PROJET 8 COMPETITION KAGGLE : AMERICAN EXPRESS - DEFAULT PREDICTION

SOUTENANCE OPENCLASSROOMS, LE 07/09/2022

ERWAN CHESNEAU

kaggle

PLAN :

I. Présentation de la compétition
II. Problématique de la taille de la base de données
III. Exploration des données
IV. Traitement des données
V. Modélisation
VI. Soumission des résultats
VII. Conclusions
VIII. Améliorations à envisager

I. PRÉSENTATION DE LA COMPÉTITION:

- Objectifs :
  - American Express souhaite améliorer ses modèles de prédiction de défauts de paiements
  - Dans un but d'amélioration de l'expérience client
  - Dans le but de diminuer les pertes financières?!
  - Récompenses de 100k$
- Données :
  - American Express met à disposition des données industriels sur les clients.
  - Les données sont anonymisées et normalisés (difficile de savoir à quoi correspond réellement les variables)
  - Les données sont séparées en train set et test set.
  - Le test set contient deux parties : une partie publique et une partie privée
  - Les données d'apprentissage = 16GO
  - Données de test = 33GO

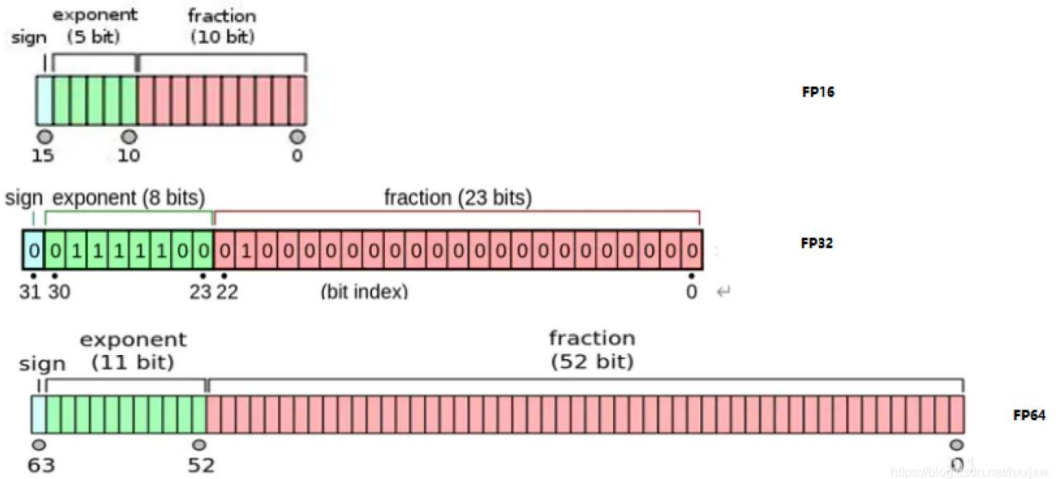
I. PRÉSENTATION DE LA COMPÉTITION:

- Démarches :
  - Traiter les données pour réduire la taille de la base de données (Eviter les problèmes mémoires)
  - Explorer les données pour se familiariser avec les variables
  - Traiter les données en vue de la modélisation
  - Effectuer des tests de modélisations
  - Effectuer la soumission des prédictions sur le jeu de test

- Des concurrents partagent leurs travaux intéressants.
  - Amélioration de leurs propositions
  - Combiner les propositions
  - Eviter de refaire un travail déjà fait (on est pas forcément meilleur que les autres…)

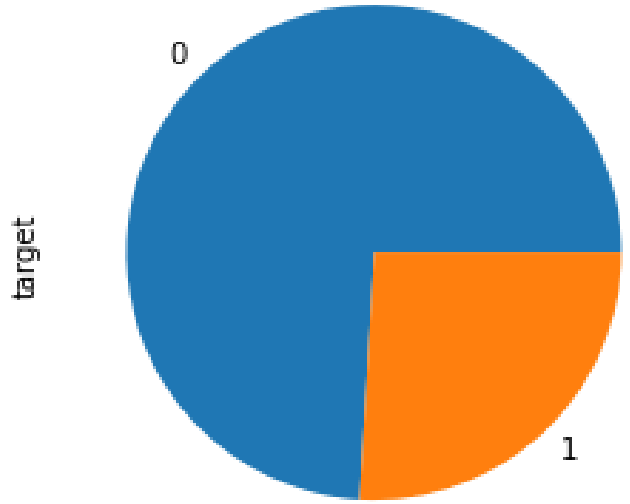
II. PROBLÉMATIQUE DE LA TAILLE DE LA BASE DE DONNÉES:

- La base de données d'apprentissage = 16GO
- La base de données de test = 33GO
  - ➔ Impossible à passer en mémoire et d'effectuer des traitements …
- Solutions envisagés :
  - Séparations des bases de données en plusieurs :
    - Découpage en horizontal : toutes les colonnes mais moins de lignes
    - Découpage vertical : toutes les lignes mais moins de variables
  - Réduire la précision : changement du type d'encodage : float64 ➔float16
- Solutions retenues : réduire la précision
  - Données transformer par @RADDAR et disponibles au format parquet
    - Format plus rapide à lire que le CSV



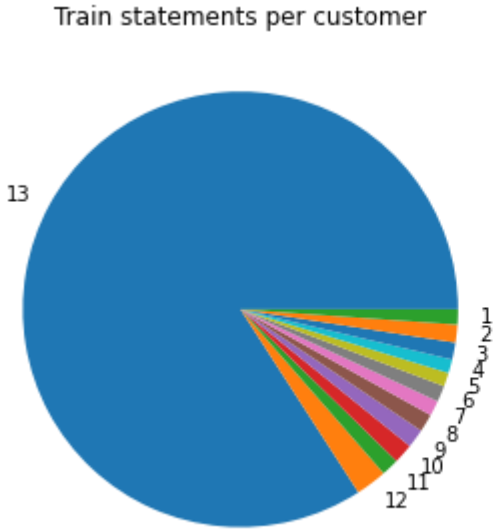
III. EXPLORATION DES DONNÉES : LABELS

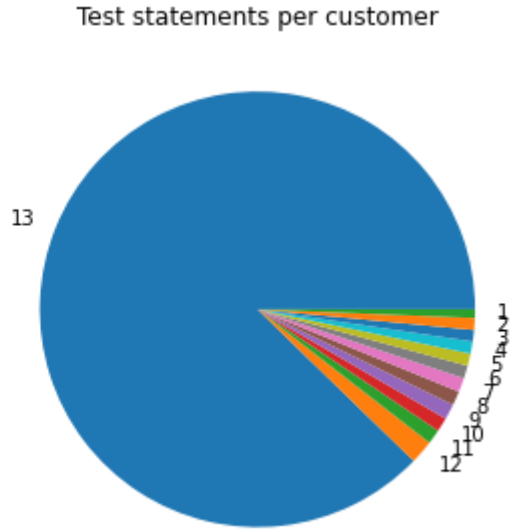
- Nombre de clients : 458913
- Aucun doublons
- Aucun label non renseigné
- Environ 75% de labels 0 : sans défaut de paiement

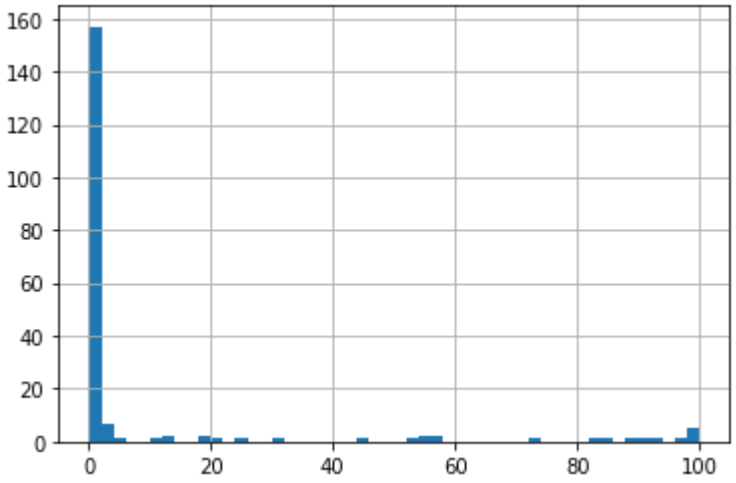


III. EXPLORATION DES DONNÉES : DATA

- Nombre de lignes : 5531451
- Nombre de clients : 458913
  - Plusieurs relevés par client
  - Entre 1 et 13 relevés par client
  - Majorité de 13 relevés
- 190 variables
  - 11 variables catégorielles
- 67 variables contient des valeurs non renseignées
  - 17 renseignées moins de 50% du temps



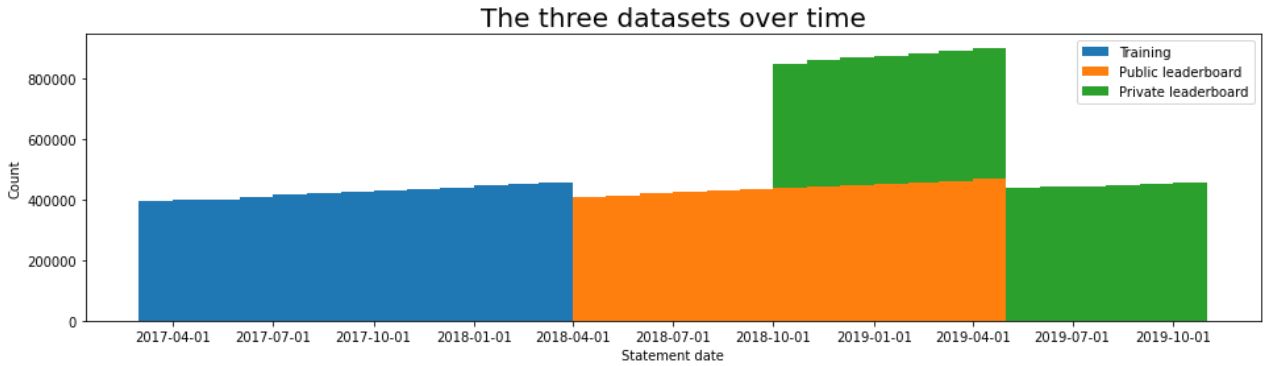


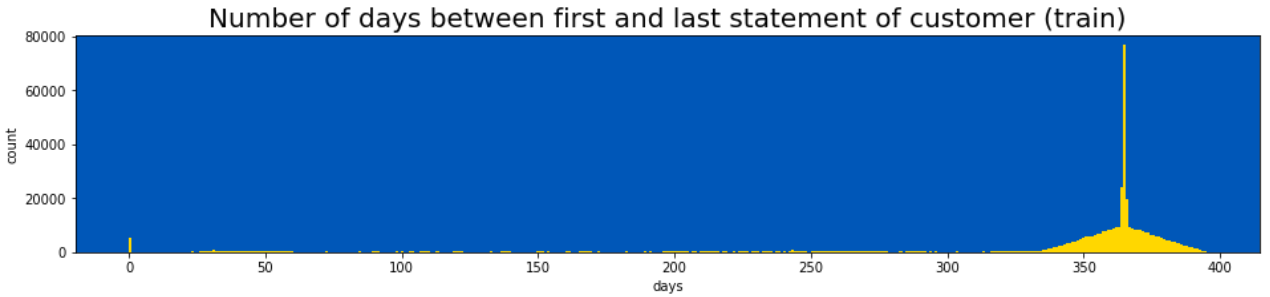


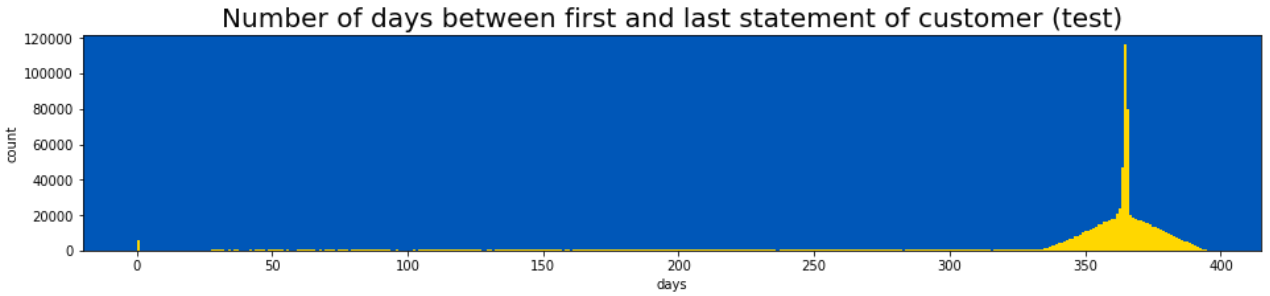
III. EXPLORATION DES DONNÉES : DATA

Date des relevés

- Les trois jeux de données ne couvrent pas les mêmes période
  - La date ne sera pas utilisée directement dans la modélisation
- La distribution du temps entre le premier et le dernier relevés est similaire entre les jeux de données
  - Environ 1 an d'ancienneté.







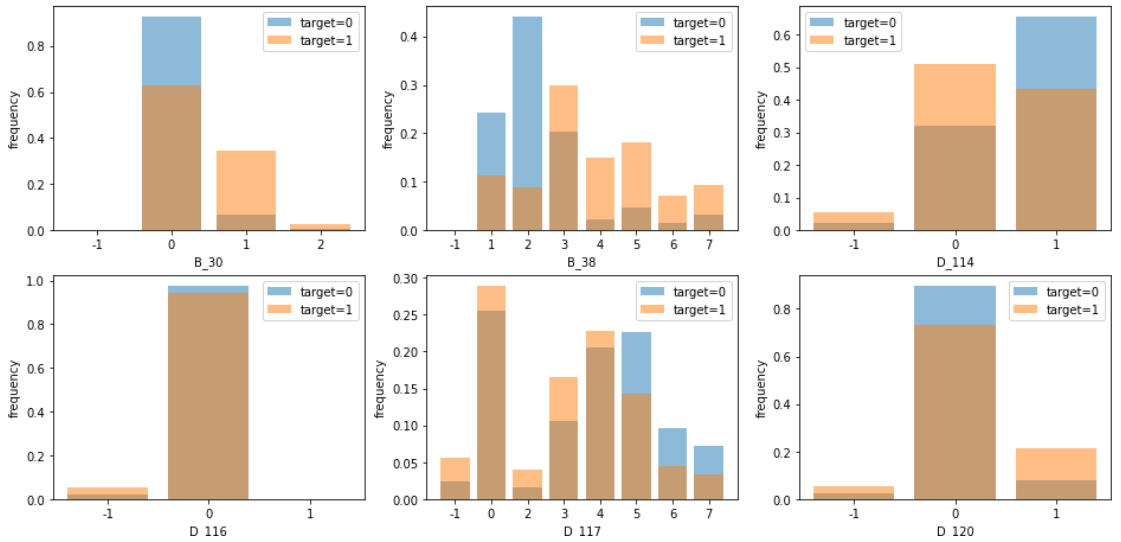
III. EXPLORATION DES DONNÉES : DATA

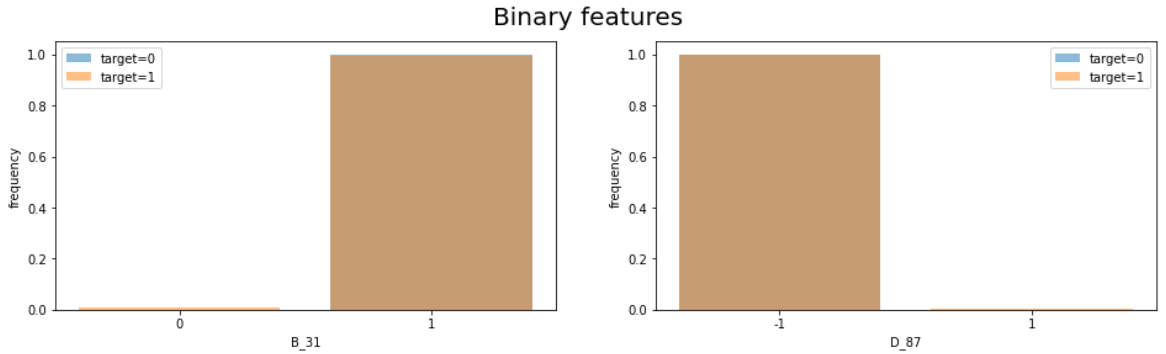
Données catégorielles

- La majorité des distributions des données catégorielles montre des différences entre les labels :
  - Ces données peuvent apporter de l'information
- Certaines variables sont peu distribuées (majoritairement la même classe)
  - Ces données apporteront moins d'information
  - A supprimer ?

Données binaires

- Les données binaires ne semble pas contenir beaucoup d'information
  - Peu renseignées et peu variables.

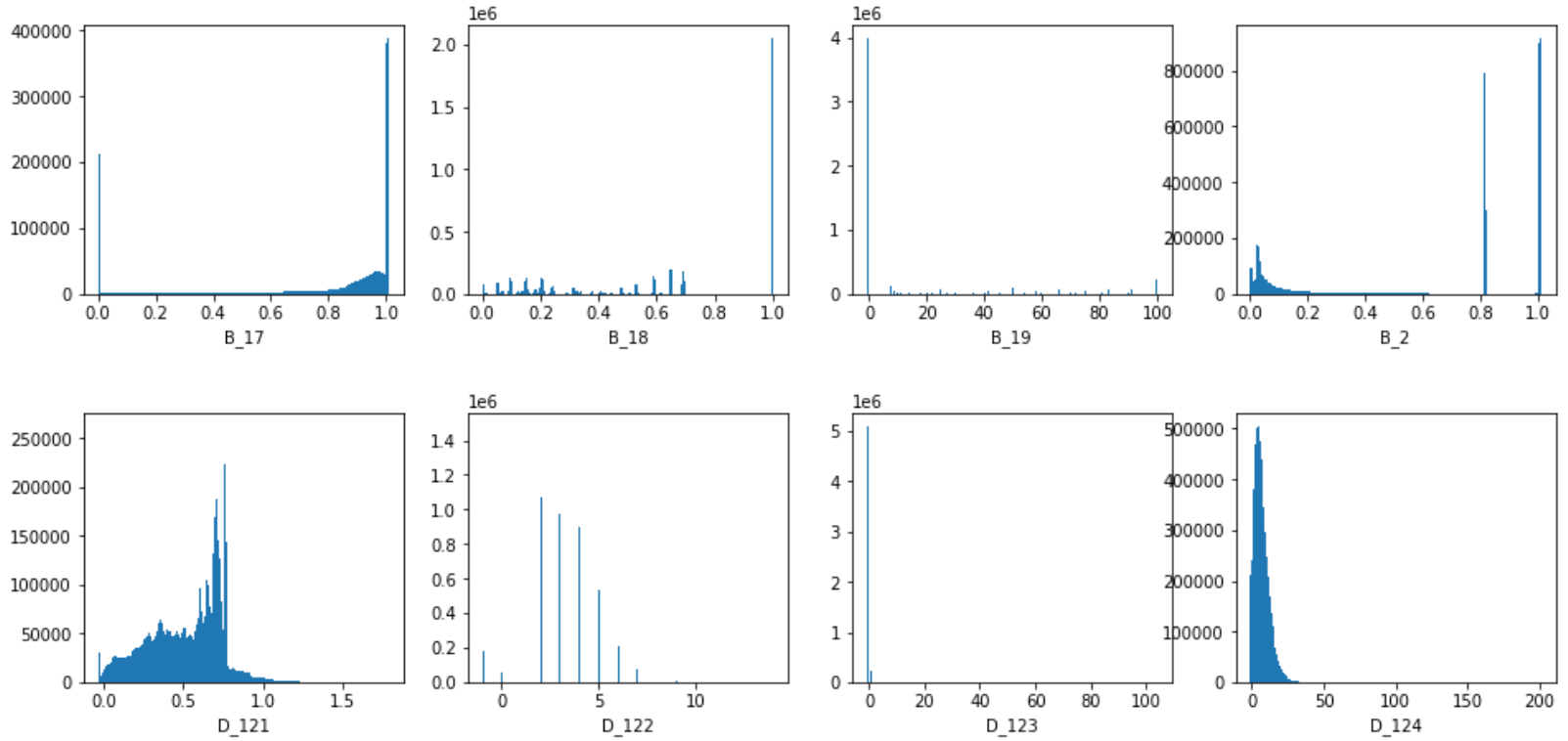




III. EXPLORATION DES DONNÉES : DATA

Données continues

- Les données ont des distributions différentes :
  - Formes, valeurs moyennes…
  - Toutes ne contiennent pas le même type d'information
- Certaines variables peuvent être considérées comme catégorielles
  - Distributions en forme de peignes
- Normalisation appliquée?
  - Pas centrée renduite
  - Pas toujours entre 0 et 1…
  - Renormalisation?



IV. TRAITEMENT DES DONNÉES

Traitement proposés par @huseyincot

- But : obtenir une base de données d'une ligne par client
  - Agréger les données
- Données continues :
  - Agrégées en calculant la moyenne, la déviation standard, la valeur minimale et maximale et la dernière valeur
  - 1 colonne ➔ 5 colonnes
- Données catégorielles :
  - Agrégées en calculant le nombre de valeur unique, la dernière entrée, le nombre d'entrée et le nombre de valeur possible
  - 1 colonne ➔ 3 colonnes
- Importante augmentation du nombre de colonne (918)
- Remplacement des NaN par -127
  - Initialement réalisé à la lecture des données, mais je fais le choix de l'effectuer après le traitement pour ne pas influer sur le traitement

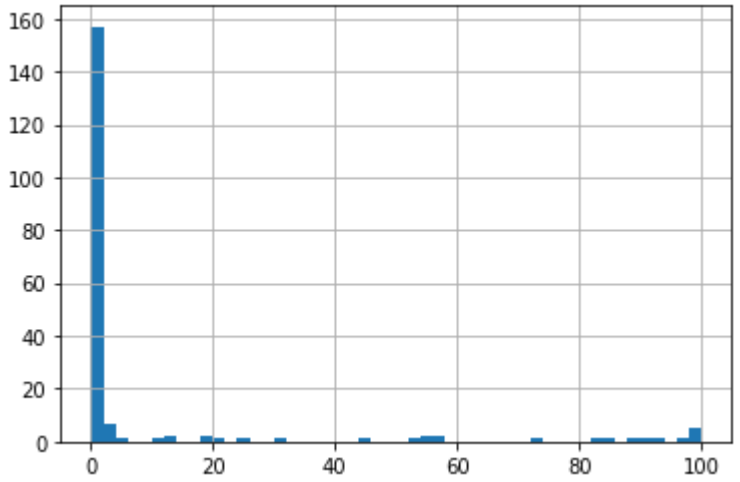
IV. TRAITEMENT DES DONNÉES

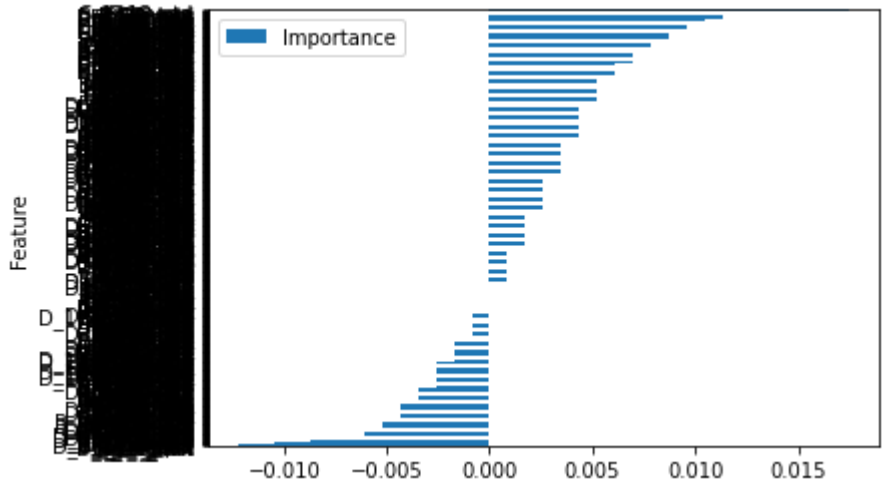
Suppression des variables faiblement renseignées

- Les variables peu renseignées ne vont pas apporter d'information
- Simplification du modèle et réduction de la taille des données
- Suppression des données renseignées moins de 80% du temps (21)
- Même traitement appliqué ensuite

Selection de features

- Un random forest est entrainé sur avec toutes les variables
- On supprime tour à tour une variable
- On compare la performance du modèle
- Suppression des importances négatives
- Effectué sur 1/400 des lignes ➔ Assez ??





V. MODÉLISATION : TOUTES LES VARIABLES

- Base de données trop grande :
  - Séparation en 5 folds (train/val)
- Pour chaque fold :
  - Apprentissage de deux modèles : XGBoost et CatBoost
- But : effectuer un modèle averaging : les erreurs ne sont pas toujours les mêmes
- Léger surapprentissage
- Tous les folds n'ont pas les mêmes performances
- CatBoost légèrement plus performant
- Performance globale calculée en concaténant tous les sets de validations

# FOLD	Train AMEX metric	Valid AMEX metric
1	0.842965907145271 1	0.7923028981785427
2	0.8391523394831826	0.7926375314496539
3	0.8435234258602844	0.7906647706308858
4	0.8373528703029521	0.7878870687695534
5	0.855473544651 1862	0.7946437200173261
Global	/	0.7915386392343893

# FOLD	Train AMEX metric	Valid AMEX metric
1	0.8599299277086581	0.7940031548283464
2	0.836942414204036	0.7945407675830432
3	0.8430667497295776	0.7913994511439725
4	0.8411339639961068	0.787824688662 5178
5	0.8430603077357024	0.795836927877813
Global	/	0.7927899116150396

V. MODÉLISATION : SELECTION DE VARIABLES

- Même démarche que précédemment
  - 5 folds, 2 modèles / fold
- Sélection de features :
  - Les variables sont supprimer avant le traitement pour la suppression des NaN
  - Les variables sont supprimer après traitement pour la méthode de sélection de variables
  - Les features finales sont sauvegardées
    - A utiliser pour la prédiction
- Les deux méthodes de sélections de features donnent des performances similaires
- La performance obtenue en utilisant toutes les features est meilleure

# FOLD	XGBoost	CatBoost
All Features	0.7915386392343893	0.7927899116150396
Without NaN	0.7888496155047777	0.7895673878803888
Features selection	0.7898008093050062	0.7898946134300707

VI. SOUMISSION DES RÉSULTATS

- Fichiers de soumissions :
  - Une colonne ID, une colonne prédiction
- Prédiction réalisées sur plusieurs modèles
- Moyenne de tous les modèles sélectionnées
- Plusieurs soumissions réalisées pour étudier la performance sur le jeu de test public
  - 1 : toutes les variables, learning rate 0.1
  - 2 : Tous les traitement,s learning rate 0.1
  - 3 : Tous les traitements, learning rate 0.05
  - 4 : tous les traitements, learning rate 0.05, XGBoost uniquement
- Amélioration par rapport au modèle de référence
- Classement : 2986 / 4876 sur le jeu de test privé

# soumission	Public score	Private score
1	0.79409	0.80255
2	0.79381	0.80203
3	0.79435	0.80214
4	0.79408	0.80207

VII. CONCLUSIONS

- Participation à ma 1ère compétition Kaggle
  - Découverte du monde de la compétition ML
- Expérience avec les très grosses bases de données
  - Utilisation d'astuces pour réduire la taille des bases de données
  - Gestion rigoureuse de l'utilisation de la mémoire
- Amélioration du score d'un autre participant
  - Mise en place d'une méthode d'assemblage de modèles fort
  - Combiner les traitements de données
- Ecrire et diffuser du code au standards PEP8
- Pour gagner, il faut travailler beaucoup !!

VIII. AMÉLIORATIONS

- Pour les prochaines compétitions :
  - Intégrer la compétition assez tôt pour avoir le temps de développer plus en détails les modèles
- Utiliser des modèles LGBM
- Utiliser des TabTransformers
- Améliorer le traitement initiale
  - Transformer des variables continues en catégorielles
- Réaliser les calculs sur une machine avec plus de mémoire…

MERCI POUR VOTRE ATTENTION !