TP-1

Transformation- manipulation de données

Soit les datasets fournis sur le répertoire du cours. Ceux-ci décrivent les données fournies par le club lender pour différents clients (voir le lien : https://data.world/jaypeedevlin/lending-club-loan-data-2007-11

On nous indique que pour le dataset de test, on veut prédire le risque de non-paiement du pret.

Étape 1 : on considère le fichier **lending_club_loans.csv**

Après étude de ce dataset, indiquer les points marquants de votre exploration. Pour chaque observation, indiquer l'opération à effectuer qui serait la plus appropriée.

De ce fait, dans cette phase d'exploration, on considère les points suivants :

- Quelles sont les descripteurs (colonnes) du dataset?
- Combien d'enregistrements (lignes) ont été fournis?
- Quel est le format des données. Par exemple, dans quel format les dates sont fournies, existe-t-il des valeurs numériques, à quoi ressemblent les différentes valeurs catégoriques ?
- Y a-t-il des valeurs manquantes ?
- Est-ce qu'il y'a des dépendances évidentes au niveau des descripteurs?

D'autres observations sur le dataset qui pourraient être pertinentes ?

Étape 2 : Étape de nettoyage

Dans cette étape, on s'intéresse à implémenter les correctifs soulignés dans l'étape 1.

De ce fait, il serait important de considérer les opérations suivantes :

- Imputing: évaluer les colonnes avec des valeurs manquantes. Par exemple, voir les colonnes:
- **title, revol_util et pub_rec_bankruptcies**

Une stratégie à employer:

Supprimer la colonne ayant plus de 1 à 2% de valeurs manquantes

Supprimer les lignes ayant des valeurs NaN

- Convertir les colonnes catégorielles en numériques. Faire attention ici aux valeurs ordinales et nominales (dummy var).
- Suppression de colonnes non adéquates pour la prédiction
- Suppression des colonnes qui ne seront intéressantes que pour prédire le statut de paiement du prêt. Les colonnes dont les valeurs seront obtenues après le prêt ne doivent pas être considérées. Par exemple, évaluer si les colonnes suivantes sont à supprimer:

- `zip_code`
- `out_prncp` –
- `out_prncp_inv`
- `total_pymnt`
- `total_pymnt_inv`
- `total_rec_prncp`
- `total_rec_int`
- `total_rec_late_fee`

- `recoveries`
- `collection_recovery_fee`
- `last_pymnt_d`
- `last_pymnt_amnt`

- Vérifier si la colonne cible est dans un format adéquat pour le modèle.
- Correction/Standardisation/Normalisation de données

Documents à remettre : Votre (**jupyter notebook et .html**) devrait contenir le code et la documentation.

Indiquer au niveau de votre notebook, les références aux questions (étape dans le document)