Analyse du comportement de crédit des clients bancaires

Nom : FAVEUR

Prénom : Kenson

Date : 02/05/2025

Table des matières

Introduction

L'accès au crédit constitue un levier essentiel pour le développement économique, tant pour les particuliers que pour les institutions financières. Pour les banques, il s'agit non seulement d'une source de revenu significative, mais également d'un facteur de fidélisation client. Dans ce contexte, la capacité à anticiper le comportement de crédit devient stratégique. Pouvoir identifier les clients les plus susceptibles d'obtenir un crédit permet :

D'améliorer l'efficacité des campagnes marketing (ciblage plus précis),

D'optimiser la mobilisation des conseillers bancaires (focus sur les profils à fort potentiel),

De réduire les risques d'octroi à des profils non éligibles,

Et de mieux comprendre les déterminants socio-économiques qui influencent l'accès au crédit.

Objectif de l'étude

L'objectif principal de cette étude est de construire un modèle prédictif fiable capable d'identifier, à partir d'un ensemble de caractéristiques clients, ceux qui sont les plus susceptibles de détenir ou de demander un crédit. Pour cela, nous utilisons une base de données fournie par une banque, contenant des informations socio-démographiques, historiques de crédit et comportementales sur un échantillon de 33 663 clients.

Jeu de données utilisé

La base de données étudiée contient 21 variables décrivant de manière détaillée le profil de chaque client. Ces variables peuvent être regroupées en plusieurs catégories :

Informations générales :

id_client : Identifiant unique du client

flag_credit : Variable cible indiquant si le client a souscrit un crédit (1 = oui, 0 = non)

age : Âge du client en années

sexe : Sexe du client

situation_familiale : Situation matrimoniale (célibataire, marié, etc.)

statut_logement : Statut d'occupation du logement (locataire, propriétaire, hébergé, etc.)

Situation financière :

revenu : Revenu mensuel déclaré par le client

nb_credits_total : Nombre total de crédits souscrits dans le passé

mt_credits_total : Montant cumulé de l'ensemble des crédits

nb_credits_actuel : Nombre de crédits actuellement en cours

mt_credits_actuel : Montant cumulé des crédits en cours

mt_echeances_actuel : Montant mensuel à rembourser pour les crédits en cours

duree_remboursement_actuel : Durée totale de remboursement des crédits en cours (en années)

Historique de crédit :

mt_premier_credit : Montant du premier crédit contracté

anc_premier_credit : Ancienneté (en jours) depuis le premier crédit

canal_premier_credit : Canal de souscription (agence, internet, téléphone, etc.)

mt_dernier_credit : Montant du dernier crédit souscrit

anc_dernier_credit : Ancienneté (en jours) du dernier crédit

canal_dernier_credit : Canal de souscription du dernier crédit

Informations de contact :

flag_tel : Indique si le numéro de téléphone est connu (1 = oui, 0 = non)

flag_email : Indique si l'adresse e-mail est connue (1 = oui, 0 = non)

Approche méthodologique

L'étude suit une démarche structurée en plusieurs étapes clés :

- Exploration initiale des données pour mieux comprendre les profils clients et identifier les premières tendances.
- Nettoyage et traitement des données (gestion des valeurs manquantes, outliers).
- Création de nouveaux indicateurs pour enrichir les variables existantes et renforcer la capacité prédictive du modèle.
- Optimisation des variables explicatives via des techniques comme la discrétisation.
- Rééquilibrage de la variable cible, fortement déséquilibrée, à l'aide de méthodes d'échantillonnage.
- Modélisation prédictive et évaluation des performances à l'aide de plusieurs critères : précision, rappel, et AUC.

Exploration et première modélisation

L'étape exploratoire permet de mieux comprendre la structure et la qualité du jeu de données, d'en identifier les principales tendances, et de guider les choix de modélisation.

Statistiques descriptives clés

Une première exploration statistique des données a été réalisée afin de mieux comprendre la structure de la population étudiée et d'identifier d'éventuelles anomalies à traiter avant toute modélisation. Voici les principaux enseignements :

Profil sociodémographique des clients

La base contient 33 663 individus. La variable cible flag_credit indique que 23 % des clients ont contracté un crédit, ce qui révèle un fort déséquilibre des classes, justifiant ultérieurement un rééquilibrage.

L'âge moyen des clients est de 41,7 ans, avec une distribution relativement symétrique autour de la médiane (42 ans). Les individus sont majoritairement âgés de 35 à 50 ans.

La répartition par sexe est relativement équilibrée, avec 55,5 % d'hommes et 44,5 % de femmes. En revanche, la situation familiale est très polarisée : plus de 60 % des clients sont célibataires,

contre 24 % de personnes mariées. On note également une part non négligeable de valeurs manquantes sur cette variable.

Situation financière et coordonnées

Le revenu moyen déclaré est d'environ 2 160 €, mais les valeurs extrêmes (jusqu'à 245 000 €) suggèrent la présence d'outliers importants. Ce point nécessitera un traitement particulier (winsorisation ou transformation). De plus, plus de 40 % des clients n'ont pas renseigné leur revenu, ce qui soulève la question de la pertinence de cette variable brute dans la modélisation.

Concernant les moyens de contact, 93 % des clients disposent d'un numéro de téléphone, mais seulement 5 % d'un e-mail. Cette faible couverture de l'e-mail limite son intérêt pour des campagnes marketing numériques.

Comportement de crédit

En moyenne, un client détient 3 crédits au total, dont 1,3 en cours. Le montant total des crédits en cours avoisine les 2 170 €, avec des montants extrêmes allant jusqu'à 27 215 €. L'échéance mensuelle moyenne est de 465 €, mais certains montants très élevés (jusqu'à 35 557 €) paraissent incohérents, ce qui indique d'éventuelles erreurs de saisie. La durée moyenne des crédits en cours est de 2 ans. Cependant, 818 valeurs sont manquantes, ce qui représente un volume non négligeable de données à traiter.

Historique des crédits

Le montant moyen du premier crédit contracté est de 1 576 €, et son ancienneté moyenne est d'environ 6 ans. On observe ici aussi des anomalies (ex : montants négatifs) et 1 485 valeurs manquantes sur l'ancienneté.

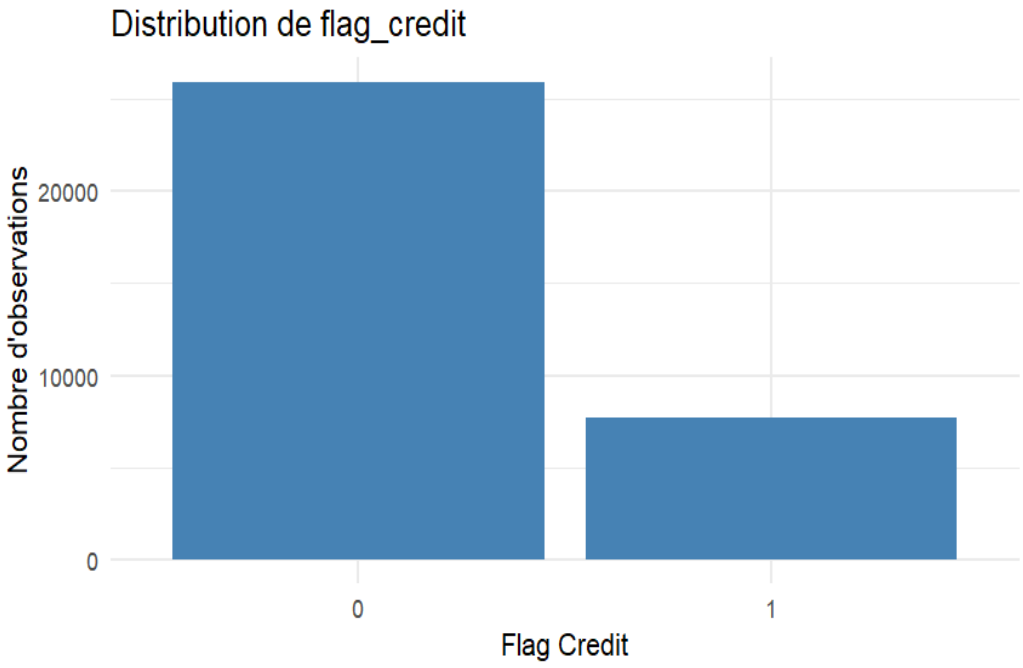
Le dernier crédit est plus récent (ancienneté moyenne de 15 mois), pour un montant moyen de 2 047 €. Très peu de données sont manquantes sur cette variable, ce qui la rend plus fiable pour la modélisation.

En ce qui concerne le canal de souscription, la majorité des crédits ont été contractés via un partenaire commercial (près de 60 %), suivis des agences bancaires et du canal mailing. Cette répartition très déséquilibrée pourrait justifier un regroupement des modalités ou une pondération lors de l'analyse.

Analyse de la variable cible

La variable cible flag_credit indique si un client a déjà contracté un crédit (1) ou non (0). L'analyse de sa distribution révèle un déséquilibre marqué dans les classes :77 % des clients n'ont jamais souscrit de crédit, seulement 23 % des clients ont contracté au moins un crédit.Ce déséquilibre est confirmé par la visualisation sous forme d'histogramme, où la classe "0" prédomine largement dans le graphe ci-dessous.

Graphe de la variable à expliquer



Modèle brut – Régression logistique

.Performance globale du modèle

- Le test du **rapport de vraisemblance** montre que le modèle est **globalement significatif** (*p-value < 0.001*), ce qui signifie qu'il apporte une meilleure prédiction que le modèle sans variables explicatives.
- Le **pseudo R² de McFadden** est de **0.117**, indiquant une qualité d'ajustement **modérée** (typique des modèles logistiques appliqués à des données sociales ou économiques).

Variables significatives du modèle final

Après sélection et regroupement pertinent des modalités, le modèle final inclut 14 variables explicatives toutes significatives au seuil de 5 %. En voici les principales :

Variables sociodémographiques :

- **Âge** : plus l'âge est élevé, plus la probabilité d'obtenir un crédit augmente légèrement ($p = 0.0004$).
- **Sexe** : les **hommes ont une probabilité significativement plus faible** d'obtenir un crédit que les femmes ($OR \approx 0.84$, $p < 0.001$).
- **Situation familiale** : les personnes **non mariées (célibataires/divorcées)** ont une probabilité significativement **plus élevée** d'obtenir un crédit que les personnes mariées ou en union libre ($p < 0.001$).
- **Statut logement** : les **locataires** sont plus susceptibles d'obtenir un crédit que les propriétaires ($p < 0.001$).

Contact et historique client :

- **flag_email** : le fait de disposer d'une adresse email augmente fortement la probabilité d'octroi d'un crédit ($OR \approx 1.96$, $p < 0.001$).
- **Ancienneté du premier crédit** : une ancienneté plus grande est associée à une **baisse** de probabilité d'obtenir un crédit actuellement ($p < 0.001$).

Historique de crédits :

- **Nombre total de crédits** : chaque crédit supplémentaire augmente la probabilité d'obtenir un nouveau crédit ($p < 0.001$).
- **Nombre de crédits actifs** : effet fortement positif également ($p < 0.001$).
- **Montant des crédits actuels** : effet significatif positif ($p < 0.001$).
- **Montant des échéances actuelles** : effet positif ($p < 0.001$).
- **Durée de remboursement actuelle** : plus la durée est longue, plus la probabilité d'octroi est élevée ($p < 0.001$).

- **Montant du dernier crédit** : un montant plus élevé du dernier crédit est **négativement** corrélé à l'obtention d'un nouveau ($p < 0.001$).

Canaux de distribution :

- **Canal du premier crédit** :

Les crédits passés par un **partenaire** sont associés à une probabilité plus faible d'en obtenir un nouveau, comparé à l'**agence** ($p < 0.001$).

- **Canal du dernier crédit** :

Le **canal "Mailing"** est associé à une probabilité plus faible d'obtention d'un crédit comparé au canal **"Direct" (agence ou partenaire)** ($p < 0.001$).

Effet des regroupements de modalités

- Le regroupement des modalités (par ex. "Non marié" vs "Marié", ou "Direct" vs "Mailing") **simplifie l'interprétation tout en conservant la significativité des effets**, comme l'indique la stabilité des coefficients et le maintien du R^2.

Évaluation des performances du modèle

Trois indicateurs principaux ont été utilisés pour évaluer les performances du modèle de classification : la précision, le rappel et l'aire sous la courbe ROC (AUC).

Précision (seuil = 0.5) : 0.556

Cela signifie que parmi toutes les prédictions positives effectuées par le modèle, 55,6 % sont correctes. Autrement dit, lorsqu'il prédit une classe positive, le modèle se trompe environ une fois sur deux. Cette précision modérée suggère une présence non négligeable de faux positifs.

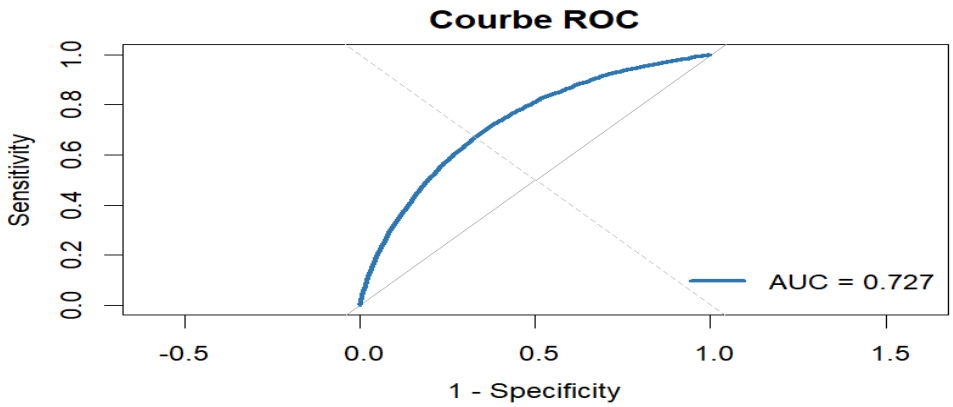
Rappel (seuil = 0.5) : 0.169

Le rappel mesure la capacité du modèle à identifier correctement les cas réellement positifs. Avec un rappel de seulement 16,9 %, le modèle parvient à détecter une faible proportion des instances positives réelles, ce qui peut être problématique si l'objectif est de capter un maximum de cas positifs (par exemple, dans des contextes de détection d'anomalies ou de diagnostics médicaux).

AUC (Aire sous la courbe ROC) : 0.7268

L'AUC mesure la capacité du modèle à discriminer entre les classes positives et négatives, indépendamment du seuil de classification. Une AUC de 0.7268 indique que le modèle possède une bonne capacité discriminante globale, bien qu'imparfaite. En d'autres termes, il classe correctement un exemple positif avant un négatif dans environ 72,7 % des cas.

Graphe de la courbe



Préparation et enrichissement des données

Avant l'entraînement du modèle, une phase essentielle de préparation et d'enrichissement des données a été réalisée afin d'assurer la qualité et la pertinence des informations utilisées. Nous avons traités les valeurs manquantes par l'imputation.

Analyse des résultats après le traitement des valeurs manquantes.

Le modèle de régression logistique estimé a pour objectif de prédire la probabilité qu'un individu détienne un crédit (flag_credit). À partir des données disponibles, le modèle s'avère **statistiquement robuste** et **globalement pertinent**, avec une **réduction notable de la déviance** (de 36 270 à 32 065) et un **AIC raisonnable** de 32 107, ce qui confirme une bonne capacité explicative.

Variables explicatives significatives

L'analyse des coefficients met en évidence plusieurs variables ayant un **impact significatif** sur la probabilité de souscrire un crédit :

- **Âge** : effet positif modéré, indiquant une probabilité légèrement plus élevée avec l'âge.

- **Sexe** : les hommes sont statistiquement moins susceptibles d'avoir un crédit que les femmes.
- **Situation familiale** : les célibataires, divorcés ou en union libre ont une probabilité plus élevée que les personnes mariées.
- **Statut locatif** : les locataires présentent une probabilité significativement plus forte d'avoir un crédit que les autres.
- **Email renseigné** : les individus ayant fourni un email sont nettement plus enclins à souscrire un crédit.
- **Comportement de crédit** (nombre, montant, échéances, durée, ancienneté) : ces variables financières ont un impact fortement significatif et intuitif sur la probabilité d'avoir un crédit.
- **Canaux commerciaux** : les canaux « Partenaire » et « Mailing » sont associés à une probabilité plus faible de crédit, suggérant une efficacité moindre de ces canaux.

Variables à impact limité ou non significatif

Certaines variables, bien que présentes, n'ont pas d'effet significatif sur la souscription d'un crédit (ex. : statut de propriétaire, situation « marié », canal du premier crédit par mailing), ce qui peut orienter leur poids dans des modèles futurs ou dans la prise de décision marketing.

Évaluation des performances du modèle

L'évaluation du modèle a été réalisée à l'aide de plusieurs métriques clés :

Précision (Accuracy) : 78,2 %

Cela signifie que le modèle prédit correctement 78 % des cas (positifs ou négatifs). C'est un bon score global, mais il ne reflète pas à lui seul la capacité à détecter les cas positifs.

Rappel (Recall ou Sensibilité) : 95,6 %

Le modèle parvient à identifier presque tous les individus ayant souscrit un crédit (positifs). Ce taux élevé est particulièrement intéressant si l'objectif est de ne pas manquer de clients potentiels pour un crédit.

AUC (Area Under Curve) : 0.735

L’aire sous la courbe ROC mesure la capacité globale du modèle à distinguer entre les clients ayant ou non souscrit un crédit. Une AUC de 0,735 indique une performance correcte, nettement meilleure qu’un modèle aléatoire (AUC = 0,5), mais avec une marge d’amélioration possible.

Création de nouveaux indicateurs

Dans le cadre de l'analyse de la solvabilité des clients, il est crucial d'examiner des variables qui permettent d'évaluer à la fois la capacité de remboursement et le risque associé à chaque emprunteur. Pour enrichir le modèle et de mieux comprendre le comportement financier des clients, cinq nouveaux indicateurs ont été créés.

1. **Taux d'endettement** : Ce ratio mesure la part des remboursements par rapport au total des crédits. Il aide à évaluer la capacité du client à rembourser ses dettes et à détecter un potentiel risque de surendettement.
2. **Montant moyen par crédit actif** : Il permet d'estimer le montant moyen des crédits d'un client. Cela donne une idée de l'intensité de l'endettement et aide à comprendre les besoins financiers du client.
3. **Durée moyenne par crédit** : Cet indicateur calcule la durée moyenne des crédits en cours. Il permet d’évaluer si un client a des engagements à court ou long terme, ce qui influence le risque associé à ses remboursements.
4. **Ratio du dernier crédit sur le total des crédits** : Il mesure la proportion du dernier crédit par rapport à l’ensemble des crédits du client. Un ratio élevé peut signaler un emprunt récent plus important, ce qui peut augmenter le risque de défaut.
5. **Ancienneté moyenne par crédit** : Cet indicateur évalue depuis combien de temps un client est engagé dans des crédits. Une ancienneté plus longue peut signaler une relation stable avec l’institution financière, ce qui réduit le risque.

Analyse des résultats du modèle

L'analyse des résultats du modèle montre l'impact des variables sur la probabilité qu'un client ait un **flag_credit** (indicateur binaire représentant si un client a un crédit en défaut ou non). Voici une interprétation des principaux résultats obtenus :

Variables significatives :

- **Age** : L'âge a un effet positif et significatif sur la probabilité d'avoir un crédit en défaut ($p < 0.01$). Cela suggère que les personnes plus âgées ont une probabilité légèrement plus élevée d'être en défaut.
- **Sexe** : Le sexe (homme vs femme) joue un rôle significatif ($p < 0.01$), avec un coefficient négatif pour les hommes, indiquant qu'ils ont une probabilité plus faible d'être en défaut par rapport aux femmes.
- **Situation familiale** : Les personnes **célibataires**, **divorcées** et **en union libre** ont une probabilité plus élevée d'être en défaut de crédit par rapport à celles mariées. Cela peut indiquer un risque plus élevé chez les personnes seules ou séparées, peut-être en raison de l'absence de soutien financier d'un partenaire.
- **Statut de logement** : Les personnes **locataires** ont une probabilité plus élevée d'être en défaut par rapport aux propriétaires, ce qui peut suggérer que la stabilité financière des propriétaires est plus forte. En revanche, les propriétaires ont également un coefficient négatif, mais cela n'est pas aussi marqué que pour les locataires.
- **Flag_email1** : La variable indicative de la réception d'emails (flag_email1) est très significative ($p < 0.01$), ce qui suggère que cette variable a un fort impact sur la probabilité d'être en défaut.
- **Crédits actuels et montant des crédits** : Les variables **nb_credits_actuel** et **mt_credits_actuel** ont un effet positif significatif ($p < 0.01$), ce qui indique que les clients avec plus de crédits actifs et des montants de crédit plus élevés ont une probabilité plus élevée d'être en défaut.
- **Montant des échéances** : Le montant des échéances actuelles (**mt_echeances_actuel**) a également un impact significatif et positif ($p < 0.01$), suggérant que des échéances plus élevées sont associées à un plus grand risque de défaut.
- **Ancienneté du premier crédit** : L'ancienneté du premier crédit (**anc_premier_credit**) a un impact significatif ($p < 0.01$). Les clients ayant un plus grand nombre d'années d'expérience avec des crédits ont une probabilité plus faible d'être en défaut.

Variables non significatives :

- **Endettement** : Bien que le taux d'endettement soit introduit dans le modèle, il ne semble pas avoir d'impact significatif sur la probabilité de défaut de crédit ($p = 0.973$). Cela peut indiquer que cet indicateur, dans son état actuel, n'ajoute pas beaucoup d'information supplémentaire pour prédire le défaut de crédit.
- **Montant moyen par crédit actif** : Bien qu'il soit significatif ($p < 0.01$), cet indicateur montre un faible impact dans le modèle, suggérant que l'effet du montant moyen par crédit n'est pas aussi fort que d'autres facteurs comme le montant total des crédits.
- **Durée moyenne par crédit et ratio du dernier crédit sur le total** : Ces deux variables sont également non significatives ($p > 0.05$), ce qui indique qu'elles n'ont pas d'impact direct sur la probabilité de défaut dans ce modèle.

Indicateurs améliorés :

L'introduction des cinq nouveaux indicateurs (tels que l'endettement, le montant moyen par crédit actif, la durée moyenne des crédits, le ratio du dernier crédit et l'ancienneté des crédits) a permis d'enrichir le modèle, mais il est notable que certaines de ces variables ne sont pas significatives, comme l'endettement et le ratio du dernier crédit. Cependant, l'**ancienneté par crédit** et le **montant moyen par crédit** sont des variables pertinentes dans ce modèle et contribuent de manière significative à la prédiction du défaut de crédit.

Ajustement du modèle :

- **Deviance** : La **devance résiduelle** du modèle est de 31,546, et la **devance nulle** est de 35,919. Cela indique une amélioration notable dans la qualité du modèle comparé à un modèle nul.
- **AIC** : L'AIC (Critère d'Information d'Akaike) est de 31,598, ce qui suggère une complexité modérée du modèle et son efficacité pour la prédiction des défauts de crédit.

Analyse des indicateurs de performances

L'analyse des indicateurs de performances se base sur trois indicateurs clés : **précision**, **rappel** et **AUC**.

Précision (Accuracy) :

La précision du modèle est de **78.13%**. Cela signifie que le modèle a correctement prédit la classe (défaut ou non défaut de crédit) dans environ 78.13% des cas. Bien que ce soit un bon résultat, la précision seule peut être trompeuse, surtout lorsque les classes sont déséquilibrées, comme c'est souvent le cas dans les modèles de classification binaire où il y a plus de non-défaillants que de défaillants. La précision ne capture pas la capacité du modèle à détecter les défaillants dans ces situations.

Rappel (Sensitivity) :

Le **rappel** (ou sensibilité) est de **95.64%**, ce qui indique que le modèle est très efficace pour détecter les clients qui sont réellement en défaut de crédit (classe positive). Un rappel élevé signifie que le modèle réussit à identifier presque toutes les instances de défaut. Cependant, un rappel élevé peut parfois être associé à un modèle qui génère beaucoup de faux positifs (c'est-à-dire classer à tort des clients non-défaillants comme défaillants).

AUC (Area Under the ROC Curve) :

L'AUC (aire sous la courbe ROC) est de **0.74**, ce qui est un indicateur assez bon de la capacité du modèle à distinguer entre les classes (défaut et non défaut). Une AUC proche de 1 indique une excellente capacité de discrimination entre les classes, tandis qu'une AUC proche de 0,5 suggère un modèle qui fait des prédictions aléatoires. Dans ce cas, l'AUC de 0.74 montre que le modèle a une capacité modérée à distinguer les clients en défaut de crédit des autres.

Discrétisation des variables

Le modèle de régression logistique discrétisé a été ajusté afin d'évaluer les facteurs influençant la probabilité de défaut de crédit. En utilisant des variables discrètes, nous avons pu analyser l'impact de différentes caractéristiques des clients et des crédits sur le risque de défaut. Les résultats obtenus permettent de mieux comprendre les comportements à risque et d’orienter les décisions stratégiques en matière de gestion du risque de crédit.

Signification des variables

Les résultats montrent que plusieurs variables sont significativement liées à la probabilité de défaut, comme l'indiquent les p-values associées aux coefficients. Les variables avec une p-value inférieure à 0.05 sont considérées comme ayant un impact notable sur le risque de défaut. Voici les principales conclusions tirées de ces variables :

Sexe (sexeM) : Le sexe masculin présente un effet significatif sur la probabilité de défaut, avec une **p-value** de **4.16e-14**, suggérant que les hommes ont une probabilité plus faible de faire défaut par rapport aux femmes. Le coefficient de **-0.256** indique une réduction modérée du risque de défaut pour les hommes.

Situation familiale : Les personnes célibataires (**p-value = 0.007**) et divorcées (**p-value = 0.014**) présentent des probabilités plus élevées de défaut que celles mariées, ce qui souligne l'impact potentiel de la stabilité familiale sur la gestion du crédit. En revanche, les catégories **union libre** et **marié** n'ont pas d'effet statistiquement significatif (respectivement **p-value = 0.321** et **0.934**).

Statut de logement : Les locataires et propriétaires ont un impact significatif sur la probabilité de défaut : **Locataire** (**p-value = 0.0044**) : Les locataires sont plus susceptibles de faire défaut par rapport aux propriétaires.

Propriétaire (**p-value < 2e-16**) : Les propriétaires sont moins susceptibles de faire défaut, suggérant que la stabilité du logement est un facteur protecteur contre le défaut de paiement.

Email valide (flag_email1) : La présence d'un email valide (p-value = **7.57e-10**) a un impact positif sur la probabilité de défaut. Cela pourrait refléter un lien entre la gestion des informations clients et la gestion des risques.

Montants et durées des crédits : Les montants de crédit et la durée des crédits ont un effet significatif sur la probabilité de défaut. En particulier :

Montants de crédits actuels (mt_credits_actuel) : Les crédits plus élevés sont associés à un risque accru de défaut, avec des p-values très significatives (**p-value < 0.001**).

Durée des remboursements (duree_remboursement_actuel) : Une durée de remboursement plus longue est également liée à un risque plus élevé de défaut, avec des coefficients et p-values indiquant un effet substantiel sur le modèle.

Coefficients des variables

Les coefficients indiquent l'ampleur de l'effet de chaque variable. Par exemple : Le coefficient **0.726** pour les crédits actuels très élevés (categorie **mt_credits_actuel_discrettrès élevée**) suggère qu'un montant de crédit plus élevé augmente significativement le risque de défaut. La **durée de remboursement très élevée** (coefficient **0.892**) augmente également la probabilité de défaut, montrant que les crédits à long terme sont plus risqués.

Amélioration du modèle

Comparé au modèle initial, ce modèle discrétisé a montré une amélioration notable en termes d'ajustement. La **deviance résiduelle** a diminué de **35919 à 31421**, ce qui reflète une meilleure adéquation du modèle aux données observées. La réduction de la deviance indique que le modèle discrétisé prédit mieux les défauts de crédit, ce qui justifie l'utilisation de variables discrètes pour mieux saisir les relations non linéaires dans les données.

Ajustement global du modèle

Le **AIC (Akaike Information Criterion)** de **31529** indique une bonne qualité d'ajustement du modèle, en tenant compte du nombre de paramètres estimés. Un AIC plus faible suggère une meilleure performance du modèle, en particulier par rapport aux modèles précédents.

Évaluation de la Performance du Modèle Discrétisé

Dans le cadre de l'évaluation de notre modèle de prédiction du défaut de crédit, trois indicateurs clés ont été utilisés pour mesurer sa performance : la **précision**, le **rappel** et l'**AUC** (Area Under the Curve). Ces métriques permettent d'analyser l'efficacité du modèle à identifier correctement les clients à risque tout en minimisant les erreurs. Voici les résultats obtenus après l'évaluation du modèle discrétisé.

Précision

La précision, mesurée à **0.7805**, indique que le modèle a correctement prédit environ **78.05%** des cas, qu'ils soient positifs ou négatifs. En d'autres termes, parmi toutes les prédictions de défaut faites par le modèle, environ **78.05%** étaient exactes. Cette métrique est particulièrement importante pour éviter les faux positifs, c'est-à-dire prêter de l'argent à des clients qui ne sont pas réellement à risque.

Rappel

Le rappel, quant à lui, est évalué à **0.9603**, ce qui signifie que le modèle a réussi à identifier **96.03%** des clients réellement à risque de défaut. Cette haute performance est cruciale, car elle montre que le modèle est efficace pour détecter les clients qui risquent de faire défaut, réduisant ainsi la possibilité de ne pas accorder de crédit à une personne à risque.

AUC (Area Under the Curve)

L'**AUC**, évaluée à **0.7432**, mesure la capacité du modèle à distinguer les clients en défaut des autres. Une AUC proche de **1** indique une bonne capacité à discriminer entre les classes, tandis qu'une valeur proche de **0.5** serait équivalente à un tirage au sort. Avec une AUC de **0.7432**, le modèle montre une bonne capacité de discrimination entre les clients à risque de défaut et ceux qui ne le sont pas, ce qui suggère une performance acceptable à bonne.

Rééquilibrage de la variable cible

Dans le jeu de données initial, la variable cible **flag_credit**, indiquant le défaut de paiement, présente un déséquilibre important entre les deux classes : les clients *non défaillants* (classe majoritaire) sont largement plus nombreux que les clients *défaillants* (classe minoritaire). Ce déséquilibre peut fortement biaiser l'apprentissage du modèle, qui aura tendance à prédire principalement la classe majoritaire et à ignorer les cas rares mais cruciaux des défauts de crédit.

Pour pallier ce problème, deux techniques de rééquilibrage ont été mises en œuvre :

Sous-échantillonnage de la classe majoritaire

Cette méthode consiste à réduire la taille de la classe majoritaire en échantillonnant aléatoirement un sous-ensemble de ses observations, de façon à équilibrer les proportions entre les deux classes. Bien que cette approche permette un bon équilibre, elle présente l'inconvénient de perdre potentiellement une grande quantité d'informations utiles contenues dans la classe majoritaire.

SMOTE (Synthetic Minority Oversampling Technique)

La méthode SMOTE génère artificiellement de nouvelles observations dans la classe minoritaire à partir d'interpolations entre les individus existants. Elle permet d'équilibrer les

classes sans perdre d'information, mais peut parfois introduire du bruit et rendre le modèle plus instable ou moins précis si mal paramétrée.

Évaluation des Modèles : Sous-échantillonnage et SMOTE

Dans le cadre de l'amélioration de la prédiction du défaut de crédit, deux approches ont été mises en œuvre pour traiter le déséquilibre de classes dans les données : le sous-échantillonnage et la méthode SMOTE. Ces techniques visent à rééquilibrer les classes pour permettre au modèle d'apprendre de manière plus efficace et d'améliorer la prédiction. Après l'entraînement des modèles sur ces techniques de rééquilibrage, voici les résultats obtenus sur le jeu de données de test.

Modèle avec Sous-échantillonnage

Le modèle appliqué avec la technique de sous-échantillonnage a donné les résultats suivants :

Précision : 0.3762

Cela signifie que le modèle a correctement classifié 37.62% des prédictions, en tenant compte à la fois des classes positives et négatives. Cette précision relativement faible suggère que le modèle a une tendance à effectuer un grand nombre de fausses prédictions, principalement dans la classe négative.

Rappel : 0.623

Le rappel de 62.3% montre que le modèle a réussi à détecter 62.3% des clients réellement à risque de défaut. Bien que ce soit un bon score en termes de détection des défauts, il pourrait être amélioré pour éviter que certains défauts passent inaperçus.

AUC : 0.7173

L'AUC de 0.7173 indique une capacité raisonnable du modèle à discriminer entre les clients à risque de défaut et ceux qui ne le sont pas. Bien que l'AUC soit un peu plus faible que l'idéal, elle reste suffisamment élevée pour que ce modèle soit jugé utile.

Modèle avec SMOTE

Pour le modèle utilisant la technique SMOTE, les résultats sont les suivants :

Précision : 0.147

La précision est particulièrement faible à 14.7%, ce qui signifie que le modèle a prédit de manière incorrecte une large majorité des observations. Cette faible précision peut résulter d'une mauvaise gestion de l'équilibre entre les classes après l'application de SMOTE.

Rappel : 0.4196

Avec un rappel de 41.96%, le modèle a identifié un peu moins de la moitié des clients réellement à risque de défaut. Cela montre que, malgré une précision faible, le modèle parvient à prédire un nombre raisonnable de cas positifs (clients à risque de défaut).

AUC : 0.7185

L'AUC de 0.7185 est très proche de celle obtenue avec le sous-échantillonnage, ce qui suggère que, même avec une faible précision, le modèle est capable de discriminer les classes assez efficacement.

Analyse comparative des deux modèles

Le modèle basé sur sous-échantillonnage a montré un meilleur rappel (62.3%) que celui basé sur SMOTE (41.96%), ce qui signifie qu'il est plus efficace pour identifier les clients à risque de défaut.

En revanche, la précision du modèle SMOTE est plus faible que celle du modèle sous-échantillonné, ce qui peut indiquer qu'il y a plus de faux positifs dans le modèle SMOTE, c'est-à-dire qu'il classifie à tort des clients comme étant à risque alors qu'ils ne le sont pas.

Les deux modèles ont des valeurs d'AUC proches, suggérant que leurs performances de discrimination sont similaires, bien que le modèle sous-échantillonné présente un léger avantage.

Synthèse comparative des performances

L'objectif de cette étude est d'améliorer la détection des défauts de crédit à l'aide de modèles de régression logistique, tout en optimisant la qualité des données d'entrée et en traitant les déséquilibres de classe. Plusieurs versions du jeu de données ont été testées, chacune apportant un enrichissement ou une transformation spécifique. Le tableau ci-dessous présente une comparaison synthétique des performances des modèles selon trois métriques clés : **précision**, **rappel** et **AUC (Area Under the Curve)**.

Tableau des indicateurs de performances

Version des données	Précision	Rappel	AUC
Données brutes	55,63 %	16,91 %	0,7268
Données nettoyées	78,16 %	95,62 %	0,7353
Avec indicateurs dérivés	78 %	96 %	0,7400
Variables discrétisées optimisées	78 %	96 %	0,7430
Équilibrage - Sous-échantillonnage	37,62 %	62,30 %	0,7173
Équilibrage - SMOTE	14,70 %	41,96 %	0,7185

Interprétation des résultats

Données brutes : Le modèle initial, construit sur les données sans traitement, montre une précision faible (55,63 %) et un rappel très limité (16,91 %), indiquant une mauvaise détection des défauts. L'AUC, bien qu'acceptable (0,7268), souligne un potentiel d'amélioration.

Données nettoyées : Le nettoyage (traitement des valeurs manquantes et aberrantes) permet une amélioration significative des performances, avec un bond du rappel à 95,62 %, tout en augmentant la précision à 78,16 %. L'AUC passe également à 0,7353.

Ajout d'indicateurs dérivés : La création de nouvelles variables enrichit le modèle et améliore légèrement l'AUC à 0,740, tout en conservant d'excellents niveaux de précision et de rappel.

Discrétisation optimisée : La transformation de variables continues en classes discrètes permet d'affiner encore la performance du modèle, atteignant un AUC de 0,743, le meilleur obtenu dans cette analyse, avec une précision stable (78 %) et un rappel élevé (96 %).

Équilibrage (Sous-échantillonnage et SMOTE) : Ces méthodes visent à corriger le déséquilibre de la variable cible. Bien qu'elles améliorent le rappel (62,30 % pour le sous-échantillonnage, 41,96 % pour SMOTE), la précision chute considérablement (jusqu'à 14,70 % avec SMOTE). Ces techniques sont utiles lorsque la détection de la classe minoritaire est prioritaire, mais leur application nécessite un compromis sur la qualité globale des prédictions.

Conclusion et Recommandations

Cette étude nous a permis de mettre en lumière les facteurs déterminants du comportement de crédit des clients d'une banque à travers une démarche rigoureuse, allant de l'exploration des données brutes à l'élaboration de modèles prédictifs performants. L'analyse statistique et la modélisation ont révélé que des variables sociodémographiques (âge, sexe, situation familiale), financières (montants et durées de crédits), ainsi que comportementales (canal de souscription, ancienneté des crédits) influencent significativement la probabilité qu'un client obtienne ou soit en défaut de crédit.

Les modèles construits, notamment ceux intégrant des variables discrétisées et enrichies, affichent des performances satisfaisantes, avec des taux de précision avoisinant les 78 %, un rappel élevé (jusqu'à 96 %) et une AUC dépassant 0,74. Ces résultats montrent l'intérêt d'un prétraitement minutieux des données et d'un enrichissement par des indicateurs métiers. Toutefois, les techniques de rééquilibrage comme le SMOTE, bien qu'utiles pour améliorer le rappel dans un contexte de classes déséquilibrées, peuvent fortement dégrader la précision.

Recommandations :

- Poursuivre l'optimisation du modèle par des approches avancées (arbres de décision, forêts aléatoires, gradient boosting) pour capter des interactions non linéaires.
- Intégrer des données comportementales dynamiques (transactions, navigation en ligne) pour affiner les prédictions.
- Envisager une segmentation client pour adapter les modèles à des profils spécifiques.

- Utiliser les modèles comme support à la prise de décision, tout en gardant une évaluation humaine pour les cas limites.

Ainsi, cette démarche constitue une base solide pour renforcer la gestion du risque de crédit et cibler plus efficacement les actions commerciales de la banque.