ETUDE PROJET TRANSACTIONS FRAUDULEUSES

Auteur : Julia OUZZINE

L'objectif du projet choisis est de savoir si les transactions sont frauduleuses ou non. Pour présenter les données, nous présenterons les différentes étapes de notre algorithme :
- Analyse des données
- Choix de l'algorithme
- Choix de la métrique

I - Analyse des données

Lors de notre analyse, nous pouvons constater que près de 99% sont des transactions non frauduleuses. Nous avons donc un dataset déséquilibré. Si nous lançons un algorithme avec tout ce dataset, nous risquons d'avoir un taux de 1%, correspondant aux données frauduleuses. Pour gérer ce déséquilibre de catégorie, nous devons utiliser la méthode SMOTE.

Trois types de graphiques sont lancés :
- celle correspondant aux différentes distributions : elle permet un traitement se fait sur les colonnes asymétriques (time et amount)
- celle correspondant aux corrélations entre deux colonnes entre elles. Ainsi nous pourrons savoir quelles colonnes nous permettrait de déterminer si la transaction est frauduleuse ou non. Nous pouvons constater que principalement toutes les colonnes entre scaled_time et V18 sont assez corrélés avec la class (0 pour non frauduleux et 1 pour frauduleux).
- celle correspondant aux boîtes à moustaches (boxplot) afin de comparer statistiquement (min, max, mean, IQ) les types de transaction

Il a fallu utilisé un dataset 50% pour chacune des catégories.

II - Choix de l'algorithme

Nous sommes en présence d'un dataset où nous connaissons quelles sont les transactions frauduleuses et non frauduleuses. Nous choisirons un algorithme supervisé. Pour la détections de fraudes, plusieurs algorithmes sont utilisés : le réseau de neurone, logistic regression, arbre de décision, random Forest.

N'ayant pas tester à ce jour logistic régression, il était intéressant d'utiliser cet algorithme pour ce projet. Avec plus de temps, il aurait été intéressant de comparer les différents résultats des algorithmes.

III - Choix de la métrique

Afin de confirmer la qualité du programme, deux outils ont été mis à disposition. ROC score et le f1-score. Par le ROC score, nous pouvons en déduire que le modèle est overfitté. L'accuracy score n'est pas adapté pour les données non équilibrés.

Nous pouvons constaté que nous avons beaucoup de falses positifs. Mais l'objectif est de détecter les fraudes. Il est donc préférable d'avoir plusieurs falses positifs que des falses négatifs.