UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Modélisation du risque de défaut sur les prêts hypothécaires

Bethuel ASSE - Gaétan DUMAS - Dimitri GUIFT - Pierre LIBERGE

Université Paris 1 Panthéon-Sorbonne
Master 2 MoSEF

27 Octobre 2024

Table des matières

1 Introduction

1.1 Contexte

Dans le domaine des prêts hypothécaires, l'analyse et la prédiction du risque de défaut de paiement sont des enjeux majeurs pour les institutions financières. Un prêt hypothécaire, ou prêt immobilier, est accordé pour l'acquisition d'un bien immobilier et est sécurisé par la propriété acquise. En cas de non-remboursement, le prêteur peut saisir et vendre ce bien pour récupérer les fonds prêtés, un mécanisme essentiel pour mieux gérer les risques associés à ce type de crédit.

L'évaluation du crédit bancaire permet de mesurer la capacité de remboursement d'un client à partir de son historique financier et de ses comportements en matière de crédit. Grâce à cette évaluation, les institutions financières peuvent adapter les conditions des prêts, notamment les taux d'intérêt, en fonction du profil de risque du client. En s'appuyant sur des critères objectifs et transparents, ce processus d'évaluation garantit une approche équitable et impartiale, où chaque client est évalué en fonction de sa situation financière réelle.

L'objectif principal de ce projet est de construire un modèle prédictif de la variable cible BAD, qui indique le risque de défaut de paiement d'un emprunteur. Un tel modèle permet d'anticiper les situations de non-remboursement et de renforcer les décisions de prêt pour limiter les risques financiers des prêteurs.

1.2 Description de la base de données

La base de données nommée HMEQ contient des informations sur 5960 prêts hypothécaires (home equity loans). Ces prêts permettent à l'emprunteur d'utiliser la valeur nette de son logement comme garantie. Les variables de cette base sont les suivantes :

- **BAD** : Variable dichotomique indiquant le statut de défaut de paiement, où 1 signifie que l'emprunteur a fait défaut et 0 qu'il n'est pas en défaut.
- **LOAN** : Montant du prêt demandé.
- **MORTDUE** : Montant restant dû sur le prêt hypothécaire principal.
- **VALUE** : Valeur actuelle de la propriété ayant servi de garantie
- **REASON** : Motif de la demande de prêt, qui peut être l'une des deux catégories suivantes :
  - **DebtCon** : Consolidation de dettes, c'est-à-dire le regroupement de plusieurs dettes en une seule.
  - **HomeImp** : Amélioration de l'habitat, visant à financer des rénovations ou réparations de la résidence.
- **JOB** : Catégorie professionnelle de l'emprunteur.
- **YOJ** : Nombre d'années d'ancienneté à l'emploi actuel de l'emprunteur.
- **DEROG** : Nombre d'incidents financiers graves, comme des saisies ou des faillites, dans l'historique de crédit.
- **CLNO** : Nombre total de lignes de crédit ouvertes par l'emprunteur.
- **DELINQ** : Nombre de lignes de crédit présentant des retards de paiement.
- **CLAGE** : Âge (en mois) de la ligne de crédit la plus ancienne.
- **NINQ** : Nombre de demandes de crédit récentes effectuées par l'emprunteur.
- **DEBTINC** : Ratio dette/revenu, calculé comme le rapport entre les paiements

mensuels de la dette et le revenu mensuel.

Ces variables fourniront des informations clés pour élaborer un modèle prédictif de la variable cible BAD, permettant ainsi d'améliorer notre compréhension des facteurs de risque associés au défaut de paiement.

1.3 Plan d'action et organisation du rapport

Dans les sections suivantes, nous commencerons par présenter notre méthodologie pour la préparation et le nettoyage des données, étape essentielle pour garantir la fiabilité de l'analyse. Nous poursuivrons avec une exploration approfondie des variables clés, visant à mieux comprendre les dynamiques et interactions présentes dans le jeu de données. Ensuite, nous développerons et calibrerons des modèle prédictifs pour identifier les facteurs de risque associés au défaut de paiement. Enfin, nous évaluerons leurs performances en utilisant différentes métriques d'évaluation.

2 Exploration et traitement des données

2.1 Exploration des données

2.1.1 Structure de la base de données et de la variable cible

La base de données utilisée pour cette analyse contient un total de 5960 observations, représentant des prêts hypothécaires. Chaque observation de la base de données représente un prêt individuel avec ses caractéristiques spécifiques et l'information relative au prêt. La variable cible de notre projet est la variable BAD, qui est binaire et indique l'état de défaut du prêt : 1 pour un prêt défaillant (défaut de paiement de l'emprunteur) et 0 pour un prêt non défaillant. Dans notre base de données, nous observons que 1189 prêts sont en défaut (BAD=1) et 4771 prêts ne le sont pas (BAD=0). Cela représente une distribution avec environ 20% des prêts en défaut et 80% des prêts sans défaut.

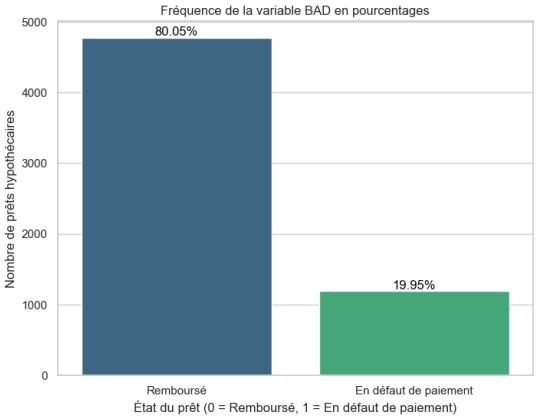


FIGURE 1 – Fréquence de la variable BAD en pourcentages

L'objectif de notre analyse est de modéliser la probabilité de défaut de paiement d'un emprunteur quant à un prêt, en utilisant les différentes variables explicatives fournies dans la base de données. Ces variables incluent, entre autres, le montant du prêt (LOAN), la valeur de la propriété (VALUE), le montant dû sur l'hypothèque existante (MORTDUE),

et d'autres informations liées au crédit et à l'emploi de l'emprunteur. Cet aperçu statistique initial est crucial pour comprendre la composition de notre base de données et pour guider les choix de modélisation et les techniques d'analyse qui seront appliqués pour prédire les risques de défaut de paiement.

2.1.2 Insights préliminaires sur la variable cible

Avant d'aborder les étapes de prétraitement et de modélisation, il est essentiel de comprendre les relations initiales entre la variable cible, BAD, et les variables explicatives. Cette analyse préliminaire permet de mettre en lumière les facteurs potentiellement influents et d'orienter les choix méthodologiques pour la suite du projet. Nous avons trouvé pertinent de commenter ce graphique qui montre les corrélations entre la variable cible BAD et les variables explicatives numériques.

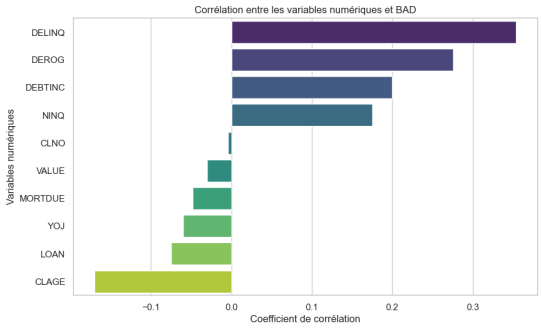


FIGURE 2 – Corrélations préliminaires des variables numériques avec le statut de défaut

Les variables DELINQ (nombre de retards de paiement), DEROG (nombre d'incidents financiers graves) et DEBTINC (ratio dette/revenu) montrent les corrélations positives les plus fortes avec BAD, ce qui est économiquement prévisible : des retards, incidents financiers et un fort endettement augmentent la probabilité de défaut. CLAGE (ancienneté des comptes) présente une légère corrélation négative avec BAD, suggérant que des comptes plus anciens sont liés à une moindre probabilité de défaut, probablement grâce à l'expérience de gestion du crédit des clients.

En revanche, CLNO (nombre total de lignes de crédit) n'a quasiment pas de lien avec BAD, indiquant que le nombre de lignes ouvertes n'influence pas directement le risque de défaut sans considération de la gestion de ces lignes. Enfin, LOAN (montant du prêt demandé) présente une faible corrélation négative avec BAD. Bien que contre-intuitif, cela pourrait refléter le fait que les emprunteurs sollicitant des montants élevés tendent à être financièrement plus stables, ou qu'ils sont soumis à des critères d'évaluation plus stricts de la part des prêteurs.

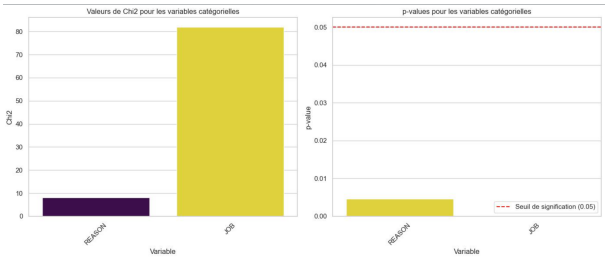


FIGURE 3 – Test de χ^2 pour les variables catégorielles : analyse d'association avec le défaut

En examinant les variables catégorielles, nous observons que le graphique précédent présente les valeurs du test du χ^2 ainsi que les p-values associées pour les variables catégorielles REASON et JOB par rapport à la variable cible BAD. La valeur de χ^2 pour la variable JOB est significativement plus élevée que celle de REASON, ce qui indique une association plus forte entre JOB et BAD. De plus, les p-values des deux variables sont inférieures au seuil de significativité $\alpha = 0,05$, ce qui suggère que REASON et JOB sont statistiquement significatives (au seuil α) pour expliquer la probabilité de défaut. Ces résultats justifient leur inclusion dans la modélisation, avec une attention particulière portée à JOB en raison de sa contribution plus importante.

2.1.3 Valeurs manquantes

Notre base de données présente un ensemble de défis liés aux valeurs manquantes, dont l'ampleur varie considérablement d'une variable à l'autre, nécessitant une attention particulière lors du prétraitement des données pour la modélisation.

Certaines variables, comme JOB et REASON, ont des taux de valeurs manquantes relativement faibles. Ce taux modéré, conforme aux tendances observées dans d'autres ensembles de données financiers, indique que la majorité des emprunteurs fournissent ces informations.
En revanche, d'autres variables, telles que DEROG et DEBTINC, présentent des pourcentages de valeurs manquantes nettement plus élevés. Cette situation est particulièrement préoccupante, car ces variables montrent de fortes corrélations avec BAD (la probabilité de défaut) et peuvent être cruciales pour évaluer le risque de défaut d'un emprunteur. En effet, un nombre élevé d'incidents financiers graves (DEROG) et un ratio dette/revenu important (DEBTINC) sont des indicateurs clés de l'instabilité financière, et leur absence dans les données pourrait nuire à la précision du modèle.

Le graphique suivant illustre la distribution des valeurs manquantes dans notre dataset, mettant en évidence les variables les plus problématiques en termes de données incomplètes :

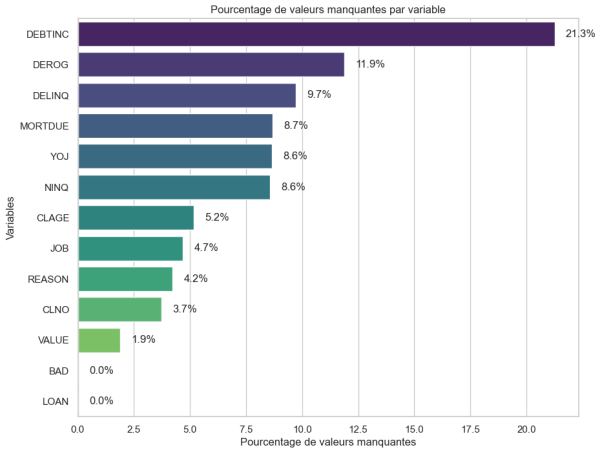


FIGURE 4 – Distribution des valeurs manquantes par variable

Seules les variables cible et explicative LOAN ne présentent aucune valeur manquante.

2.1.4 Tests de normalité

Les tests de normalité jouent un rôle crucial dans la détection et le traitement des données aberrantes (outliers). En effet, certains modèles statistiques et méthodes de prétraitement, comme l'imputation par moyenne ou la détection d'outliers via les Z-scores, supposent que les variables suivent une distribution normale. Avant d'appliquer ces techniques, il est donc nécessaire de vérifier la normalité des variables à l'aide de tests statistiques appropriés.

Nous avons appliqué trois tests de normalité sur nos variables :

— Test de Kolmogorov-Smirnov
— Test de Shapiro-Wilk
— Test de Jarque-Bera

Ces tests partagent l'hypothèse nulle (H_0) selon laquelle les variables suivent une distribution normale. L'hypothèse alternative (H_1) stipule que la distribution de la variable diffère significativement d'une distribution normale.

Interprétation de la p-value : Si la p-value est inférieure à 0,05, nous rejeterons l'hypothèse nulle (H_0) au profit de l'hypothèse alternative (H_1). Cela signifie que la variable testée ne suit pas une distribution normale. Dans notre cas, comme toutes les p-values sont bien en dessous de 0,05 et très proches de zéro, l'hypothèse de normalité est rejetée pour toutes les variables.

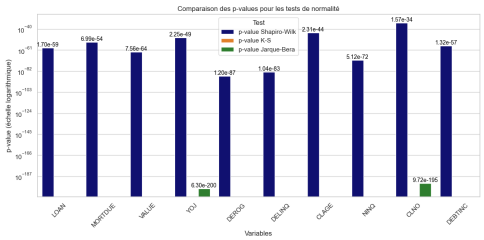


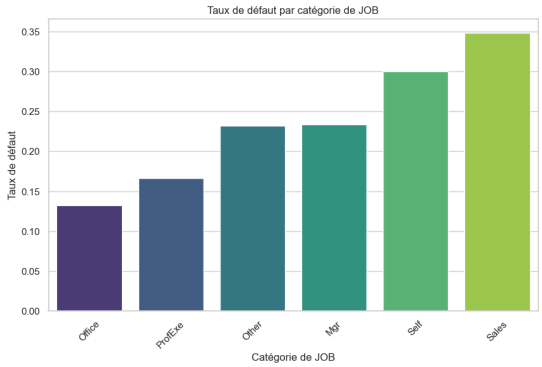
FIGURE 5 – Distribution des valeurs manquantes par variable

Les trois tests confirment que les variables analysées ne sont pas normalement distribuées.

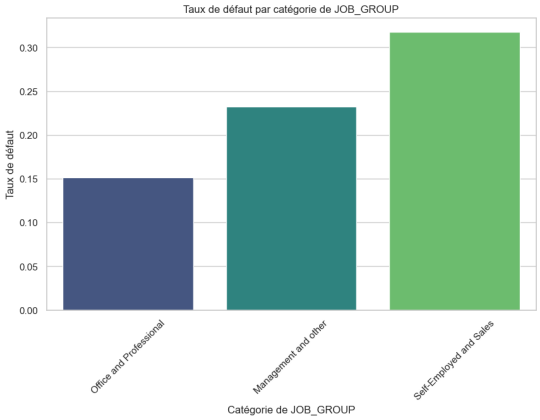
2.2 Préparation des données

2.2.1 Regroupement des modalités de variables catégorielles

Le regroupement des modalités de la variable catégorielle 'JOB' a été envisagé pour simplifier le modèle et potentiellement améliorer sa performance. En analysant les taux de défaut par catégorie, il apparaît que certaines professions affichent des taux similaires, suggérant une opportunité de réduction du nombre de catégories. L'axe des ordonnées du graphique de gauche ci-dessous représentent les taux de défaut moyen pour chaque modalité de la variable JOB. Ce taux représente la proportion de prêts qui ont abouti à un défaut (BAD = 1) parmi tous les prêts accordés aux individus d'une même catégorie professionnelle . Voici comment l'on pourrait interpréter l'un de ces taux : 30.05% des prêts où le demandeur est auto-entrepreneur (Self ou Self-employed) ont abouti à un défaut. Le graphique de droite montre le regroupement effectué en fonction de la similitude entre taux de défaut moyen.



(a) Taux de défaut moyen par catégorie de JOB avant regroupement



(b) Taux de défaut moyen par catégorie de JOB après regroupement

FIGURE 6 – Comparaison des taux moyens de défaut par catégorie de JOB avant et après regroupement

Cette stratégie de regroupement peut réduire le bruit dans les données, minimiser le risque de surajustement (overfitting) et faciliter l'interprétation des résultats, tout en rendant le modèle potentiellement plus robuste pour les nouvelles applications.

2.2.2 Séparation du jeu de données

La séparation du jeu de données en ensembles d'entraînement et de test a été réalisée avant toute transformation, en suivant une répartition classique de 80% pour l'entraînement et 20% pour le test. Cette méthode est largement adoptée pour évaluer de manière fiable la capacité du modèle à généraliser sur des données non vues. Le `train_test_split` de la librairie `scikit-learn` a été utilisé avec un `random_state` fixe pour garantir la reproductibilité des résultats. Cette approche prévient le risque de fuite d'information entre les ensembles de données et assure que les évaluations de performance reflètent véritablement l'efficacité du modèle dans des conditions réalistes.

2.2.3 Encodage des variables catégorielles

Le OneHotEncoder de scikit-learn a été utilisé pour transformer les variables catégorielles en variables indicatrices. Cette technique permet de convertir les variables catégorielles en un format numérique compatible avec les algorithmes de machine learning et d'éviter toute hiérarchisation artificielle que pourrait induire un encodage ordinal. En outre, cela facilite la modélisation en traitant chaque catégorie comme une variable distincte, ce qui améliore la capacité du modèle à apprendre des distinctions fines entre les différentes catégories. Cette méthode est particulièrement efficace pour les variables catégorielles avec un nombre limité de catégories, d'où le regroupement des modalités qu'on a effectué plus haut.

2.2.4 Traitement des valeurs manquantes

Dans cette section, nous avons appliqué différentes techniques d'imputation pour gérer les valeurs manquantes dans notre dataset. Trois approches principales ont été utilisées, chacune adaptée aux types de variables concernées et à leur distribution :

- **Imputation par la mode pour les variables catégorielles**
  Les variables catégorielles **REASON** et **JOB_GROUP** ont été imputées par la valeur la plus fréquente (mode). L'imputation par la mode est justifiée ici car, pour des variables catégorielles, il est courant de remplacer les valeurs manquantes par la modalité la plus répandue, réduisant ainsi l'impact de la perte d'informations tout en maintenant la cohérence des données.
- **Imputation par la médiane pour certaines variables continues**
  Les variables continues **DEBTINC**, **MORTDUE**, et **VALUE** ont été imputées par la médiane. Cette méthode est préférée à la moyenne pour les variables dont la distribution est asymétrique ou qui présentent des outliers. La médiane, étant moins sensible aux valeurs extrêmes, permet d'éviter les biais que l'imputation par la moyenne pourrait introduire dans de tels cas.
- **Imputation par MICE (Multiple Imputation by Chained Equations)**
  Pour les autres variables continues (**DEROG**, **DELINQ**, **YOJ**, **CLAGE**, **NINQ**, **CLNO**), nous avons utilisé l'algorithme **MICE**. Cette méthode d'imputation itérative permet d'imputer les valeurs manquantes en prenant en compte les relations entre les différentes variables du dataset. L'itération multiple permet de modéliser les incertitudes liées aux valeurs manquantes en utilisant les informations disponibles dans les autres variables, offrant ainsi une meilleure estimation des valeurs manquantes.

2.2.5 Création de nouvelles variables

Pour améliorer la qualité de l'analyse et la performance du modèle, deux nouvelles variables ont été créées :

- **LOAN_VALUE_ratio** : Ce ratio entre le montant du prêt (**LOAN**) et la valeur de la propriété (**VALUE**) capture le niveau relatif d'endettement de l'emprunteur. Un ratio élevé peut indiquer un endettement plus important, signalant potentiellement une plus grande difficulté de remboursement. Nous avons par la suite supprimé les variables initiales **LOAN** et **VALUE** pour éviter la multicolinéarité entre elles et **LOAN_VALUE_ratio**.

— **DEBTINC_missing** : Cette variable binaire indique la présence de valeurs manquantes pour **DEBTINC**. L'idée est de tester si le fait qu'une valeur soit manquante est en soi révélateur de comportements associés au risque de défaut. Nous avons pensé qu'une absence de données sur le ratio d'endettement pourrait parfois être volontaire de la part d'un emprunteur, reflétant ainsi une situation financière potentiellement instable ou un profil plus risqué. Cette absence pourrait donc servir d'indicateur supplémentaire dans l'évaluation du risque de défaut.

3 Régression logistique

3.1 Pré-traitement spécifique pour le modèle Logit

3.1.1 Vérification des hypothèses

La régression logistique repose sur plusieurs hypothèses fondamentales qui garantissent la validité de l'interprétation des coefficients et la robustesse des résultats obtenus. Les hypothèses principales incluent : (1) **la relation linéaire entre les variables prédictives continues et le log-odds de la variable cible**, (2) **l'indépendance des observations**, (3) **l'absence de multicolinéarité excessive entre les variables prédictives**, et (4) **l'absence de valeurs aberrantes influentes** qui pourraient fausser les résultats. Nous supposons dans notre étude que l'hypothèse (2) est valide. Vérifier ces hypothèses est essentiel pour s'assurer que le modèle répond aux exigences théoriques de la régression logistique, ce qui assure la fiabilité des coefficients ainsi que la qualité de nos prédictions.

3.1.1.1 Hypothèse de linéarité

L'hypothèse de linéarité postule que chaque variable prédictive continue doit entretenir une relation linéaire avec le log-odds de la variable cible pour que les coefficients du modèle de régression logistique soient interprétables et que les résultats soient fiables. Elle a été testée en appliquant le test de Box-Tidwell. Dans le cadre du test de Box-Tidwell, l'hypothèse de linéarité entre chaque variable continue x et le *log-odds* de la variable cible est testée en introduisant un terme d'interaction sous la forme de $x \times \log(x)$. Ce terme permet de vérifier si la relation entre x et le *log-odds* est bien linéaire.

Une **p-value élevée** (≥ 0.05) associée à ce terme d'interaction indique que la relation linéaire est respectée, tandis qu'une **p-value faible** suggère une déviation de la linéarité, nécessitant potentiellement une transformation ou une discrétisation de la variable pour améliorer l'ajustement du modèle.

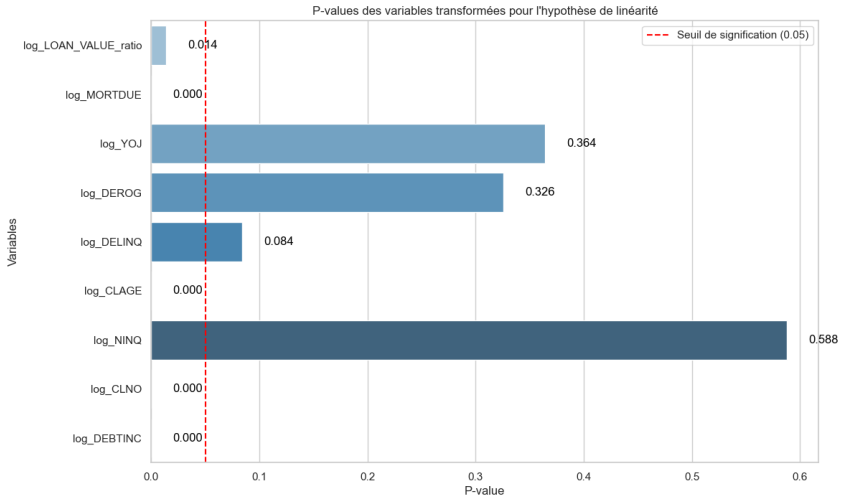


FIGURE 7 – P-values des variables du test de Box-Tidwell

Seules les variables **YOJ**, **DEROG**, **DELINQ** et **NINQ** respectaient notre hypothèse tandis que les autres l'a violaient. Pour gérer ces violations de linéarité, nous avons adopté une approche hybride en fonction de la nature des variables :

- **Les variables qui ne respectaient pas l'hypothèse de linéarité** ont été discrétisées en utilisant la méthode du Weight of Evidence (WoE) à l'aide de la bibliothèque optbinning. Le WoE attribue à chaque catégorie une valeur numérique : si le WoE est positif, cela signifie que cette catégorie a un risque plus élevé de défaut par rapport à la moyenne ; si le WoE est négatif, cela indique un risque de défaut plus faible. En résumé, plus le WoE est élevé, plus la catégorie est associée au risque de défaut, et inversement. Cela a permis de capturer les relations non linéaires sans recourir à des transformations plus complexes.
- **Les variables continues qui respectaient l'hypothèse de linéarité** ont été standardisées à l'aide de StandardScaler pour améliorer la stabilité et la convergence du modèle.
- **Les variables binaires** ont été laissées telles quelles, car elles ne nécessitent ni transformation ni standardisation.

3.1.1.2 Détection et traitement des points influents

Dans un premier temps, nous avons choisi de ne pas supprimer ni de transformer directement ces points influents et de procéder à la modélisation de notre variable cible malgré leur présence. Cette décision repose sur plusieurs considérations méthodologiques.

Dans un second temps, pour évaluer l'impact potentiel de ces points influents, nous avons adopté une approche complémentaire en les supprimant avant de relancer la modélisation et d'observer les variations de performance du modèle ainsi que son ajustement.

Les résultats de ces deux approches seront présentés dans la sous-section **3.2.2** ≪**Modélisation et évaluation du modèle** ≫.

3.1.1.3 Application de la méthode de sur-échantillonnage SMOTE

Initialement, nous avons construit un modèle simple pour observer les performances du modèle sur les données d'origine et évaluer la capacité de celui-ci à traiter la classe minoritaire (qui est de 20%). Cela devait nous permettre de mieux comprendre les biais

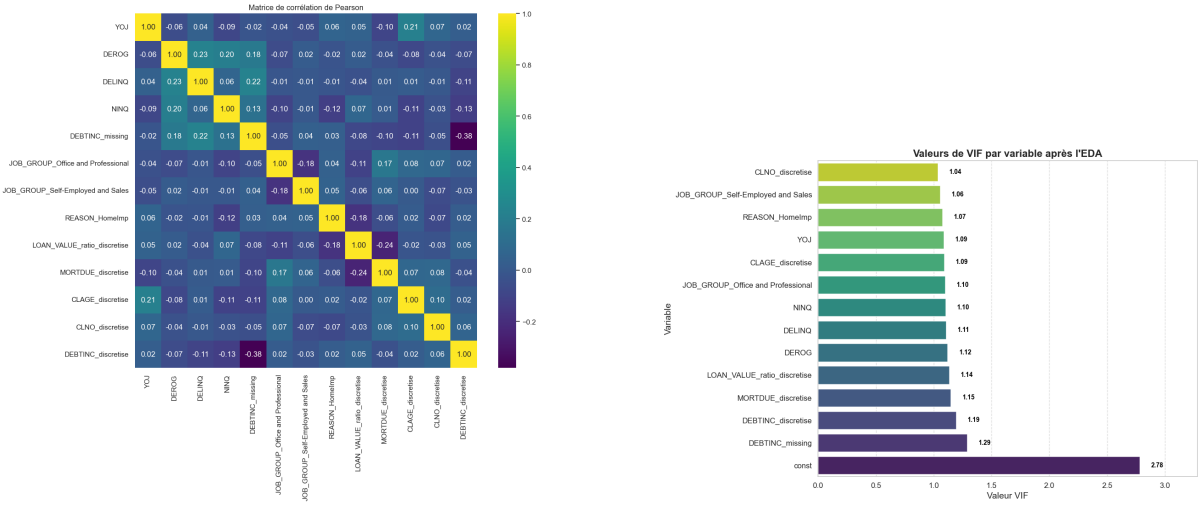
potentiels du modèle et d'évaluer les performances sans ajustements artificiels. Cependant, en raison du déséquilibre dans notre variable cible, ce premier modèle pouvait présenter un biais en faveur de la classe majoritaire, risquant de sous-estimer les défauts de paiement.

C'est la raison pour laquelle nous avons ensuite décidé d'appliquer la méthode SMOTE (Synthetic Minority Over-sampling Technique), une méthode de sur-échantillonnage pour équilibrer les classes et construire un modèle capable de mieux prédire les observations de la classe minoritaire. L'approche SMOTE génère de nouvelles instances synthétiques pour la classe minoritaire, augmentant ainsi sa représentativité et favorisant un apprentissage plus équilibré. L'application de SMOTE s'est déroulée avant la phase de modélisation et s'est effectuée uniquement sur l'ensemble d'entraînement pour garantir que le modèle n'ait pas accès aux données synthétiques lors de l'évaluation, préservant ainsi l'intégrité et la validité des résultats de test.

La comparaison des deux modèles nous a permis d'évaluer l'impact du déséquilibre de classe sur les performances globales et d'optimiser notre prédiction en choisissant l'approche la plus adaptée.

3.1.1.4 Absence de multicolinéarité

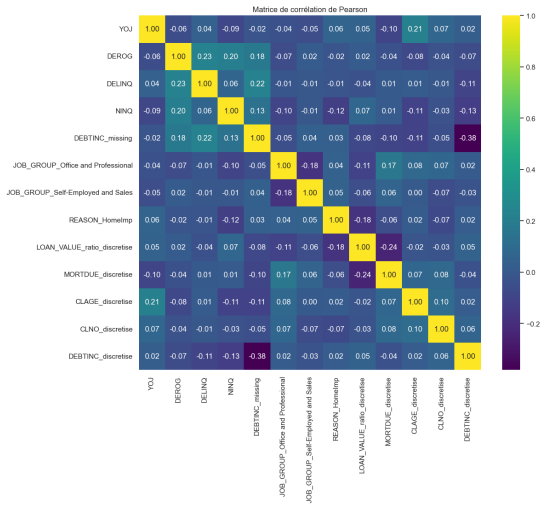
La multicolinéarité entre les variables prédictives a été évaluée à l'aide du facteur d'inflation de la variance (VIF), et aucune corrélation excessive n'a été observée entre les variables incluses dans le modèle. En général, des valeurs de VIF supérieures à 5 indiqueraient une multicolinéarité préoccupante, mais ici, toutes les valeurs sont proches de 1, ce qui confirme l'indépendance relative des variables et suggère que les estimations des coefficients ne sont pas biaisées par des redondances entre variables.



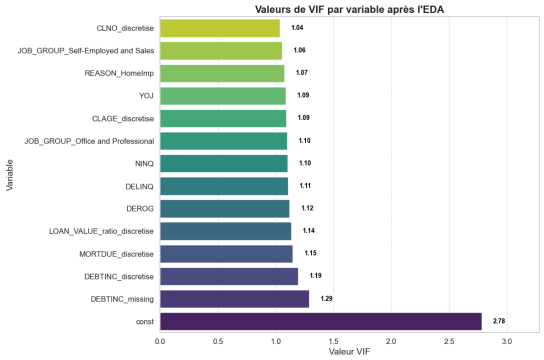
(a) Matrice de corrélation de Pearson après l'EDA

(b) Valeur de la VIF par variable après l'EDA

FIGURE 8 – Analyse de la multicolinéarité



(a) Matrice de corrélation de Pearson après l'EDA (avec SMOTE)



(b) Valeur de la VIF par variable après l'EDA (avec SMOTE)

FIGURE 9 – Analyse de la multicolinéarité (avec SMOTE)

3.2 Modélisation sans SMOTE

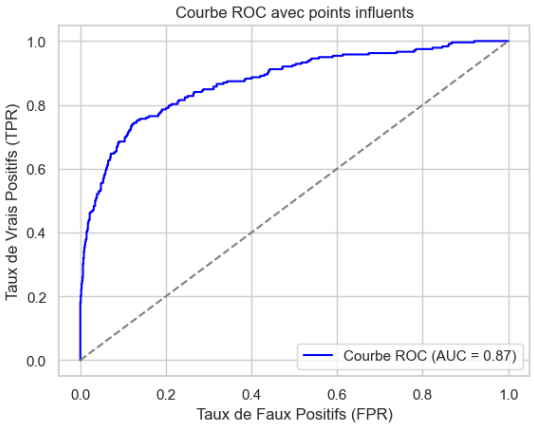
3.2.1 Sélection de variables via LASSO

Dans cette partie, nous avons utilisé la régularisation L1 (Lasso) pour sélectionner les variables les plus pertinentes afin de réduire le modèle tout en conservant une performance optimale. L'algorithme Lasso met à zéro les coefficients des variables moins importantes, facilitant ainsi la sélection automatique des variables les plus influentes. Pour affiner le modèle, nous avons appliqué une recherche en grille (GridSearchCV) afin de déterminer la meilleure valeur du paramètre de régularisation C, garantissant un équilibre entre biais et variance. La liste finale des variables sélectionnées est donc composée de celles ayant des coefficients significativements non nuls, reflétant leur importance pour la prédiction de la variable cible.

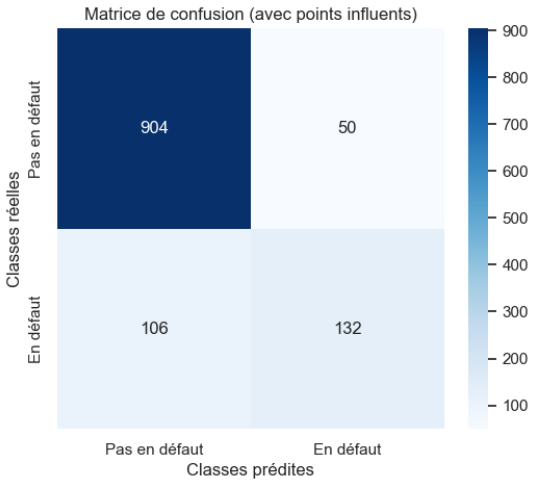
3.2.2 Modélisation et évaluation du modèle

3.2.2.1 Avec présence des points influents

Après la sélection des variables, une régression logistique a été appliquée sur le dataset d'entraînement sans supprimer les points influents, afin d'évaluer leur impact sur les performances du modèle. Les prédictions ont ensuite été testées sur l'ensemble de test, et les métriques telles que l'accuracy, le F1 score et l'AUC-ROC ont été calculées pour mesurer la performance.



(a) Courbe ROC-AUC avec présence de points influents



(b) Matrice de confusion avec présence de points influents

FIGURE 10 – Métriques du modèle Logit avec points influents

Ce rapport de classification révèle des performances satisfaisantes pour le modèle de prédiction de risque de crédit. L'accuracy de 86,9 % montre que le modèle classifie correctement les observations dans 87 % des cas. Cependant, en se focalisant sur la classe d'intérêt (1), c'est-à-dire les emprunteurs en défaut, on observe une précision (precision) de 73 % et un rappel (recall) de 55 %. Cela signifie que, parmi les emprunteurs prévus en défaut, 73 % sont réellement en défaut, mais le modèle ne détecte que 55 % des cas de défaut dans l'ensemble des emprunteurs réellement en défaut (cf A.1.1).

Le F1-score de 0,63 pour la classe de défaut montre que le modèle pourrait encore être amélioré pour mieux détecter les emprunteurs à risque sans augmenter significativement les faux positifs. L'AUC-ROC de 0,87 indique une bonne capacité générale du modèle à distinguer entre les classes « défaut » et « non-défaut », ce qui est essentiel pour une gestion efficace du risque de crédit.

Pour évaluer plus en profondeur la qualité du modèle, nous avons également effectué un test d'ajustement. Le test de Hosmer-Lemeshow, en particulier, permet de vérifier si les probabilités prédites par le modèle sont bien calibrées par rapport aux observations réelles, en mesurant la concordance entre les valeurs attendues et observées dans différents sous-groupes de la population. Les résultats du test sont consignés dans le graphe suivant :

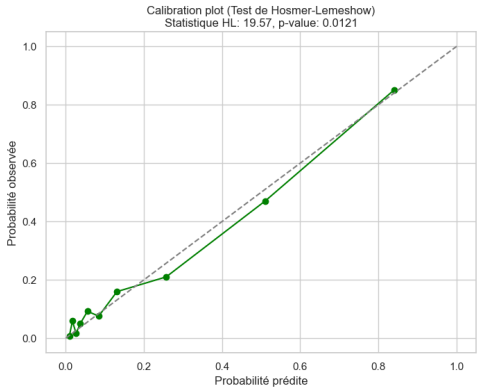
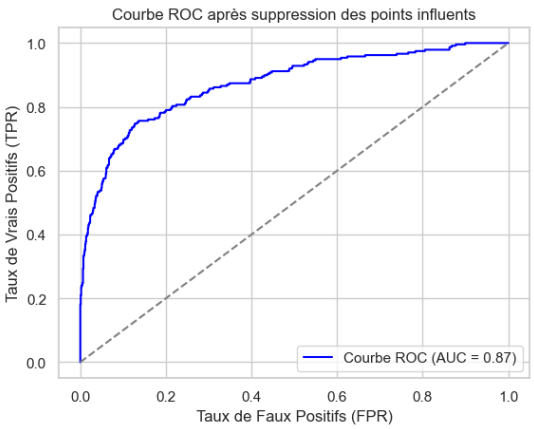


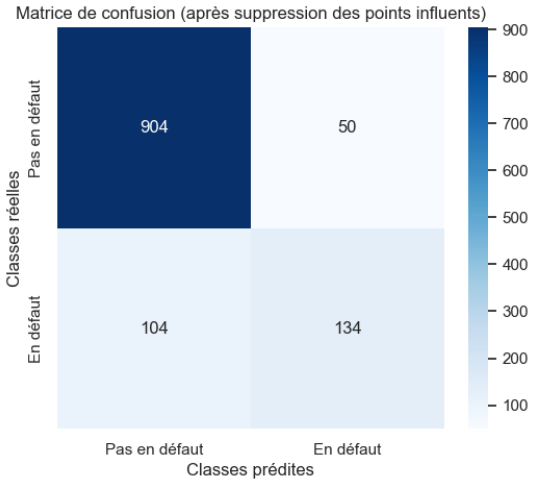
FIGURE 11 – Test de Hosmer-Lemeshow pour l'ajustement du modèle

L'ajustement du modèle, représenté par une statistique de Hosmer-Lemeshow de 19.57 et une p-value de 0.0121 donc inférieure à 0.05, indique que les probabilités prédites s'écartent des observations réelles, suggérant un ajustement imparfait. Face à ce manque d'ajustement, nous avons ensuite choisi d'effectuer la modélisation sans les points influents pour observer si cela améliore la calibration du modèle.

3.2.2.2 Après suppression des points influents



(a) Courbe ROC-AUC après suppression des points influents



(b) Matrice de confusion après suppression des points influents

FIGURE 12 – Métriques du modèle Logit après suppression des points influents

Après suppression des points influents, les métriques de performance montrent un léger ajustement du modèle. L'accuracy reste à 0,87, indiquant une bonne proportion de prédictions correctes, tandis que le F1 score de 0,64 pour la classe "défaut" montre une légère amélioration dans l'équilibre entre rappel et précision pour cette classe. L'AUC-ROC de 0,87 reflète une bonne capacité de discrimination globale (cf A.1.2).
La matrice de confusion révèle une légère réduction des faux positifs et faux négatifs, ce qui indique une meilleure stabilité du modèle sans points influents.

Enfin, le test de Hosmer-Lemeshow donne une p-value de 0,6088 ($\geq$ 0.05), ce qui signifie que le modèle est finalement bien ajusté aux données après suppression des points influents.

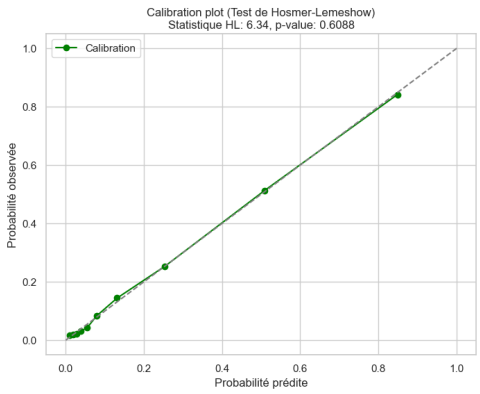


FIGURE 13 – Test de Hosmer-Lemeshow après suppression des points influents

3.2.2.3 Bilan de la modélisation Logit : avec points influents vs sans points influents

En somme, la suppression des points influents n'a pas radicalement changé les performances de classification du modèle vu que les métriques ont très légèrement augmenté, mais elle a amélioré son ajustement aux données (meilleure calibration des prédictions). Cela suggère que, bien que les points influents n'impactent pas fortement les principales métriques de performance, ils peuvent introduire des anomalies dans l'ajustement global du modèle. Pour une application dans le cadre du risque de crédit, l'option sans points influents offre un modèle plus calibré et potentiellement plus fiable pour les décisions.

3.3 Modélisation avec SMOTE

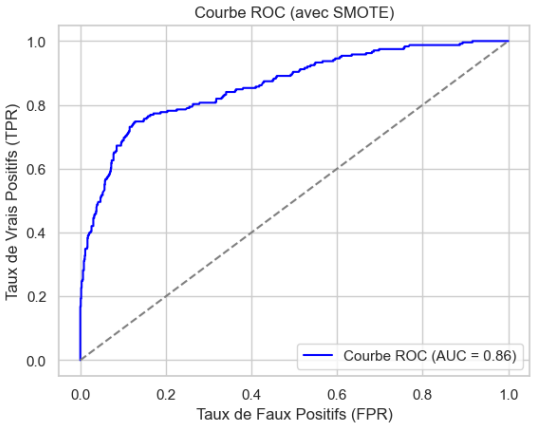
Pour cette modélisation Logit avec sur-échantillonnage, les points influents ont été retirés dès le prétraitement, car leur maintien aurait conduit SMOTE à amplifier ces valeurs extrêmes, augmentant ainsi le bruit dans les données synthétisées. En les excluant, nous avons pu garantir un sur-échantillonnage plus représentatif et minimiser l'influence des anomalies.

3.3.1 Sélection de variables avec LASSO

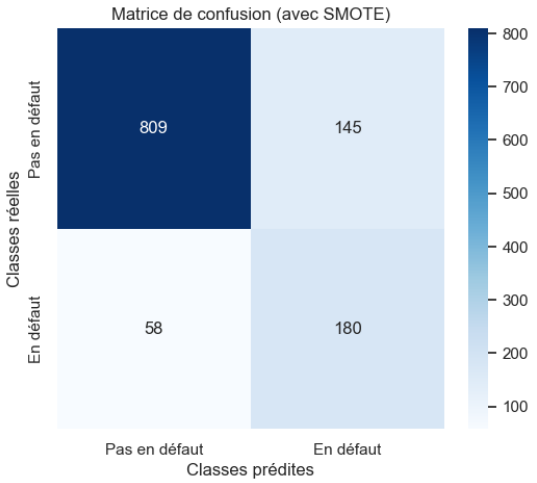
La sélection des variables pour le modèle Logit avec SMOTE suit exactement la même méthode de régularisation L1 (Lasso) que celle utilisée sans sur-échantillonnage. Cependant, contrairement au modèle sans SMOTE, le modèle avec sur-échantillonnage a retenu uniquement un sous-ensemble de variables, excluant celles jugées moins pertinentes.

3.3.2 Modélisation et évaluation du modèle

L'AUC de 0.86 montre une bonne capacité du modèle à différencier entre les emprunteurs en défaut et ceux qui ne le sont pas. Toutefois, elle est légèrement inférieure à celle du modèle sans sur-échantillonnage, indiquant que SMOTE a amélioré la balance des classes mais avec un léger impact sur cette capacité de distinction (cf A.2). Quant à la matrice de confusion, on observe une bonne précision pour la classe majoritaire (pas de défaut) avec 809 vraies prédictions, mais aussi un nombre plus élevé de fausses prédictions (145) comparé au modèle sans SMOTE. La classe minoritaire (En défaut) est mieux capturée avec 180 bonnes prédictions et seulement 58 erreurs. SMOTE a donc aidé à mieux identifier les cas des emprunteurs en défaut en comparaison au modèle sans sur-échantillonnage, mais au prix d'une augmentation des erreurs pour la classe majoritaire des emprunteurs solvables.



(a) Courbe ROC-AUC du modèle Logit avec SMOTE



(b) Matrice de confusion du modèle Logit avec SMOTE

FIGURE 14 – Métriques du modèle Logit avec sur-échantillonnage (SMOTE)

Quant au test de Hosmer-Lemeshow, la p-value de 0.4896 indique un bon ajustement du modèle aux données, ce qui confirme que l'introduction de SMOTE n'a pas dégradé la calibration du modèle. La probabilité observée est proche de la probabilité prédite, montrant que le modèle est bien calibré après sur-échantillonnage.

En conclusion, l'utilisation de SMOTE a permis une meilleure détection des cas en défaut (classe minoritaire), tout en maintenant une bonne performance générale, bien que légèrement moins précise pour la distinction entre classes. L'ajustement reste adéquat selon le test de Hosmer-Lemeshow, ce qui confirme la stabilité du modèle même après rééquilibrage des classes.

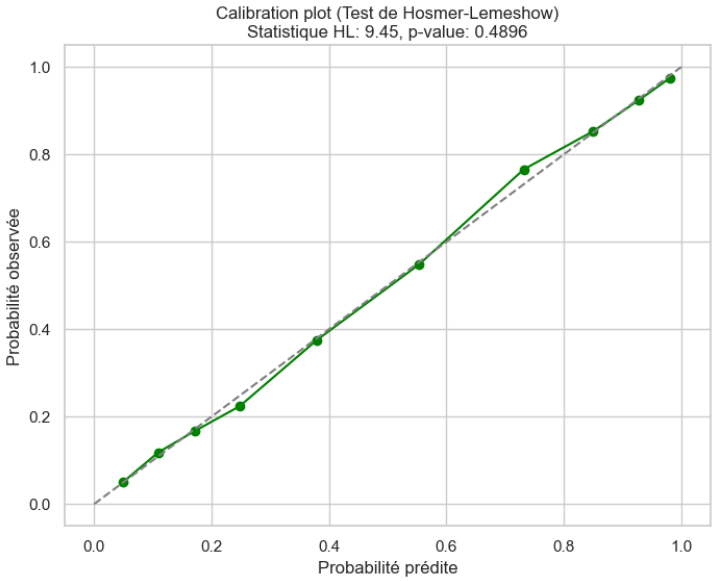


FIGURE 15 – Test de Hosmer-Lemeshow avec sur-échantillonnage (SMOTE)

3.4 Comparaison des trois approches de modélisation

- **Modélisation sans suréchantillonnage avec points influents** : Cette approche a montré des performances globalement bonnes avec un AUC de 0.87, mais un test d'ajustement de Hosmer-Lemeshow indiquant un léger manque de calibration. La présence des points influents a vraisemblablement introduit un certain bruit dans les prédictions, rendant le modèle moins robuste.
- **Modélisation sans suréchantillonnage après suppression des points influents** : En retirant les points influents, nous avons obtenu des métriques de performance similaires (AUC toujours à 0.87), mais une meilleure calibration selon le test de Hosmer-Lemeshow. Cela indique que le modèle est plus aligné aux données réelles sans ces points influents, offrant un ajustement plus fiable sans pour autant affecter la capacité de distinction.
- **Modélisation avec suréchantillonnage (SMOTE) sans points influents** : Avec SMOTE, l'objectif était de mieux capturer la classe minoritaire (en défaut). Cette approche a légèrement réduit l'AUC à 0.86 mais a permis d'améliorer le rappel pour la classe minoritaire, équilibrant ainsi mieux les prédictions. Le test de Hosmer-Lemeshow a également montré un bon ajustement, suggérant que le modèle est bien calibré malgré le rééquilibrage des classes.

Conclusion partielle : Dans l'ensemble, la suppression des points influents a permis d'améliorer la robustesse et la calibration du modèle, tandis que l'ajout de SMOTE a équilibré les prédictions pour les classes minoritaires. Ces approches, combinées de manière judicieuse, contribuent à un modèle mieux adapté et mieux calibré pour la prédiction des risques de défaut.

Le modèle ayant le Log-Likelihood le plus élevée est le modèle sans points influents et sans SMOTE (cf A.3.3). Cela suggère que ce modèle capte le plus efficacement la structure sous-jacente des données.

L'analyse des odds ratios de ce modèle pour certaines variables clés permet de mettre en lumière les facteurs de risque associés au défaut de paiement.
L'odds ratio attaché à la variable DELINQ est égal à 2,1540. Cela signifie précisément que la chance relative de faire défaut sur sa dette est environ 2,5 fois plus élevée pour un emprunteur avec une ligne de crédit en retard que pour un emprunteur sans ligne en retard, conditionnellement aux facteurs pris en compte dans le modèle. Cette relation souligne l'importance des antécédents de crédit sur le risque de défaut, chaque retard supplémentaire augmentant significativement la probabilité de non-remboursement.

L'odds ratio attaché à la variable DEBTINC_missing est égal à 9,6474. Cela signifie précisément que la chance relative de faire défaut sur sa dette est quasiment 10 fois plus élevée pour un emprunteur dont le ratio dette/revenu n'est pas mentionné que pour un emprunteur avec un ratio connu, conditionnellement aux facteurs pris en compte dans le modèle. Ce résultat souligne l'importance cruciale de cette métrique, où l'absence d'informations peut être un indicateur de risque accru pour les prêteurs.

L'odds ratio attaché à la variable JOB_GROUP_Office and Professional est égal à 0,6537. Cela signifie précisément que la chance relative de faire défaut sur sa dette est 0,65 fois moins élevée pour un emprunteur appartenant à ces catégories professionnelles que pour un emprunteur appartenant à une autre catégorie, conditionnellement aux facteurs pris en compte dans le modèle. Cela suggère que les individus travaillant dans des environnements de bureau peuvent être perçus comme ayant une meilleure stabilité

financière, contribuant ainsi à un risque de défaut plus faible.

En résumé, les odds ratios pour DELINQ, DEBTINC_missing et JOB_GROUP_Office and Professional révèlent des dynamiques clés dans l'évaluation du risque de défaut. Un historique de paiements défaillants et des informations manquantes sur la situation financière sont des indicateurs forts de risque, tandis qu'une profession stable semble offrir une protection contre le défaut de paiement. Ces insights peuvent orienter les décisions des prêteurs dans leur processus d'évaluation du risque crédit (Voir A.4).

4 Random Forest

4.1 Théorie

Un arbre de décision est un algorithme de machine learning très utile pour résoudre des problèmes de régression et de classification. Cet algorithme populaire se base, comme son nom l'indique, sur une structure arborescente qui permet de prendre des décisions en segmentant les données en sous-ensembles de plus en plus homogènes, facilitant ainsi l'interprétation des décisions.

Toutefois, les arbres de décision présentent certaines limites, notamment une sensibilité au surapprentissage, une variance élevée et une dépendance aux données d'entraînement. Pour pallier ces faiblesses, on utilise souvent les forêts aléatoires (Random Forests) comme alternative. Les forêts aléatoires combinent plusieurs arbres de décision : chaque arbre est entraîné sur un sous-ensemble aléatoire des données (sélectionné grâce à la méthode de bootstrap), et les prédictions de tous les arbres sont ensuite combinées pour obtenir une prédiction finale plus robuste et précise.

Dans notre cas, pour mettre en place ce modèle de machine learning, les forêts aléatoires offrent plusieurs avantages par rapport à d'autres techniques comme la régression logistique. Contrairement à cette dernière, elles n'exigent pas d'hypothèses strictes de linéarité entre les variables indépendantes et la variable cible, ni de conditions sur l'absence de multicolinéarité entre les variables explicatives, car le modèle y est naturellement robuste.

4.2 Modélisation

4.2.1 Application de la classification des emprunteurs selon leur risque de défaut via un modèle de Random Forest

Pour classifier les emprunteurs en fonction de leur risque de défaut, nous avons d'abord appliqué le modèle de Random Forest sur l'ensemble des variables disponibles. Cette étape initiale nous a permis de mesurer l'importance des variables (features) afin d'identifier celles qui sont les plus pertinentes pour la prédiction.

Après l'entraînement et les tests effectués sur des bases incluant toutes les variables, nous avons obtenu les résultats suivants :

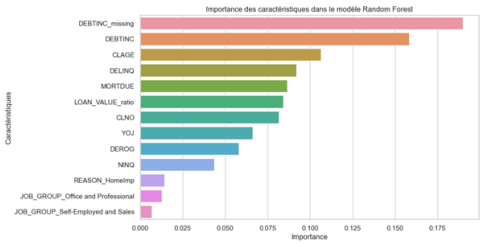
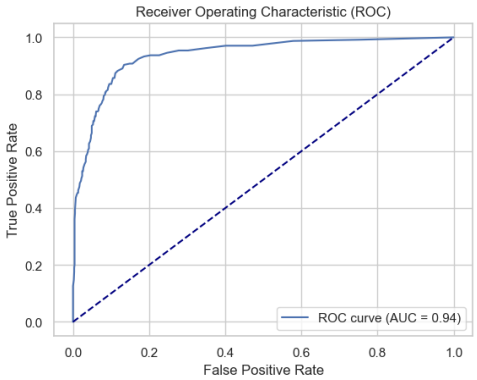


FIGURE 16 – Pertinence des features

Nous avons donc effectué un test afin de voir les variables les plus déterminantes dans notre modélisation Random Forest, nous voyons par exemple que la variable DEBTINC est très importante dans notre modélisation tandis que JOB_GROUP_Self-Employed and Sales ne l'est presque pas du tout.

4.2.2 Évaluation des performances du modèle de Random Forest

Pour optimiser les performances de notre modèle Random Forest, nous avons effectué une optimisation des hyperparamètres via la méthode de GridSearch, notamment le nombre d'arbres, le nombre de caractéristiques par division, la profondeur maximale de chaque arbre, et le nombre maximum de nœuds. Cette démarche a permis de maximiser les performances globales du modèle.



Avant l'optimisation des hyperparamètres, notre modèle affichait un AUC, ou «Area Under the Curve de 0,95. Cependant, après l'application de l'optimisation via « GridSearchCV », cet AUC a légèrement diminué, atteignant 0,94. Il serait facile de conclure que cette optimisation n'est pas bénéfique, étant donné la baisse de l'AUC. Toutefois, il est important de considérer que l'optimisation des hyperparamètres ne vise pas uniquement à améliorer la performance du modèle, mais également à prévenir les risques de sous-apprentissage ou de surapprentissage.
En effet, en ajustant les paramètres du modèle, nous augmentons sa robustesse et sa capacité de généralisation sur des données inconnues. De plus, une bonne optimisation peut réduire le temps d'entraînement, ce qui est un avantage significatif dans le cadre de projets impliquant des ensembles de données volumineux. Ainsi, même si l'AUC a légèrement diminué, les bénéfices liés à la robustesse et à la généralisation du modèle en font un processus d'optimisation précieux.

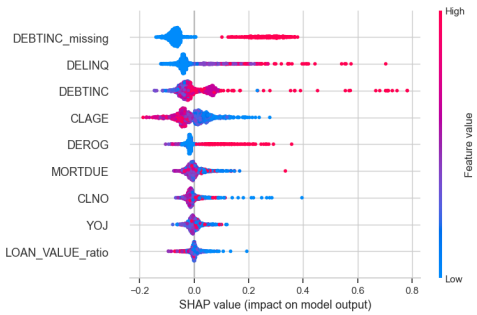
De plus, cette métrique de 0,94 indique une excellente capacité de discrimination entre les classes d'emprunteurs susceptibles de faire défaut de remboursement (BAD=1) et ceux

jugés solvables (BAD=0). L'AUC, est une mesure de performance basée sur la courbe ROC, qui trace le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs. Un AUC de 1 correspond à une séparation parfaite, tandis qu'un AUC de 0,5 indiquerait un modèle qui n'a pas de capacité discriminante et se contente de prédictions aléatoires. Dans notre cas, cela montre que notre modèle est capable de différencier avec une grande précision les emprunteurs à risque de ceux qui sont fiables.

En termes de précision, le modèle atteint un score de 0,77, ce qui signifie que 77 % des emprunteurs identifiés comme risqués (BAD=1) sont effectivement en défaut, avec un faible taux de faux positifs. En revanche, le rappel est de 0,65, indiquant que 35 % des emprunteurs en défaut sont à tort classés comme solvables, ce qui représente une limitation du modèle en termes de couverture des cas de défaut. Bien que l'AUC du modèle soit excellent, ce rappel modéré suggère qu'il reste une marge pour mieux capter tous les emprunteurs en situation de risque (cf A.5).

4.2.3 Shapley Values

Les Shapley Values sont une méthode d'explication des modèles d'apprentissage automatique, dérivée de la théorie des jeux. Dans le cadre des modèles de Random Forest, ils attribuent une importance à chaque caractéristique en mesurant sa contribution à la prédiction. Pour une observation donnée, les Shapley Values calculent la contribution moyenne de chaque variable, en tenant compte de toutes les combinaisons possibles de caractéristiques. Cela permet d'évaluer non seulement l'impact individuel d'une caractéristique, mais aussi ses interactions avec d'autres variables. Cette approche offre une interprétation cohérente des décisions du modèle, essentielle dans des domaines nécessitant transparence et fiabilité.



La variable DELINQ, qui représente le nombre de délinquances passées, exerce un impact positif sur la probabilité de défaut : plus le nombre de délinquances est élevé, plus le risque de défaut augmente. Inversement, un faible nombre de délinquances diminue cette probabilité.

Le ratio dette/revenu (DEBTINC) suit une logique similaire : un ratio élevé accroît le risque de défaut, tandis qu'un ratio faible est associé à une probabilité réduite. De plus, l'absence de valeur pour ce ratio, indiquée par DEBTIN_missing, a un impact significatif, augmentant les prévisions du modèle, suggérant que l'absence de cette donnée pourrait être liée à un risque accru de défaut.

L'ancienneté de la ligne de crédit (CLAGE) présente un effet plus modéré. En règle générale, une ancienneté élevée semble légèrement atténuer le risque de défaut.

La présence d'antécédents financiers négatifs, mesurée par DEROG, constitue également un indicateur crucial du risque de défaut, car un plus grand nombre d'incidents financiers

accroît cette probabilité.

Le montant de l'hypothèque en cours (MORTDUE) a une influence faible sur le risque de défaut ; des montants élevés l'augmentent légèrement.

Le nombre de lignes de crédit ouvertes (CLNO) présente un effet variable : un nombre réduit de lignes peut légèrement augmenter le risque de défaut.

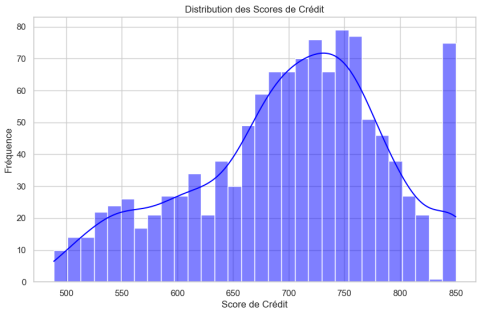
Enfin, l'ancienneté dans l'emploi (YOJ) exerce une influence minime. Une faible ancienneté est associée à un risque accru, tandis qu'une ancienneté élevée réduit légèrement ce risque.

Pour resumer, les variables les plus influentes dans ce modèle de prédiction de défaut sont DEBTINC_missing, DELINQ, et DEBTINC. En général, des valeurs élevées de dettes ou des antécédents de délinquance constituent des indicateurs d'un risque de défaut accru, ce qui est conforme à l'interprétation classique des indicateurs financiers.

4.3 Grille de Score

Afin de conclure notre projet et de mettre en application les résultats obtenu, nous avons décidé de créer des faux individus en associant à chaque observation un prénom ainsi qu'un nom. Par la suite, chaque individu aura donc un score associé à ses caractéristiques propres.

Pour la création de ce score, nous avons associé à chaque individu une probabilité de faire défaut sur sa dette. Plus la probabilité est proche de 1 plus le risque de défaut est élevé est donc plus le score associé à la ligne sera proche de 500 qui est le score minimum dans notre cas.
Le score est modélisé par la probabilité associé à chaque individu multiplié par une constante qui sera commune à tous.
Nous avons appliqué cette méthode à environ 20% de notre base de donnée pour éviter de surcharger nos calculs et notre rendu.
Nous obtenons une distribution des scores de crédit sur laquelle nous voyons qu'il y a une densité assez importante d'individus entre 700 et 750 de score ce qui représente une probabilité de faire défaut compris entre 0.10 et 0.01.Il y a aussi un pic assez conséquent quand le score est de 850 ce qui signifie qu'il y a également une bonne partie des individus qui ne feront presque sûrement pas défaut.



Ainsi, nous avons sauvegardé un nouveau fichier que nous avons nommé *hmeq_score* qui reprend exactement les mêmes données de base en y ajoutant pour chaque observation un prénom, un nom, un score et de facto une probabilité de faire défaut.

5 Conclusion

Ce projet de modélisation du risque de défaut sur les prêts hypothécaires visait à développer une approche prédictive permettant d'évaluer la probabilité de défaut de remboursement d'un emprunteur. En adoptant une démarche rigoureuse d'analyse et de traitement des données, nous avons exploité les informations de la base HMEQ pour construire et évaluer des modèles robustes, en intégrant des variables financières et sociodémographiques significatives.

Dès l'étape d'exploration des données, nous avons identifié des variables essentielles, telles que le ratio dette/revenu (DEBTINC), le nombre de retards de paiement (DELINQ), et la présence d'incidents financiers passés (DEROG), qui montrent une forte corrélation avec le risque de défaut. Le prétraitement des données a constitué une étape cruciale pour garantir la qualité des analyses. Nous avons traité les valeurs manquantes de manière adaptée, par exemple, avec l'imputation par la médiane pour les variables continues sensibles aux valeurs extrêmes. De plus, nous avons appliqué une standardisation avec StandardScaler et effectué des regroupements de modalités afin de renforcer la pertinence des variables.

Trois approches de modélisation ont été testées et comparées : la régression logistique avec et sans suréchantillonnage (SMOTE), ainsi que l'algorithme de Random Forest. La régression logistique, à la fois avec et sans points influents, a démontré une bonne capacité de prédiction, mais l'application de SMOTE a amélioré le rappel pour la classe minoritaire. Cependant, c'est le modèle de Random Forest qui a offert la meilleure performance, avec un AUC de 0,94 après optimisation, montrant une forte capacité à discriminer entre les classes de risque. L'utilisation de SMOTE a également permis une meilleure représentation des emprunteurs en défaut, bien que cela ait parfois eu pour effet d'augmenter les erreurs de prédiction pour la classe majoritaire.

L'explicabilité des modèles a également été au cœur de notre démarche. En recourant aux valeurs de Shapley, nous avons pu interpréter les contributions individuelles des variables clés, confirmant que des variables comme DELINQ, DEBTINC, et DEROG étaient parmi les plus influentes. Par cette analyse, nous avons apporté une transparence cruciale à la modélisation, ce qui est essentiel pour une application dans le domaine du risque de crédit.

En conclusion, le modèle final offre une solution fiable pour anticiper les risques de défaut sur les prêts hypothécaires, permettant aux institutions financières de prendre des décisions de prêt plus éclairées et d'adapter leurs stratégies en fonction des profils de risque. Ce travail ouvre également des perspectives d'améliorations futures, notamment par l'intégration de données comportementales additionnelles ou l'exploration de techniques d'apprentissage plus avancées, telles que les réseaux neuronaux profonds, pour renforcer encore la précision du scoring de crédit.

A Annexe

A.1 Métriques modèle Logit sans SMOTE

A.1.1 Avec points influents

— **Accuracy** : 0.8691
— **F1 Score** : 0.6286
— **AUC-ROC** : 0.8732

TABLE 1 – Rapport de classification

Classe	Précision	Rappel	F1-Score	Support
0	0.90	0.95	0.92	954
1	0.73	0.55	0.63	238
Accuracy			0.87	1192
Macro avg	0.81	0.75	0.77	1192
Weighted avg	0.86	0.87	0.86	1192

A.1.2 Sans points influents

— **Accuracy** : 0.8708
— **F1 Score** : 0.6351
— **AUC-ROC** : 0.7553

TABLE 2 – Rapport de classification sans points influents

Classe	Précision	Rappel	F1-Score	Support
0	0.90	0.95	0.92	954
1	0.73	0.56	0.64	238
Accuracy			0.87	1192
Macro avg	0.81	0.76	0.78	1192
Weighted avg	0.86	0.87	0.86	1192

A.2 Métriques modèle Logit avec SMOTE

— **Accuracy** : 0.8297
— **F1 Score** : 0.6394
— **AUC-ROC** : 0.8613

Table 3 – Rapport de classification

Classe	Précision	Rappel	F1-Score	Support
0	0.93	0.85	0.89	954
1	0.55	0.76	0.64	238
Accuracy			0.83	1192
Macro avg	0.74	0.80	0.76	1192
Weighted avg	0.86	0.83	0.84	1192

A.3 Tables des régressions logistiques

A.3.1 Modèle avec SMOTE

Dep. Variable :	BAD	**No. Observations :**	7634
Model :	Logit	**Df Residuals :**	7625
Method :	MLE	**Df Model :**	8
Date :	Sun, 27 Oct 2024	**Pseudo R-squ. :**	0.3892
Time :	18 :25 :54	**Log-Likelihood :**	-3231.8
converged :	True	**LL-Null :**	-5291.5
Covariance Type :	nonrobust	**LLR p-value :**	0.000

	coef	**std err**	**z**	**P> |z|**
const	-0.6343	0.042	-15.088	0.000
YOJ	-0.0860	0.034	-2.517	0.012
DEROG	0.3243	0.036	9.069	0.000
DELINQ	0.8057	0.040	20.359	0.000
NINQ	0.0650	0.030	2.179	0.029
DEBTINC_missing	2.1896	0.074	29.773	0.000
JOB_GROUP_Office and Professional	-0.6686	0.069	-9.658	0.000
CLAGE_discretise	-1.2282	0.066	-18.614	0.000
DEBTINC_discretise	-0.5885	0.038	-15.316	0.000

A.3.2 Modèle avec points influents

Dep. Variable :	BAD	**No. Observations :**	4768
Model :	Logit	**Df Residuals :**	4754
Method :	MLE	**Df Model :**	13
Date :	Sun, 27 Oct 2024	**Pseudo R-squ. :**	0.4037
Time :	18 :32 :59	**Log-Likelihood :**	-1420.6
converged :	True	**LL-Null :**	-2382.3
Covariance Type :	nonrobust	**LLR p-value :**	0.000

	coef	std err	z	P> |z|
const	-2.2452	0.084	-26.854	0.000
YOJ	-0.0546	0.053	-1.026	0.305
DEROG	0.3586	0.047	7.573	0.000
DELINQ	0.7448	0.053	14.185	0.000
NINQ	0.1480	0.045	3.316	0.001
DEBTINC_missing	2.2285	0.106	21.052	0.000
JOB_GROUP_Office and Professional	-0.3493	0.110	-3.166	0.002
JOB_GROUP_Self-Employed and Sales	0.6350	0.199	3.185	0.001
REASON_HomeImp	0.1493	0.108	1.377	0.169
LOAN_VALUE_ratio_discretise	-1.0775	0.246	-4.377	0.000
MORTDUE_discretise	-0.9054	0.201	-4.496	0.000
CLAGE_discretise	-1.0218	0.098	-10.461	0.000
CLNO_discretise	-0.8903	0.162	-5.504	0.000
DEBTINC_discretise	-0.5955	0.069	-8.568	0.000

A.3.3 Modèle sans points influents

Dep. Variable :	BAD	**No. Observations :**	4427
Model :	Logit	**Df Residuals :**	4413
Method :	MLE	**Df Model :**	13
Date :	Sun, 27 Oct 2024	**Pseudo R-squ. :**	0.4098
Time :	18 :35 :19	**Log-Likelihood :**	-1297.2
converged :	True	**LL-Null :**	-2198.0
Covariance Type :	nonrobust	**LLR p-value :**	0.000

	coef	std err	z	P> |z|
const	-2.2327	0.087	-25.679	0.000
YOJ	-0.0588	0.056	-1.058	0.290
DEROG	0.3436	0.052	6.670	0.000
DELINQ	0.7673	0.055	13.852	0.000
NINQ	0.1466	0.047	3.125	0.002
DEBTINC_missing	2.2667	0.111	20.398	0.000
JOB_GROUP_Office and Professional	-0.4251	0.117	-3.643	0.000
JOB_GROUP_Self-Employed and Sales	0.5516	0.209	2.641	0.008
REASON_HomeImp	0.1433	0.114	1.255	0.209
LOAN_VALUE_ratio_discretise	-1.1889	0.258	-4.609	0.000
MORTDUE_discretise	-0.8790	0.212	-4.153	0.000
CLAGE_discretise	-0.9925	0.103	-9.654	0.000
CLNO_discretise	-0.8972	0.170	-5.282	0.000
DEBTINC_discretise	-0.5912	0.073	-8.133	0.000

A.4 Odds ratios

Variable	Odd-Ratio
const	0.107238
YOJ	0.942903
DEROG	1.409977
DELINQ	2.154046
NINQ	1.157893
DEBTINC_missing	9.647386
JOB_GROUP_Office and Professional	0.653728
JOB_GROUP_Self-Employed and Sales	1.735952
REASON_HomeImp	1.154073
LOAN_VALUE_ratio_discretise	0.304555
MORTDUE_discretise	0.415189
CLAGE_discretise	0.370659
CLNO_discretise	0.407714
DEBTINC_discretise	0.553642

Table 4 – Odds Ratios des Variables du Modèle Logistique sans points influents et sans SMOTE

A.5 Métriques du Random Forest

	Precision	Recall	F1-Score	Support
0	0.92	0.96	0.94	954
1	0.81	0.66	0.73	238
Accuracy			0.90	1192
Macro Avg	0.86	0.81	0.83	1192
Weighted Avg	0.90	0.90	0.90	1192