nexialog CONSULTING

Scoring d'octroi par Machine Learning interprétable ?

Nexialog Consulting, Paris, France

20 février 2023

Résumé

Le scoring est un outil d'aide à la décision à destination des banques permettant d'anticiper la capacité de remboursement d'un emprunteur. Dans cette note, deux méthodologies de classement des prêts selon leur probabilité de défaut sont comparées sur une base de données emprunteur. Premièrement, l'approche classique de la construction d'une grille de score - basée sur un modèle de regression logistique - permet une identification et une interprétabilité claire des variables contributrices au risque. Secondement, les méthodes ensemblistes - Random Forest et XgBoost - offrent de meilleures performances prédictives sur nos données mais leur complexité peut limiter leur portée opérationnelle. En effet, contrairement à l'approche classique de scoring, l'identification de la contribution de chaque variable au risque de défaut et la décision d'octroi nécessitent le recours au modèle prédictif et à des techniques d'interprétabilité qui peuvent être complexes à mettre en œuvre.

En tant qu'activité déterminante des banques commerciales, l'octroi de crédit nécessite une anticipation des risques, pérennisant ainsi la solidité du système financier. Ce risque de crédit représente la perte potentielle qu'une banque s'attend à recevoir lorsqu'un emprunteur est insolvable, c'est-à-dire, qu'il est en incapacité de rembourser son prêt selon les conditions contractuelles. Afin de limiter le risque, les modalités d'emprunts et les informations personnelles de l'emprunteur sont étudiées avec soin afin d'évaluer la faisabilité de l'octroi.

A partir des caractéristiques du client et du prêt, le *credit scoring* est ensemble de méthodes statistiques visant à ordonner le risque emprunteur selon différentes catégories ou facteurs de risque. Grâce à cet outil d'aide à la décision, qui se doit d'être précis, les banques peuvent anticiper la capacité de remboursement d'un emprunteur. La décision finale de l'octroi bancaire ne dépend pas uniquement du score attribué grâce à la modélisation statistique mais plus globalement d'un ensemble de systèmes experts mis en place par la banque. Ainsi, d'importants efforts sont réalisés depuis ces dernières années pour améliorer la quantité et de la qualité de la donnée, et ainsi pérenniser le recours à la modélisation.

La base de données *Open Source* traitée dans cette note provient du site *Kaggle* et s'intule *Credit Risk Dataset*. Elle concerne des données issues d'une banque américaine octroyant des crédits à faible montant à ses clients. Il s'agit de prêts liés à un besoin de financement médical, professionnel, personnel ou éducatif ou faisant l'objet d'une consolidation de plusieurs emprunts. Leurs montants n'excèdent que rarement les quelques dizaines de milliers d'euros. La base de données est constituée de 32 581 observations et de 12 variables et contient des informations propres aux contrats comme le montant emprunté, le taux d'intérêt associé, l'objet du prêt, une note donnant la qualité du prêt en fonction de l'historique bancaire du client, la durée écoulée depuis le début de l'emprunt. Également, elle donne des informations propres aux clients comme son âge, son revenu, sa durée d'expérience professionnelle ou sa condition d'hébergement actuelle (locataire, propriétaire, ...). La variable à modéliser est le défaut prenant la modalité 1 si le client est en défaut et 0 sinon. Notons que la granularité de la base de données se positionne au niveau contrat. Ainsi, un même client peut apparaître sur plusieurs lignes pour différents contrats.

Deux méthodologies de classement des prêts selon leur probabilité de défaut sont proposées dans cette note (dans d'autres cas, ce sont les clients qui sont scorés, ce détail est dépendant de la granularité des données). Premièrement, une approche classique de construction de la grille de score est proposée. Son interprétabilité permet une identification claire des variables contributrices au risque, appelées également *Risk Drivers*. Secondement, des méthodes ensemblistes avec de meilleures performances prédictives seront employées. Pour ces modèles, dits "black-box", la contribution de chaque variable à la probabilité de défaut nécessite des techniques d'interprétabilité plus complexes à mettre en œuvre. La valeur de Shapley sera abordée afin de fournir une interprétabilité à ces méthodes.

La Section 1 aborde les premiers traitements de la base de données : *label encoding*, analyse des données manquantes et extrêmes, création de variables et discrétisation. Dans les Section 2 et 3, l'approche classique est présentée avec respectivement la procédure de modélisation "step by step" et les étapes de la construction de la grille de score. Cette méthode sera finalement challengée en Section 4 par des méthodes ensemblistes donnant de meilleurs résultats de prédiction et la valeur de Shapley qui permet l'identification des *Risk Drivers*.

1 Traitement de la base de données

Composée de 11 variables explicatives, la base de données étudiée a déjà fait l'objet d'un pré-traitement. En principe, une base de données composée d'une centaine de variables serait attendue. Une sélection de variables métiers ou statistiques serait nécessaire pour réduire sa dimensionnalité.

1.1 *Label Encoding*

L'approche du *Label Encoding* consiste à transformer une variable catégorielle en une variable discrète quantitative en fonction des valeurs prises par la variable expliquée. Chaque modalité est donc encodée conditionnellement au taux de risque qu'elle présente, permettant une relation ordinale entre la variable d'intérêt et la variable encodée.

Prenons l'exemple suivant :

X	Moy. par label	X enc.
A	0.45	1
B	0.60	2
C	0.15	0

TABLE 1 – Exemple explicatif du Label Encoding

La variable catégorielle notée X contient trois modalités : "A", "B" et "C". A chacune de ces modalités est associé un taux de risque moyen. Avec un taux de 60%, les emprunteurs ayant la caractéristique "B" présentent en moyenne le risque le plus élevé. Au sein de la variable transformée "X enc.", la valeur attribuée à ce sous-groupe d'emprunteurs est donc la plus forte. A contrario, avec un taux de 15%, les emprunteurs ayant la caractéristique "C" présentent en moyenne le risque le plus faible. La valeur attribuée à ce sous-groupe est donc de 0 *(Table 1)*.

Le choix de cette méthodologie est motivé par les avantages suivants :

- Premièrement, certains algorithmes de sélection de variables ou de modélisation ne fonctionnent que sur des variables quantitatives, comme ceux considérés dans cette note. Notons qu'en cas de grande dimensionnalité, lorsque l'algorithme de sélection de variables n'admet que des données encodées, cette étape doit être réalisée avant son exécution.

- Les modèles de prédiction apprennent plus aisément lorsqu'un ordre logique ressort des variables discrètes.

- Contrairement à l'approche du *"One Hot Encoding"* décrite en Section 3.1, son utilisation évite l'augmentation de la dimensionnalité [3].

1.2 Données manquantes et aberrantes

1.2.1 Données manquantes

En règle générale, la présence de valeurs manquantes dans un jeu de données est un problème assez fréquent, cela s'explique par diverses raisons comme des erreurs de saisie, un refus de partage d'informations ou un oubli de l'opérateur. Généralement, les données manquantes se caractérisent par une valeur nulle, par une valeur notée "NaN" ou encore par une valeur extrême préalablement définie.

Dans un premier temps, il s'agit d'identifier la présence de données manquantes, ainsi que leurs origines, afin de choisir la méthode d'imputation. L'imputation correspond au remplissage des valeurs manquantes à partir d'estimations. Soulignons que la proportionnalité des données manquantes au sein de la variable doit être prise en considération.

Dans notre cas d'étude, deux variables présentent des valeurs manquantes : la "durée d'expérience professionnelle" et le "taux d'intérêt" avec respectivement 2.7% et 9.6% de valeurs manquantes. En ce qui concerne leur traitement, plusieurs choix sont possibles :

- Les valeurs des variables "note du prêt en fonction de la qualité historique de l'emprunteur" et "taux d'intérêt", ayant une relation forte, une imputation par la note pourrait être utilisée. En effet, en fonction de la note échelonnée de "A" à "G", le taux d'intérêt observé est croissant.

- Étape primordiale pour la construction d'une grille de score, la discrétisation, présentée plus en détail *(en section 1.4)*, permet de regrouper les modalités de chaque variable par taux de défaut les plus proches. Cette méthode est privilégiée dans cette note, car contrairement à l'imputation, elle ne présente pas de biais d'estimation.

- D'autres méthodes d'imputation comme les KNN [1], la méthode MICE [4] ou encore les MissForest pourraient être employées [6]. Pour une synthèse des méthodes permettant de traiter les données manquantes, voir la note Nexialog [2].

1.2.2 Outliers

Les valeurs aberrantes sont des observations rares au sein d'une variable provenant d'erreurs de saisie ou d'effets exceptionnels. Elles doivent faire l'objet d'un traitement spécifique car elles faussent les liens observés dans les données, pouvant ainsi biaiser les sorties des modèles.

Le traitement des outliers est une étape importante et délicate car il s'agit de traiter une information qui fausse nos observations mais qui reste tout de même un renseignement. Dans cette étude, la discrétisation *(en Section 1.4)* permet de réduire l'effet de ces valeurs grâce au regroupement des valeurs prises par la variable. D'autres méthodes paramétriques (comme le Zscore [5]) ou non paramétriques (comme le DBSCAN [7]) peuvent être utilisées.

1.3 Création de variable

L'absence de repère temporel, tel que la date de début de contrat, empêche l'ajout de variables macroéconomiques. Ainsi, la conjoncture économique au moment de l'octroi du prêt est donc difficilement identifiable. En principe, en période de ralentissement économique, le défaut d'un client peut en partie être expliqué par un effet extérieur.

Toutefois, la variable du taux d'intérêt, renseignée dans plus de 90% des cas, permet d'intégrer cette dimension macro-économique. La méthodologie que nous avons retenue est la suivante :

- Grâce à des données *Open Source* de la FED *(Federal Reserve Economic Data)*, la première étape consiste à identifier les taux d'intérêt aux USA en période de ralentissement économique. Entre 1995 et 2021, deux périodes de récessions sont identifiées : durant la crise financière mondiale de 2007 à 2009 et la seconde de 2019 à fin 2021, période de la pandémie Covid-19 *(voir Figure 1)*.

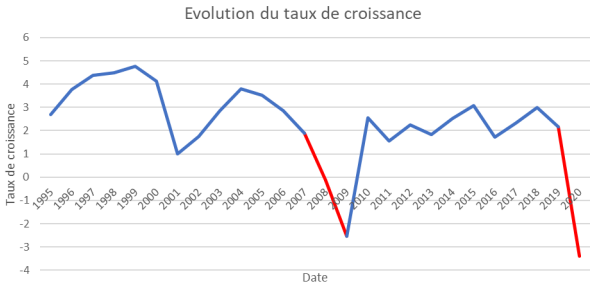

FIGURE 1 – Evolution du taux de croissance aux USA depuis 1995 avec mise en lumière des phases de ralentissement et de récession

- Ensuite, la moyenne des taux d'intérêt de consommation lors des périodes où le taux de croissance est négatif est calculée. Elle est de 13.7%.
- Une variable intermédiaire notée Z donne le rapport entre le taux d'intérêt du prêt (p_i pour le prêt i) et la moyenne des taux d'intérêt de consommation lors des périodes de crise :

$$z_i = \frac{p_i}{13.7} \tag{1}$$

- Finalement, la variable finale "downturn" est formée :

$$\text{downturn} = \begin{cases} 1 & \text{si } z_i >= 1\,; \\ 0 & \text{sinon.} \end{cases} \tag{2}$$

La variable vaut 1 lorsque le prêt est octroyé en période où les taux sont très forts, moment d'incertitude macro-économique.

D'autres types de croisement auraient pu être créés comme la variable LTI (Loan to Income) déjà présente ou la variable LTV (Loan to Value), mais nous ne disposons pas de l'information sur la valeur du bien financé.

1.4 Discrétisation

La discrétisation des variables continues est une méthode statistique visant à transformer une variable continue en une variable discrète et ordinale. Premièrement, la distribution en vingtile ou en décile d'une variable continue est calculée. Dans notre étude, la variable est divisée en 20 groupes composés chacun de 5% des observations. Ensuite, le défaut (rapport entre le nombre de prêts en défaut et le nombre de prêts total) par vingtile est calculé et tracé graphiquement. Afin de séparer la variable en plusieurs modalités ordinales, un ou plusieurs points de rupture sont identifiés visuellement.

Prenons l'exemple de la variable "revenu de l'emprunteur" que nous avons présenté en figure 2.

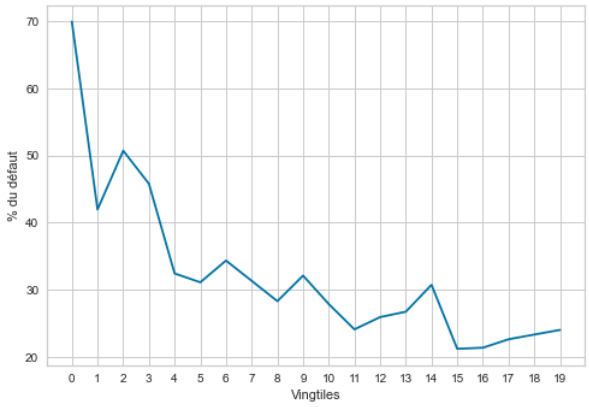

FIGURE 2 – Pourcentage du défaut par vingtile de la variable "revenu de l'emprunteur"

Trois sous-groupes sont identifiés avec un premier point de rupture au niveau du quatrième vingtile et un second au niveau du quinzième vingtile :

- Les vingtiles situés avant le premier point de rupture présentent le taux de défaut le plus élevé et porteront ainsi la modalité "2".
- Les vingtiles situés entre le premier et le second point de rupture présentent un taux de défaut moyennement élevé et porteront la modalité "1".
- Les vingtiles situés après le second point de rupture présentent un taux de défaut plus faible et porteront la modalité "0".

Lorsque l'on discrétise une variable, plusieurs conditions sont à respecter :

- Afin de garder une certaine aisance de lecture, le nombre de modalités choisi ne doit pas excéder 4 ou 5 (sauf exception). A titre d'exemple, la variable "durée d'expérience professionnelle" discréditée présente deux modalités *(voir Figure 3 et 4)*.

- Les modalités créées doivent contenir minimum 5% des observations et leurs répartitions doivent être stables dans le temps. Pour la variable "durée d'expérience professionnelle", la répartition de la modalité 0 varie entre 70% et 80% et celle de la modalité 1 entre 10% et 30% *(voir Figure 3)*.

- Les modalités doivent être discriminantes avec un taux de défaut stable au cours du temps. Pour la va-

riable “durée d’expérience professionnelle”, au sein de la modalité 0, le % d’observation en défaut est d’en moyenne 31% sur la période et de 38% pour la modalité 1. Par extension, les croisements du % de défaut d’une modalité par rapport à une autre doivent être évités, même si un léger croisement dit conjoncturel peut être accepté lorsqu’il ne perdure pas dans le temps *(voir Figure 4)*.

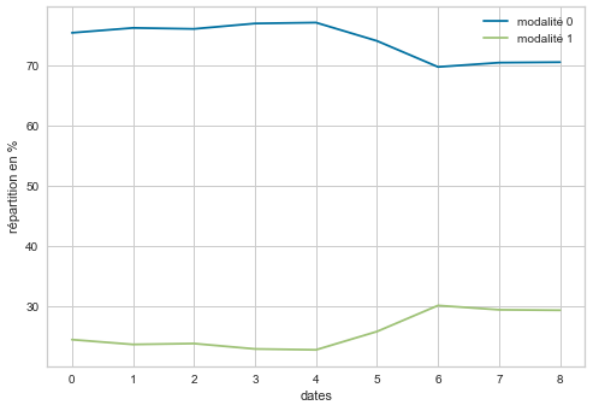

FIGURE 3 – Evolution de la répartition (en %) des observations par modalité de la variable ”durée d’expérience professionnelle”

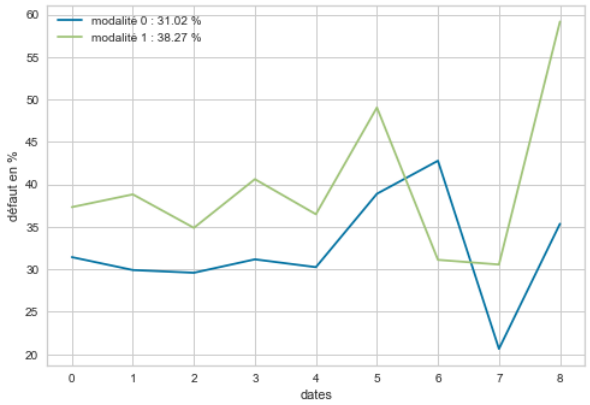

FIGURE 4 – Evolution de la répartition (en %) du défaut par modalité de la variable ”durée d’expérience professionnelle”

La discrétisation des variables est une étape importante lors de la construction d’un score car elle présente divers avantages. Tout d’abord, elle permet une meilleure lisibilité de la grille de score et simplifie ainsi son interprétation. D’autre part, cette méthode est statistiquement plus robuste par rapport à l’imputation. Effectivement, elle traite les valeurs manquantes en conservant l’information et sans introduction de biais. Également, la répartition par vingtile, revient à lisser la série et réduit le bruit, c’est-à-dire, l’impact des fluctuations d’échantillonnage et des valeurs aberrantes. Finalement, produire des variables à modalités stables en répartition et en défaut au cours du temps permet au score de garder sa capacité prédictive dans le temps.

2 Modélisation

La variable à modéliser étant le statut du prêt prenant la modalité 1 si l’emprunt est en défaut et 0 sinon, la régression logistique binaire est une technique naturellement adaptée à notre problème. Pour identifier le meilleur modèle de régression logistique, nous avons procédé par un algorithme itératif.

La première étape permettant une sélection de variable, consiste, grâce au test du khi-deux, à ordonner les variables selon leur niveau de corrélation à la variable cible. Seules les variables les plus corrélées avec la variable d’intérêt sont retenues dans la construction du modèle.

Ensuite, le modèle est formé pas à pas en priorisant l’ordre établi par les résultats de corrélation. Par exemple, la note donnant la qualité du prêt en fonction de l’historique bancaire du client “loan_grade” est la plus corrélée au défaut suivie suivi par le ratio entre la dette et le revenu “loan_percent_income”. Un premier modèle avec la variable “loan_grade” puis un autre avec les variables “loan_grade” et “loan_percent_income” seront formés et ainsi de suite.

A chaque itération, un critère d’information bayésien est calculé. Lorsque l’ajout d’une variable supplémentaire permet de minimiser ce critère, la variable est retenue. Également, la statistique du V de Cramer est estimée. Si la variable ajoutée est fortement corrélée avec une variable déjà présente dans le modèle, alors elle est retirée.

2.1 Corrélation des variables qualitatives

2.1.1 Test du khi-deux et le V de Cramer

Le test du khi-deux permet de révéler une potentielle corrélation entre deux variables qualitatives alors que le V de Cramer donne l’ampleur de la corrélation.

Le test consiste à comparer la distribution jointe théorique espérée sous l’hypothèse nulle à celle effectivement observée. Les hypothèses sont les suivantes :

$$\begin{cases} H_0 : \text{Indépendance des deux variables discrètes} \\ H_1 : \text{Existence d'un lien entre les deux variables} \end{cases} \tag{3}$$

La statistique de test est la suivante :

$$\sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(P-1)(Q-1) \tag{4}$$

où O_{ij} désigne l’effectif observé des individus présentant la modalité i de la première variable et la modalité j de la seconde variable ;

E_{ij} : désigne l’effectif attendu en cas d’indépendance des individus présentant la modalité i de la première variable et la modalité j de la seconde variable ;

P : représente le nombre de modalités de la première variable ;

Q : représente le nombre de modalités de la seconde variable.

Lorsque la p-value est inférieure à $\alpha = 5\%$, l'hypothèse nulle d'indépendance des deux variables qualitatives est rejetée.

Le V de Cramer mesure l'intensité de la corrélation indépendamment du nombre de modalités de la variable et de son effectif.
Sa formule est la suivante :

$$V_{\text{cramer}} = \sqrt{\frac{\chi^2}{N \cdot \min(P-1, Q-1)}} \tag{5}$$

où N désigne le nombre total d'observations et où χ^2 est donné dans l'équation 4 ;

Ainsi, plus le V de Cramer est important, plus la corrélation est importante. Généralement, lorsqu'il est supérieur à 0.7, la corrélation est considérée comme forte entre les deux variables.

2.2 Les critères d'information

Il existe plusieurs critères d'information. Le critère d'Akaike (AIC), le Bayesian Information Criteria (BIC) et le Hannan–Quinn sont parmi les critères les plus populaires.

L'objectif des critères d'information est de garantir au mieux la parcimonie du modèle, c'est-à-dire, d'obtenir la meilleure prédiction à l'aide d'un nombre minimal de variables explicatives. L'augmentation du nombre de variables est donc pénalisée.

La pénalisation de l'augmentation du nombre de paramètres est la principale distinction de ces critères. La formule générale de la pénalisation est la suivante :

$$C_M(N, p_M) = -2 \cdot L_M + p_M \cdot g(N) \tag{6}$$

où L_M est la log vraisemblance du modèle M, N la taille de l'échantillon, p_M le nombre de paramètres du modèle et enfin g(N) la fonction qui décrit l'ampleur de la pénalisation. Seul g(N) diffère d'un critère à un autre :

— Pour l'AIC : g(N) = 2 ;
— Pour le BIC : g(N) = log(N).

Le critère d'information employé dans cette note est le BIC. Plus ce critère est faible, plus l'ajout de la variable est judicieux.

2.3 Résultats de la procédure "step by step"

Avec une p-value inférieure à 5%, l'ensemble des variables sont corrélées à la variable cible. Elles seront donc toutes incluses dans la procédure itérative de sélection de modèle. Ces variables sont classées de la plus corrélée à la moins corrélée. A la première itération, le modèle reprend seulement la variable "loan_grade". Le critère d'information est de 29948. A la deuxième, itération, les variables "loan_grade" et "loan_percent_income" sont les deux explicatives du modèle. A cette itération, le critère d'information étant plus faible, les deux variables sont retenues dans le modèle. Ces étapes sont répétées jusqu'à ce que la variable la moins corrélée à la variable cible soit testée.

Variable	Stat.	P-value	BIC
loan_grade	5609	0.00%	29948
loan_percent_inc.	5018	0.00%	26687
downturn	3263	0.00%	26437
person_income	2788	0.00%	25240
person_home_owner.	1908	0.00%	24182
loan_intent	499	0.00%	23683
person_emp_length	252	0.00%	23637
person_age	53	0.00%	23645

TABLE 2 – Résultats de la procédure d'incrémentation

Finalement, la variable "person_age" est exclue du modèle car son ajout augmente la valeur du critère d'information BIC, n'améliorant pas la précision du modèle. Les autres variables sont retenues car elles ne sont corrélées entre elles selon le V de Cramer *(voir Table 2)*.

A titre d'information, les variables du "montant du prêt" et "taux d'intérêt" ont été retirées car elles sont d'un trop corrélées à d'autres variables déjà présentes. Rappelons, par exemple, que la variable "downturn" est construite à partir des valeurs du taux d'intérêt.

2.4 Estimation du modèle

Un modèle final constitué de l'ensemble des variables maximisant la parcimonie ou minimisant les critères d'information est créé. L'estimation du modèle permet d'obtenir les coefficients de la régression et de s'assurer de leur significativité. Les hypothèses du test de significativité sont les suivantes :

$$\begin{cases} H_0 : \text{La variable est non significative.} \\ H_1 : \text{La variable est significative.} \end{cases} \tag{7}$$

La statistique de test est la suivante :

$$U \equiv \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \sim \mathcal{N}(0,1) \tag{8}$$

Avec : $\hat{\beta}_i$: le coefficient estimé de la variable i dans le modèle de régression logistique retenu et $s(\hat{\beta}_i)$: l'erreur standard estimée de la variable i ;

L'ensemble des variables sont significatives au seuil de 5%. Cette constatation est renforcée par les bornes inférieures (BI) et les bornes supérieures (BS) des rapport de cotes (odd ratio). Effectivement, la valeur 1

Variables	Coeff.	p-value	BI OR	BS OR
loan_grade	2.45	0.00%	10.8	12.5
loan_percent_inc.	0.84	0.00%	2.2	2.4
person_income	0.55	0.00%	1.6	1.8
person_home_owner.	0.70	0.00%	1.9	2.1
downturn	0.50	0.00%	1.5	1.8
loan_intent	0.43	0.00%	1.4	1.6
person_emp_length	0.14	0.00%	1.1	1.2

TABLE 3 – Résultats du modèle de régression logistique

n'est pas comprise dans l'intervalle de confiance à 95% des rapports de côtes, validant ainsi la significativité de notre modèle *(voir les deux dernières colonnes Table 3)*.

2.5 Les critères de performance

2.5.1 ROC curve et indice de Gini

L'évaluation des performances se fait par la courbe ROC (Receiving Operating Characteristic) et la métrique AUC (Area Under the Curve). La courbe ROC est une représentation graphique de la relation entre le taux de vrais positifs et le taux de faux positifs 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.

Le taux de vrais positifs (TVP) se note :

$$\text{TVP} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{9}$$

et s'interprète comme le nombre d'emprunts classés correctement en défaut parmi l'ensemble des emprunts en défaut.

Le taux de faux positifs (TFP) se note :

$$\text{TFP} = \frac{\text{FP}}{\text{FP} + \text{VN}} \tag{10}$$

et s'interprète comme le nombre d'emprunts non défaillants classés en défaut parmi l'ensemble des emprunts non défaillant.

Mesurant l'air sous la courbe ROC, la métrique AUC est un score qui calcule la qualité de précision d'un modèle indépendamment de tous seuils de classification. Plus l'AUC est proche de 1, plus le modèle est précis. Un AUC proche de 0.5 signifie que les prédictions sont proches de l'aléatoire *(courbe rouge de la Figure 5)*. Un AUC proche de 0 signifie que le modèle prédit l'exact inverse de la classe attendue. Dans notre étude, cela signifierait que lorsqu'un prêt est non défaillant, ce dernier est identifié en défaut, et inversement.

L'indice de Gini, indicateur de référence largement utilisé dans le domaine, permet d'ajuster l'AUC afin de le rendre plus interprétable. Comme l'AUC, plus l'indice de Gini est proche de 1, plus le modèle est précis. Cependant, il permet d'attribuer la valeur 0 lorsque le modèle est aussi performant que l'aléatoire et la valeur -1 lorsque le modèle est inversé. Sa formule est la suivante :

$$\text{GINI} = 2 \cdot \text{AUC} - 1 \tag{11}$$

La courbe ROC associée au modèle de régression logistique est donnée en Figure 5 :

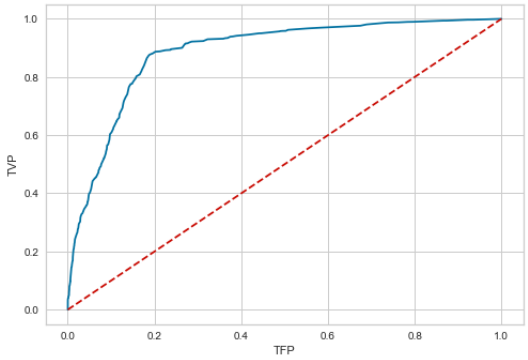

FIGURE 5 – Courbe ROC du modèle de régression logistique (en bleue), courbe ROC de prédiction aléatoire (en rouge)

Dans notre cas d'étude, le modèle fournit une meilleure précision que l'aléatoire comme le souligne la courbe bleue placée au-dessus de la courbe rouge, l'AUC supérieure à 50% et l'indice de Gini supérieur à 0. Sans surprise, le taux de vrai positif croît fortement lorsque le taux de faux positifs est plus faible *(voir Figure 5)*. Finalement, indépendamment du seuil de classification choisi, le modèle de régression logistique a une performance prédictive AUC de 89% *(voir Table 4)*.

2.5.2 Métriques de déséquilibre

Malgré une part importante de cas de défaut au sein de l'historique de données, l'équilibre entre positifs et négatifs dans la variable cible n'est pas atteint. Ainsi, des métriques prenant en compte d'éventuels déséquilibres de classes sont intéressantes à aborder. C'est notamment le cas du recall, équivalent au taux de vrais postifs :

$$\text{recall} = \text{TVP} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{12}$$

et de la pécision :

$$\text{precision} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{13}$$

qui s'interprète comme le nombre d'emprunts classés correctement en défaut parmi l'ensemble des emprunts qui ont été prédits comme défaillants.

Grâce à ces deux métriques, le F1-score, représentant la moyenne harmonisée entre la précision et le recall, peut être calculé. Son objectif est donc de trouver un équilibre entre les deux mesures :

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{14}$$

Lorsqu'un modèle est parfaitement prédictif, le F1-Score est de 100%. Un modèle équivalent à l'aléatoire présente un F1-Score de :

$$\text{F1-score} = 2 \cdot \frac{\text{TP}}{\text{TP} + 1} \tag{15}$$

avec TP le taux de positif dans les données.

Sachant que le taux d'emprunts en défaut dans la base de données est d'environ 21,82%, pour que le modèle soit informatif, le F1-Score doit être supérieur à 35,82%. Ainsi, la métrique étant de 60% pour notre modèle, il peut être considéré comme informatif *(voir Table 4)*.

AUC	Gini	F1-Score
89%	78%	60%

TABLE 4 – Métriques de performance du modèles de régression logistique

3 Construction de la grille de score

3.1 Calcul des pondérations

La grille de score est un outil d'aide à la décision qui permet d'attribuer une note à une demande de prêt en fonction de différents critères. Grâce à cette dernière, la banque peut évaluer le niveau de risque pris lors de l'octroi d'un prêt. A chaque dossier est donc attribuée une note comprise entre 0 (représentant la note minimale) et une note maximale choisie (100, 1000 ou encore 10000). Dans cette étude, 1000 est retenu comme note maximale.

La formation du score repose sur plusieurs notations intermédiaires. La première étape de la méthodologie consiste à modéliser la variable binaire du défaut en fonction des variables transformées par la technique du *One-Hot-Encoding*. Cette technique consiste à représenter des variables discrètes sous forme de vecteurs binaires. Une variable par modalité est obtenue prenant la valeur 0 si l'observation n'a pas cette caractéristique et 1 sinon.

Après l'estimation des coefficients associés à chaque vecteur binaire lors de la régression logistique, nous pouvons calculer les pondérations à associer à chaque modalité de manière à construire la grille de score. Finalement, la note associée à la modalité j de la variable i est calculée comme suit :

$$N_i^j = \frac{1000 \cdot \left| \max\left(x_i^1, \ldots, x_i^p\right) - x_i^j \right|}{\sum_{i=1}^{k} \left(\max\left(x_i^1, \ldots, x_i^p\right) - \min\left(0, x_i^1, \ldots, x_i^p\right) \right)} \tag{16}$$

où :

- x_i^j le coefficient estimé de la modalité j de la variable i ;
- $\max\left(x_i^1, \ldots, x_i^p\right)$ le coefficient maximum de la variable i ;
- $\min\left(0, x_i^1, \ldots, x_i^p,\right)$ le coefficient minimum de la variable i ;
- p le nombre de modalités de la variable i ;
- k le nombre de variables dans le modèle.

D'après la formule (16), moins le défaut est sensible à la modalité, plus grand sera le score pour cette modalité.

Finalement, le score attribué à chaque prêt ou emprunteur (cela dépend de la granularité des données) représente la somme de chaque note intermédiaire attribué aux modalités.

3.2 Grille de score finale

La grille de score attendue est présentée en Table 5 :

Variable (i)	Mod. (j)	Notes	Tx defaut	Répart.
loan_grade	A	417	13.1%	65.2%
	B ou C	82	67.9%	30.1%
	D,E,G	0	83.8%	3.8%
loan_person_inc	< 14%	244	24.1%	52.8%
	[14% ;28%]	234	31.1%	31.8%
	>28%	0	64.57%	16.0%
downturn	Non	55	22.5%	80.4%
	Oui	0	74.5%	19.6%
person_income	> 79k	129	22.4%	25.2%
	[35k ;79k]	109	30.3%	55.7%
	< 35k	0	53.7%	19.1%
person_home_own.	Propriétaire	87	23.1%	49.2%
	Autres	0	42.1%	50.0%
loan_intent	Pro, Educ, Perso	53	29.6%	54.3%
	Autres	0	36.6%	45.6%
person_emp_length	>=2 ans	15	31.0%	75.7%
	<2 ans	0	38.3%	24.3%

TABLE 5 – Scores associés à chaque modalité

Les notes attribuées à chaque modalité sont cohérentes avec le pourcentage de prêts en défaut. A titre d'exemple, plus le revenu de l'emprunteur est élevé, plus la note attribuée est élevée. Selon la Table 5, 25.2% de l'ensemble des prêts sont demandés par des emprunteurs ayant un revenu supérieur à 79 000 dollars, parmi eux 22.4% sont en défaut.

Afin d'illustrer le score final, imaginons qu'un prêt est demandé par un emprunteur étant propriétaire de son logement, ayant plus de 2 ans d'expérience professionnelle et demandant un prêt pour un projet professionnel. Sa note finale sera de : $87 + 15 + 53 = 155$.

3.3 Étude du caractère discriminant

Par des réprésentations graphiques, l'objectif de cette section est d'étudier le caractère discriminant de la note attribuée à chaque prêt.

3.3.1 Densité conditionnelle

La densité conditionnelle permet de visualiser la répartition des notes selon le statut du contrat. Ainsi, deux courbes de distributions distinctes représentent graphiquement l'écart des notes entre les prêts en défaut (courbe verte) et les prêts non défaillants (courbe bleue) *(voir Figure 6)*.

Sur notre échantillon, les prêts en défaut ont globalement une note plus faible, contrairement aux prêts non défaillants qui ont une distribution concentrée à droite, tirant la moyenne vers le haut. Un écart visuel entre les deux courbes indique que le score est discriminant.

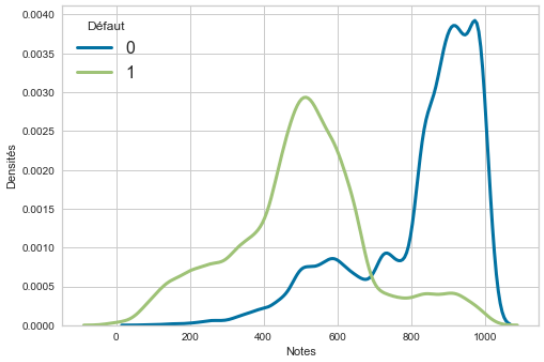

FIGURE 6 – Densité conditionnelle

3.3.2 Croisement du taux de défaut et de la note finale

Pour rappel, le taux de défaut représente la proportion de prêts en défaut parmi l'ensemble des prêts. Dans cette partie, le taux de défaut (en ordonnée) par vingtile de la note finale (en abscisse) est tracé graphiquement.

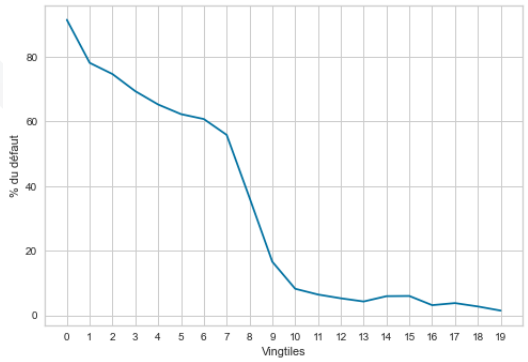

FIGURE 7 – Évolution du taux de défaut de la note par vingtile

Nous constatons en Figure 7 que le taux de défaut décroît bien en fonction des vingtiles, signifiant que plus la note est élevée, plus le taux de défaut est faible. Par conséquent, les scores semblent discriminants et rendent compte des différents profils de risque. Notons, cependant, qu'une décroissance plus régulière aurait été préférable.

3.4 Création des Classes Homogènes de Risque (CHR)

L'objectif final du scoring est de regrouper les individus dans un nombre limité de classes homogènes en fonction de leur risque de défaut. Les taux de défaut des classes doivent être significativement différents. Egalement, il est préférable que chaque classe représente au moins 5% de la population totale. Plusieurs méthodes de regroupement (ou de clustering) peuvent être intéressantes pour ainsi construire ces CHR.

Le clustering appartient à la famille des algorithmes d'apprentissage non supervisés. En ce sens, contrairement à l'apprentissage supervisé, il s'agit de regrouper les observations sans variable cible. Ces méthodes sont appliquées sur l'échantillon des scores.

3.4.1 K-means

L'algorithme des K-means appartient à la famille de la classification non supervisée. La première étape consiste à déterminer le nombre de K clusters de l'algorithme, qui s'obtient généralement par la méthode *Elbow*.

L'idée derrière cette méthode est de fixer le meilleur niveau de K permettant de maximiser la variance inter-classe, c'est-à-dire, la séparation des classes et de minimiser l'inertie intra-classe garantissant la similarité des observations au sein de chaque cluster. L'objectif est d'arbitrer entre un K trop grand (donc pas assez de généralisation) et un K trop petit (regroupant trop d'erreurs).

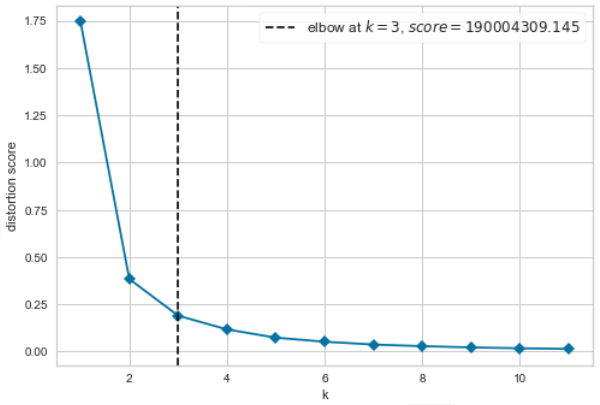

FIGURE 8 – Graphique d'Elbow

La méthode *Elbow* permet de tracer la variance (en ordonnée) en fonction du nombre de clusters (en abscisse). Le "coude" correspond au point de rupture où la variance ne baisse plus significativement, ici le nombre de clusters choisi est de trois (K=3).

Une fois le nombre de K cluster fixé, les K centre de classe sont tirés aléatoirement parmi nos observations. Les observations sont regroupées avec pour règle de décision de minimiser les distances euclidiennes par rapport aux K centres de classe. L'opération est réitérée jusqu'à convergence de l'algorithme, soit jusqu'à ne plus voir de modification des clusters. La distance euclidienne, critère à minimiser, est présentée ci-dessous :

$$D\left(x_i, c_k\right) = \sqrt{\sum_{k=1}^{k} \left(x_i - c_k\right)^2} \qquad (17)$$

où x_i est l'observation i et c_k le centre du cluster k.

3.4.2 Classification Ascendante Hiérarchique (CAH)

Tout comme les K-means, le CAH est un algorithme de classification non supervisé. Son principe est assez simple. Dans l'état initial le nombre de clusters donc de centre (ou centroïde) est égal au nombre d'observations. Ensuite, un regroupement des centroïdes les plus proches est réalisé à l'aide de la minimisation d'un critère. Comme les K-means, il s'agit généralement de la distance euclidienne. L'opération de regroupement est réitérée jusqu'à converger vers un nombre préétabli de cluster. Le nombre de clusters doit être fixé au préalable (K=3). Encore une fois, l'objectif est de maximiser l'inertie interclasse et de minimiser l'inertie intraclasse.

3.4.3 Discrétisation par vingtile

Comme évoqué dans la Section 1.4, la discrétisation est une méthode statistique visant à transformer une variable numérique en une variable discrète. Ainsi, le taux de défaut par vingtile de la note finale est calculé et tracé graphiquement *(Figure 7)*.

Trois sous-groupes sont identifiés avec un premier point de rupture au niveau du sixième intervalle et un second entre le neuvième et dixième intervalle.

- Les vingtiles situés avant le premier point de rupture présentent le taux de défaut le plus élevé et seront ainsi placés dans la classe "3".
- Les vingtiles situés entre le premier et le second point de rupture présentent un taux de défaut moyennement élevé et seront placés dans la classe "2".
- Les vingtiles situés après le second point de rupture présentent un taux de défaut plus faible et et seront placés dans la classe "1".

3.4.4 Algorithme de clustering final

Grâce aux trois méthodes présentées précédemment, des Classes Homogènes de Risque sont formées. En fonction de la méthode, les bornes de notation attribuées à chaque cluster varient. A titre d'exemple, l'algorithme des K-means classe les prêts avec une note inférieur à 443 parmi les plus risquée, tandis que cette borne est de 304 selon la méthode CAH *(voir Table 6)*.

Les bornes étant différentes d'une méthode à une autre, l'algorithme présentant la meilleure performance prédictive est sélectionné. Pour se faire, nous effectuons trois régressions logistiques afin de prédire le défaut. Chaque modèle est composé d'une seule variable explicative correspondant aux classes retenues selon l'une des trois méthodes.

Méthodes	Classes	Labels	Bornes
k-means	3	Risque élevé	<443
	2	Risque moyen	[443 ;741[
	1	Risque élevé	≥ 741
CAH	3	Risque élevé	<304
	2	Risque moyen	[304 ;683[
	1	Risque élevé	≥ 683
Vingtiles	3	Risque élevé	<576
	2	Risque moyen	[576 ;770[
	1	Risque élevé	≥ 770

TABLE 6 – CHR en fonction de la méthode employée

Métrique	k-means	CAH	Vingtiles
AUC	85%	87%	86%

TABLE 7 – Evaluation de la petinence des classes en fonction des méthodes de clustering

Finalement, le modèle avec les CHR fournit par l'algortihme CAH semble avoir les meilleures performances prédictives. Effectivement, son AUC de 87% est plus élevé que pour les modèles concurrents *(voir Table 7)*.

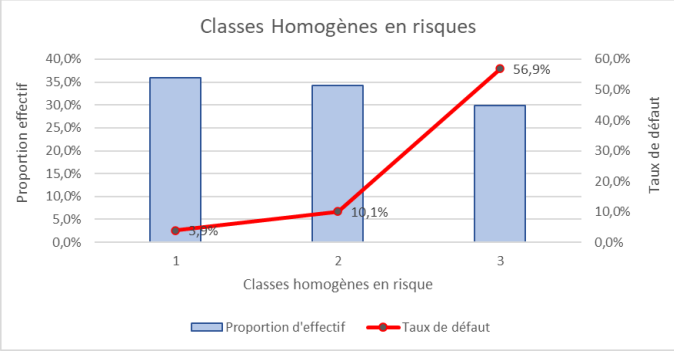

FIGURE 9 – Stabilité des CHR - méthode CAH

Les CHR sont composées d'au moins 30% des observations totales. La classe présentant le plus de prêt en défaut est la classe 3. L'ensemble des prêts de cette classe ont un score inférieur à 304. La classe 1, avec les meilleures notes, à un pourcentage de défaut bien inférieur, soit de 3.9% *(voir Table 6 et Figure 9)*.

4 Les modèles challenger

L'objectif de cette partie est de présenter deux modèles ensemblistes alternatifs à la régression logistique : la Random Forest (algorithme de bagging) et l'eXtrême Gradient Boosting (XgBoost, algorithme de boosting). L'idée est de pouvoir améliorer la performance prédictive du modèle de régression logistique.

Contrairement aux modèles linéaires, ces méthodes ensemblistes tiennent compte de la non-linéarité inhérente des données, et peuvent ainsi capter davantage d'informations. Cependant, l'inconvénient de ces méthodes plus complexes est l'interprétabilité, c'est ce que l'on appelle l'effet "boîte noire". Pour pallier cette limite, la méthode de Shapley, très populaire, est abordée dans cette note.

4.1 Traitement de la base de données

Les variables explicatives des modèles abordés dans cette partie sont celles selectionnées lors de la procédure itérative de selection de variable abordée en Section 2 *(voir Table 3)*. Cependant, leur prétraitement diffère. Elles sont non discrétisées.

Afin d'optimiser notre approche et de permettre à l'algorithme de converger plus rapidement, les variables sont centrées et réduites. La formule de la standardisation est la suivante :

$$X_{stand} = \frac{X - \mu}{\sigma} \tag{18}$$

avec X la variable ; μ et σ représentent respectivement sa moyenne et son écart-type empirique.

4.2 Comparaison des performances prédictives

Modèles	ROC AUC	Gini	F1-Score
Logistic	89%	78%	60%
Random Forest	92%	84%	82%
XgBoost	93%	86%	83%

TABLE 8 – Comparaison des performances des différents modèles

En table 8, nous constatons que les performances sont meilleures que celles de la régression logistique avec une hausse de la ROC AUC de 4 points de pourcentage. Cette hausse peut s'expliquer par l'abscence de discrétisation d'une part, mais aussi car ces algorithmes captent les effets de dépendance non linéaires dans les données.

Finalement, l'algorithme de boosting est celui qui obtient les meilleures performances. C'est donc avec le XgBoost que les profils de risque seront créés.

4.3 Interprétation des modèles grâce à la valeur de Shapley

Provenant initialement de la théorie des jeux, la valeur de Shapley consiste à calculer la contribution marginale d'une variable dans la prédiction suite à sa variation. Intuitivement, il s'agit donc de faire varier un facteur et d'observer le changement dans la précision du modèle.

La contribution de la variable correspond donc à la différence entre la prédiction sortie et la prédiction moyenne. L'opération est réitérée sur l'ensemble des variables de sorte à former une moyenne des contributions de chaque variable à la prédiction moyenne globale.

En résumé, il s'agit d'une méthode :

— locale, car elle calcule la contribution individuelle dite valeur de Shapley ;

— agnostique, car elle est applicable à tout type de modèle ;

— a posteriori, car elle se calcule après estimation du modèle.

Les Figures 10, 11 et 12 illustrent graphiquement la contribution des variables dans la performance prédictive de chaque modèle :

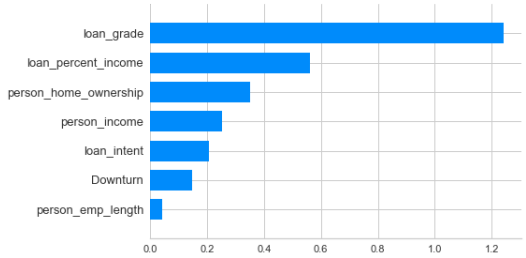

FIGURE 10 – Shapley Value - Régression Logistique

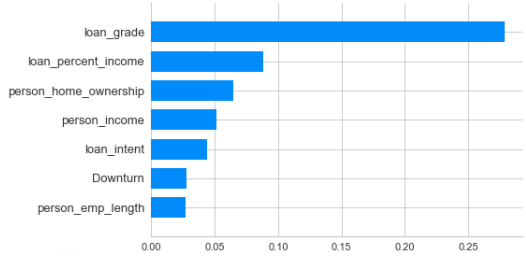

FIGURE 11 – Shapley Value - Random Forest

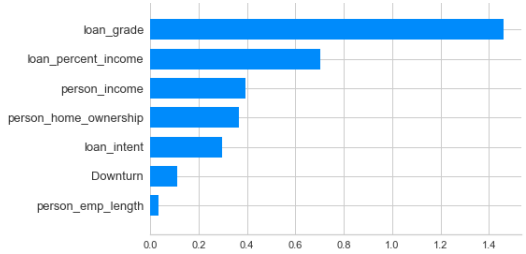

FIGURE 12 – Shapley Value - XgBoost

L'ordre des variables les plus contributrices est identique pour le modèle de régression logistique et le modèle Random Forest. Pour les trois modèles, la variable la plus explicative correspond à la note accordée au prêt en fonction de la qualité de l'emprunteur ("loan_grade") : un prêt avec une note plus faible (en raison d'un défaut, d'un recouvrement passé ...) contribue donc à la probabilité qu'un prêt soit en défaut. Contrairement aux deux autres modèles, le XgBoost, présentant la meilleure performance prédictive, place

le revenu en troisième position. Un individu avec un faible revenu aurait donc une probabilité de défaut plus élevée.

4.4 Création des profils de risque

La création des profils de risque, analogue des CHR présentées en section 3.4, permet de regrouper les prêts selon leur probabilité de défaut. Les probabilités de défaut prédites par l'algorithme du XgBoost sont donc classifiées. L'ensemble des méthodes de clustering abordées précédemment peuvent être utilisées *(voir Section 3.4)*. Ainsi, par la méthode CAH, chaque groupe discrétisé représente un profil de risque homogène :

Proba. Defaut	Profils de risque
$0\% \leq Risque < 20\%$	Risque faible
$20\% \leq Risque < 69\%$	Risque modéré
$20\% \leq Risque < 91\%$	Risque élevé
$91\% \leq Risque \leq 100\%$	Risque très élevé

TABLE 9 – Profils de risque - XgBoost

Lorsque la probabilité de défaut prédite est de 10%, le prêt présente un risque faible. Il sera donc non défaillant. *A contrario*, lorsque la probabilité est de 95%, le prêt est fortement risqué et sera donc classé en défaut *(voir Table 9)*.

4.4.1 Étude de cas avec un individu

Afin d'analyser l'impact de chaque variable sur la prédiction de la probalité de défaut, le modèle est appliqué à un nouveau prêt :

Variables	Valeurs
loan_grade	B
loan_person_inc	30%
person_income	35000
person_home_own.	Locataire
Downturn	Oui
loan_intent	Education
person_emp_length	$\geq 2ans$

TABLE 10 – Caractéristiques du prêt sélectionné au hasard

La *Figure 13* présente l'impact de chaque variable dans la prédiction de la probabilité de défaut pour un prêt donné *(exemple de la Table 10)*. La valeur de Shapley est calculée variable par variable. Elle mesure la différence entre les sorties du modèle de référence $E[f(X)]$ et les sorties du modèle prédit $f(x)$.

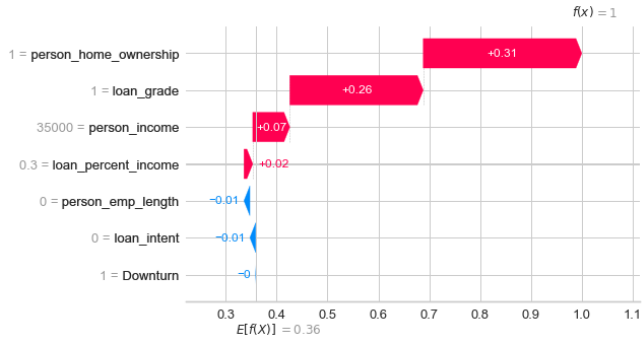

FIGURE 13 – Valeurs de Shapley pour un prêt donné

Contrairement aux variables bleues, les variables rouges augmentent la probalité de défaut. La variable "loan_grade", étant la plus contributrice du taux de défaut selon l'algorithme du XgBoost *(voir Figure 12)*, est seulement la deuxième contributrice pour ce prêt. Cela s'explique par le fait que ce prêt ne se situe pas dans les notes les plus mauvaises selon la grille de score *(voir Table 5)*. La variables "downturn" fixée à 1, indiquant que le prêt est accordé en période de ralentissement économique est peu contributrice pour le modèle et par conséquent pour cette nouvelle observation.

Finalement, pour ce prêt la probabilité de défaut prédite est de 90%, signifiant que ce prêt présente un risque élevé *(voir Table 9)*. Egalement, la sortie du modèle $f(x) = 1$ indique que le prêt est en défaut.

5 Conclusion

Même si la régression logistique permet de fournir un outil interprétable et industrialisable, les performances de cet algorithme sont généralement plus faibles que celles des algorithmes ensemblistes.

Ces derniers soulèvent cependant une complexité de mise en œuvre importante. En effet, contrairement à l'approche classique, l'identification de la contribution de chaque variable au risque de défaut nécessite le recours au modèle et à des techniques d'interprétabilité plus lourdes comme le calcul de la valeur de Shapley.

La combinaison des méthodes, permettant un arbitrage entre interprétabilité et performance, pourrait être envisagée. La grille de score, fournit pas la méthode classique, pourrait être utilisée sur les taux de défaut très faibles et les taux de défaut très élevés car ce sont ceux qui sont les plus facilement détectables. Pour les prêts avec un taux de défaut est plus modéré, l'emploi de méthodes ensemblistes couplé à l'utilisation des valeurs de Shapley permettrait d'améliorer la précision.

Références

[1] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms : a critical evaluation. *National Library of Medicine*, 2016.

[2] Nexialog Consulting. Traitement des données manquantes dans le milieu bancaire. *working paper*, 2022. Lien vers l'article.

[3] John T. Hancock and Taghi M. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7(28), 2020.

[4] Constantine Frangakis Melissa J. Azur, Elizabeth A. Stuart and Philip J. Leaf. Multiple imputation by chained equations : What is it and how does it work ? *International Journal of Methods in Pyschiatric Research*, 2011.

[5] Majid Sarmad. Robust data analysis for factorial experimental designs : Improved methods and software. *Statistics Department of Mathematical Sciences University of Durham England*, 2006.

[6] Daniel J. Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 2011.

[7] Supriyanto Wibisono, Anwar and Amin. Multivariate weather anomaly detection using dbscan clustering algorithm. *Journal of Physics : Conference Series*, 2021.

Nexialog Consulting est un cabinet de conseil spécialisé en Banque et en Assurance. Organisés autour de 3 domaines d'activité - Risques Bancaires, Financiers & Assurantiels - nous intervenons au sein des équipes métiers afin de les accompagner depuis le cadrage jusqu'à la mise en œuvre de leurs projets. Associant innovation et expertise, le savoir-faire de notre cabinet a permis de consolider notre positionnement sur ce segment et de bénéficier d'une croissance forte et régulière.

Les besoins de nos clients étant en constante évolution, nous nous adaptons continuellement pour proposer le meilleur accompagnement. Le département R&D de Nexialog Consulting se donne pour objectif de proposer des solutions innovantes à des problématiques métier ou d'actualité. Pour cela, nous nous appuyons sur des librairies internes et sur le travail de nos consultants. Le pôle R&D Nexialog a également pour mission de former les collaborateurs sur l'évolution des techniques et la réglementation en lien avec leur activité.

Site web du cabinet : https://www.nexialog.com
Publications : https://www.nexialog.com/publications-nexialog/

Contacts

Ali BEHBAHANI
Associé, Fondateur
Tél : + 33 (0) 1 44 73 86 78
Email : abehbahani@nexialog.com

Christelle BONDOUX
Associée, Directrice commerciale
Tél : + 33 (0) 1 44 73 75 67
Email : cbondoux@nexialog.com

Areski COUSIN
Directeur scientifique
Email : acousin@nexialog.com