MODÉLISATION INTERPRÉTATION DES CAS DE FRAUDE

Marie-Lou BAUDRIN ~ Abdoul-Aziz BERRADA
Cécile BRISSARD ~ Théo LORTHIOS

DATA CHALLENGE 2021 - GRP. 1 ~ PARIS I

CHOIX DES VARIABLES & MODÈLES UTILISÉS

NETTOYAGE DE LA BASE DE DONNÉES :

✓ Création d'une variable donnant longueur du prêt :

$$\text{LOAN\_LENGTH} = \frac{100}{\text{NEW\_AMT\_PAYMENT\_RATE}}$$

✓ Suppression de la variable TOTALAREA_MODE : LIVINGAREA_AVG nous paraît plus pertinente en donnant le lieu de vie.

➢ Ces deux variables explicatives sont très corrélées entre elles et peu corrélées à la TARGET.

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) :

✓ Procédé de création artificiel de données ;

✓ Le but est d'augmenter le nombre d'observations pour avoir une base de training plus importante et plus équilibrée.

✓ Nous avons choisi les paramètres par défaut : 35% de 1 – 65% de 0 dans la TARGET.

TRAIN / TEST :

✓ TEST = 20% du data frame avec observation choisies de façon aléatoire ;

✓ TRAIN = les 80% restant.

MODELISATION :

✓ KNN (K-nearest neighbors)

✓ Random Forest

✓ Logistic Regression

✓ XgBoost

MODÈLE	PRÉCISION	RECALL
KNN	0.1387	0.6291
Random Forest	0.1459	0.23
Logistic	0.1541	0.55
XgBoost	0.1672	0.27

RESULATS & INTERPRETATIONS

POPULATION STABILITY INDEX (PSI) :

- ✓ Métrique permettant de mesurer l'ampleur du changement de distribution d'une variable entre deux échantillons ou au cours du temps.
- ✓ Permet de diagnostiquer d'éventuels problèmes de performance des modèles.

INTERPRÉTATIONS :

- ✓ Sur la base du tableau donnant la précision et le recall, le modèle le plus performant est KNN : Recall = 0.6291
- ✓ Le recall nous importe plus que la précision car il est composé des faux négatifs : c'est la part des prédictions correctes parmi toutes les prédictions réalisées.
- ✓ La probabilité de défaut (cf. tableau et graphique à droite) décroît car les clients avec les scores PSI les plus élevés sont les plus solvables.

CLASSES	EFFECTIF	PROBABILITÉS DE DÉFAUT
[600,700[	139	0.330935
[700,750[	702	0.192308
[750,800[	1782	0.108305
[800,850[	1879	0.051091
[850,1000[	998	0.030060