UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Master 2 Modélisations Statistiques et Financières
Université Paris 1 Panthéon Sorbonne

Challenge Crédit Logement

Modélisation Probabilité de recouvrement complet

CREDIT LOGEMENT

Mounira Arbouch, Johanna Roll, Maysa Slimani et Mélanie Yakoub.

Résumé

Dans ce rapport, nous cherchons à modéliser la probabilité de recouvrement complet des dossiers de nos clients dès leur passage en DET, en partenariat avec Crédit Logement. Après une phase préliminaire de compréhension des variables et de nettoyage des données, deux approches de modélisation ont été explorées : la régression logistique pour estimer la probabilité de succès du recouvrement complet, et le modèle de durée Cox pour examiner la durée jusqu'au recouvrement complet. Les résultats ont montré l'efficacité des deux méthodes pour prédire le recouvrement complet des dossiers, fournissant ainsi des outils précieux pour optimiser les stratégies de recouvrement et la gestion des risques chez Crédit Logement.

Contents

Introduction

$\text{RW} = \left(LDG \times N\left(\sqrt{\frac{1}{1-R}}G(PD) + \sqrt{\frac{R}{1-R}}G(0.999)\right) - PD \times LGD\right) \times \left(\frac{1+(M-2.5)b}{1-1.5b}\right) \times 12.5 \times 1.06$

L'anticipation du recouvrement des créances sur une période prolongée revêt une importance cruciale pour les institutions financières telles que Credit Logement. En effet, la gestion d'un dossier en situation de recouvrement peut engendrer des coûts significatifs tels que les frais de gestion administrative, les frais judiciaires, ainsi que les pertes financières liées au non-remboursement des créances. Par conséquent, comprendre et anticiper le comportement des clients tout au long de leur processus de remboursement est essentiel pour garantir la stabilité financière et la pérennité de l'entreprise.

Dans cette optique, notre étude vise à modéliser la probabilité de recouvrement complet d'un dossier après son passage en déchéance de terme, sur une période de 10 ans, de 2007 à 2018. L'objectif est de déterminer les critères favorisant le maintien de la stratégie amiable de recouvrement, afin d'optimiser les performances en matière de recouvrement.

Pour répondre à cet objectif, nous avons tout d'abord procédé à une analyse exploratoire de notre base de données, en nous assurant de la qualité des données en imputant les valeurs manquantes et en sélectionnant les variables pertinentes (1). Ensuite, nous avons testé deux approches complementaires (2) : la régression logistique pour modéliser la probabilité de recouvrement complet à partir de la déchéance de terme, puis le modèle de durée Cox qui prend en compte les variables évolutives dans le temps pour modéliser la durée jusqu'au recouvrement complet. Après avoir évalué les performances des modèles, nous avons créé des scores de recouvrement pour chaque individu, basés sur les résultats des modèles, afin de fournir à Credit Logement un outil efficace pour optimiser ses stratégies de recouvrement.

Chapter 1

Compréhension des variables et de la base de données

1.1 Description de la base de données

Pour traiter le sujet 2, nous disposons d'une base de données comprenant 739 626 observations et 93 variables. Ces données retracent le parcours complet des clients, depuis leur entrée en défaut jusqu'au remboursement (ou non) intégral de leur créance. Au total, nous avons étudié 13 235 clients, avec des durées de suivi variables. Pour mener notre analyse, nous avons défini une période temporelle allant de 2007 à 2018, ce qui nous a permis de prédire la probabilité de recouvrement de la créance sur une période de 10 ans. Cette approche nous a donné une perspective suffisamment étendue pour évaluer les performances des modèles de prédiction dans des conditions réalistes.

Notre base de donnée est composée de données temporelles (nos variables d'octroi) et statiques. De fait, il était primordial pour nous de prendre en compte les potentiels changement au cours du dossier d'un client. Pour se faire, nous avons décidé de tester 2 modèles différents : un modèle de Régression Logistique, efficace pour ce genre de modèle mais qui ne prends pas en compte la dimension temporelle, et un modèle de Cox.

Avant d'entamer notre analyse, nous avons pris soin de nettoyer et de préparer au mieux notre base de données. Dans cette optique, nous avons restreint notre sélection aux individus au moment de leur entrée en défaut, identifiée par la variable "fam_PCD" égale à 1. Suite à ça, nous avons identifié certaines incohérences dans les données. N'ayant pu trouver de motif ou de cohérence justifiant ces valeurs aberrantes, nous avons pris la décision de les exclure de notre base. Cette démarche visait à éviter tout biais dans notre analyse.

Dans un premier temps, nous avons décidé d'inclure une variable "duration" qui retrace la durée de chaque client dans la base de données, de son entrée en défaut à sa dernière date d'arrêté. En faisant cela, nous avons remarqué que certains clients contenaient des dates de sortie de défaut plus récente que la date d'entrée. Nous avions donc une durée dans la base négative, ce qui est impossible. De plus, nous avions certains clients qui voyaient leur target passer à 1 en pleins milieu de leur dossier. Les garder ajouterai un biais significatif étant donnée que la target n'est pas fixe.

Pour le modèle de régression logistique, nous devions conserver seulement 1 seule ligne par client. Pour se faire, nous avons sélectionné la première ligne de chaque client, afin d'avoir toutes ses informations à l'entrée, et extrait la dernière ligne de la target afin de savoir si il a ou non recouvré toute sa créance. Pour le modèle de Cox, deux tests ont été fait : un sur la data en entière et un sur les premières lignes seulement.

1.2 Analyse exploratoire des données

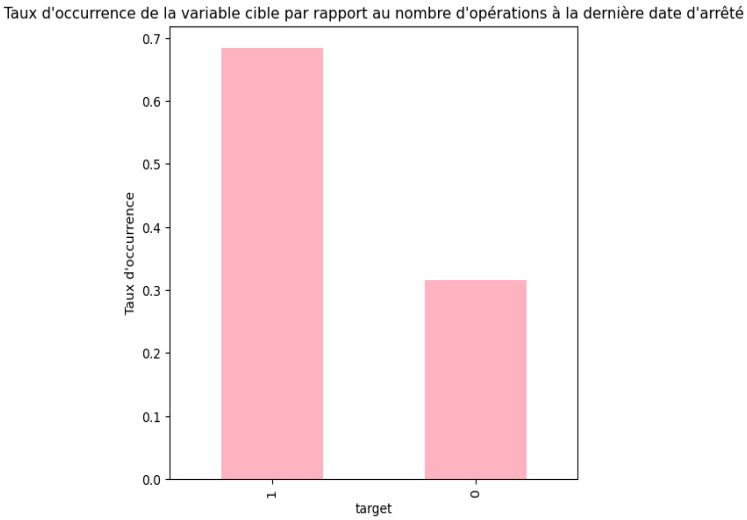

Figure 1.1: Taux d'occurence de la target

Notre target binaire est équilibré en terme de classes, avec environ 53% des observations étiquetées comme appartenant à la classe 1 et 47% à la classe 0 (en se basant sur la dernière lignes de chaque client). Cette répartition suggère que notre ensemble de données est approprié pour une modélisation de classification, sans biais marqué en faveur d'une classe particulière.

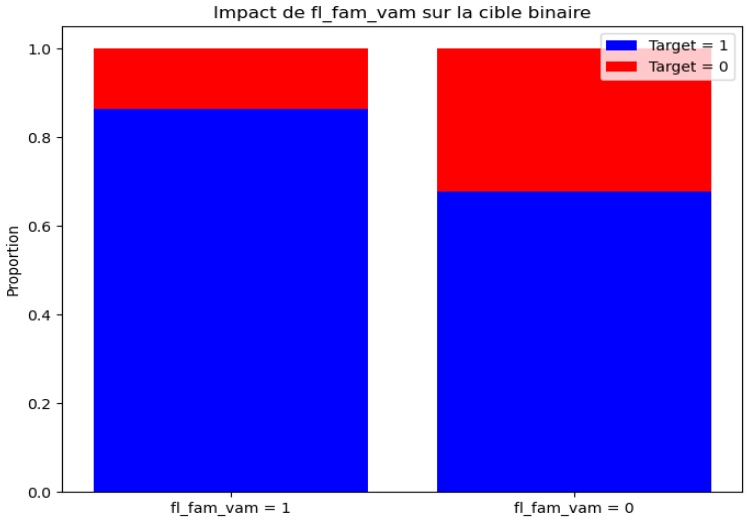

Figure 1.2: Impact d'une procédure à l'amiable sur notre target

Il était également pertinent de s'intéresser à l'impact d'une procédure à l'amiable sur la probabilité du client à recouvrir totalement sa créance. Nous voyons une différence significative entre les deux modalités, avec une probabilité de 30% de ne pas recouvrir totalement sa dette pour les procédures qui ne sont pas faites à l'amiable.

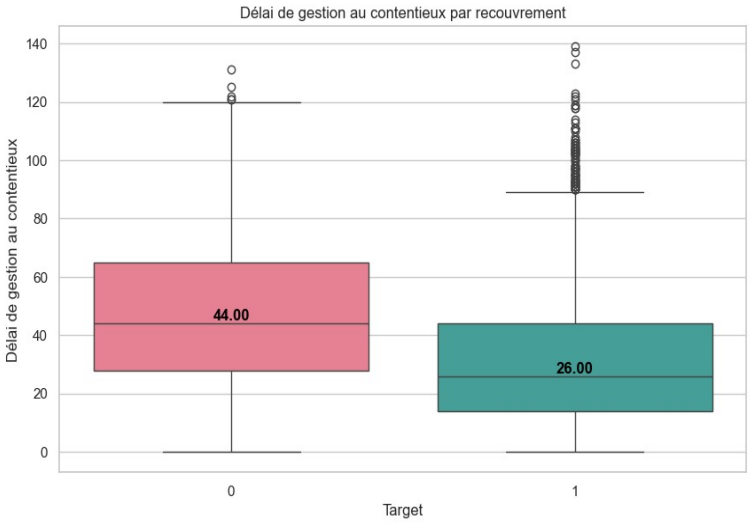

Figure 1.3: Délai de gestion au contentieux par modalité de la target

Le délai de gestion a également un impact significatif sur la probabilité de recouvrir totalement sa créance. En effet, un délai de gestion plus long pourrai signifier un risque pour le client de ne pas recouvrir. Ces Box-plots traduisent parfaitement cette idée, avec un délai de gestion maximum d'envion 120 mois pour les personnes n'ayant pas recouvert leur dette contre 90 pour ceux qui l'ont fait. La médiane des deux groupes est même presque divisée par deux entre ces derniers, avec un nombre de mois médian de 44 contre 26 pour les 1.

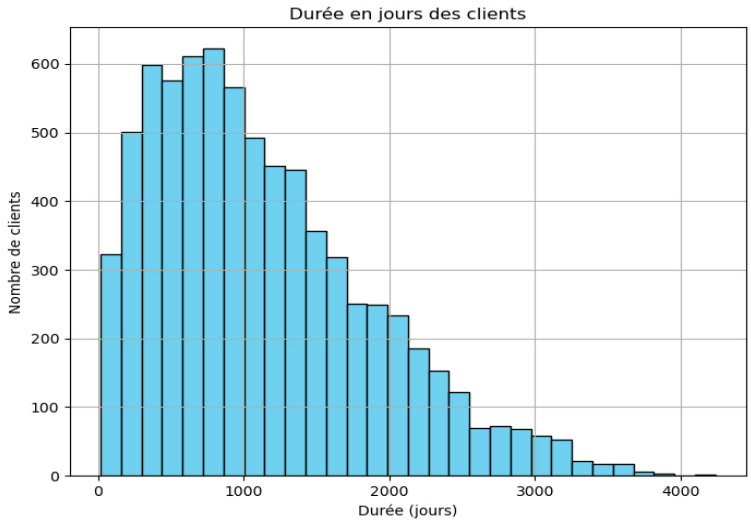

Figure 1.4: Durée des clients depuis leur arrivée en défaut jusqu'à leur dernière date d'arrêté

Sur un total de 7 442 observations (clients entrant entre 2007 et 2012), la moyenne de durée dans la base est d'environ 1 133 jours, avec un écart-type de 752 jours, indiquant une certaine variabilité dans les clients. Les valeurs de cette variable s'étendent de 16 jours (le temps le plus court) à 4 240 jours (le temps le plus long), ce qui souligne la diversité des situations rencontrées.

1.3 Sélection de Variables

1.3.1 Multicolinéarité

La multicolinéarité se produit lorsque deux ou plusieurs variables indépendantes dans un modèle statistique sont fortement corrélées, ce qui peut rendre difficile l'interprétation des coefficients et accroître la sensibilité du modèle aux petites variations dans les données. En d'autres termes, elle peut rendre les estimations des coefficients moins fiables, car les variables corrélées fournissent une information similaire au modèle.

Dans un premier temps, nous avons appliqués une matrice de corrélation à nos données afin de visualiser la présence, ou non, de multicolinéarité. Lorsqu'une paire de variables présentait une corrélation supérieure à 30%, nous avons examiné laquelle de ces variables était plus fortement corrélée avec la variable cible. Par la suite, nous avons conservé la variable qui avait la corrélation la plus élevée avec la variable cible et nous avons supprimé l'autre variable, afin de réduire le risque de multicolinéarité dans notre modèle.

Après avoir appliqué notre sélection de variable, nous avons fait une deuxième vérification de multicolinéarité et aucune de nos variables n'en présentaient.

1.3.2 Méthodes de sélections

Dans le but d'obtenir le modèle le plus explicatif possible, nous avons commencé par supprimer les variables contenant plus de 30% de valeurs manquantes. Cette démarche vise à éviter tout biais potentiel qui pourrait résulter de l'imputation des valeurs manquantes. Après cette étape, notre ensemble de données était composé de 73 variables. À partir de cet ensemble, nous avons pu effectuer une sélection judicieuse des variables à inclure dans notre modèle.

Pour se faire, nous avons utiliser plusieurs méthodes : le V de Cramer pour les variables qualitatives, et le test Kruskal-Wallis pour les variables continues. Malgrè l'efficacité de ces méthodes de selection, nous avons décidé d'utiliser la méthode de Recursive Feature Elimination (RFE). Son objectif est d'évaluer l'importance de chaque variable en éliminant progressivement les moins importantes (moins informatives) à chaque entrainement du modèle. Cependant, nous devions connaître quel était le nombre optimal de features à inclure dans notre modèle afin de maximiser au mieux leur métriques. Pour cela, nous avons utilisé la validation croisée sur 5 plis grâce à l'option "RFECV" de scikit-learn, utilisée pour la sélection de caractéristiques récursive avec validation croisée. Nous avons alors affiché le graphique qui trace le score de validation croisée moyen (roc_auc) en fonction du nombre de caractéristiques testées. Il est important de préciser que toutes les méthodes testés pour la selection de nos variables

avaient des résultats relativement similaire, ce qui témoigne de la fiabilité de notre sélection.

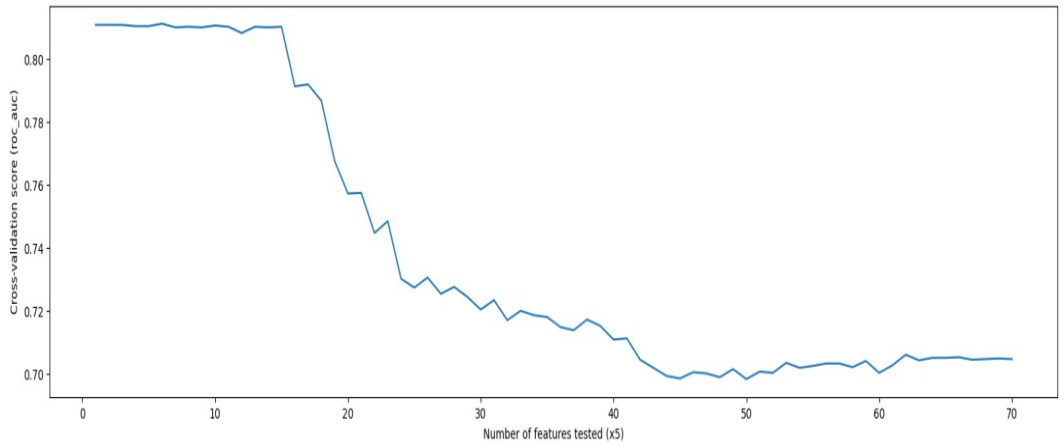

Figure 1.5: Nombre de features optimales par RFE

Le nombre de variable optimale se trouve donc entre 1 et 15. Voici les variables qui ont été selectionnées minutieusement par l'algorithme, mais aussi celles qui ont été ajouté manuellement grâce à leur sens métier :

- DURATION : Durée du client de l'entrée en défaut à sa dernière date d'arrêté.
- ANC_PRO_MAX_PFI : nombre de mois d'ancienneté professionnelle de l'emprunteur.
- DLS_GES_DEF : Délai de gestion depuis l'entrée au contentieux.
- FL_FAM_VAM : Flag indiquant si une vente amiale a été faite ou non.
- DLS_MEP_ENTREE_DEF : Délais écoulés entre les dates de mise en place et d'entrée en défaut.
- MAX_AGE_CTX : Age maximal à l'entrée en défaut.
- CUM_DEC : Cumul des décaissements.
- MT_EAD_DNE_CHR_d : EAD à l'entrée en premier CHR observée sur l'arrêté dernier connu.
- NB_EMPR_OPE : Nombre d'emprunteurs au niveau de l'opération.
- CD_MTF_ENE_CTX : Motif du défaut dernier connu au niveau de l'opération.
- CD_NAT_EMP1 : Nationnalité de l'empreunteur 1.
- REGROUP_NATUR_OP : Nature de l'opération.
- CD_SITFAM_EMP1 : Situtation familiale de l'emprunteur 1.
- CD_DEST_PFI : Destination du bien.
- CD_CSP_EMP1 : CSP de l'emprunteur 1 de l'opération.

1.4 Preprocessing des données

1.4.1 Regroupement des modalités

Afin de minimiser le bruit (variabilité aléatoire non systématique des données qui peut perturber la précision du modèle) dans nos modèles, nous avons regrouper des modalités dans certaines variables lorsque cela était possible. En effet, certaines modalités peuvent ne pas avoir de relation significative avec la variable cible, introduisant ainsi du bruit. De plus, si certaines modalités ont une fréquence d'occurrence très faible dans les données d'entraînement, le modèle peut avoir du mal à généraliser à de nouvelles occurrences de ces modalités. Cela peut conduire à des prédictions moins fiables pour ces modalités moins fréquentes.

Notre démarche a pour but de simplifier au maximum nos variables afin de faciliter l'interprétation de ces dernières pour la suite de l'analyse. Nous avons donc regrouper de manière manuelle les modalités qui, globalement, étaient similaire et avaient un sens. Par exemple, pour la nationnalité de l'emprunteur, nous avons décidé de regrouper nos modalités en 3 : les Français, les Européens et les étrangers. Pour le motif de défaillance du client, nous avons pu faire 5 regroupements : les motifs liés à la finance, à la Santé et au décès, à une procédure légale et tout autres problèmes. Nos 6 variables catégorielles ont été modifiés de façon a n'avoir à la fin qu'un maximum de 5 modalités.

1.4.2 Imputation

A partir de la sélection de variable faites, nous nous sommes occupées de traiter leurs valeurs manquantes. Les méthodes d'imputation varient en fonction du type de variable et du contexte des données, mais le but ici est d'approcher au mieux chaque variable tout en préservant la structure générale des données. Après nos regroupements, seulement les variables ANC_PRO_MAX_PFI, CD_DEST_PFI et CD_MTF_ENE_CTX avaient des valeurs manquantes. Nous avons donc décidé d'imputer par un "0" ANC_PRO_MAX_PFI car pour nous, une valeur manquante représente ici une valeur non renseigné volontairement et peut être que le client ne s'est potentiellement pas senti concerné par cette question. Pour CD_MTF_ENE_CTX, nous avons ajouté les valeurs manquantes à la catégorie "100 : Non identifié ". Pour CD_DEST_PFI, les NaN ont été rajoutés à la catégorie "NR". Enfin, CD_MTF_DFC a été regrouper dans la catégorie "Divers".

1.4.3 Encodage

L'encodage des variables qualitatives est une étape importante de la préparation des données dans le cadre du Machine Learning. Dans le cadre de ce projet, nous avons eu deux approches distinctes : faire du one-hot-encoding pour notre modèle de Regression Logistique, et du Label Encoding pour notre modèle de Cox. Chaque variables catégorielle à été réduite à maximum 5 modalités afin de faciliter la comparaison entre les groupes.

Pour la regression logistique, la méthode de one-hot-encoding nous à permis, lors du calcule des odds ratio, de comparer nos coefficients par rapport à notre variable de référence qui n'apparait pas dans ces resultats. Le one-hot-encoding a été utilisé ici seulement parce qu'il n'y avait pas un nombre trop conséquent de modalités, ce qui aurait pu entraîner de la multicolinéarité.

Pour Cox, l'approche a été faite différemment. Nous avons privilégié le Label Enconding en instaurant une notion d'ordre entre les labels. Cette notion d'ordre a été determiné par la contribution de chaque modalité sur la target = 1. En d'autre terme, à partir de nos modalités regroupés, nous avons commencé par regarder l'impact de chacune d'entre elles sur la modalité "1" de "fl_recup_tot_y", en regardant leur pourcentage d'apparition. Après cela, nous avons instaurer la notion d'autre de celle qui impacte plus la target=1 (label = 1) à celle qui l'impact le moins (label = dernier chiffre). Les groupes, étant là aussi définit de manière cohérents, nous ont permis de facilité l'interprétation des coeffiients de nos variables.

Chapter 2

Processus de Modélisation

2.1 La Regréssion Logistique

Un modèle de régression logistique permet de prédire la probabilité qu'un événement arrive (target = 1) ou non (target = 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas. L'estimation des paramètres du modèle se fait généralement par la méthode du maximum de vraisemblance, visant à maximiser la probabilité d'observer les données réelles en prenant en compte les paramètres estimés (on maximise le logarithme de la vraisemblance). La fonction de vraisemblance pour un modèle de régression logistique est définie comme le produit des probabilités des observations étant générées par le modèle. La forme de la vraisemblance dépend de la distribution de probabilité choisie pour le problème de régression logistique, qui est généralement la distribution de Bernoulli dans le cas d'une classification binaire. Elle est définie comme :

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = y_i | X_i = x_i, \beta)$$

Où :

- $L(\beta)$ est la fonction de vraisemblance.
- n est le nombre d'observations.
- Y_i est la variable de réponse (recouvrement : 0 ou 1) pour la i-ème observation.
- X_i représente le vecteur de caractéristiques (features) pour la i-ème observation.
- β est le vecteur des coefficients du modèle.
- $P(Y_i = y_i | X_i = x_i, \beta)$ la fonction logistique telle que :

$$P(Y_i = 1 | X_i = x_i, \beta) = \frac{e^{(\beta_0 + \sum_{i=1}^{n} \beta_i * X_i + \epsilon)}}{1 + e^{(\beta_0 + \sum_{i=1}^{n} \beta_i * X_i + \epsilon)}}$$

Dans le cadre de notre projet, la régression logistique était un choix pertinent étant donné que nous devions modéliser la probabilité qu'un client ait totalement recouvert sa dette ou non.

A la suite, nous avons divisé nos données en base d'entraînement et en base de test (train test split). Le principe de base du train-test split est de décomposer l'ensemble des données de manière aléatoire. Une partie servira à entraîner le modèle de Régression Logistique. L'autre partie, quant à elle, permet de réaliser le test de validation. Cela permet de vérifier si le modèle peut généraliser ses prédictions à de nouvelles données et contribue à prévenir le surajustement. Nous avons ensuite pris la décision de normaliser nos données. En effet, la normalisation des données est un processus couramment utilisé en machine learning pour mettre à l'échelle les valeurs de différentes variables dans un même intervalle. L'objectif principal de la normalisation est de rendre les données comparables entre elles et plus facilement interprétables par les algorithmes d'analyse et de modélisation. Dans le cas de notre étude, nous avons trouvé judicieux d'utiliser RobustScaler afin de normaliser les données de manière robuste, offrant une solution adaptée à la présence d'outliers.

Dans un premier temps, une fois le preprocessing des données effectués, nous avons initialisé notre modèle. Tout d'abord, nous avons défini un seuil (threshold). Ce dernier est une valeur définie (ici, nous l'avons fixé à 0.7) qui définit si le client a recouvert ou non. Si la probabilité prédite de la classe positive dépasse ce seuil, la prédiction est positive ; sinon, elle est négative. Dans notre contexte, si la probabilité de recouvrement prédite par l'algorithme est supérieure à 0.7, la prédiction finale est "le client a recouvert" (classe positive), sinon c'est "il n'a pas recouvert" (classe négative). Pour le choix du seuil, de nombreux tests ont été faits et le 0.7 a été judicieusement choisie, permettant d'ajuster l'équilibre entre les faux positifs et les faux négatifs en maximisant notre AUC. Dans le cadre de notre modèle, le seuil de 0.7 nous a permis d'avoir une AUC à 0.72. Nous nous sommes surtout penché sur l'interprétation des faux positifs, c'est à dire des clients que nous avions prédit qu'ils allaient recouvrir alors qu'ils ne l'ont pas fait. Notre modèle a alors mal classé 232 clients.

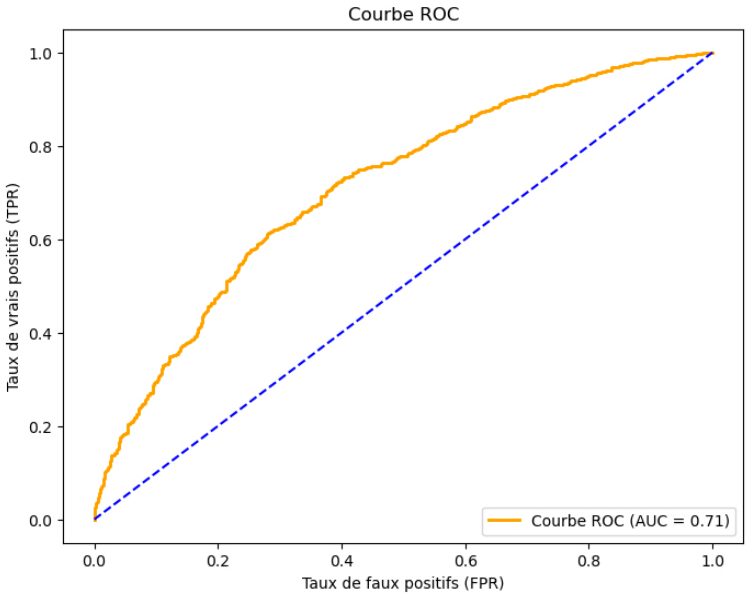

Figure 2.1: Courbe de ROC

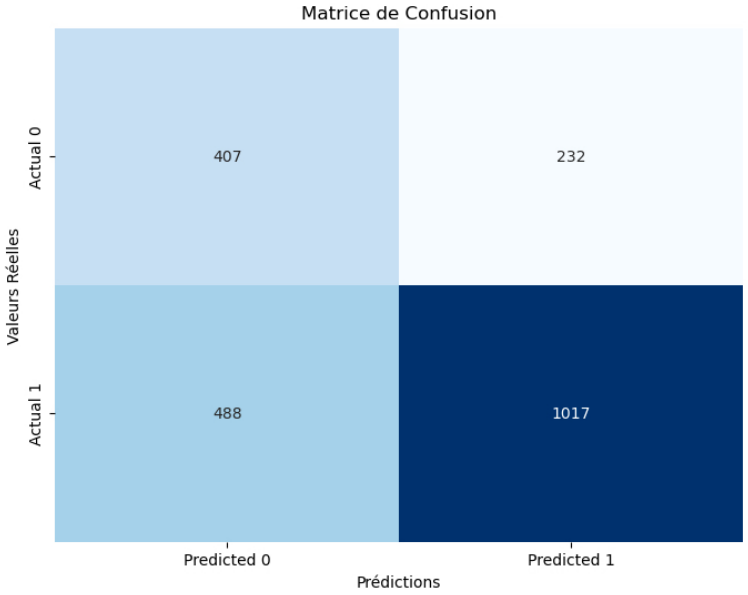

Figure 2.2: Matrice de confusion

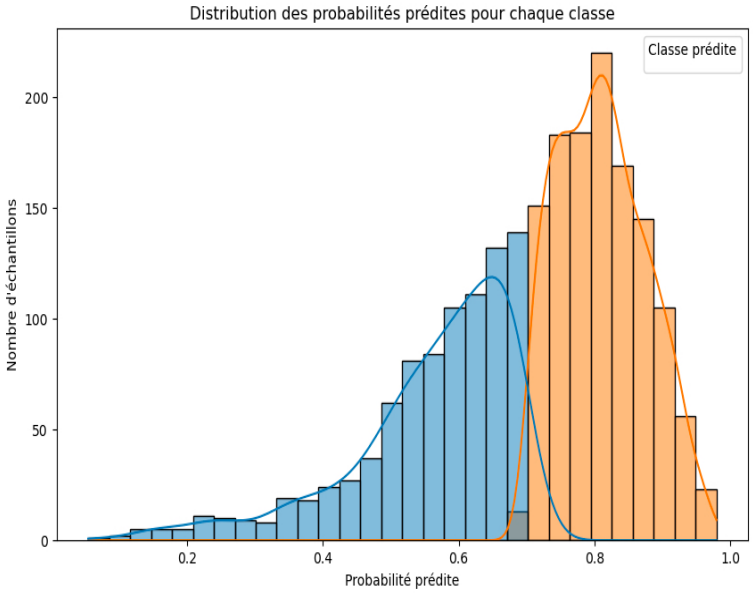

Figure 2.3: Distribution des probabilités prédites pour chaque classe

2.1.1 Interprétation des résultats :

L'analyse des coefficients de la régression logistique nous permet de déterminer les variables les plus influentes sur la probabilité de recouvrement complet, ce qui est crucial pour prédire et optimiser les stratégies de recouvrement. Notre modèle a donner de bonne performances predictives avec une AUC de 0.71

Nous observons que certaines variables contribuent positivement à la probabilité de recouvrement complet, telles que le nombre d'emprunteurs, le nombre de mois d'ancienneté professionnelle, le délai entre la date de mise en place du prêt et la date d'entrée en défaut, l'âge maximal à l'entrée en défaut, l'exposition nette de crédit à l'entrée en première charge hypothécaire, ainsi que certaines catégories de biens comme les biens de type mixte, les résidences principales ou secondaires et les propriétés résidentielles. D'autre part, certaines variables contribuent négativement à la probabilité de recouvrement complet, telles que le délai entre la date d'entrée en défaut et la date de sortie, le cumul des décaissements et certaines catégories de biens comme les investissements locatifs.

Plus précisément :

- Le motif du dernier défaut connu au niveau de l'opération CD_MTF_ENE_CTX 1.07 augmente la probabilité de recouvrement complet de 7% lors du passage d'un motif à un autre, cependant, cette estimation n'est pas robuste en raison du déséquilibre des modalités, et l'augmentation n'est pas constante entre ces modalités.

- Le nombre d'emprunteurs au niveau de l'opération, représenté par la variable nb_empr_ope, montre qu'une augmentation d'une unité du nombre d'emprunteurs augmente la probabilité de recouvrement complet de 86%, toutes choses étant égales par ailleurs. Ceci est intuitif car plus il y a d'emprunteurs impliqués, plus il y a de ressources potentielles pour assurer le recouvrement complet.

- Le nombre de mois d'ancienneté professionnelle (maximum sur les emprunteurs), indiqué par ANC_PRO_MAX_PFI, démontre qu'une augmentation d'une unité (un mois) d'ancienneté professionnelle induit une hausse de 20% de la probabilité que l'emprunteur soit recouvert complètement, ce qui est cohérent avec l'intuition car une expérience professionnelle plus longue est généralement associée à une stabilité financière accrue.

- La variable nb_empr_ope, qui représente le délai en mois entre la date d'entrée en défaut et la date de sortie, ou entre la date d'entrée en défaut et la date d'arrêté, montre qu'une augmentation d'un mois de ce délai diminue la probabilité que le dossier soit complètement recouvert de 25% TCEPA. Ceci est intuitif car plus l'emprunteur reste en défaut pendant une période prolongée, plus les intérêts s'accumulent, rendant le recouvrement complet plus compliqué.

- Le drapeau de vente à l'amiable, représenté par fl_fam_vam, déclenché à la date DET, augmente de 121% la probabilité d'être recouvert complètement par rapport à ne pas l'être. Cette observation semble intuitive car les dossiers non déclenchés pour une vente à l'amiable sont généralement moins susceptibles d'être recouverts complètement.

- Le délai en mois entre la date de mise en place du prêt et la date d'entrée en défaut, représenté par dls_mep_entree_def, montre qu'une augmentation d'un mois entre ces dates augmente de 4% la probabilité de recouvrement complet du dossier TCEPA. Cela s'explique intuitivement par une entrée en défaut plus tardive par rapport à la date de prêt, suggérant une plus grande solvabilité. Cependant, cette augmentation est légère en raison d'un biais potentiel, car elle dépend également du montant du prêt.

- L'âge maximal à l'entrée en défaut (max_age_ctx) : Une augmentation d'un mois de l'âge de l'emprunteur à l'entrée en défaut augmente la probabilité de recouvrement complet de 20% TCEPA, car un emprunteur plus âgé peut être perçu comme ayant une stabilité financière plus établie et une meilleure compréhension des obligations de remboursement.

- Le cumul des décaissements (cum_dec) : Une augmentation d'une unité du cumul des décaissements diminue la probabilité de recouvrement complet du dossier de 10% TCEPA. Ceci s'explique aisément par le fait que plus le montant décaissé cumulé augmente, plus le temps de recouvrement augmente et plus l'individu restera en défaut plus longtemps.

- L'exposition nette de crédit (EAD) à l'entrée en première charge hypothécaire (CHR), observée sur le dernier arrêté connu (MT_EAD_DNE_CHR_d) : Une augmentation d'une unité du montant de l'EAD augmente la probabilité de recouvrement complet d'un dossier de 16% TCEPA. Cela peut indiquer une plus grande capacité financière de l'emprunteur ou une garantie plus solide associée au prêt.

- La destination du bien (CD_DEST_PFI) : - Le fait que le dossier concerne un bien de type mixte par rapport à un bien d'investissement locatif augmente la probabilité de recouvrement complet de 46% TCEPA. - Le fait que le dossier concerne un bien de type XX, ZZ ou NR (autres) par rapport à un bien d'investissement lo-

catif augmente la probabilité de recouvrement complet de 3% TCEPA. - Le fait que le dossier concerne un bien de type résidence principale ou secondaire par rapport à un bien d'investissement locatif augmente la probabilité de recouvrement complet de 42% TCEPA. - Le fait que le dossier concerne un bien de type propriété résidentielle par rapport à un bien d'investissement locatif augmente la probabilité de recouvrement complet de 22% TCEPA. Cela suggère que les biens résidentiels offrent une meilleure garantie pour les prêteurs en cas de défaut.

- Le motif de défaillance (CD_MTF_DFC) : - Le fait d'avoir un motif financier, de procédures légales ou de santé/décès par rapport à d'autres motifs diminue respectivement la probabilité de recouvrement complet d'un dossier de 18%, 82% et 5% TCEPA.

- La nationalité de l'emprunteur 1 (CD_NAT_EMP1) : - Le fait que l'emprunteur soit d'origine française ou étrangère par rapport à être européen diminue respectivement la probabilité de recouvrement complet de 57% et 66% TCEPA. Cela pourrait être dû à des facteurs tels que la stabilité économique ou l'accès aux ressources financières.

- La nature des opérations : - Quand il s'agit d'une nature travaux ou autres par rapport à l'acquisition, cela diminue la probabilité de recouvrement complet, tandis que lorsque c'est un motif de construction, la probabilité augmente.

- Concernant l'impact des catégories socio-professionnelles (CSP) de l'emprunteur 1 : La probabilité de recouvrement complet augmente de 38% lorsque l'emprunteur 1 est fonctionnaire par rapport à d'autres catégories, mais diminue de 40% lorsqu'il est indépendant. Il ne semble pas y avoir de différence entre les emprunteurs qui travaillent dans le privé et d'autres catégories sur la probabilité de recouvrement complet TCEPA.

2.2 Modèle de Durée: COX

Le modèle de Cox, ou "modèle continu semi-paramétrique à risques proportionnels", est un modèle de régression en temps continu. L'objectif est de modéliser le logarithme du rique instantané en fonction d'un ensemble de variables explicatives x dont la valeur peut éventuellement varier au fil du temps.

$$h(t \mid \mathbf{X}) = h_0(t) \times \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

- $h(t \mid \mathbf{X})$ est la fonction de risque instantané pour un individu avec un ensemble de valeurs spécifiques pour les covariables explicatives $\beta_1, \beta_2, \ldots, \beta_p$.
- $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ les covariables au temps t.
- $h_0(t)$ est la fonction de risque de base qui représente le risque de l'événement pour un individu en l'absence d'effet des covariables.
- $\beta_1, \beta_2, \ldots, \beta_p$ sont les coefficients de régression associés aux covariables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, respectivement.

Le modèle de Cox postule que les risques sont proportionnels entre individus quelle que soit la durée depuis le début de l'étude. Formellement, pour deux individus quelconques i et j, avec X étant l'ensemble des covariables, l'hypothèse de proportionnalité peut être exprimée comme suit :

$$\mathrm{h}_i(t) = \frac{h_j(t)}{h_i(t)} = \text{constante}, \quad \forall t$$

2.2.1 Train Test Split

Nous avons procédé en divisant les données en ensembles d'entraînement et de test en tenant compte de l'appartenance des données à un même client. Cette approche est préférable à une division aléatoire car les opérations d'un même client se prosuivent dans le temps. En divisant les données par client, on s'assure que le modèle est évalué sur des clients qu'il n'a pas vus lors de l'entraînement, ce qui permet d'obtenir une estimation plus réaliste de ses performances en conditions réelles. De plus, en veillant à ce que les données de test suivent chronologiquement les données d'entraînement, on évite les problèmes de fuite d'informations du futur lors de la modélisation. En résumé, cette approche garantit une évaluation plus robuste du modèle et une meilleure simulation de son déploiement dans un environnement opérationnel.

2.2.2 Modélisation

Avant d'entamer la modélisation , nous avons fait une séléction de variables en se basant sur la sélection faite avec RFE et en rajoutant des variables qui évoluent dans le temps. Dans ce processus, nous utilisons le module lifelines de Python pour entrainer la régression de Cox.

La première étape consiste à créer une instance du modèle de régression de Cox à l'aide de la classe CoxPHFitter sur la dataframe avec les premières lignes pour chaque identifiant comme dans la regression logistique. Ensuite, nous ajustons le modèle initial en fournissant les données d'entraînement ainsi que les variables pertinentes, en spécifiant la durée jusqu'à l'événement et la colonne indiquant si l'événement s'est produit ou non. Après avoir ajusté le modèle, nous examinons les coefficients et les p-values associées pour chaque variable pour évaluer leur significativité statistique.

Nous avons utiliser ensuite la classe CoxTimeVaryingFitter du module Lifelines pour ajuster un modèle de régression de Cox qui prend en compte les variables qui évoluent dans le temps. Nos données de train et de test sont définit comme dans la partie (2.2.1 Train Test Split).

2.2.3 Résultats

Après avoir estimé du modèle de durée Cox, qui a pris en compte les variables évoluant dans le temps, nous sommes maintenant en mesure d'interpréter ces résultats pour identifier les variables ayant le plus d'impact sur la probabilité de recouvrement complet. Cette analyse nous aidera à élaborer des scores pour nos clients, ce qui nous permettra de prendre des décisions adaptées pour limiter les pertes de recouvrement pour Credit Logement.

Nous pouvons remarquer à l'aide du rapport des risques et du graphique representant les signfificativité des coefficients, les variables qui contribuent le plus à expliquer la durée de recouvrement complet sont : nb_empr_ope , CD_NAT_EMP1, CD_DEST_PFI, fl_fam_vam, fam_ETH, fam_sim, , fam_AEP.

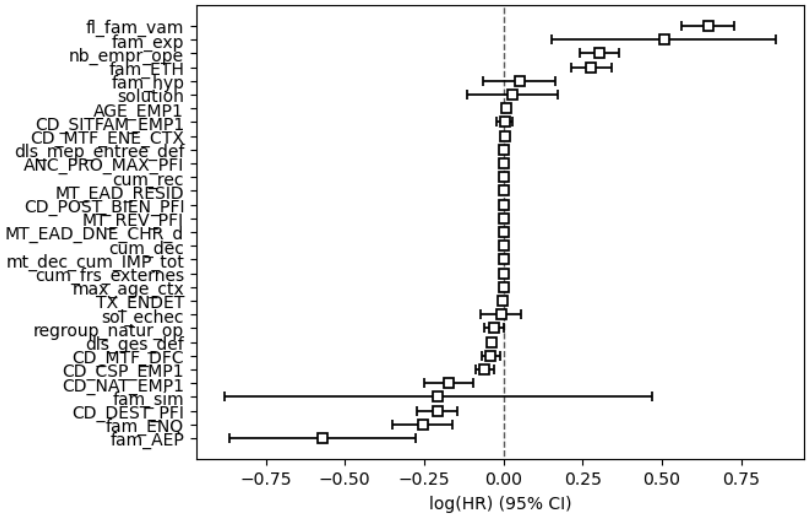

Figure 2.4: Significativité des rapport de risque du Modele de durée Cox

En particulier, nous examinons les courbes de survie pour certaines variables :

- La fonction de survie pour la variable fl-fam-vam montre que les dossiers vendus à l'amiable ont une courbe inférieure à ceux qui ne le sont pas, ce qui suggère que la probabilité de ne pas être recouvert à l'amiable diminue avec le temps. De même, pour la variable fam-ETH lorsque elle vaut un, indiquant une procédure ETH.

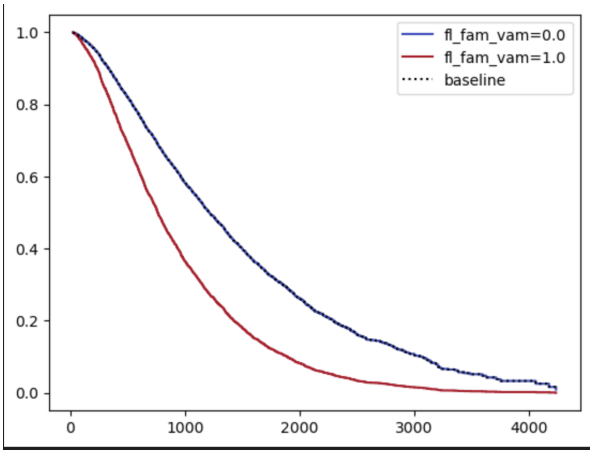

Figure 2.5: Figure : Fonction de survie de la variable fl_fam_vam

- Les courbes de survie pour les emprunteurs français sont inférieures à celles des emprunteurs européens, mais les deux sont supérieures à la fonction de survie globale de la base de données et à celle des étrangers. L'absence de croisement entre les courbes

(modalités) suggère que la variable est significative, mais il convient de prendre en compte le déséquilibre des classes.

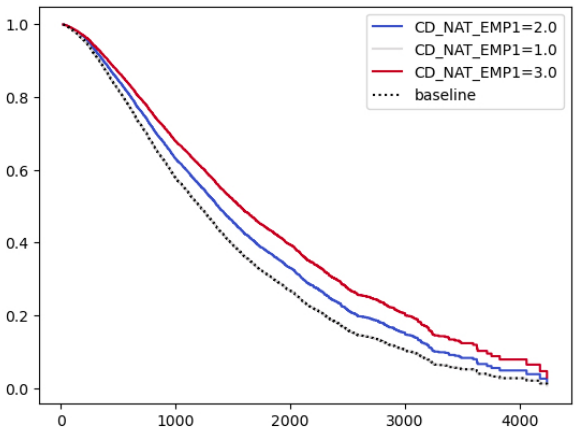

Figure 2.6: Figure : Fonction de survie de la variable CD_NAT_EMP1

- Pour la variable CD-SITFAM-EMP1, les courbes de survie sont confondues pour toutes les modalités de situation familiale de l'emprunteur 1 (autres, célibataire, marié, séparé, veuf), ce qui indique que cette variable n'est pas significative.

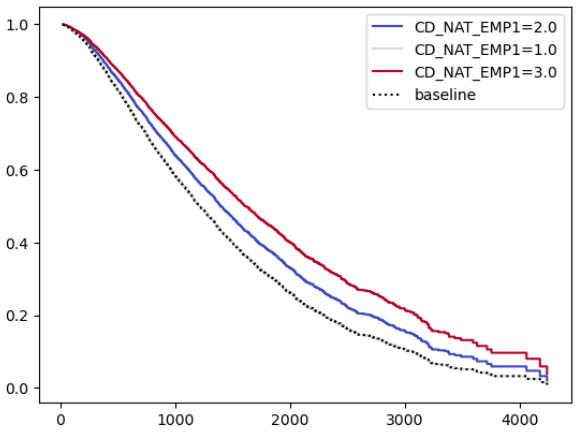

Figure 2.7: Figure : Fonction de survie de la variable CD_SITFAM_EMP1