Projet : Classification dans un contexte déséquilibré
Une application à la fraude bancaire
M2 BIBD et M2 SISE

Ce projet sera effectué par groupe de deux ou trois étudiants maximum et porte sur l'apprentissage dans un contexte de données déséquilibrées avec une application plus précise sur la détection de fraudes.

Le travail sera à rendre par mail à **guillaume.metzler@univ-lyon2.fr** **avant le 9 Janvier 2022**. Ce travail se composera d'un dossier retraçant vos démarches et résultats entrepris pour traiter le sujet, de plus amples explications vous sont données ci-dessous. Vous êtes libres d'utiliser le langage de votre choix pour effectuer ce travail, R ou Python. Choisissez celui avec lequel vous êtes le plus à l'aise.

Vous pouvez télécharger les données à l'adresse ci-dessous

https://filez.univ-lyon2.fr/9dzynu

1 A propos des données

Les données sur lesquelles vous allez travailler sont des données réelles, elles sont issues d'une enseigne de la grande distribution ainsi que de certains organismes bancaires (FNCI et Banque de France). Chaque ligne représente une transaction effectuée par chèque dans un magasin de l'enseigne quelque part en France, elles ne sont pas brutes et plusieurs variables sont déjà des variables créées, *i.e.* sont issues du *feature engineering*, nous avons un ensemble de 23 variables qui ont la signification suivante :

- **ZIBZIN** : identifiant relatif à la personne, *i.e.* il s'agit de son identifiant bancaire (relatif au chéquier en cours d'utilisation)
- **IDAvisAutorisAtionCheque** : identifiant de la transaction en cours
- **Montant** : montant de la transaction
- **DateTransaction** : date de la transaction
- **CodeDecision** : il s'agit d'une variable qui peut prendre ici 4 valeurs
  0 : la transaction a été acceptée par le magasin
  1 : la transaction et donc le client fait partie d'une **liste blanche** (bons payeurs). Vous n'en rencontrerez pas dans cette base de données

2 : le client fait d'une partie d'une **liste noire**, son historique indique cet un mauvais payer (des impayés en cours ou des incidents bancaires en cours), sa transaction est alors automatiquement refusée

3 : client ayant était arrêté par le système par le passé pour une raison plus ou moins fondée

- **VérifianceCPT1** : nombre de transactions effectuées par le même identifiant bancaire au cours du même jour
- **VérifianceCPT2** : nombre de transactions effectuées par le même identifiant bancaire au cours des trois derniers jours
- **VérifianceCPT3** : nombre de transactions effectuées par le même identifiant bancaire au cours des sept derniers jours
- **D2CB** : durée de connaissance du client (par son identifiant bancaire), en jours. Pour des contraintes légales, cette durée de connaissance ne peut excéder deux ans
- **ScoringFP1** : score d'anormalité du panier relatif à une première famille de produits (ex : denrées alimentaires)
- **ScoringFP2** : score d'anormalité du panier relatif à une deuxième famille de produits (ex : électroniques)
- **ScoringFP3** : score d'anormalité du panier relatif à une troisième famille de produits (ex : autres)
- **TauxImpNb_RB** : taux impayés enregistrés selon la région où a lieu la transaction
- **TauxImpNB_CPM** : taux d'impayés relatif au magasin où a lieu la transaction
- **EcartNumCheq** : différence entre les numéros de chèques
- **NbrMagasin3J** : nombre de magasins différents fréquentés les 3 derniers jours
- **DiffDateTr1** : écart (en jours) à la précédente transaction
- **DiffDateTr2** : écart (en jours) à l'avant dernière transaction
- **DiffDateTr3** : écart (en jours) à l'antépénultième transaction
- **CA3TRetMtt** : montant des dernières transactions + montant de la transaction en cours
- **CA3TR** : montant des trois dernières transactions
- **Heure** : heure de la transaction
- **FlagImpaye** : acception (0) ou refus de la transaction (1)

Vous disposez donc d'un jeu de données comprenant 10 mois de transactions du "2017-02-01" au "2017-11-30". A vous de voir si toutes ces informations sont nécessaires ou non pour établir le modèle.

On définira les ensembles de la façon suivante :

- **Apprentissage :** transactions ayant eu lieu entre le "2017-02-01" et le "2017-08-31".
- **Test** : transactions ayant eu lieu entre le "2017-09-01" et le "2017-11-30"

Attention, ne validez pas un modèle avec des transactions antérieures aux transactions utilisées pour apprendre votre modèle !

2 Travail à effectuer

La variable à prédire est la variable *FlagImpaye*, il s'agit d'une variable qui ne peut prendre que deux valeurs possibles : 0 la transaction est acceptée et considérée comme "normale", 1 la transaction est refusée car considérée comme "frauduleuse".

Nous avons vu que plusieurs critères peuvent être utilisées pour évaluer la performance d'un modèle comme l'Accuracy, la précision, le rappel, la F-mesure ou encore l'aire sous la courbe ROC (AUC ROC). Dans le cas présent, vous allez chercher à établir le modèle vous permettant d'obtenir les meilleurs résultats en classification en terme de F-mesure F dont la définition est rappelée ci-dessous :

$$F = \frac{2TP}{2TP + FN + FP},$$

où un TP est une fraude prédite fraude par votre modèle, un FN est une fraude non identifiée comme tel par votre modèle, un FP est une transaction non frauduleuse mais identifiée comme frauduleuse par votre modèle et enfin un TN est une transaction non frauduleuse

Vous devrez pour cela effectuer un travail en trois temps que vous présenterez dans votre rapport :

- vous commencerez par une analyse synthétique des données à l'aide d'outils statistiques élémentaires : vous présenterez quelques graphes pertinents et intéressants s'il y a lieu ainsi que les grandes caractéristiques du jeu de données, sélectionnez les informations intéressantes.
- vous dresserez ensuite votre protocole expérimentale qui présentera la ou les méthodes sélectionnées pour répondre à la tache demandée :
  (i) quelles sont les données sélectionnées pour apprendre et valider votre modèle et la méthode utilisée. Est-ce que vous appliquer des méthodes d'échantillonnage sur vos données (over-sampling ou under-sampling par exemple)
  (ii) vous présenterez le ou les algorithmes (supervisés ou non supervisés) utilisées ainsi que les éventuels hyper-paramètres à tuner lors du processus d'apprentissage. Dans le cas où la méthode est classique, soyez synthétique dans la présentation. Dans le cas où vous utilisez une méthode raffinée, présentez ces spécificités.
  (iii) indiquez si vous effectuez un éventuellement traitement a posteriori des scores/probabilités ou résultats obtenus
- présentez ensuite les résultats obtenus sous forme graphique ou de tableaux et analysés ces résultats et comparez éventuellement les différentes méthodes en essayant d'expliquer pourquoi certaines méthodes semblent mieux fonctionner que d'autres sur certains aspects et concluez

Idéalement, le travail effectué devrait comprendre au moins 5 procédures ou méthodes différentes vues en cours : une méthode peut être un algorithme de classification seul ou encore couplé ou non à une méthode d'échantillonnage par exemple. Je vous donne ci-dessous une liste non exhaustive des méthodes que vous pourriez utiliser pour votre travail.

- **Pre-Process sur les données** : utilisation d'algorithmes d'over-sampling (random - SMOTE - Adasyn - ...) ou encore des approches d'under-sampling (random - Tomek

Link - Edited Nearest Neighbout - NearMiss - ...) vous pouvez aussi utiliser des méthodes dites *cost-sensitive* qui vont accorder plus de poids aux exemples d'une classe donnée, voire même des poids spécifiques à chaque exemple.

- **Algorithmes** : vous pourrez utiliser des algorithmes non supervisés comme des méthodes de clustering (k-means, clustering hiérarchique ou encore les auto-encodeurs) pour détecter les fraudes. Vous disposez également d'un large éventail d'algorithmes de classification supervisés que vous pouvez utiliser : decision trees, random forests, gradient boosting, nearest-neighbors, réseaux de neurones profonds, SVM (linéaire ou non ...), analyse discriminante, boosting, ...
- **Post-traitement :** combinez les résultats issus de différents modèles (bagging) afin de créer un modèle potentiellement plus puissant.

3 Conseils et Bonus

N'hésitez pas à regarder sur internet quelques exemples d'utilisations des algorithmes susmentionnés et votre objectif sera de les adapter au contexte des données (consulter des sites comme Kaggle - MachineLearningMastery ou encore Medium qui seront pour vous une bonne source d'inspiration). Vous verrez que toutes les méthodes ne sont pas forcément applicables à ce type de données : si tel est le cas, n'hésitez pas à préciser dans votre rapport pourquoi une méthode n'a pas fonctionné selon vous.

Enfin, si vous avez le temps et l'envie, vous pouvez ensuite essayer d'établir un modèle qui va essayer de maximiser le chiffre d'affaire d'une enseigne et dont le calcul dépend du montant m de la transaction

- si on accepte une bonne transaction (TN) : le chiffre d'affaire **généré** est égal au montant de la transaction $f(m) = m$
- si on accepte une mauvaise transaction (FN) : le chiffre d'affaire **perdu** est proportionnel au montant à $f(m) = m\left(1 - \exp\left(\frac{1}{m}\right)\right)$. Plus le montant de la transaction est élevé, plus la perte est importante.
- lorsque vous refusez une bonne transaction (FP) : vous **générez** un chiffre d'affaire égal à 80% du montant de la transaction, $f(m) = 0.8m$
- lorsque vous refusez une transaction frauduleuse, le chiffre d'affaire est nul $f(m) = 0$