Prédiction Conforme

Vincent TARDIEUX

Contents

December 7, 2024

Prédiction Conforme

Vincent TARDIEUX

December 7, 2024

1 Description du jeu de données et du prétraitement appliqué

Le jeu de données choisit est disponible sur Kaggle, il correspond à des demandes de crédit et contient notamment le montant du crédit demandé (colonne choisie pour la régression) et également la décision finale sur le crédit (colonne choisie pour la classification).

Le prétraitement appliqué fut assez simple :

- On commence par enlever les colonnes non importantes (En se basant notamment sur la matrice de confusion [1]) et également les colonnes textuelles (catégoriques).
- On normalise nos colonnes entre 0 et 1 à l'aide d'un minmax afin d'avoir des colonnes d'une importance similaire.

Notre jeu de données contient 75829 lignes avec 11 dimensions après prétraitement.

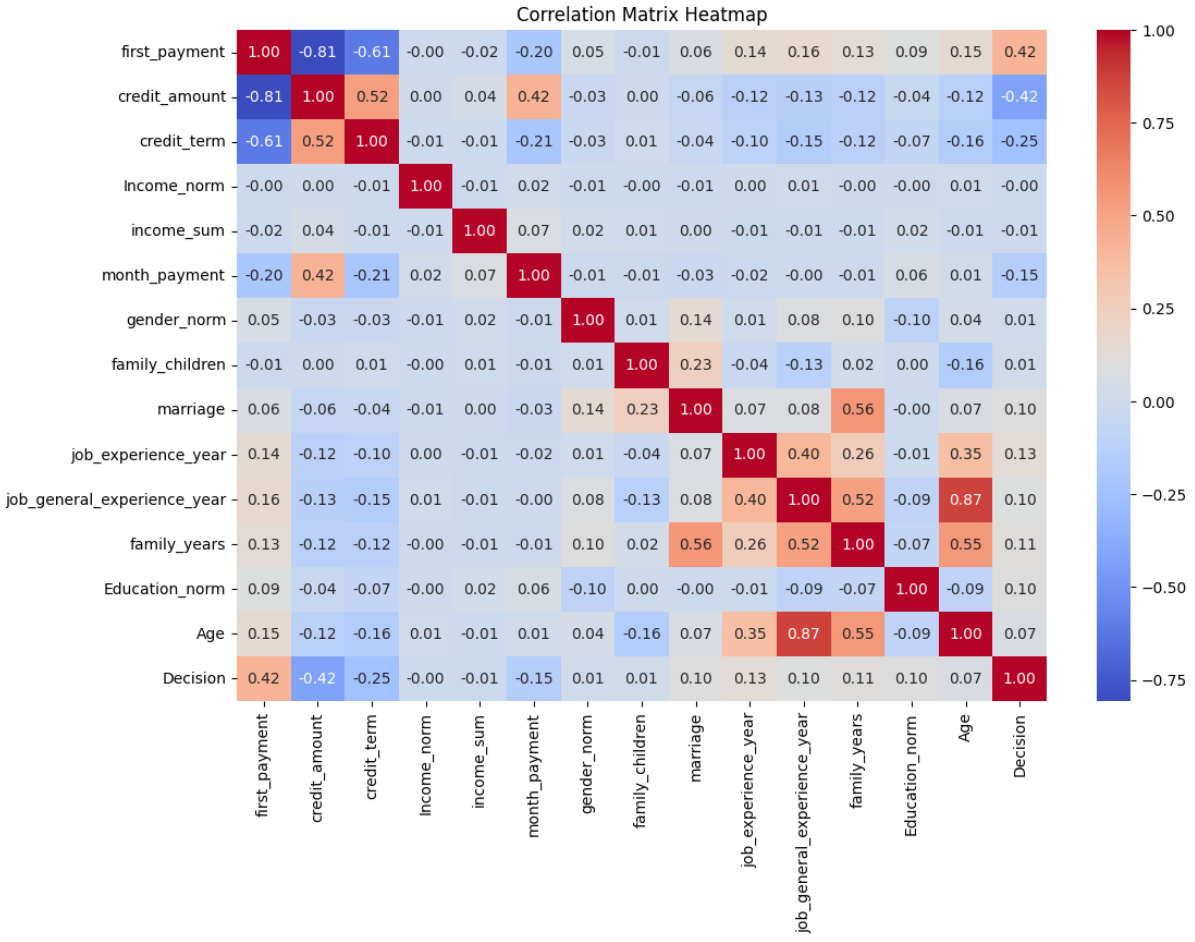

Figure 1: Matrice de confusion de notre jeu de données

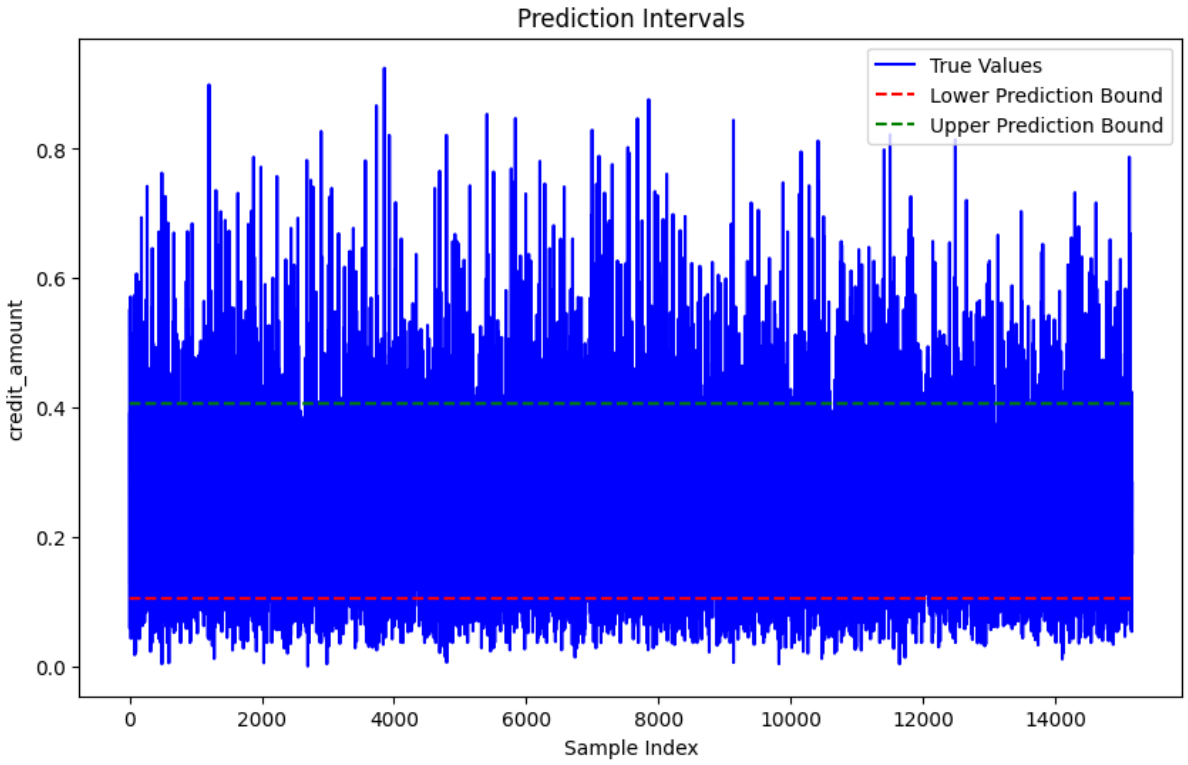

Figure 2: Régression quantile de la quantité de crédit en fonction des intervalles de prédiction

2 Régression quantile

2.1 Choix du jeu de données (pertinence)

Le jeu de données utilisé contient des informations détaillées sur des prêts bancaires, telles que les revenus, les dépenses, l'expérience professionnelle, et d'autres variables pertinentes. Ce jeu de données est plutôt pertinent puisqu'il permet d'effectuer une régression sur la quantité de crédit demandé mais également de la prédiction si oui ou non le crédit à été accepté.

Les données sont bien structurées et représentent un problème réel lié à la finance. L'objectif principal est d'exploiter ces données pour établir une relation entre les caractéristiques des clients et la quantité de crédit demandée, ce qui est essentiel pour modéliser le comportement des emprunteurs et assister les institutions financières dans leur prise de décision.

2.2 Mise en place du modèle

Pour la tâche de régression quantile, nous avons utilisé le modèle **QuantileRegressor**. Ce modèle est bien adapté pour générer des intervalles de prédiction robustes, car il permet de prédire directement les quantiles supérieurs et inférieurs de la variable cible.

Les paramètres utilisés incluent :

- `quantile`: 0.1 pour la borne inférieure (resp 0.9 pour la borne supérieure)
- `solver`: 'highs'

2.3 Interprétation

A l'aide de la figure [2], on constate nos bornes inférieures et supérieures étant respectivement à 0.1 et 0.4. De plus on constate un taux de couverture de 80.13%, ce qui est logique puisque l'on a pris les quantiles 10% et 90% comme limites.

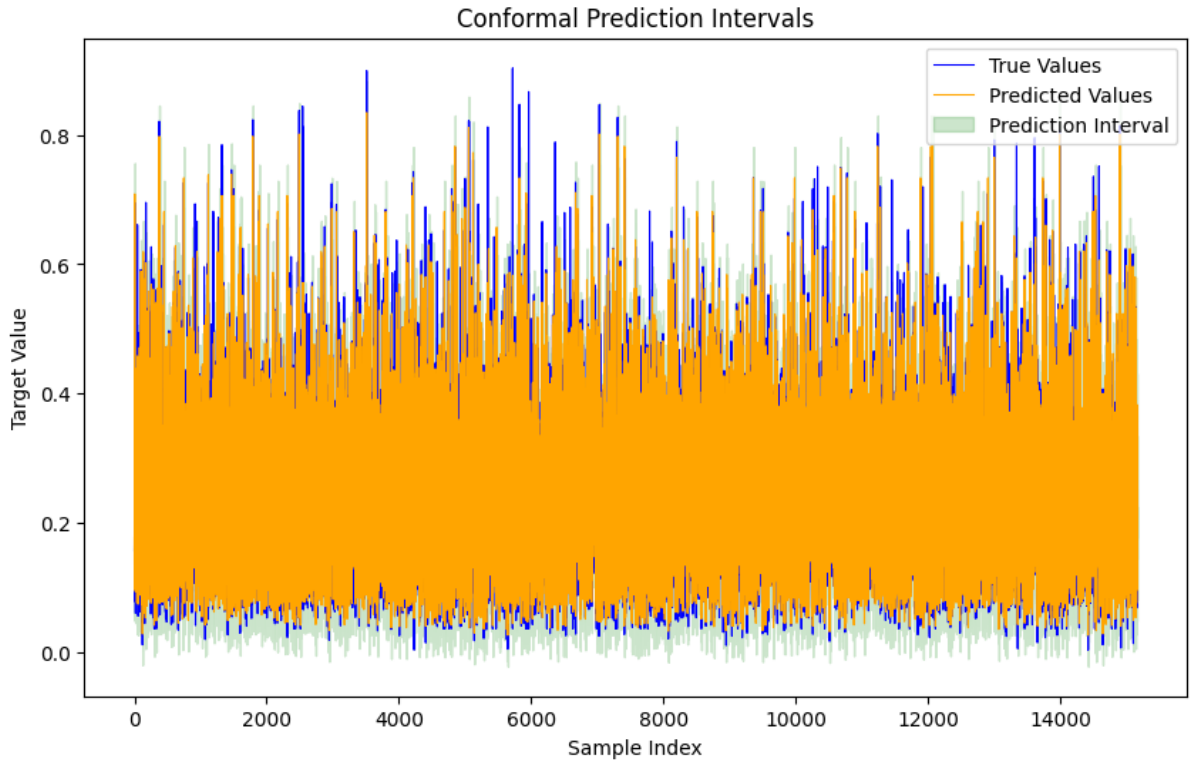

Figure 3: Régression quantile de la quantité de crédit en fonction des intervalles de prédiction

3 Prédiction conforme

3.1 Choix du jeu de données (justification + description du problème)

Le problème de prédiction conforme consiste à générer des intervalles de prédiction garantissant une couverture statistique spécifique. Ce jeu de données est pertinent car il contient des caractéristiques influentes pour les prédictions de régression et de classification, telles que l'expérience professionnelle et le nombre d'enfants des emprunteurs.

3.2 Mise en place du modèle de régression

Nous avons mis en œuvre une méthode de prédiction conforme en utilisant le **GradientBoostingRegressor** pour ajuster les quantiles. Cela permet de générer des intervalles de confiance qui respectent une probabilité de couverture définie, tout en s'adaptant à la distribution des données.

3.3 Interprétation de la régression

Les intervalles de prédiction conformes montrent une couverture adéquate sur l'ensemble de test, atteignant le niveau attendu (**90.47%**). Cela démontre que le modèle est capable de capturer efficacement l'incertitude des prédictions tout en limitant les erreurs. De plus sur la figure 3 on constate que les données prédites semblent coller sur notre vérité terrain.

3.4 Mise en place du modèle de classification

Pour la classification, nous avons utilisé un modèle basé sur des forêts aléatoires (**RandomForestClassifier**), ajusté pour distinguer si un crédit est accepté (`1`) ou refusé (`0`).

3.5 Comparaison des algorithmes

3.5.1 Présentation des algorithmes

- **SCP (Split Conformal Prediction)** : Cet algorithme divise les données en deux ensembles distincts. Le premier est utilisé pour entraîner le modèle prédictif, tandis que le second est réservé pour calibrer les intervalles de prédiction. Cette séparation garantit que les intervalles de prédiction ne sont pas biaisés par le modèle.
- **CV+ (Cross Validation Plus)** : CV+ est une extension de la validation croisée classique. Elle génère des intervalles en effectuant des validations croisées sur plusieurs sous-ensembles des données, offrant des intervalles robustes et réduisant les variations dues à un sous-échantillonnage spécifique.
- **JK+ (Jackknife+)** : Cette méthode utilise des rééchantillonnages (sans remplacement) pour ajuster les prédictions. Contrairement à CV+, elle offre des garanties plus fortes sur la couverture des intervalles, surtout pour des ensembles de données de petite taille.
- **FCP (Full Conformal Prediction)** : Contrairement à SCP, FCP utilise l'ensemble des données pour calibrer les intervalles en calculant les scores de non-conformité pour chaque point. Cette méthode offre une couverture optimale, mais elle est plus coûteuse en calculs, en particulier sur de grands ensembles de données.

3.5.2 Pourquoi les utiliser dans ce contexte ?

Chaque algorithme a été choisi pour des raisons spécifiques liées à la tâche de classification ou de régression dans ce projet :

- **SCP** : Idéal lorsque les données sont volumineuses, car il divise le problème en sous-ensembles indépendants. Cela le rend rapide et simple à mettre en œuvre, mais il nécessite suffisamment de données pour éviter des calibrations imprécises.
- **CV+** : Pratique dans des contextes où les données sont limitées ou les modèles sont sujets à des variations élevées. La validation croisée réduit les erreurs dues à un mauvais partitionnement.
- **JK+** : Offre une couverture robuste, même pour des ensembles de données plus petits. Il est particulièrement utile pour évaluer les incertitudes dans les prédictions de classification.
- **FCP** : Fournit les intervalles de prédiction les plus précis en exploitant l'intégralité des données. Cela le rend idéal pour garantir des niveaux de couverture élevés, au prix d'un temps de calcul plus important.

Ici notre jeu de données est relativement **petit** (75829 lignes avec 11 dimensions), nous nous attendons donc à avoir de meilleurs résultats avec le JK+ plutôt que le SCP.

3.6 Interprétation des résultats

	SCP	CV+	JK+	FCP
Accuracy	94.78	89.76	89.92	90.26
Precision	81.14	83.91	83.72	98.02
Recall	99.52	98.00	83.53	98.09
Average prediction probability	82.01	86.04	86.04	85.89

Table 1: Comparaison des métriques pour les différents algorithmes.

La comparaison des algorithmes dans le tableau [1] met en évidence les forces et les faiblesses de chacun :

- **SCP** montre une grande précision (94.78%), mais des probabilités de prédiction moyennes plus faibles (82.01%). Cela peut indiquer que le découpage des données a limité l'information disponible pour la calibration.
- **CV+** offre un bon compromis entre précision et rappel, mais la réduction des variations introduites par la validation croisée le rend plus fiable sur des ensembles de données variés.
- **JK+** présente des métriques similaires à CV+, avec un rappel légèrement plus bas, reflétant des ajustements potentiellement conservateurs pour garantir une couverture robuste.
- **FCP** surpasse les autres en précision (98.02%) et en rappel (98.09%).

3.7 Conclusion

En conclusion, le choix de l'algorithme dépend fortement du contexte. SCP et CV+ sont préférables pour des ensembles de données volumineux ou lorsque les ressources sont limitées, tandis que JK+ et FCP sont idéaux pour des analyses nécessitant des garanties robustes ou une précision maximale. Ici nous souhaitons prioriser une robustesse et une précision maximale puisque nous voulons savoir avec précision si nous acceptons le crédit ou non. On peut donc conclure que le FCP est l'algorithme le plus adapté à notre problème.