MACHINE LEARNING FOR CREDIT RISK

Pour réaliser notre projet nous devions appliquer plusieurs techniques de machine learning sur un jeu de données imposé par notre client. Le rendu final devait être un modèle de machine learning fonctionnel ayant une précision intéressante et où les données traitaient bien du problème de risque de défaut de crédit.

Le travail a donc été décomposé en 6 grandes étapes en suivant la méthodologie Data Scientist :

1) **La compréhension du problème :** Pour répondre au problème, il nous fallait des données bancaires de clients ayant, ou non, des retards de paiements. De plus certaines autres données comme l'âge ou les revenus mensuels des clients étaient nécessaires pour avoir un modèle pertinent.
2) **Outils :** Pour faire du machine learning deux langages sont majoritairement utilisés python et R.
   Nous avons choisi le python car il est plus répandu et permet donc l'accès à plus de bibliothèques et de fonctionnalités.
3) **La sélection des données :** Nous avions un dataset contenant 150 000 lignes (chaque ligne est un client) et 11 colonnes contenant des informations clés telles que : le revenu moyen, l'âge, l'historique de défaut (0 ou 1, ce sera notre colonne cible: variable que nous voudrions prédire), des informations sur le dépendance financière (famille, prêt), de combien de jours on-t-il déjà été en retard (30-59, 60-89 , plus de 90 jours)
4) **Nettoyage et préparation des données :** Nous avons donc effectué un travail classique : Repérer les valeurs manquantes et soit les supprimer soit les remplacer : nous les avons remplacées par la médiane des valeurs que nous avions.
   Trouver des valeurs aberrantes et les modifier
   Changer les types de variables si nécessaires : une date doit être une date et non un nombre.
   Utiliser la matrice de corrélation entre toutes les colonnes et notre colonne cible (défaut ou non) afin de voir les variables qui sont peu corrélées à notre cible et les retirer.
5) **Modeling :** Nous avons donc testé différents algorithmes de classification pour créer des modèles de machine learning. Nous avons testé 9 modèles différents grâce à des indicateurs de performance.
6) **Résultats :** Nous avons donc choisi le modèle QDA par rapport à son AUC plus élevé : 0.68 (meilleure AUC) et une bonne précision : 0.907 soit 90.7%