Modifications effectuées.

- Les variables avec plus de **85%** de la **même modalité** ont été transformées en **variable binaire** (avec 0 la modalité la plus représentée)
- Réduction du nombre de modalités afin d'alléger le nombre de groupes dans le modèle, donc la complexité :

Démarche

- Regroupement des modalités lorsque l'effectif est inférieur à 5%
- Le but est d'éviter les modalités sous représentées

Exemple : variable ‘emp_length’



Notre base finale, comportant les variables dummy, contient **40 variables**

Modèles nativement interprétables.

Régression logistique.

Transformation des variables continues en **catégorielles** :

- Facilite la gestion des **valeurs extrêmes**
- Exigences réglementaires vis à vis de **l'interprétabilité**

La fonction **ContinuousOptimalBinning** sur Python permet d'avoir une discréétisation optimale :

- Cette méthode utilise un arbre de décision **CART** pour générer des prébins
- L'algorithme CART divise la variable continue de manière itérative en choisissant le point de coupure qui **maximise la séparation entre les catégories de la variable cible**

Le tableau de droite permet d'obtenir la valeur du V de cramer entre certaines variables discréétisées et la cible.

V de Cramer	
Variable	Valeur
tot_hi_cred_lim	0.114
term	0.114
loan_amnt	0.112
bc_util	0.100
home_ownership	0.096
num_il_tl	0.091
verification_status	0.079
mo_sin_old_rev_tl_op	0.077

Régression logistique.

- L'ensemble de nos variables sont **significatives** après suppression de **pub_rec**, **il_util** et **mo_sin_old_il_acct**.
- Les variables discrétisées ont toujours **la plus petite** classe en référence, ce qui implique que la borne inférieure de la classe de référence commence à **0**.
- Les variables initialement catégorielles prennent, dans la mesure du possible, le **non-événement comme référence**.

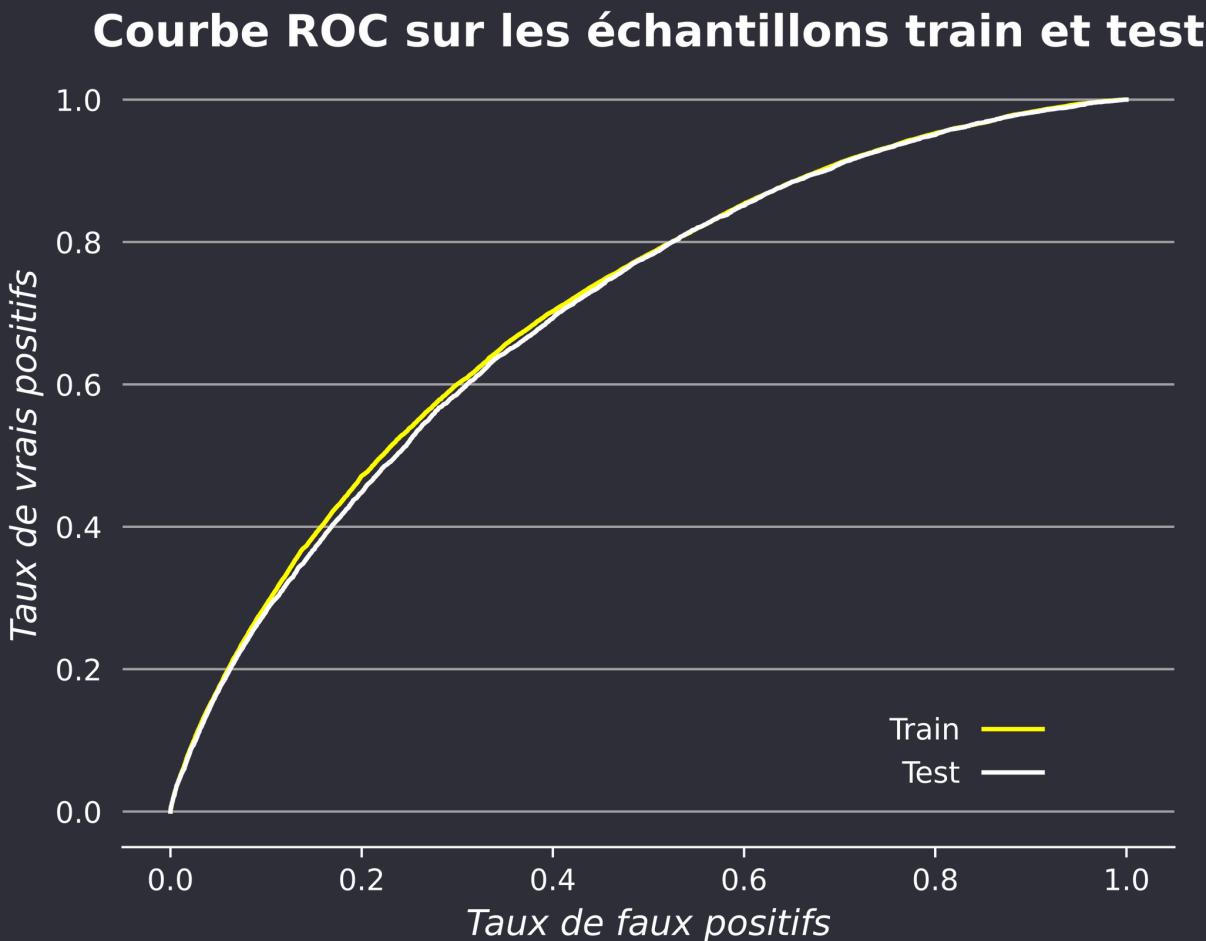
Ce tableau permet de s'assurer que nos variables sont **globalement significatives** :

Test de Wald	
Variable	P-value
tot_hi_cred_lim	<.0001
term	<.0001
loan_amnt	<.0001
bc_util	<.0001
home_ownership	<.0001
num_il_tl	<.0001
verification_status	<.0001
mo_sin_old_rev_tl_op	<.0001

Extrait des coefficients estimés :

Estimation via MLE		
Variable	Classe	Valeur
loan_amnt	[2937.50 ; 4562.50[0,3752
loan_amnt	[4562.50 ; 7462.50[0,6327
loan_amnt	[7462.50 ; 9612.50[0,8836
loan_amnt	[9612.50 ; 14512.50[0,9754
loan_amnt	[14512.50 ; 19387.50[1,0885
loan_amnt	[19387.50 ; 23712.50[1,3109
loan_amnt	[23712.50 ; inf[1,5525
delinq_2yrs	Yes	0,2

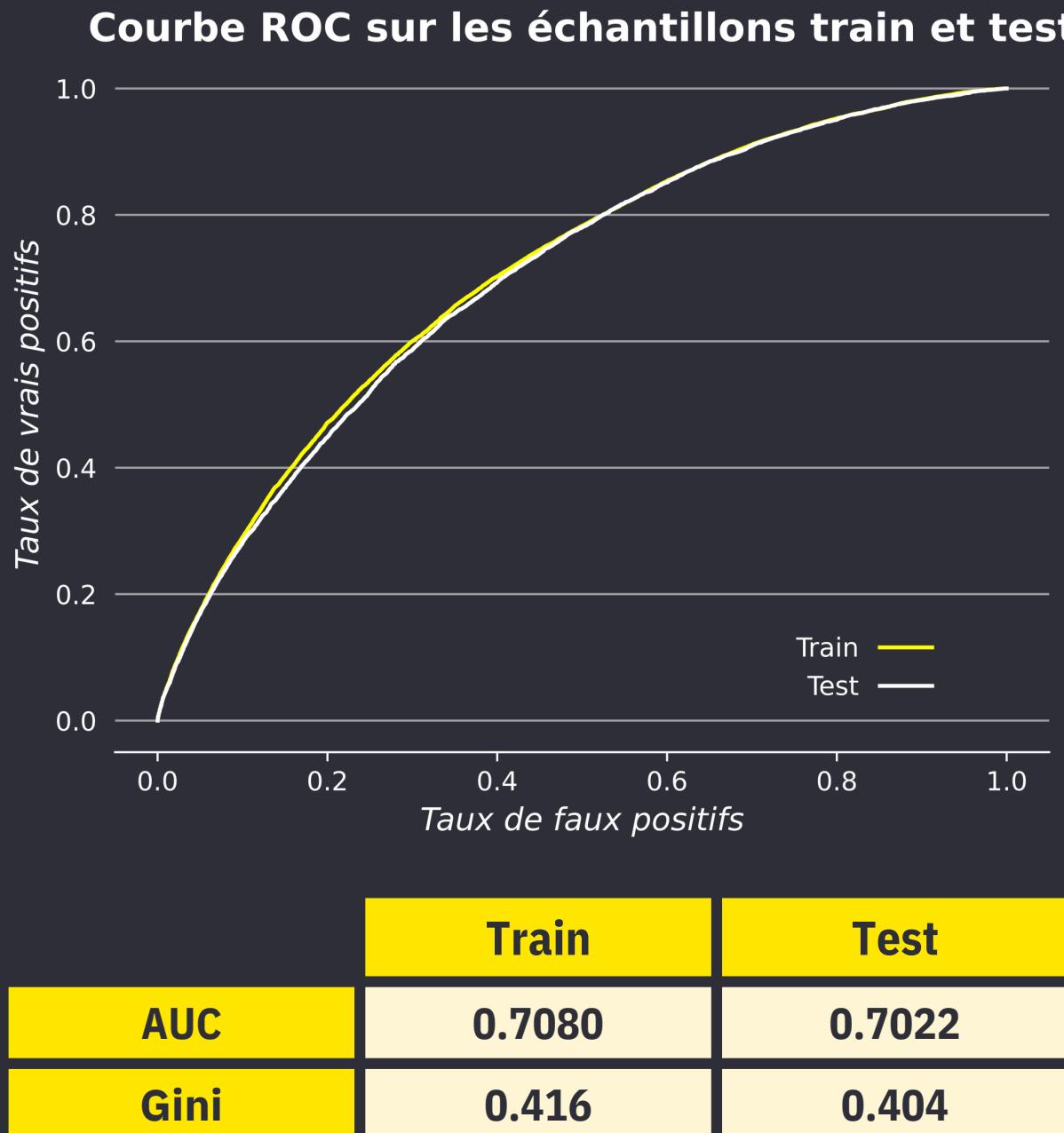
Régression logistique.



- La régression logistique produit des résultats statistiquement significatifs, avec des performances similaires sur les échantillons de test et d'apprentissage, indiquant **l'absence de surajustement**.

	Train	Test
AUC	0.7080	0.7022
Gini	0.416	0.404

Régression logistique.



- La régression logistique produit des résultats statistiquement significatifs, avec des performances similaires sur les échantillons de test et d'apprentissage, indiquant **l'absence de surajustement**.

Matrice de confusion au cut-off égalisant le taux de vrais positifs et de vrais négatifs
prédictions

		0	1
réalité	0	24767 (64,74%)	13492
	1	3394	6231 (64,74%)

→ **accuracy = 0,6474**

Régression logistique.

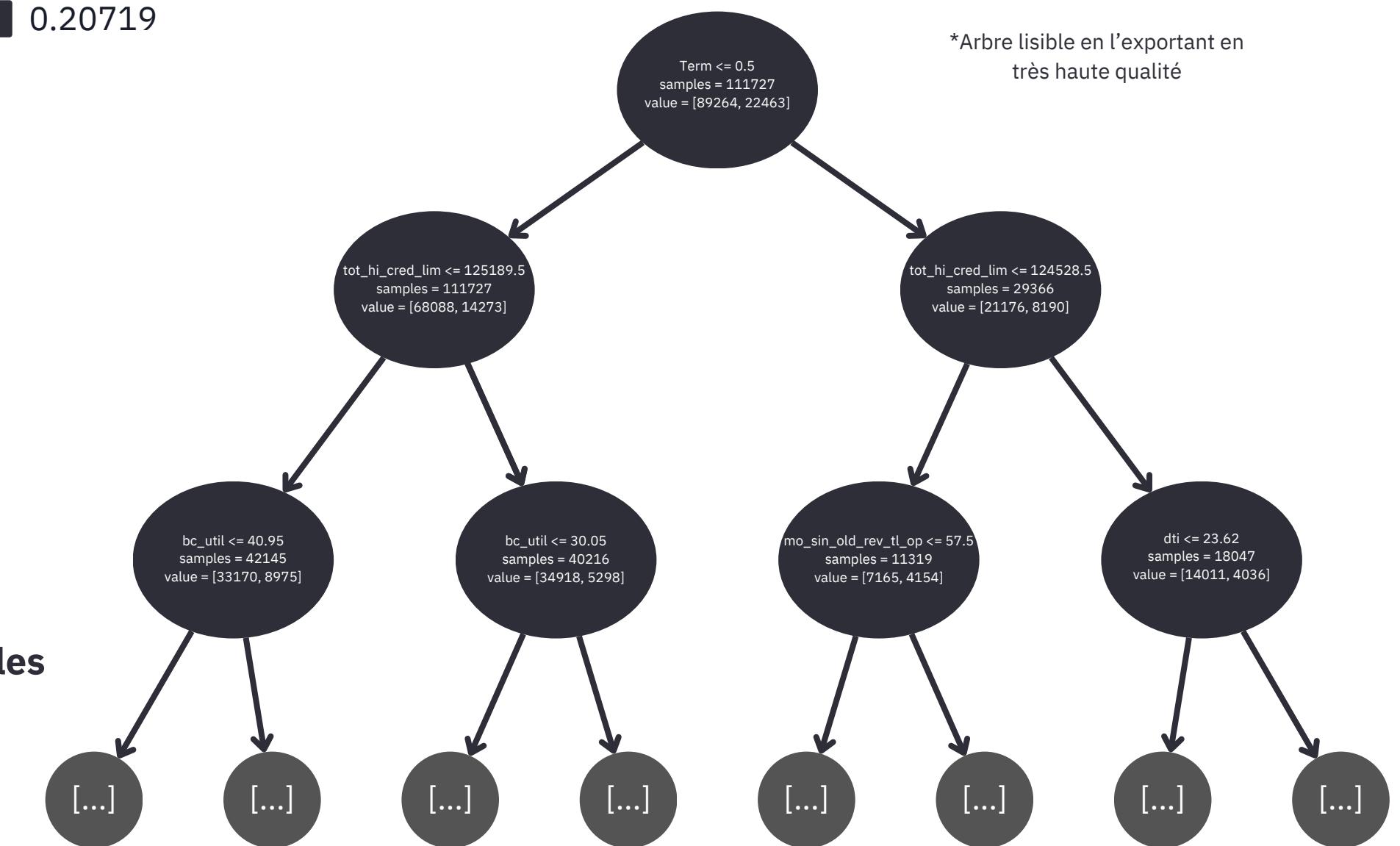
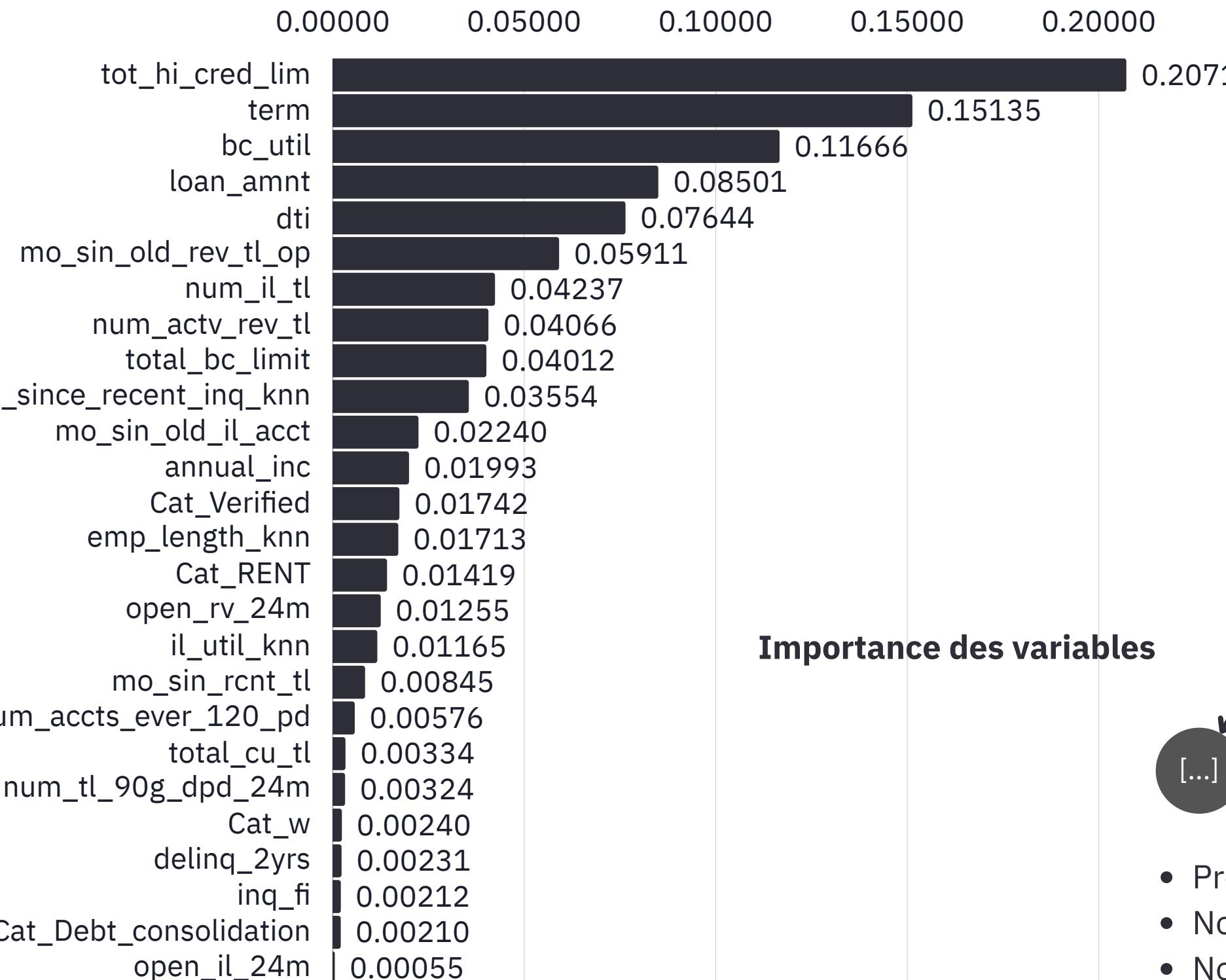
- La régression logistique est **nativement interprétable** :
cotes, rapport de cote ou élasticité

Estimation du rapport de cotes		
Variable	Duel	Valeur
loan_amnt	[23712.50 ; inf[vs [0 ; 2937.50[4,723
home_ownership	OWN vs RENT	0,897
tot_hi_cred_lim	[555197.50 ; inf[vs [0 ; 25094.00[0,394
title	Autres vs Frais_de_biens_ou_materiels	1,128
application_type	Individual vs Joint_App	1,208

Toutes choses égales par ailleurs :

- Montant du prêt (**loan_amnt**) : un client ayant un prêt équivalent à 23 712.50 \$ et plus, a **4,723 fois plus de chance de faire défaut** qu'un client avec un prêt inférieur à 2 937.50 \$.
- Type d'habitation (**home_ownership**) : un client étant propriétaire a **10,3% de chance en moins de faire défaut** qu'un client locataire.

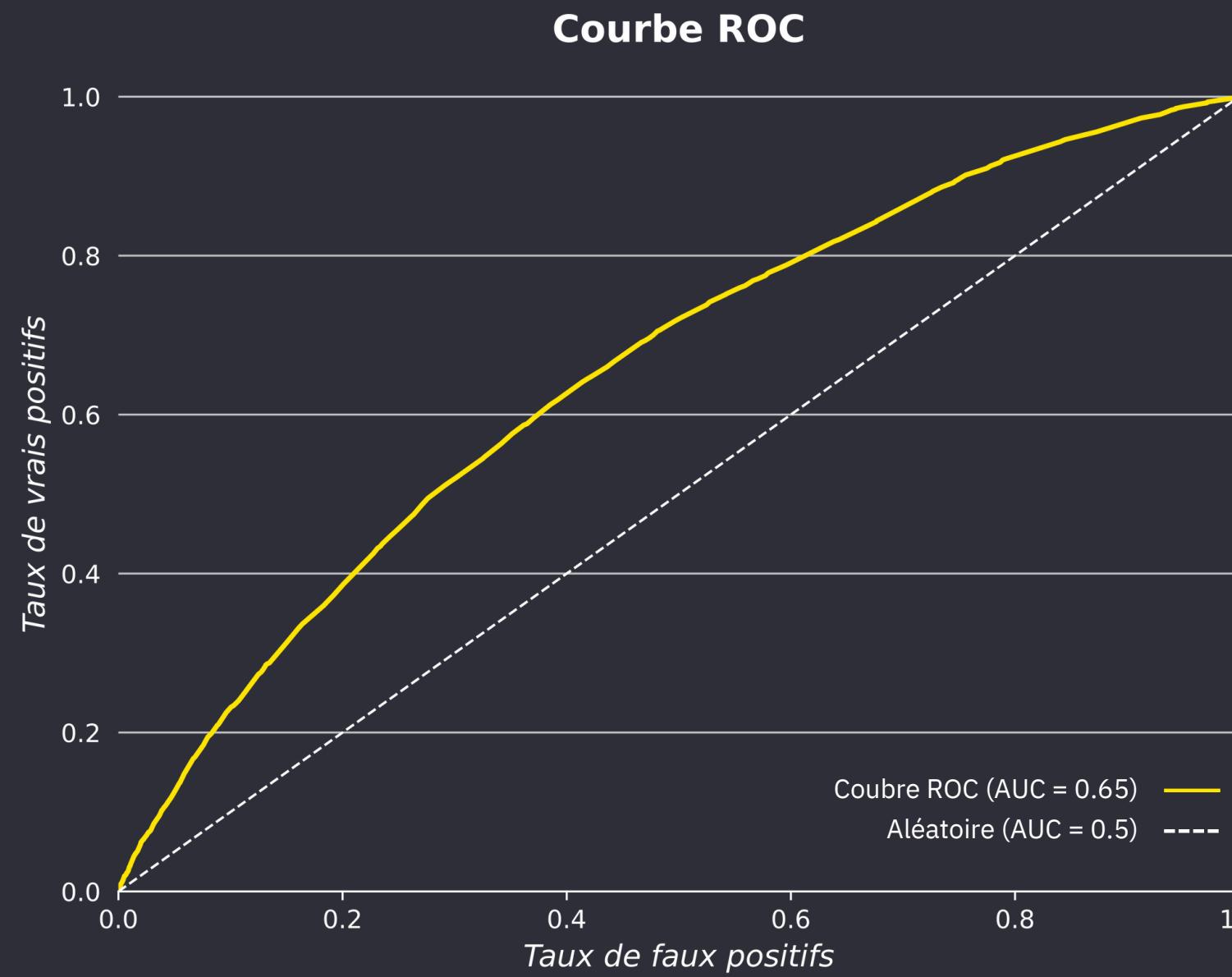
Arbre de classification.



- Profondeur de l'arbre : **8**
- Nombre minimal d'échantillons requis pour diviser un noeud : **5**
- Nombre minimal d'échantillons requis pour une feuille : **7**

*Arbre lisible en l'exportant en très haute qualité

Arbre de classification.



Matrice de confusion au cut-off égalisant le taux
de vrais positifs et de vrais négatifs

réalité

prédiction

		0	1
0	23500 (61,41%)	14764	
1	3712	5908 (61,41%)	

accuracy = 0,6141

Arbre de classification.

Chemin classant le plus de **0** (non défaut),
autrement dit le type de client **le moins risqué** :



Arbre de classification.

Chemin classant le plus de **0** (non défaut),
autrement dit le type de client **le moins risqué** :

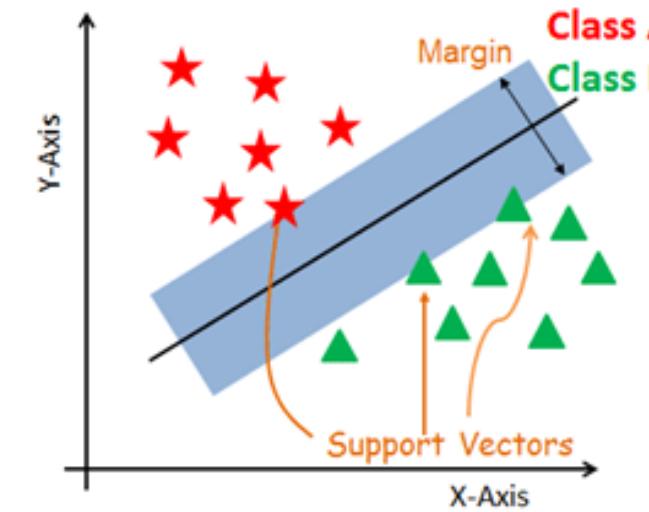
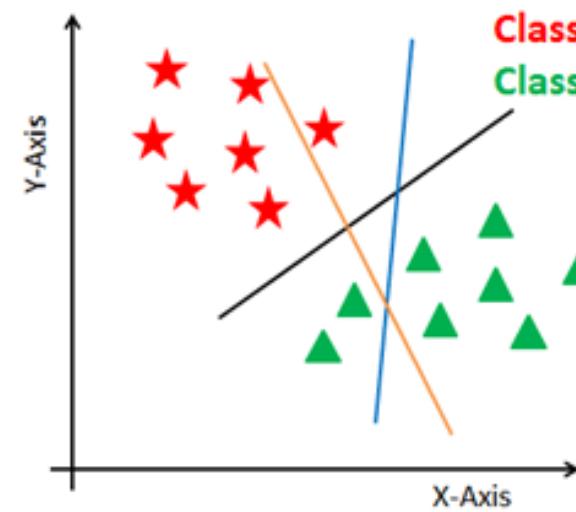


Chemin classant le plus de **1** (défaut),
autrement dit le type de client **le plus risqué** :



Modèles black box.

Support Vector Machine (SVM).

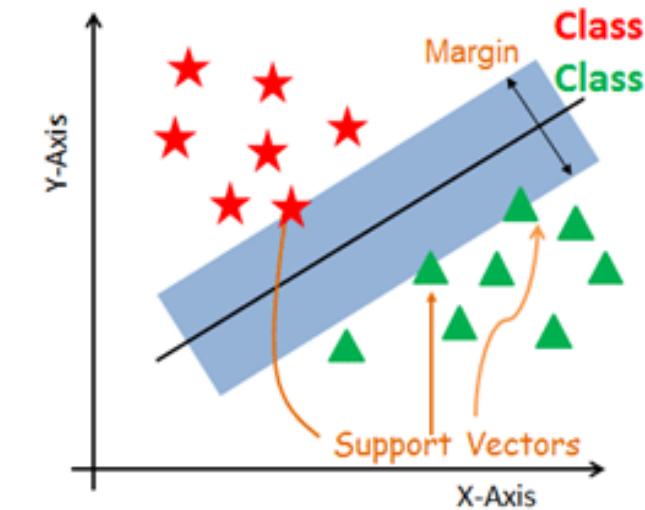
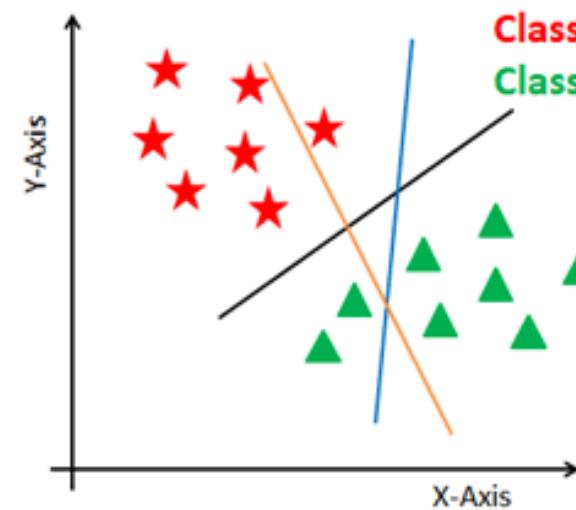


Objectif : séparer l'ensemble de données en minimisant l'erreur de classification

- Marge : distance entre les deux points de chaque classe les plus proches
- Vecteurs de support : les points qui maximise la marge

↪ L'objectif est de sélectionner un hyperplan avec la marge maximale possible

Support Vector Machine (SVM).



Objectif : séparer l'ensemble de données en minimisant l'erreur de classification

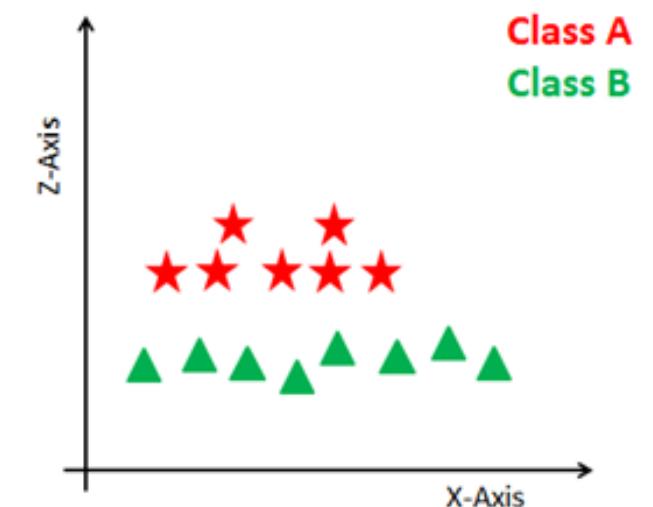
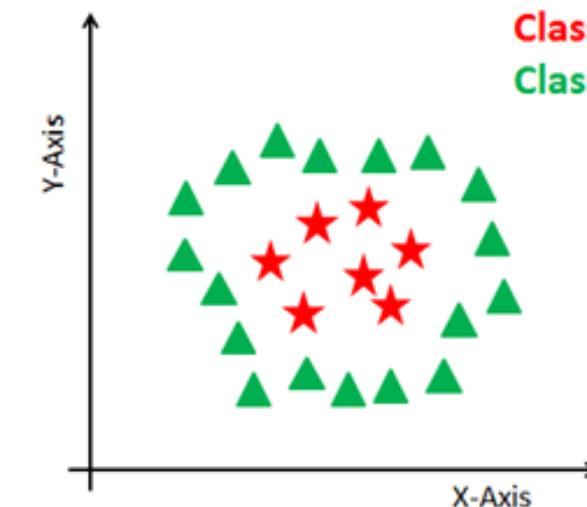
- Marge : distance entre les deux points de chaque classe les plus proches
- Vecteurs de support : les points qui maximise la marge

↪ L'objectif est de sélectionner un hyperplan avec la marge maximale possible

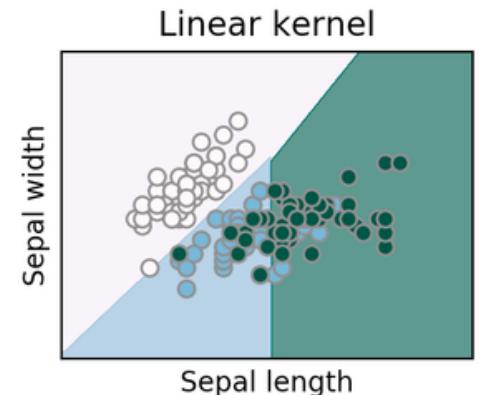
Certains problèmes **ne peuvent pas** être résolus à l'aide d'un hyperplan linéaire (figure de gauche)

Le SVM utilise un noyau (kernel trick) pour **transformer** l'espace d'entrée en un espace de **dimension supérieure requise** (figure de droite)

↪ Un problème **non séparable** devient un problème **séparable**

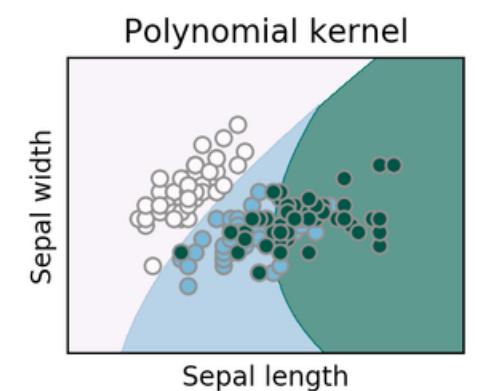


Support Vector Machine (SVM).



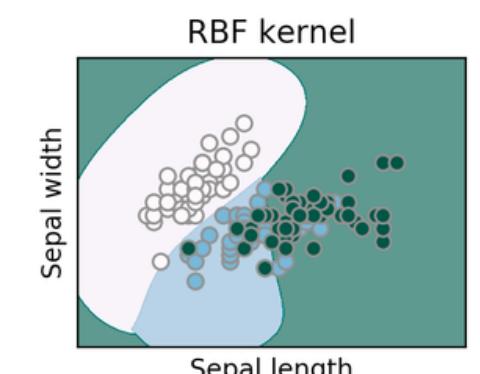
Linear Kernel :

Il est utilisé pour des problèmes où les données sont linéairement séparables.



Polynomial Kernel :

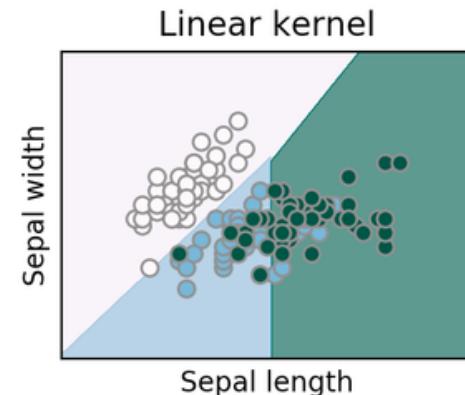
Il introduit des termes polynomiaux permettant de distinguer des espaces d'entrée courbés ou non linéaires.



Radial Basis Function Kernel :

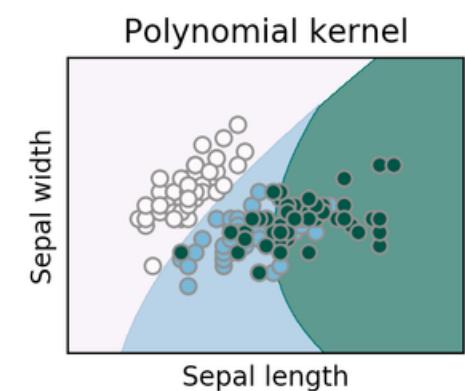
Il transforme les données dans un espace de dimension infinie.

Support Vector Machine (SVM).



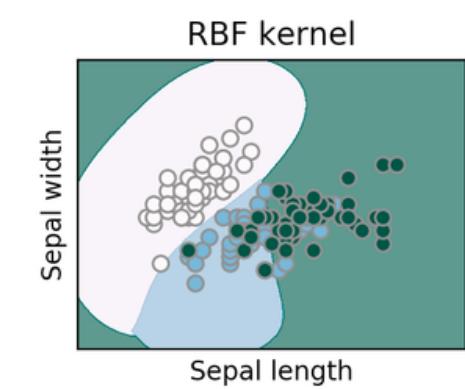
Linear Kernel :

Il est utilisé pour des problèmes où les données sont linéairement séparables.



Polynomial Kernel :

Il introduit des termes polynomiaux permettant de distinguer des espaces d'entrée courbés ou non linéaires.



Radial Basis Function Kernel :

Il transforme les données dans un espace de dimension infinie.

Sélection des **hyperparamètres** via validation croisée pour :

- **C** : paramètre de pénalisation. Il contrôle le compromis entre la classification correcte des exemples d'entraînement et la création d'une frontière de décision lisse.
- **Gamma** : entre 0 et 1, règle la portée d'influence de chaque observations d'entraînement. Une valeur élevée de gamma peut entraîner un surajustement en ajustant trop précisément l'ensemble d'apprentissage.
- **Degree** : il représente le degré du polynôme.
- **Coef0** : utilisé pour influencer le terme indépendant dans les noyaux polynomiaux.

Linear : C

Rbf : C, Gamma

Polynomial : C, Gamma, Degree, Coef0

Support Vector Machine (SVM).

1/ Performance vs. Ressources ?

- Les modèles plus complexes et gourmands en ressources peuvent conduire à une meilleure précision de prédiction.
- Mais entraîne une augmentation significative des besoins en termes de puissance de calcul, de mémoire, etc.

2/ Performance vs. Coût ?

- Il est important de considérer le coût des ressources par rapport aux gains potentiels de performance.

Support Vector Machine (SVM).

1/ Performance vs. Ressources ?

- Les modèles plus complexes et gourmands en ressources peuvent conduire à une meilleure précision de prédiction.
- Mais entraîne une augmentation significative des besoins en termes de puissance de calcul, de mémoire, etc.

2/ Performance vs. Coût ?

- Il est important de considérer le coût des ressources par rapport aux gains potentiels de performance.

Utilisation de Intel® Extension for Scikit-learn

Accélération des applications Scikit-learn tout en conservant une conformité totale avec les API et algorithmes Scikit-Learn. Intel® Extension for Scikit-learn apporte une **accélération jusqu'à 600 fois** dans une variété d'applications.

Comment l'appliquer sur Python ?

```
from sklearnex import patch_sklearn  
patch_sklearn("SVC")
```

Support Vector Machine (SVM).

Normalisation des variables quantitatives :

- Permet de réduire la complexité des modèles SVM est sensible à l'échelle des caractéristiques

Matrice de confusion au cut-off égalisant le taux de vrais positifs et de vrais négatifs

Meilleurs performances de la validation croisée

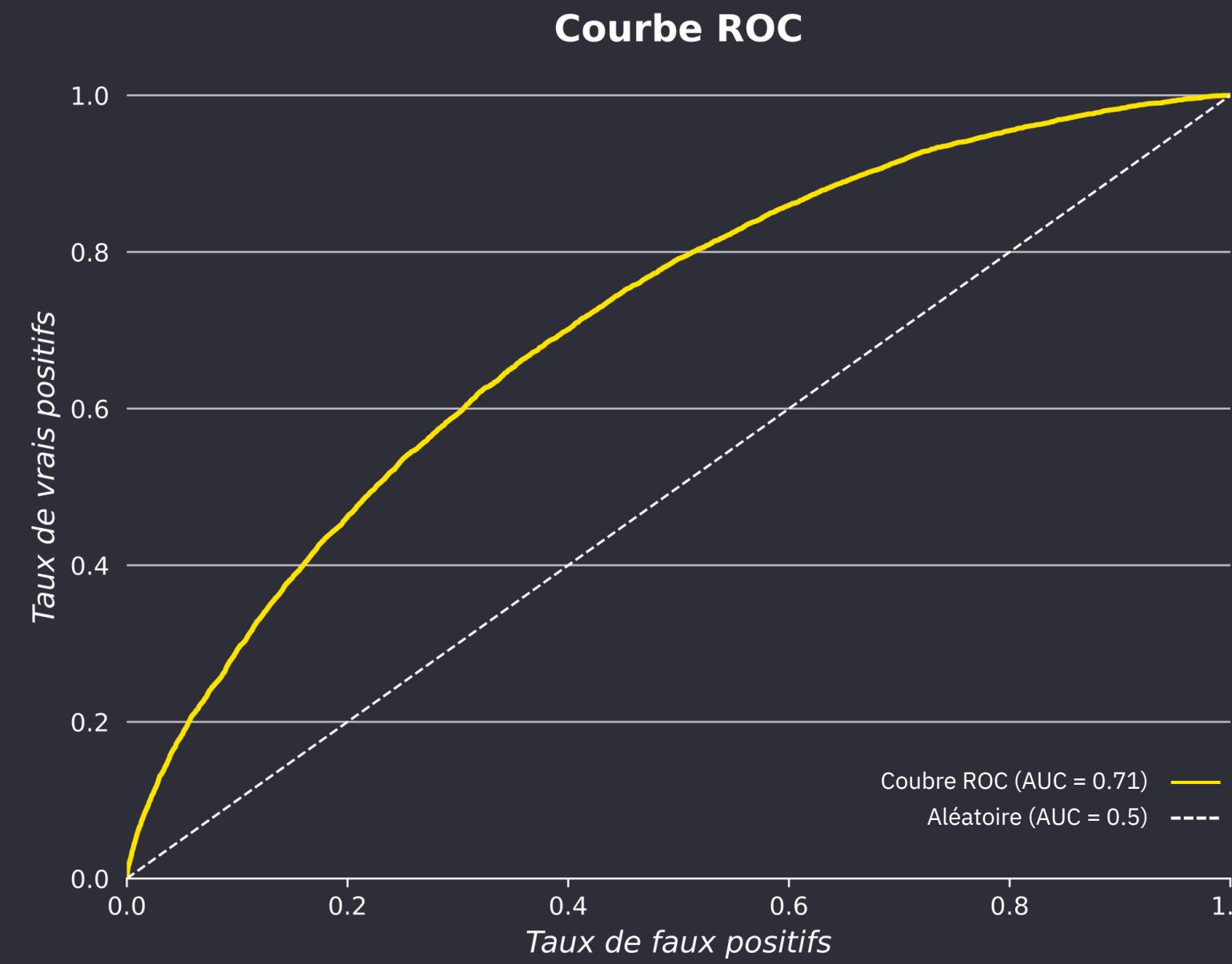
Kernel	C	Gamma	Degree	Coef0	AUC
Linear	12	.	.	.	0,578
RBF	10	0,1	.	.	0,6102
Polynomial	10	0,1	3	0	0,624

réalité

		prédictions	
		0	1
0	0	22 474 (58.74%)	15 687
	1	3961	5630 (58,49%)

accuracy = 0,5869

XGBoost.



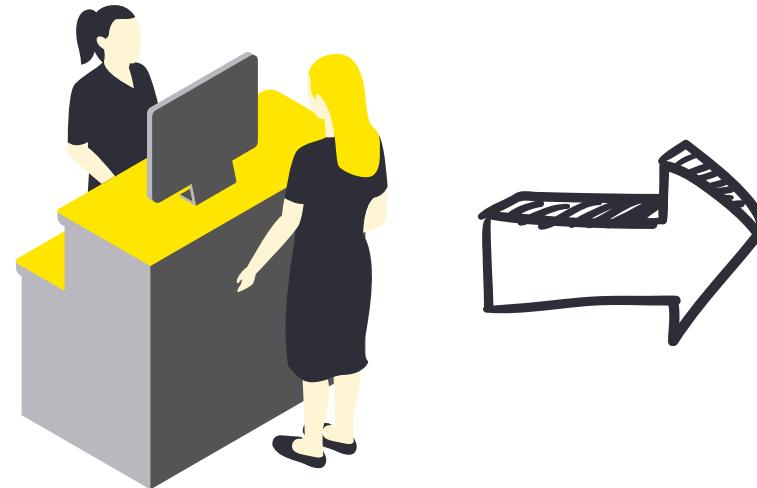
- Nombre d'arbres : **140**
- Profondeur : **4**
- Taux d'apprentissage : **0.1**

prédiction

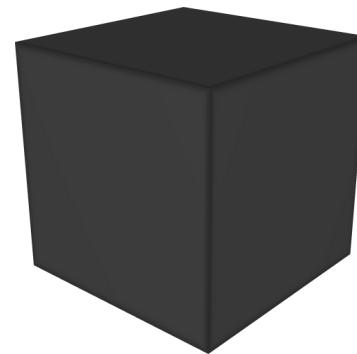
		0	1
réalité	0	24906 (65,13%)	13334
	1	3363	6280 (65,12%)

accuracy = **0,6513**

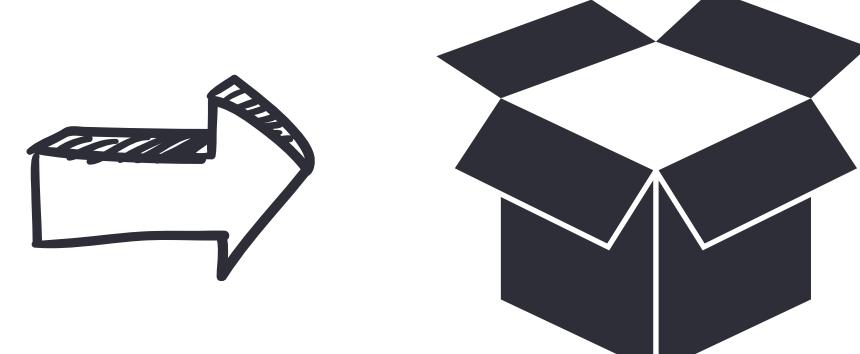
Machine Learning Interprétable.



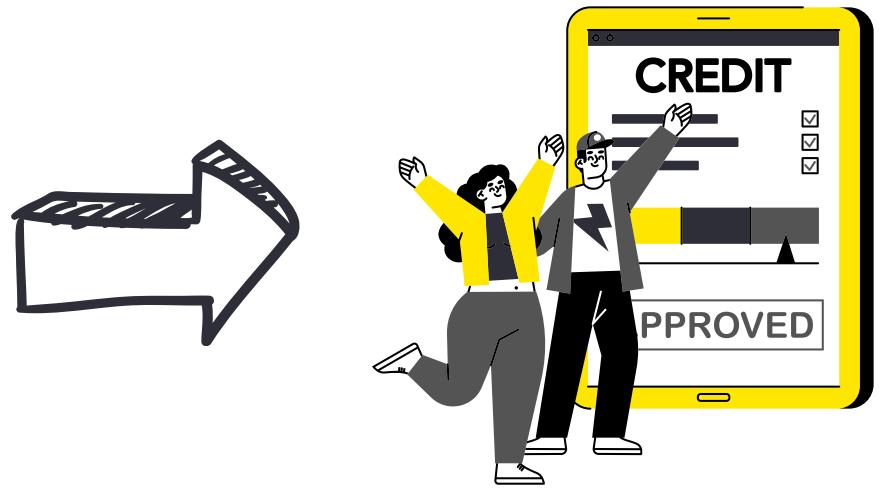
**Demande de
crédit**



**Machine
Learning**



**Méthodes
d'interprétations**



**Décision
explicable**

Machine Learning Interprétable.

Motivations :

- Nécessaire pour l'explicabilité de la banque afin d'assurer une acceptabilité des ces méthodes avancées auprès des régulateurs et d'éviter des **biais discriminatoires**.
- Les modèles de Machine Learning peuvent servir de **benchmark** afin de comparer leurs prédictions à celles des modèles interprétables.

Méthodes :

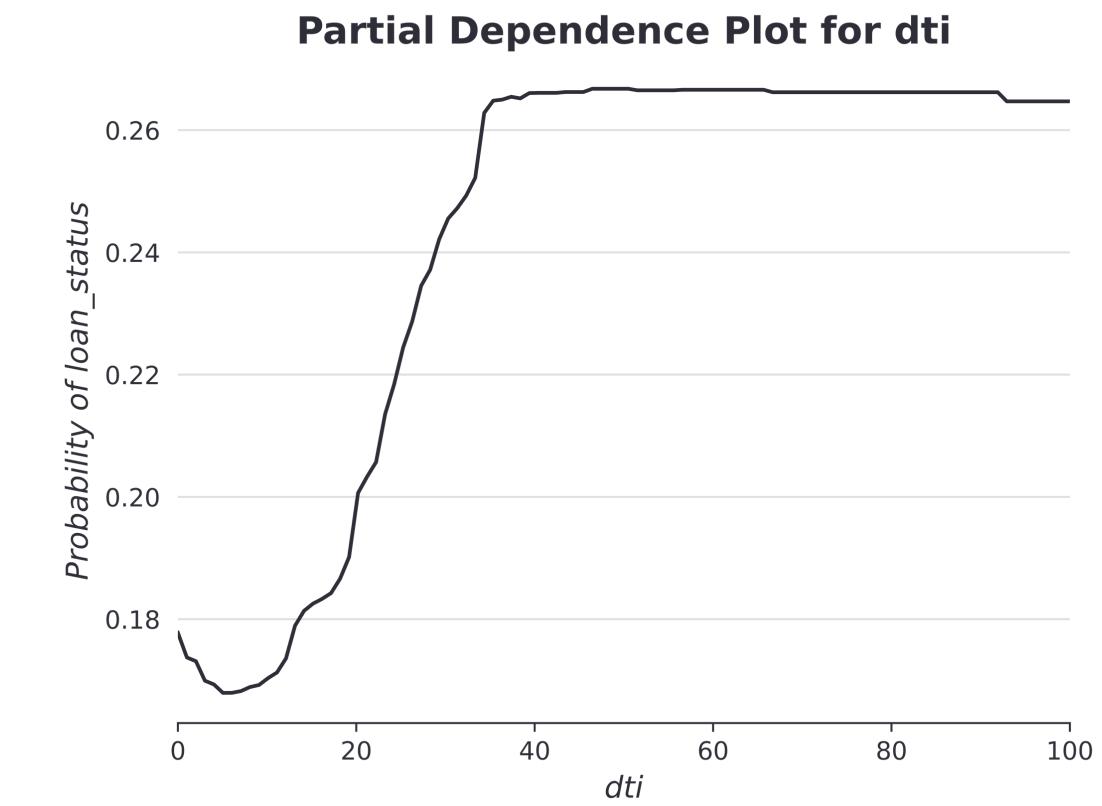
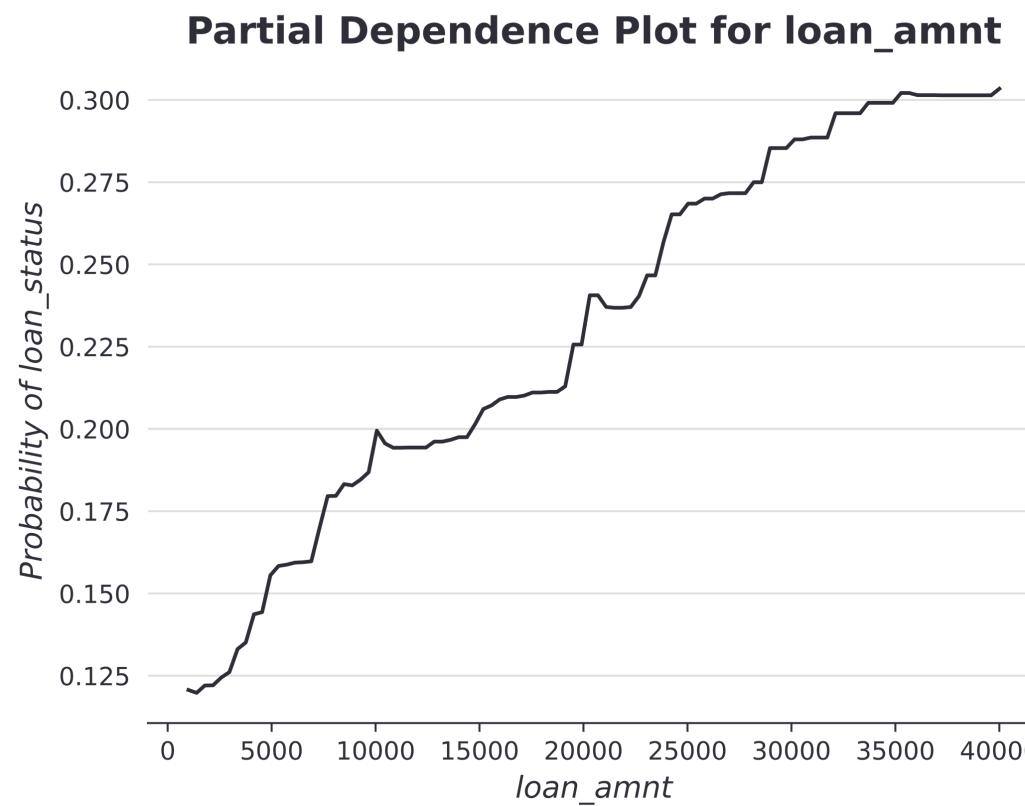
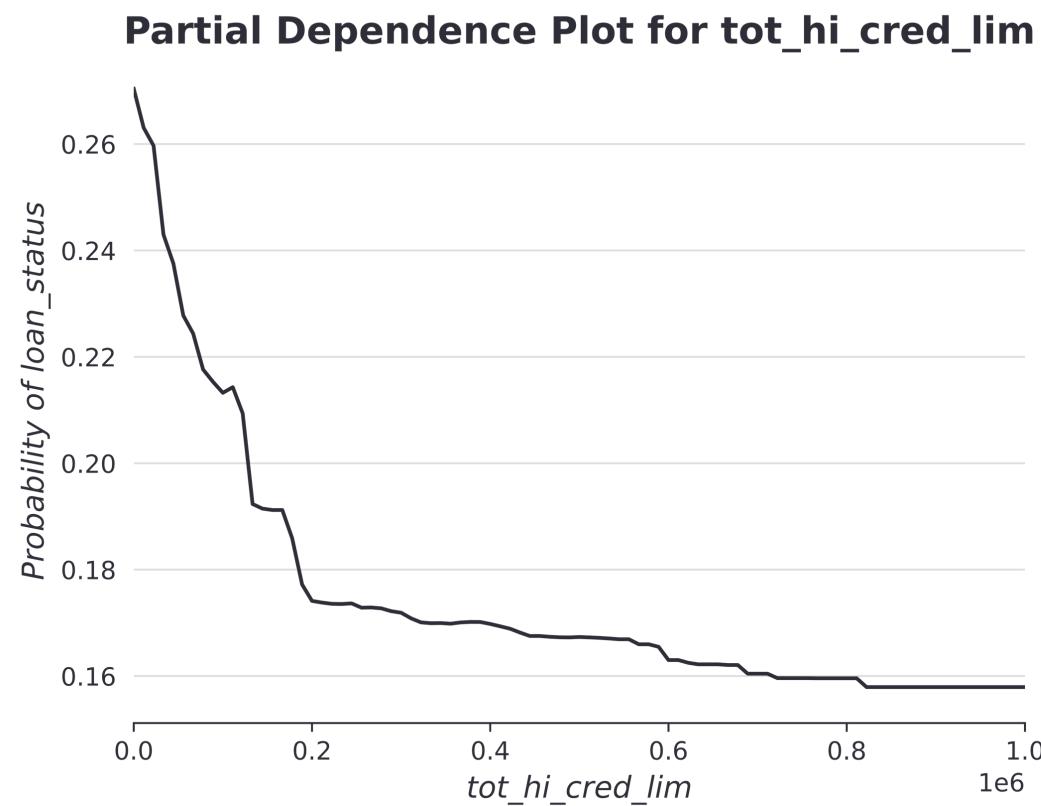
- Les méthodes **globales** décrivent le comportement moyen d'un modèle d'apprentissage automatique en utilisant des valeurs attendues basées sur la distribution des données.
- Contrairement aux méthodes globales, les méthodes **locales** examinent les détails individuels et fournissent des explications spécifiques à une observation donnée.

Machine Learning Interprétable.

Les Partial Dependence Plot montrent comment la variation d'une caractéristique particulière affecte la prédiction du modèle, tout en maintenant les autres caractéristiques constantes.

- **Avantage** : faciles d'interprétation
- **Inconvénient** : sensibles aux corrélations entre les variables (cependant, ici, elles ont été préalablement traitées)

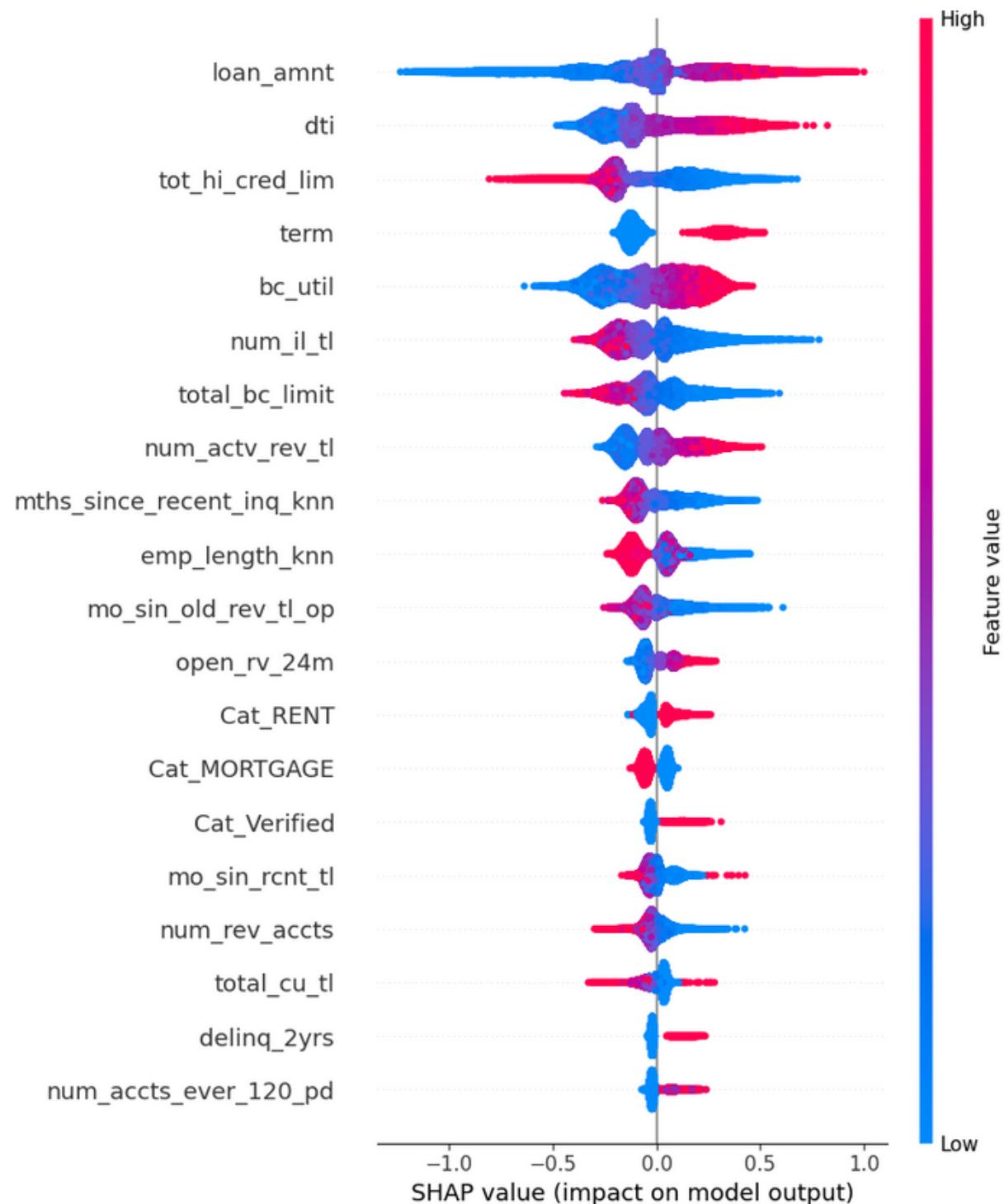
3 exemples de PDP via XGBoost :



Machine Learning Interprétable.

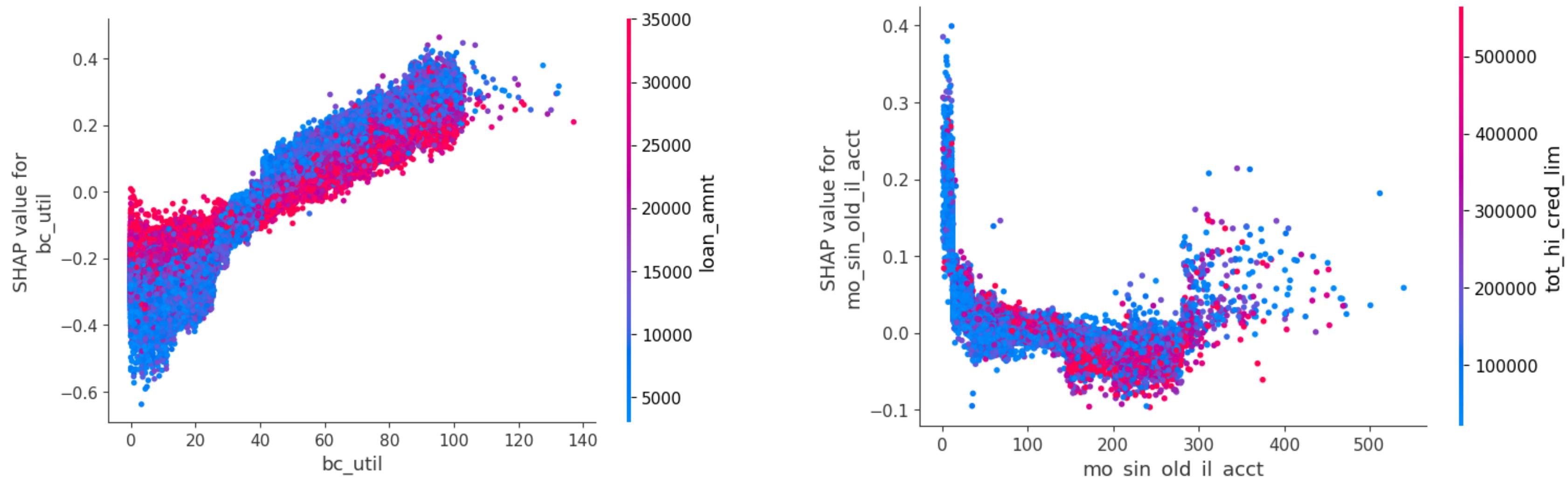
Shapley additive explanations :

- En apprentissage automatique, elle représentent la **contribution individuelle** de chaque caractéristique à la prédiction d'un modèle. Elles permettent de comprendre comment chaque élément influence la sortie du modèle.
- Le **SHAP Summary Plot** résume l'importance des caractéristiques et leurs contributions aux prédictions du modèle. Chaque point représente la contribution d'une caractéristique pour une instance, **ordonnée par importance**, fournissant une vue d'ensemble avant d'examiner les détails avec les graphiques de dépendance SHAP.



Machine Learning Interprétable.

Les **SHAP dependence plots** en interaction sont des graphiques qui aident à voir comment **l'influence d'une caractéristique sur les prédictions** d'un modèle peut changer en fonction des valeurs d'une **autre caractéristique**. Ils sont utiles pour comprendre **comment différentes variables interagissent dans un modèle prédictif**.



Conclusion

- 1 Le **XGBoost**, en tant que méthode black box, surpassé le SVM en performances. De même, la **régression logistique** se révèle plus efficace que l'arbre de classification parmi les méthodes interprétables.
- 2 Dans l'ensemble, le XGBoost affiche des performances **légèrement supérieures** à la régression logistique.
- 3 L'ensemble des modèles présentés permettent d'avoir les **mêmes explications** vis à vis de l'impact des différentes variables sur la probabilité de défaut.
- 4 Les méthodes de Machine Learning Interprétables éclairent les prédictions des modèles par rapport à nos données **sans nécessairement refléter la réalité**.
- 5 Il est essentiel de considérer le **coût** associé à l'**utilisation** de méthodes black box. Les **gains marginaux** en performances doivent être **évalués** en fonction de ce coût afin de déterminer la **rentabilité** de ces méthodes pour les institutions financières.

Opinion | IA, surveillance des marchés et credit scoring



L'intelligence artificielle est utile en matière financière, mais elle a ses limites, expliquent Bertrand Hassani, Iris Lucas et Fabrice Riva. Car elle peut, par exemple, contribuer à perpétuer des inégalités dans l'octroi de crédit liées au genre ou à l'origine ethnique.

Les Echos



BARRAUD
Lorenzo

barraudlorenzopro@gmail.com



MIRZA
Simon

simon.mrza@gmail.com



VIEIRA DE BARROS
Mathias

mathias.vieiradebarros@gmail.com

