Confiance de classe pour la prédiction de dette en gestion immobilière

Soundouss Messoudi, Sébastien Destercke, Sylvain Rousseau

Confiance de classe pour la prédiction de dette en gestion immobilière

Soundouss Messoudi[1] Sébastien Destercke[1] Sylvain Rousseau[1]
[1] HEUDIASYC - UMR CNRS 7253

Université de Technologie de Compiègne,
57 avenue de Landshut, 60203 COMPIEGNE CEDEX - FRANCE
prénom.nom@hds.utc.fr

Résumé :

La prédiction des locataires susceptibles de tomber en situation d'endettement est un enjeu majeur pour les bailleurs sociaux dans l'immobilier. Il est encore plus important pour eux de limiter le nombre de personnes faussement prédites comme endettées pour éviter d'engager des coûts inutiles (en temps et en argent), par exemple en envoyant des agents pour éviter la dette. Dans cet article, nous adaptons la prédiction conformelle mondrian pour contrôler le taux d'erreur de cette classe, tout en gardant un niveau de confiance choisi par le propriétaire du bien social, ou plus généralement par l'utilisateur. Cette petite adaptation nous permet non seulement de répondre à notre exigence d'application, mais ouvre également des questions intéressantes concernant les niveaux de confiance différenciés.

Mots-clés :

Prédiction conformelle inductive, prédiction conformelle mondrian, confiance par classe, classification de la dette, immobilier.

Abstract:

The prediction of tenants likely to fall into a debt situation is a key issue for social property owners in real estate. It is even more important for them to limit the number of people falsely predicted to be in debt to avoid incurring unnecessary costs (in time and money), for instance by sending agents to prevent the debt. In this paper, we adapt mondrian conformal prediction to control the error rate of this class, while keeping a level of confidence chosen by the social property owner, or more generally by the user. This small adaptation not only allows us to answer our application requirement but also opens up interesting questions regarding differentiated confidence levels.

Keywords:

inductive conformal prediction, mondrian conformal prediction, class-wise confidence, debt classification, real estate.

1 Introduction

Le logement social est un logement destiné aux personnes à revenus modestes qui auraient du mal à trouver un logement sur le marché privé. Il est accordé sous conditions de revenus ou de composition familiale. Même si ces loyers sont inférieurs aux loyers moyens du secteur géographique, le logement social reste victime de loyers impayés. Les faits générateurs expliquant ces impayés proviennent de raisons prévisibles (précarité de l'emploi, prêts à la consommation multiples, etc.) ou imprévisibles (budget serré, problèmes de santé, changement de situation familiale comme une naissance ou un divorce ...). Aussi, le nombre de ménages endettés est susceptible d'augmenter avec la crise sanitaire due à la pandémie de la COVID-19 [6]. Dans chaque cas, les bailleurs sociaux doivent étudier la situation particulière du ménage en difficulté de paiement et engager des conseillers sociaux et des contentieux pour les aider à résoudre leurs problèmes avant de passer à la phase d'expulsion.

Ainsi, il est essentiel pour les bailleurs sociaux d'anticiper l'endettement de certains locataires, et surtout de limiter le nombre de locataires mal classés comme endettés. D'un côté, cela permet d'éviter de payer des frais inutiles, ou de perdre le temps de l'agent social qui autrement aurait pu être utilisé au profit de locataires vraiment dans le besoin. D'un autre côté, s'il est important de maintenir une bonne précision globale, la précision sur la classe de dette n'est pas si importante, car la classification erronée d'un locataire comme n'ayant pas de dette ne fait que retarder la procédure de recouvrement. Pour contrôler tous ces facteurs, nous proposons une approche de confiance par classe basée sur la prédiction conformelle mondrian.

La prédiction conformelle mondrian est une variante de la prédiction conformelle, un cadre qui fournit une garantie statistique en donnant un

ensemble de classes dans le cas de la classification et un intervalle de prédiction dans le cas de la régression. L'une des caractéristiques souhaitables des prédicteurs conformels est la validité, c'est-à-dire que le taux d'erreur ne dépasse pas une erreur de probabilité choisie par l'utilisateur. Dans le cas des prédicteurs conformels mondrian, la validité est vérifiée dans des catégories du jeu de données au lieu du jeu de données global. Le principe de la prédiction conformelle et sa forme mondrian pour la classification dans le cadre inductif sera rappelé dans la section 2.

Notre travail utilise la classification conformelle mondrian pour obtenir une confiance de classe appliquée à la prévision de la dette dans la gestion immobilière. Par confiance de classe, nous entendons que nous ne cherchons pas à obtenir la même précision pour toutes les classes, par exemple pour des raisons sensibles aux coûts. Notre approche est décrite dans la section 3. Les expériences et leurs résultats sont présentés dans la section 4, où nous adaptons une mesure de non-conformité mondrian standard pour contrôler l'erreur de classe la plus importante pour le bailleur social, tout en garantissant une confiance globale.

2 Prédiction conformelle mondrian

Dans le cas de la classification, la prédiction conformelle est un cadre qui fournit une garantie statistique de la couverture de la vraie classe en prédisant un sous-ensemble de toutes les classes avec un niveau de confiance défini par l'utilisateur. Ce cadre a d'abord été développé pour un environnement transductif en ligne [4, 11] qui a besoin de former le modèle à chaque fois qu'une prédiction est recherchée, puis a été adapté au paramètre inductif [9] pour l'adapter aux tâches coûteuses en temps d'entraînement surtout lorsqu'il y a beaucoup de données (par exemple : les réseaux de neurones). L'approche inductive permet donc de résoudre ce problème en conservant un ensemble d'exemples d'entraînement pour la calibration au lieu de les utiliser tous pour entraîner l'algorithme sous-jacent. Ces méthodes classiques de prédiction conformelle garantissent un niveau de confiance global. Une autre approche appelée prédiction conformelle mondrian [12] a été proposée pour obtenir une garantie statistique sur un sous-ensemble du jeu de données basé sur une condition. Cette section présente cette méthode et quelques travaux connexes en classification.

2.1 Prédiction conformelle inductive (ICP) pour la classification

Soient $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n) \in \mathbf{Z}$ des paires successives constituant les exemples, avec $x_i \in \mathbf{X}$ un objet et $y_i \in \mathbf{Y} = \{C_1, \ldots, C_p\}$ sa classe. Soit $\mathbf{Z}$ échangeable (une condition plus faible que l'i.i.d.). On peut prédire $y_{n+1} \in \mathbf{Y}$ pour tout nouvel objet $x_{n+1} \in \mathbf{X}$ selon les étapes de la prédiction conformelle inductive :

1. Diviser l'ensemble de données d'origine $\mathbf{Z}$ en un *ensemble d'apprentissage propre* $\mathbf{Z}^{tr}$ avec $|\mathbf{Z}^{tr}| = m$ et un *ensemble de calibration* $\mathbf{Z}^{cal}$ avec $|\mathbf{Z}^{cal}| = n - m = q$.
2. Entraîner un *algorithme sous-jacent* de classification $h : \mathbf{X} \to \mathbf{Y}$ sur $\mathbf{Z}^{tr}$ pour obtenir la *mesure de non-conformité* $f(z)$. La mesure de non-conformité standard en classification est :

$$f(z) = 1 - \hat{P}_h[y \mid x]. \qquad (1)$$

3. Appliquer $f(z)$ à chaque exemple z_i de $\mathbf{Z}^{cal}$ pour obtenir les scores de non-conformité $\alpha_1, \ldots, \alpha_q$.
4. Choisir un *niveau de signifiance* $\epsilon \in (0, 1)$ pour avoir un ensemble de prédictions avec un *niveau de confiance* de $1 - \epsilon$.
5. Pour un nouvel exemple x_{n+1}, calculer un score de non-conformité pour chaque classe $C_k \in \mathbf{Y}$:

$$\alpha_{n+1}^{C_k} = f((x_{n+1}, y = C_k)). \qquad (2)$$

6. Pour chaque classe $C_k \in \mathbf{Y}$, calculer la p-value :

$$p_{n+1}^{C_k} = \frac{|\{i \in 1, \ldots, q : \alpha_{n+1}^{C_k} \leq \alpha_i\}|}{q}. \quad (3)$$

7. Construire l'ensemble de prédictions :

$$\Gamma^{\epsilon} = \{C_k \in \mathbf{Y} : p_{n+1}^{C_k} > \epsilon\}. \quad (4)$$

L'ensemble de prédictions peut être un singleton lorsque le prédicteur est sûr, un ensemble avec plus d'une classe en cas d'ambiguïté et un ensemble vide $\emptyset$ lorsque le modèle ne sait pas ou n'a pas vu d'exemple similaire pendant l'entraînement. Les deux propriétés souhaitables dans les prédicteurs conformels sont (a) *la validité*, c-à-d le taux d'erreur ne dépasse pas ϵ pour chaque niveau de confiance choisi $1 - \epsilon$, et (b) *l'efficacité*, signifiant les ensembles de prédiction sont aussi petits que possible.

2.2 Prédiction conformelle mondrian (MCP)

Comme mentionné ci-dessus, la caractéristique de validité souhaitée dans un classifieur conformel garantit que le niveau de confiance global est maintenu sur l'ensemble du jeu de données en choisissant un taux d'erreur ϵ qui ne doit pas être dépassé. Il n'y a cependant aucune garantie sur un sous-ensemble du jeu de données ou sur des catégories spécifiques de l'ensemble de données. La prédiction conformelle mondrian fournit cette validité de sous-ensemble basée sur des catégories telles que les catégories conditionnelles de classe ou d'attribut [12]. Dans ce cas, chaque catégorie a sa propre garantie individuelle basée sur le niveau de signification individuel choisi. Par exemple, un classifieur conformel conditionnel de classe donnera une validité sur chaque classe, tandis qu'un classifieur conformel conditionnel d'attribut garantira le taux d'erreur pour chaque catégorie de l'attribut choisi utilisé pour diviser l'ensemble de données. Nous nous concentrons sur la forme conditionnelle de classe des prédicteurs conformels mondrian car c'est ce que nous utilisons dans notre article.

La différence entre un classifieur conformel inductif et un classifieur conformel conditionnel de classe réside dans le calcul de la p-value (3), dans laquelle, au lieu de prendre tous les scores de non-conformité α_i dans le jeu de calibration, nous ne considérons que ceux liés aux exemples appartenant à la même classe que nous testons hypothétiquement pour l'objet x_{n+1}. La valeur p devient :

$$p_{n+1}^{C_k} = \frac{|\{i \in 1, \ldots, q : y_i = C_k, \alpha_{n+1}^{C_k} \leq \alpha_i\}|}{|\{i \in 1, \ldots, q : y_i = C_k\}|}. \quad (5)$$

Le MCP conditionnel de classe est principalement utilisé lorsque les données sont déséquilibrées, afin de maintenir le même taux d'erreur même pour la classe minoritaire.

3 Confiance de classe : notre approche

Cette section présente notre approche pour une confiance par classe utilisant la prédiction conformelle mondrian, dans le cas d'un problème binaire.

Notre objectif principal dans cet article et l'application associée est de contrôler le taux d'erreur d'une classe donnée, tout en préservant un taux d'erreur global relativement faible. Une spécificité de ce cas est que le taux d'erreur de la classe restante est d'une importance marginale, et c'est dans une certaine mesure ce qui compte vraiment pour cette seconde classe, c'est l'efficacité, c'est-à-dire la capacité à identifier certains échantillons qui lui appartiennent.

En pratique, cela signifie que nous devons spécifier différents niveaux de signifiance pour les classes et pour le jeu de données global. Pour ce faire, nous adaptons la prédiction conformelle mondrian conditionnelle aux classes.

Soient $\epsilon_g \in (0, 1)$ le taux d'erreur global pour tout le jeu de données, $\epsilon_0 \in (0, 1)$ celui spécifié

pour l'étiquette $y = 0$, ce qui signifie que la personne n'est pas endettée, et $\epsilon_1 \in (0, 1)$ celui lié à la classe $y = 1$, c'est-à-dire que la personne est endettée. ϵ_0 est donc la variable sur laquelle on souhaite avoir un contrôle fort (afin de ne pas envoyer d'agents sociaux inutiles aux locataires). On a

$$\epsilon_g = \epsilon_0 \mathbb{P}(y = 0) + \epsilon_1 \mathbb{P}(y = 1). \quad (6)$$

Avec ϵ_g et ϵ_0 choisis par l'utilisateur, et à condition que nous ayons des estimations raisonnables de $\mathbb{P}(y = 0)$ et $\mathbb{P}(y = 1)$, on peut calculer ϵ_1 avec :

$$\epsilon_1 = \frac{\epsilon_g - \epsilon_0 \mathbb{P}(y = 0)}{\mathbb{P}(y = 1)}. \quad (7)$$

En définissant ϵ_g et ϵ_0, et puisque $\epsilon_1 \in (0, 1)$, il faut respecter la condition :

$$\epsilon_0 \mathbb{P}(y = 0) < \epsilon_g < \epsilon_0 + (1 - \epsilon_0)\mathbb{P}(y = 1), \quad (8)$$

sinon nous obtiendrions un ϵ_1 irréalisable.

Cela nous permet d'avoir des ϵ individuels pour chaque classe qui garantiront un niveau de confiance global pour le jeu de données. En dehors de cette étape, les autres étapes du MCP conditionnel de classe restent les mêmes.

4 Évaluation

4.1 Jeu de données

Les données utilisées dans notre article sont fournies par Sopra Steria, société informatique française bien connue qui propose des logiciels de gestion locative immobilière à ses clients, principalement des bailleurs sociaux. L'origine de notre ensemble de données provient d'un entrepôt de données d'un de ses clients qui contient des enregistrements historiques mensuels de l'activité des locataires de janvier 2018 à décembre 2019. Ces données ont été anonymisées conformément au règlement général sur la protection des données (RGPD) de l'UE afin de protéger les données privées des locataires.

La procédure d'extraction des données a porté sur les informations relatives aux locataires, leur situation personnelle (âge, état matrimonial, ...), leur situation financière (emploi, salaire, ...), leur bien locatif (nombre de pièces, situation géographique,. ..), et les opérations de paiement liées au loyer (loyer, factures, montants encaissés, ...). Sur la base de ces transactions de paiement mensuel, nous avons ajouté une nouvelle variable en calculant le montant de la dette cumulée, qui est une somme de la différence entre le montant encaissé et le montant de la facture pour chaque mois. Ainsi, une personne est considérée comme endettée si le montant de sa dette cumulée est supérieur ou égal au double du montant mensuel brut de son loyer. Sur la base de cette valeur, nous avons pu ajouter une classe booléenne «endettée» qui est égale à 0 si la personne n'est pas endettée, et à 1 si elle l'est. Ensuite, nous avons créé nos exemples en prenant une période de trois mois avant l'occurrence d'une dette (ou au hasard en cas d'absence d'occurrence de dette dans les archives historiques). Notons que pour les personnes endettées, et puisque pour chaque exemple nous prenons un historique de 3 mois sur une période de presque 2 ans, nous pouvons avoir plus d'un exemple pour chaque personne correspondant à des périodes de temps différentes, dans lesquelles cette personne peut être endettée ou non. Cette stratégie d'extraction a été choisie pour avoir un plus grand jeu de données.

Le jeu de données obtenu contient 28566 exemples, 44 variables et une classe booléenne avec 1 étant endetté et 0 non endetté. Il est également fortement déséquilibré puisque seulement $7,89\%$ des locataires sont endettés et comporte de nombreuses valeurs manquantes, en raison du fait que certaines données sont recueillies au moyen d'enquêtes annuelles.

4.2 Cadre expérimental

Dans notre étude, nous nous sommes concentrés sur deux expériences principales, la première étant une comparaison entre ICP et MCP avec le même ϵ_g, et la seconde

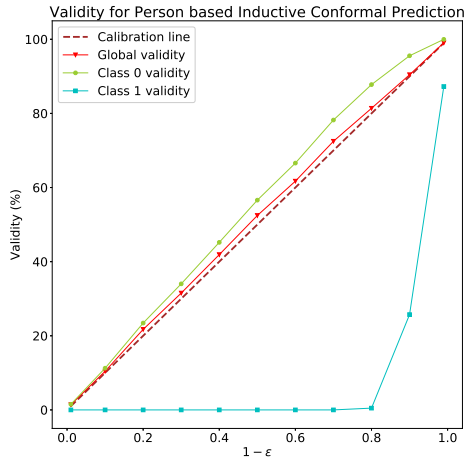

Figure 1 – Résultats de validité pour ICP.

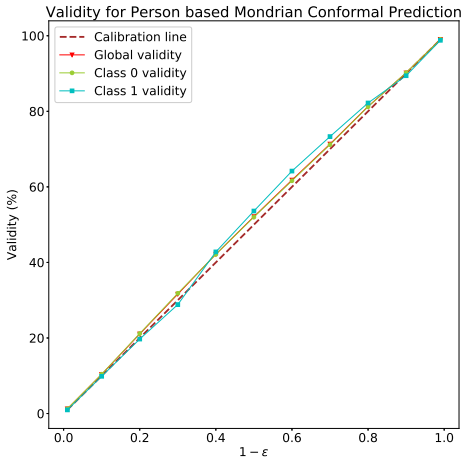

Figure 2 – Résultats de validité pour MCP.

étant une comparaison entre notre approche et MCP par rapport au contrôle de ϵ_g et ϵ_0.

Nous avons choisi l'algorithme de gradient boosting ≪LightGBM≫ comme algorithme sous-jacent car il peut gérer les valeurs manquantes et les variables tant catégorielles que numériques. Nous avons utilisé la mesure de non-conformité standard $f(z) = 1 - \hat{P}_h[y \mid x]$. Nous avons également découpé le jeu de données, avec un ensemble de test égal à 20% du jeu de données, et avec un ensemble de calibration égal à 20% des exemples d'entraînement. Le découpage a été fait selon les personnes. Ainsi, les exemples sont répartis en fonction de l'ID du locataire, car nous pouvons avoir de nombreux exemples pour la même personne. Cela signifie que les exemples de l'ensemble de test sont ceux de personnes que l'algorithme n'a pas vues auparavant dans les phases d'entraînement et de calibration. Cela semble être un scénario plus réaliste, car dans la pratique, les locataires à prédire seront de nouveaux clients, différents des anciens utilisés pendant l'entraînement.

La première expérience a été menée sur la base des étapes de l'ICP comme décrit dans la section 2, pour des valeurs de ϵ allant de 0.1 à 0.9, où $\epsilon_g = \epsilon_0 = \epsilon_1$ dans le cas du MCP.

La deuxième expérience a été menée en suivant les étapes de notre approche de confiance par classe comme décrit dans la section 3, avec des valeurs différentes de ϵ_g et ϵ_0 afin de limiter le nombre de personnes mal classées qui sont prédites comme endettées alors qu'en fait elles ne le sont pas. Nous avons comparé les résultats avec une approche MCP.

4.3 Résultats

Cette section présente les résultats de nos expériences, en étudiant en particulier la différence entre ICP et MCP, ainsi que la validité et l'efficacité de l'approche proposée.

Pour notre première expérience, et pour vérifier la validité de ICP et MCP, nous calculons la précision globale et la précision de chaque classe, et les comparons avec la ligne de calibration. Cette ligne représente le cas où le taux d'erreur est exactement égal à ϵ pour un niveau de confiance $1 - \epsilon$, ce que nous cherchons à obtenir dans un prédicteur conformel exactement valide. Les résultats sont présentés dans les figures 1 et 2.

Dans le cas de l'ICP (Figure 1), les résultats montrent que la validité globale du jeu de données est atteinte. Cependant, elle n'est pas

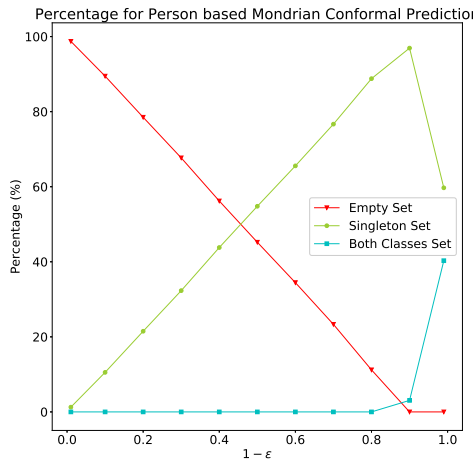

Figure 3 – Résultats de l'efficience pour MCP.

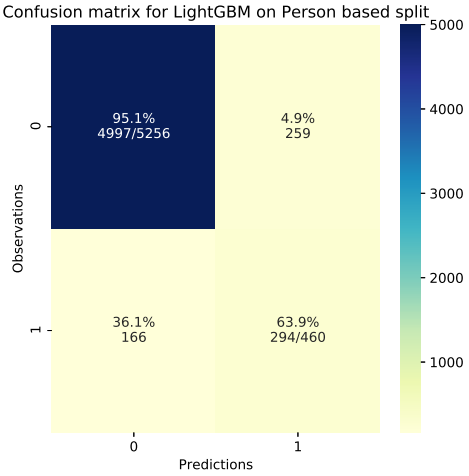

Figure 4 – Matrice de confusion du LightGBM.

respecté pour les classes, surtout pour la classe minoritaire 1 correspondant aux locataires endettés. Cela montre le problème d'avoir un jeu de données déséquilibré en termes de catégories ou de classes, et comment la zone la moins représentée de l'espace des observations souffre le plus lorsqu'une méthode de prédiction conformelle simple est utilisée. En effet, une très mauvaise validité pour la classe minoritaire peut être compensée par une validité légèrement conservatrice pour la classe majoritaire. Ce problème est résolu dans le cas du MCP (Figure 2), qui donne de meilleurs résultats de validité qui sont presque exactement valables pour le jeu de données global et aussi pour chaque classe individuelle, y compris la classe minoritaire 1.

Pour évaluer l'efficacité de l'ICP et du MCP, nous avons calculé le pourcentage de singletons, d'ensembles vides $\emptyset$ et d'ensembles $\{0, 1\}$ à partir de toutes les prédictions des exemples de test. Ayant eu des résultats similaires dans les deux cas, nous présentons les résultats de MCP uniquement dans la figure 3.

Pour les résultats d'efficacité, on remarque que lorsque ϵ diminue, le pourcentage d'ensembles vides prédits $\emptyset$ diminue. Il n'est même plus prédit (à $\epsilon = 0,1$). Inversement, le contraire est observé avec le pourcentage de singletons qui croît constamment à mesure que ϵ diminue jusqu'à $\epsilon = 0,2$. A partir de ce moment, on remarque un effet miroir entre le pourcentage de singletons et le pourcentage d'ensembles $\{0, 1\}$, qui était nul jusqu'à présent, et qui augmente alors que le pourcentage de singletons diminue, avec des valeurs plus grandes dans le cas du MCP comparé à ICP. Cela peut s'expliquer par le fait que dans MCP, le niveau de confiance est garanti pour chaque classe ainsi que pour le jeu de données global, ce qui signifie que le modèle prédit plus d'ensembles $\{0, 1\}$ pour avoir une prédiction plus fiable, même pour la classe minoritaire.

Pour la deuxième expérience, nous avons utilisé notre approche de confiance par classe et spécifié différentes valeurs pour les niveaux de signifiance ϵ_g et ϵ_0. À des fins de comparaison, la figure 4 illustre la matrice de confusion de l'algorithme sous-jacent avec un seuil de $0,3$. Les figures 5 et 6 montrent la matrice de confusion avec $\epsilon_g = 0,05$ et $\epsilon_0 = 0,01$ et avec $\epsilon_g = \epsilon_0 = 0,01$, correspondant à un classifieur MCP classique.

En plus des pourcentages et quantités de données tombant dans chaque cellule, nous avons ajouté la proportion de prédictions single-

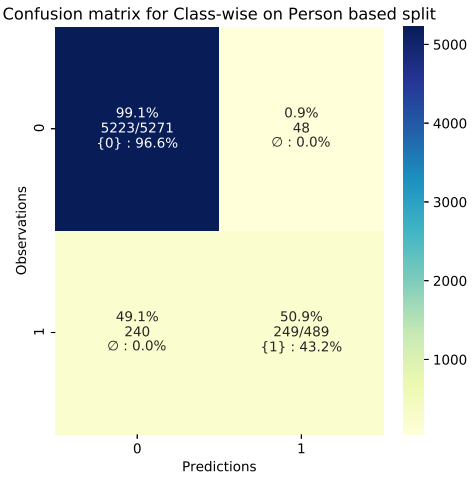

Figure 5 – Matrice de confusion pour la confiance de classe avec $\epsilon_g = 0.05$ et $\epsilon_0 = 0.01$.

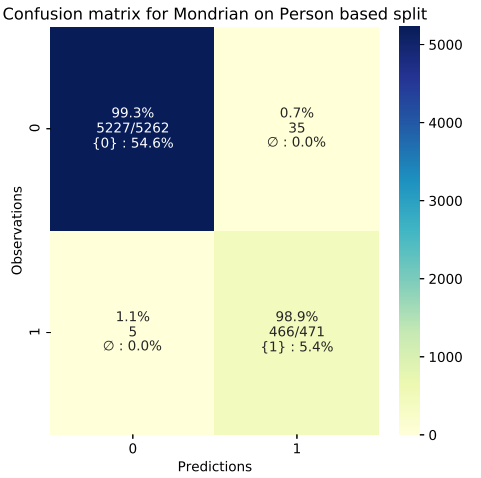

Figure 6 – Matrice de confusion pour la confiance de classe avec $\epsilon_g = \epsilon_0 = 0.01$.

tons dans le cas de prédictions correctes (c.-à-d. le rapport de $\{0\}$ ou $\{1\}$ parmi les prédictions), ou la proportion de prédictions d'ensembles vides en cas de prédictions incorrectes.

Pour notre approche, la figure 5 montre que le taux d'erreur pour la classe 0 est approximativement égal au ϵ_0 choisi, et que le taux d'erreur pour la classe 1 est également d'environ $0,52$, le résultat de l'équation (7) lorsque $\epsilon_0 = 0,01$ et $\epsilon_g = 0,05$. Aussi, la validité globale est approximativement égale à 95 % telle qu'elle a été choisie par les bailleurs sociaux, ce qui montre que notre méthode atteint une confiance de classe tout en gardant une confiance globale, toutes deux choisies par l'utilisateur. De même, la figure 6 montre les résultats attendus pour le choix plus classique $\epsilon_g = \epsilon_0$. Cependant, une différence frappante entre les deux est le nombre de personnes précisément reconnues qui seront endettées ou non endettées, c.-à-d. le pourcentage de singletons. Pour les personnes endettées, dans les matrices de la figure 6, cela équivaut à 25 personnes, alors que dans la matrice de la figure 5, cela équivaut à 108 personnes. Donc, en effet, la précision sur la première classe a considérablement chuté dans notre schéma, même avec une petite marge entre ϵ_g et ϵ_0, mais l'avantage clair est que nous avons pu détecter beaucoup plus (environ 4 fois plus) de locataires problématiques, permettant plus de prévention. Pour les personnes non endettées, 2854 personnes sont prédites avec précision dans le cas du MCP classique, alors qu'avec notre approche par classe, cela équivaut à 5056 personnes, ce qui signifie que les experts auront beaucoup moins de $\{0, 1\}$ cas à vérifier manuellement lors de l'utilisation de notre méthode. Ceci peut s'expliquer par le fait que, contrairement à notre approche, toutes les valeurs ϵ sont égales dans le MCP classique, donc $\epsilon_1 = 0,01$, ce qui conduit à plus de $\{0, 1\}$ ensembles prédits afin d'assurer la validité de 99% pour la classe minoritaire. Par conséquent, cela a un impact sur la classe 0, en diminuant le pourcentage de personnes non endettées prédites avec précision.

5 Conclusion

Dans cet article, nous avons utilisé la prédiction conformelle, et en particulier la variante mondrian conditionnelle de classe, pour avoir plus de contrôle sur le taux d'erreur d'une certaine classe, tout en préservant une confiance globale. Nous avons appliqué notre méthode au domaine

de l'immobilier afin d'aider les bailleurs sociaux à limiter le nombre de personnes faussement prédites comme endettées, car ces erreurs de classification sont coûteuses. Les résultats montrent l'intérêt de cette méthode sur notre jeu de données.

D'un point de vue méthodologique, notre contribution est plutôt modeste mais contribue à résoudre un problème que nous jugeons important : celui de l'identification de la manière dont les différents taux d'erreur devraient être obtenus/choisis en considérant des cadres conformels avec de tels taux d'erreur multiples. En effet, si le choix de taux d'erreur égaux est courant dans les études systématiques, on ne peut guère s'attendre à ce que toutes les cibles aient besoin de la même précision dans les applications réelles. Ceci s'applique bien sûr dans les frameworks où la prédiction conformelle mondrian est en jeu, mais aussi dans les paramètres multi-cibles tels que la régression multi-variée [7] ou la classification multi-classes [5]. Pour étendre notre approche actuelle à plus de deux classes dans le cas mondrian, il faudrait considérer les contraintes reliant les différents degrés de confiance, en adaptant l'équation (8) au cas multi-classes. L'établissement de cibles multiples serait moins problématique, car les degrés de confiance sur les différentes cibles ne sont pas limités les uns par les autres. Il faudrait cependant pouvoir construire le lien entre le degré de confiance global et chaque degré de confiance individuel, en utilisant par exemple des cadres basés sur la copule [8].

Concernant notre application, les perspectives incluent le traitement des valeurs manquantes pour améliorer les résultats de classification, car pour le moment nous utilisons un modèle qui les gère naturellement. Nous aimerions également explorer d'autres mesures de non-conformité autres que celle standard utilisée dans cet article. De plus, il serait intéressant de travailler sur la répartition temporelle, en appliquant des méthodes de prédiction conforme qui traitent des données temporelles telles que des séries temporelles [2].

Références

[1] Candès, Emmanuel J and Lei, Lihua and Ren, Zhimei. Conformalized Survival Analysis. *arXiv preprint arXiv :2103.09763, 2021*.

[2] Chernozhukov, Victor and Wüthrich, Kaspar and Yinchu, Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. *Conference On Learning Theory, 2018*, PMLR, pp 732–749.

[3] Devetyarov, Dmitry and Nouretdinov, Ilia and Burford, Brian and Camuzeaux, Stephane and Gentry-Maharaj, Aleksandra and Tiss, Ali and Smith, Celia and Luo, Zhiyuan and Chervonenkis, Alexey and Hallett, Rachel and others. Conformal predictors in early diagnostics of ovarian and breast cancers. *Progress in Artificial Intelligence, 2012*, Springer, volume 1, number 3, pp 245–257.

[4] Gammerman, Alex and Vovk, Volodya and Vapnik, Vladimir. Learning by transduction. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998*, pp 148–155.

[5] Lambrou, Antonis and Papadopoulos, Harris. Binary relevance multi-label conformal predictor. *Symposium on Conformal and Probabilistic Prediction wIth Applications, 2016*, Springer, pp 90–104.

[6] Manville, Michael and Monkkonen, Paavo and Lens, Michael and Green, Richard. COVID-19 and Renter Distress : Evidence from Los Angeles. *2020*.

[7] Messoudi, Soundouss and Destercke, Sébastien and Rousseau, Sylvain. Conformal multi-target regression using neural networks. *Conformal and Probabilistic Prediction and Applications, 2020*, PMLR, pp 65–83.

[8] Messoudi, Soundouss and Destercke, Sébastien and Rousseau, Sylvain. Copula-based conformal prediction for Multi-Target Regression. *Pattern Recognition, Accepted for publication, 2021*.

[9] Papadopoulos, Harris. Inductive conformal prediction : Theory and application to neural networks. *Tools in artificial intelligence, 2008*, IntechOpen.

[10] Toccaceli, Paolo and Gammerman, Alexander. Combination of inductive mondrian conformal predictors. *Machine Learning, 2019*, Springer, volume 108, number 3, pp 489–510.

[11] Vovk, Vladimir and Gammerman, Alex and Shafer, Glenn. Algorithmic learning in a random world. *Springer Science & Business Media, 2005*.

[12] Vovk, Vladimir and Lindsay, David and Nouretdinov, Ilia and Gammerman, Alex. Mondrian confidence machine. *Technical Report, 2003*.

[13] Yang, Fan and Wang, Hua-zhen and Mi, Hong and Cai, Wei-wen and others. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics, 2009*, BioMed Central, volume 10, number 1, pp 1–14.

[14] Zhang, Shiwei and Zhang, Xiuzhen and Lau, Jey Han and Chan, Jeffrey and Paris, Cecile. Less is More : Rejecting Unreliable Reviews for Product Question Answering. *arXiv preprint arXiv :2007.04526, 2020*.