Note méthodologique n°4 - Pistes d'amélioration

Limites de l'approche retenue

Le travail de modélisation effectué repose sur un certain nombre d'hypothèses discutables et de choix arbitraires. Le projet est d'autre part tellement complet et même complexe qu'il faut assez rapidement faire des arbitrages afin de respecter des contraintes temporelles. Pour un seul data scientist, c'est potentiellement un travail d'une année ou plus.

Quelques pistes d'amélioration

Les pistes d'amélioration sont nombreuses. En voici quelques-unes.

Feature engineering et imputation des valeurs manquantes

Afin d'améliorer les performances du modèle, une meilleure compréhension des variables serait nécessaire. Le retraitement des valeurs manquantes a en particulier été un travail très délicat. Il est difficile de savoir pourquoi ces données manquent: absence de données ou données manquantes? Par exemple, pour les crédits antérieurs, un NaN peut vouloir dire que le client n'a jamais contracté de crédit ou bien que l'information n'est pas renseignée. Le retraitement sera bien différent en fonction du cas à traiter. Par défaut, j'ai choisi d'imputer par le mode en respectant la distribution des classes 0 et 1 mais ce choix est contestable.

Feature selection

Le choix des variables explicatives est un sujet encore plus délicat puisque la sélection des variables est dépendante du modèle utilisé. Au départ de l'analyse, le modèle optimal est inconnu donc une phase itérative s'avère nécessaire pour réduire le nombre de features dans la phase initiale de l'analyse. En plus, pour un même modèle, plusieurs options de calcul sont disponibles, donnant toutes des résultats très différents en termes de features sélectionnées. Au final, il est important de pouvoir valider les choix de chaque feature avec des experts du domaine étudié. Ayant une formation d'économiste avec une expérience en milieu bancaire, j'ai pu utiliser ma propre expérience pour ce projet.

Variables cachées

Des opérations de réduction de dimension de type ACP ou autres méthodes non linéaires pourraient être utilisées pour sélectionner de meilleures variables explicatives, voire découvrir des variables cachées.

Modèles testés

Le nombre d'algorithmes de modélisation testés est limité. Nous avons sélectionné les plus adaptés à la problématique et pris en compte nos contraintes de calculs (PC sans GPU, pas d'accès au cloud).

Optimisation des hyperparamètres et de la fonction de coût

Enfin, il y a une marge d'amélioration de la performance des modèles en optimisant les hyperparamètres ainsi que la fonction de coût. C'est un travail fastidieux et très gourmand en ressources de calcul. J'ai commencé ce travail mais, en dépit du temps consacré, les gains en termes de performance sont faibles. Certaines librairies d'optimisation telles que Optuna permettent probablement d'obtenir de meilleures résultats que le classique GridSeachCV de Scikit-Learn. Pour cette phase d'optimisation des hyperparamètres, il faudrait également avoir accès à des machines plus puissantes équipées de GPU et utiliser des ressources cloud.