P7\_Nettoyage.ipynb

Étapes :

1. Joindre les différentes tables, feature engineering par les opérations d’agrégation distinctes pour les features numériques et catégoriques.

2. Éliminer les features colinéaire filtrés par les coefficients de corrélation de Pearson. (df.corr())

3. Éliminer les features qui contiennent plus que 75% de valeurs manquants

4. Utiliser un ML modèle pour faire une première sélection de features : enlever tous les features de zéro importance, garder seulement les top 95% features selon l’attribut ‘feature\_importance’.

NOTE : 1. Tous les relevés mensuels comme bureau\_balance.csv devrait être d’abord groupé par l’ID Bureau, ensuite regroupé par l’ID Client. Par exemple, ‘client\_bureau\_balance\_MONTHS\_BALANCE\_count\_mean’ signifie que l’on calcul d’abord le comptage de variable ‘MONTHS\_BALANCE ‘ groupé SK\_ID\_BUREAU, puis la moyenne groupé SK\_ID\_CURR. Le nom du feature original est écrit en majuscule. A gauche du nom de feature sont les tables jointes. A droit du nom de feature sont les opérations d’agrégation adoptées.

2. Tous les tableaux clientèle comme application.csv devrait être groupé par le client ID