Note méthodologique N°3 - Interprétabilité

Concept de base de l'interprétabilité

L'objectif de l'interprétabilité est de mieux comprendre comment se comporte un modèle de prévision, d'identifier les variables les plus importantes ainsi que leur influence sur les prédictions. Différentes méthodes disponibles en python sont disponibles.
Dans le cadre de ce projet, nous avons utilisé les librairies Lime (local interpretable model-agnostic explanations) et Shap. Pour l'interprétabilité globale du modèle, nous avons analysé les valeurs de Shapley. Pour l'interprétabilité locale, une combinaison de Shapley et Lime.
Pour expliquer un score de crédit à ses clients, un conseiller doit être capable de comprendre le fonctionnement du modèle ainsi que l'influence des différentes variables sur la prévision. L'interprétabilité locale consiste à expliquer, pour un ou des crédits en particulier, comment chaque variable a contribué au résultat. L'interprétabilité globale consiste, elle, à mieux comprendre le fonctionnement global du modèle et d'identifier les variables les plus importantes ainsi que leur influence sur le résultat.

Interprétabilité globale

Nous avons utilisé les librairies Lime et Shap, à la fois pour sélectionner les variables et mieux comprendre le fonctionnement du modèle sélectionné.
Tout d'abord, il est possible d'analyser l'importance des features d'un modèle XGBoost en utilisant le paramètre 'feature_importances_' disponible dans Scikit-Learn. Un graphique est disponible dans la présentation du projet ainsi que dans le Notebook 12.
La librairie python Shap permet d'analyser l'impact des variables via le calcul des valeurs de Shapley avec les fonctions 'summary_plot' et 'Beeswarm'. Les graphiques 'dependance_plot' permettent aussi d'avoir un aperçu détaillé des valeurs de Shapley en fonction des valeurs de chaque variable. Ce type de graphiques, dont nous avons mis des exemples dans le Notebook 12, indique aussi le lien avec la variable la plus corrélée.

Interprétabilité locale

La librairie Shap permet également d'analyser l'impact de chaque variable sur une prévision en particulier. La fonction 'force_plot' visualise l'influence de chaque variable sur le score final. A noter qu'il est possible d'analyser simultanément plusieurs crédits avec cette fonction. Un exemple est également disponible dans le notebook 12.
Les fonctions 'bar_plot' et 'waterfall_plot' permettent d'afficher le même type d'information mais sous une forme différente. Nous avons intégré ces fonctionnalités dans le Dashboard.
Sur l'ensemble de ces graphiques, les variables affichées en bleu contribuent positivement à une acceptation du crédit lorsque leur valeur monte. A l'inverse, les variables en rouge contribuent à un refus du crédit. Elles continuent à augmenter la probabilité d'appartenir à la classe 1.
La librairie Lime permet enfin de faire exclusivement de l'interprétabilité locale. Son utilisation est plus intuitive que Shap car elle affiche clairement la probabilité d'avoir un crédit accordé (classe 0) ou refusé (classe 1). On y voit directement l'importance de chaque variable dans le modèle ainsi que sa contribution à chaque classe.