Implémentez un modèle de scoring

Projet 7 du parcours
« Data Scientist » d'OpenClassrooms

Mark Creasey

Sommaire

Implémentez un modèle de scoring

01 Présentation de la problématique

Mission – Implémenter un modele de scoring

La société financière « Prêt à dépenser » propose:

- de **crédits à la consommation**
- pour les personnes ayant **peu ou pas de tout d'historique** de prêt

Basée sur les données financières

- Données internes
- Données externes

Critères de succès

- **Transparence de la décision** sur l'octroi du crédit
- **Deploiement** d'un dashboard permettant de visualiser les informations clients pour
- **Permettre l'interprétation** la décision fait par la modèle

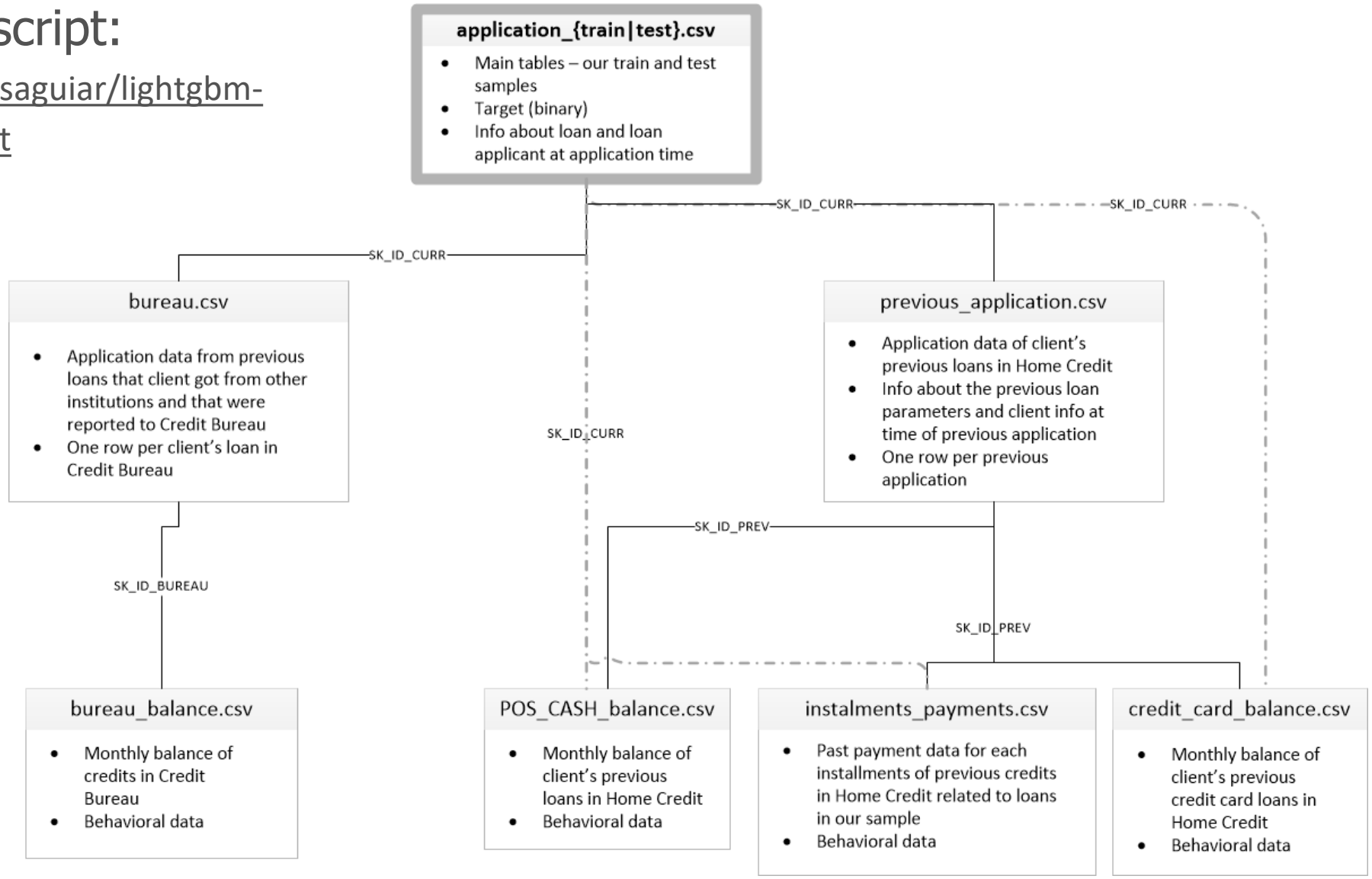
02 Les données

Nettoyage, exploration

Données financières sous 7 tables

Feature engineering script:

- https://www.kaggle.com/jsaguiar/lightgbm-with-simple-features/script



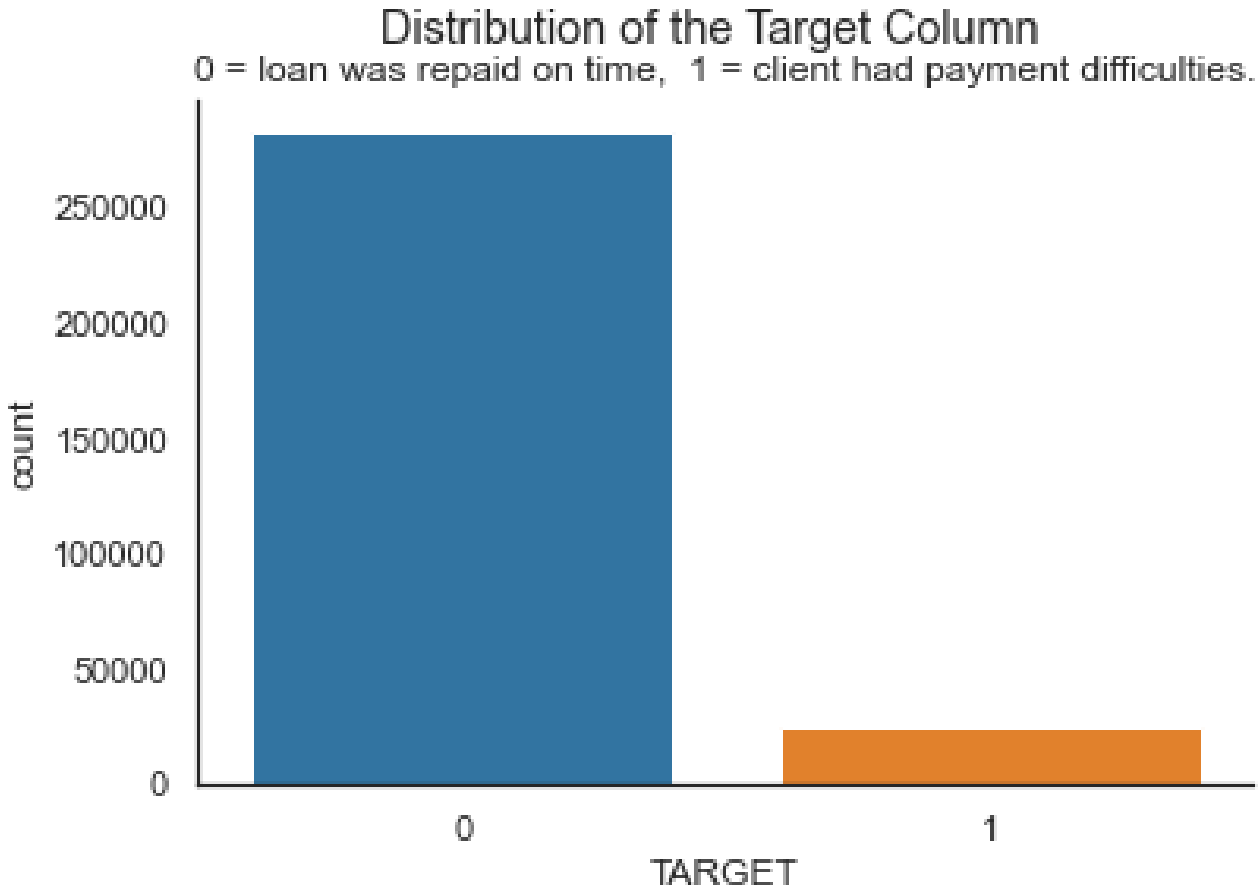
Distribution de la variable cible très déséquilibré

Prédiction que tous les clients sont bons

→ précision de 93%

→ on aura identifié aucun client défaillant.

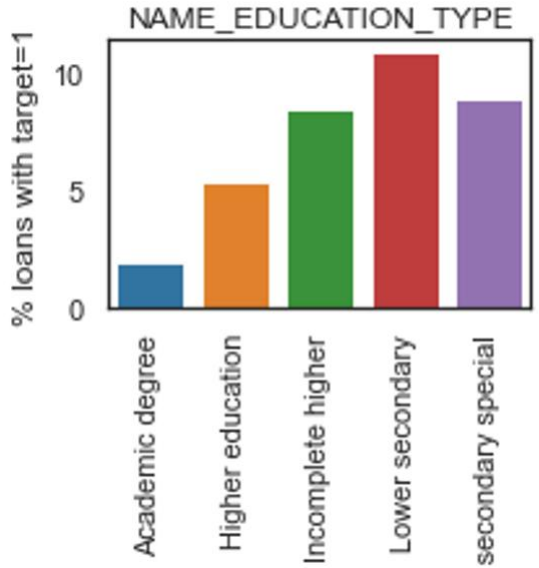
- 0 – clients non-défaillant
- 1 – clients défaillants

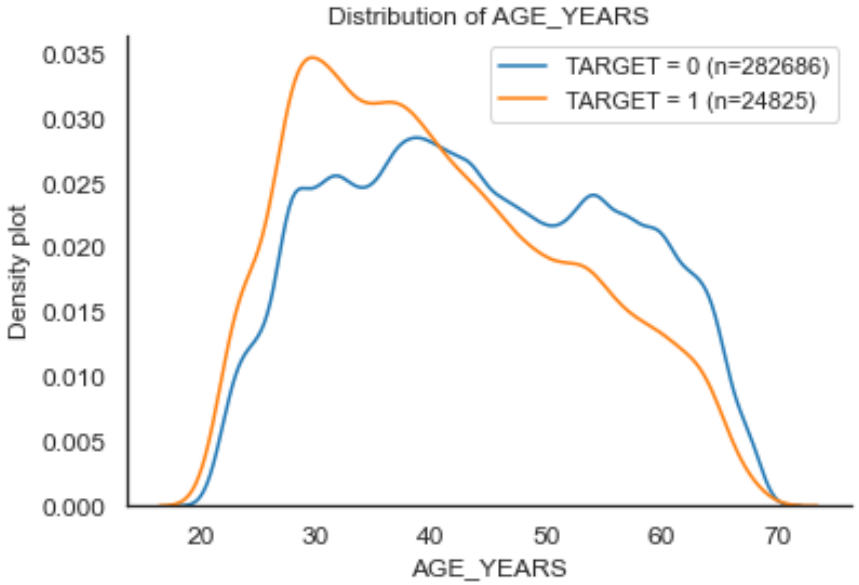


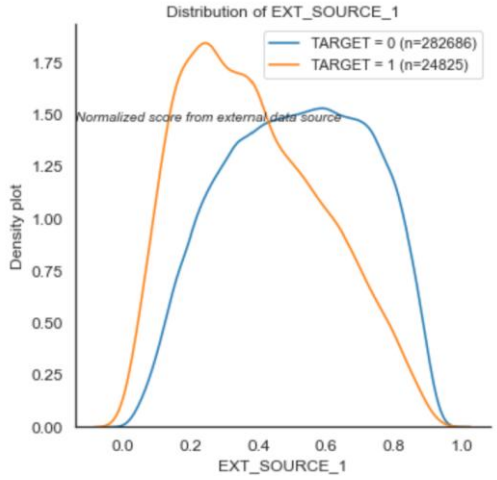
Exploration des features importants

Facteurs de risque (entre autres)

- Les hommes
- < 40 ans
- Bas niveau d'éducation
- Un score <0.5 dans des sources externes
  - EXT_SOURCE_1
  - EXT_SOURCE_2
  - EXT_SOURCE_3







03. Modélisation

Sampling : Re-équilibrage des classes

- Aucun
- Cost- sensitive
- Random under/over sample
- SMOTE
- SMOTE Tomek links

Métriques d'evaluation

Précision

- Quelle portion des clients prédit comme défaillant sont du vrai classe défaillant?

$$\frac{TP}{TP + FP}$$

Recall

- Quelle portion du vrai classe sont présent dans le cluster prédit ?

$$\frac{TP}{TP + FN}$$

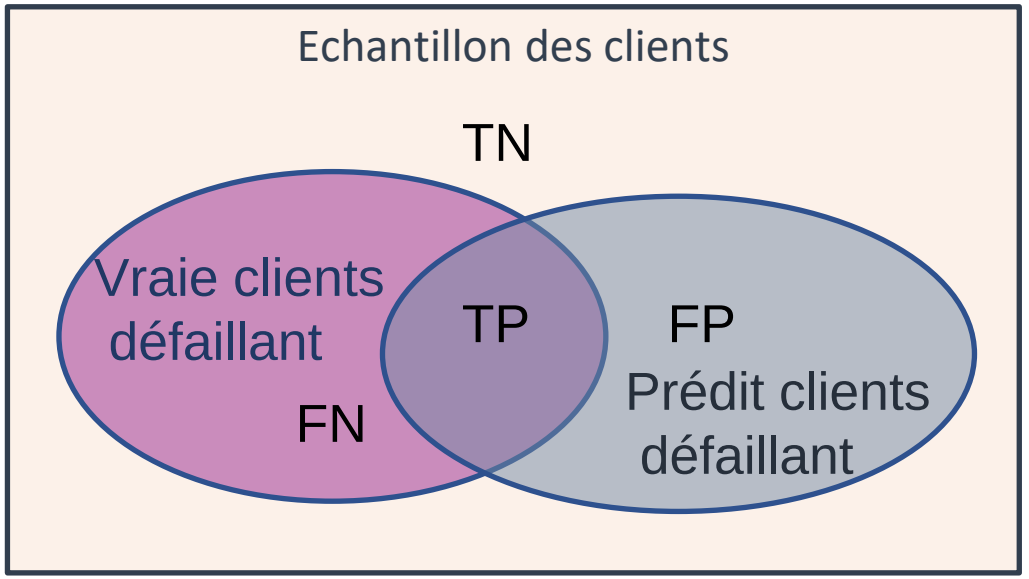
F1 Score

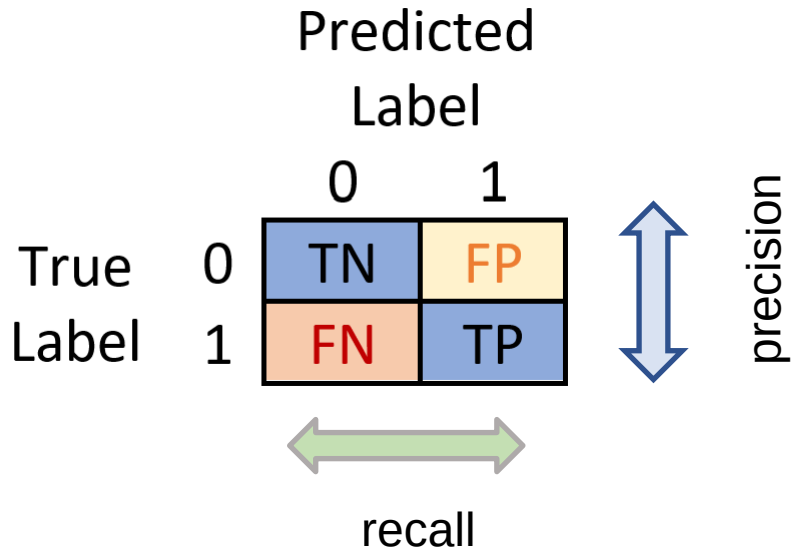
- accuracy « equilibré » :

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

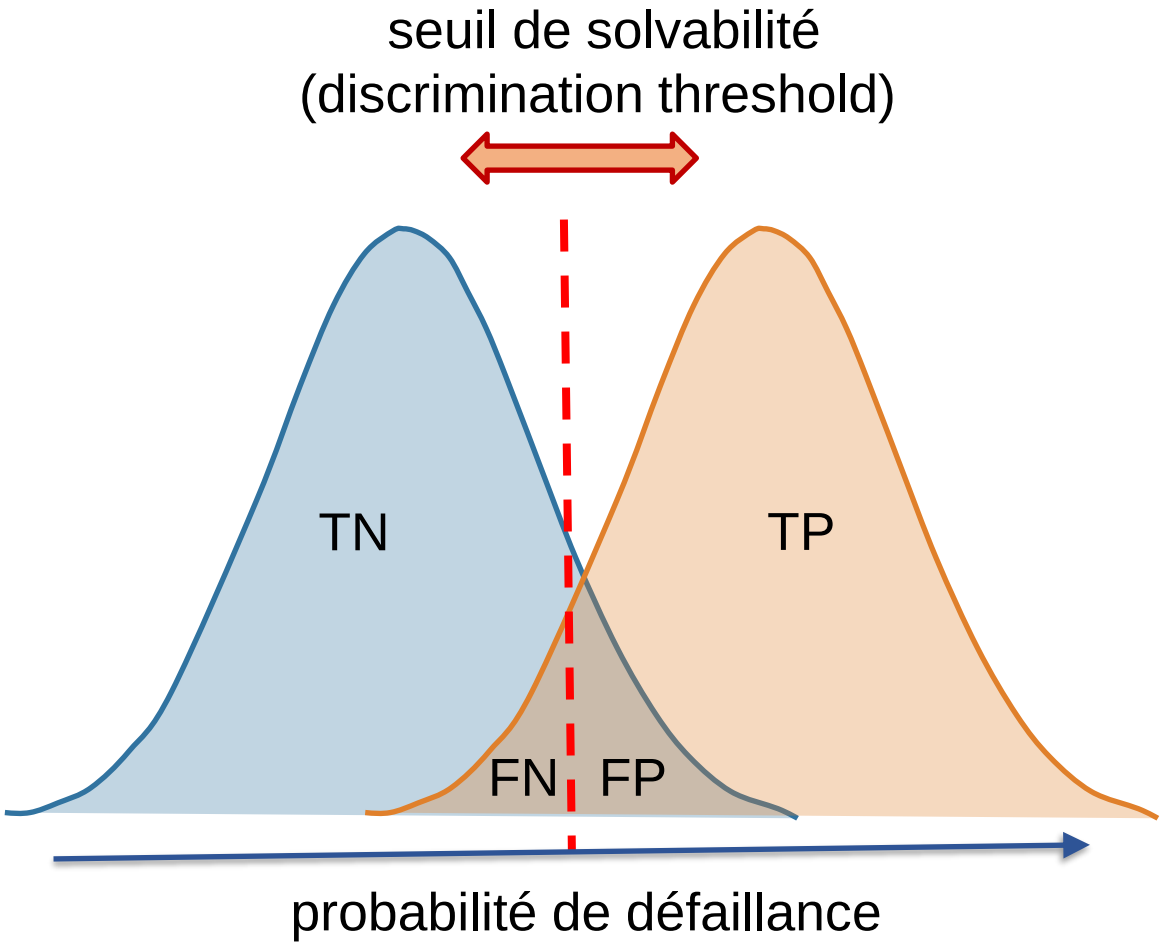
F(beta) score

- Pèse plus sur recall

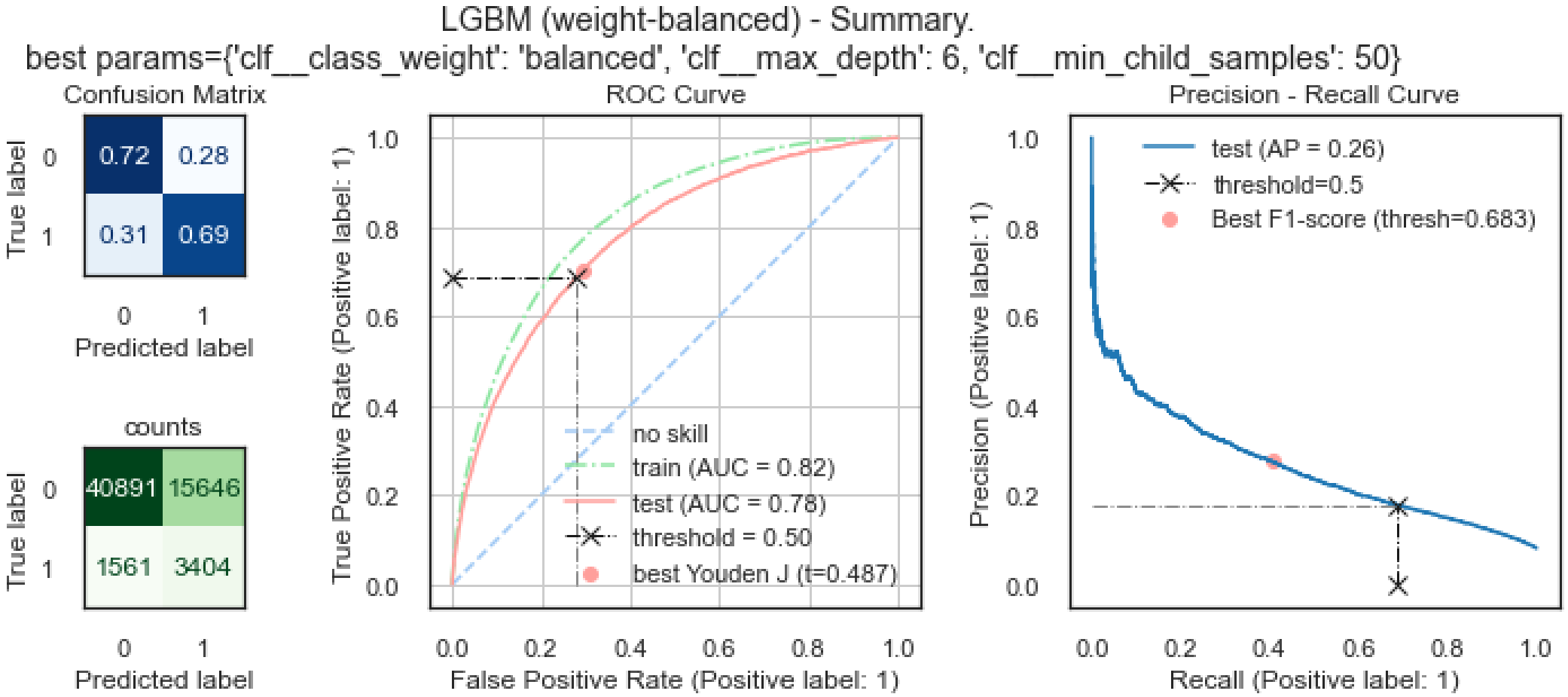




La fonction cout metier



Evaluation des modèles



Les modèles (classifiers)

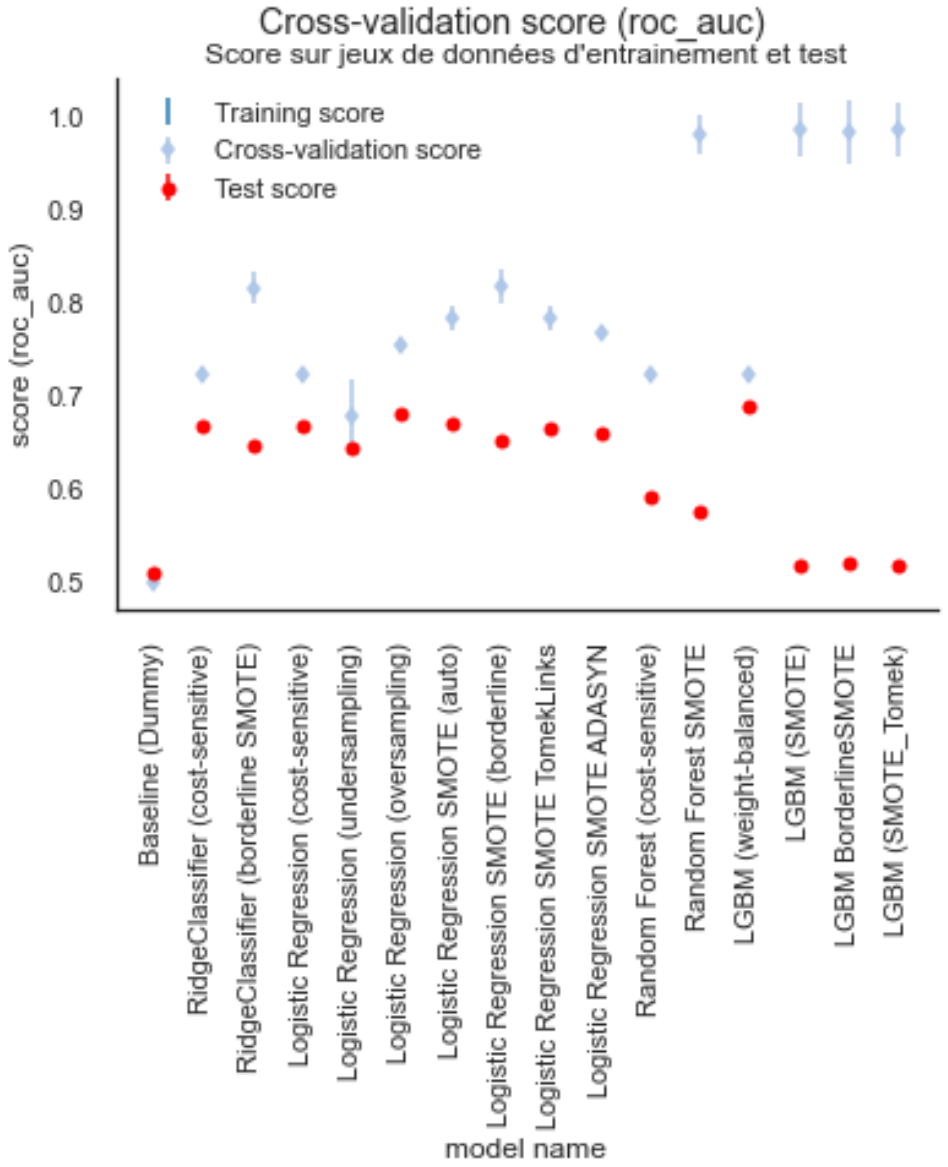
Pipelines imblearn

- Preprocess
- Sampling
- Feature Selection
- Classification

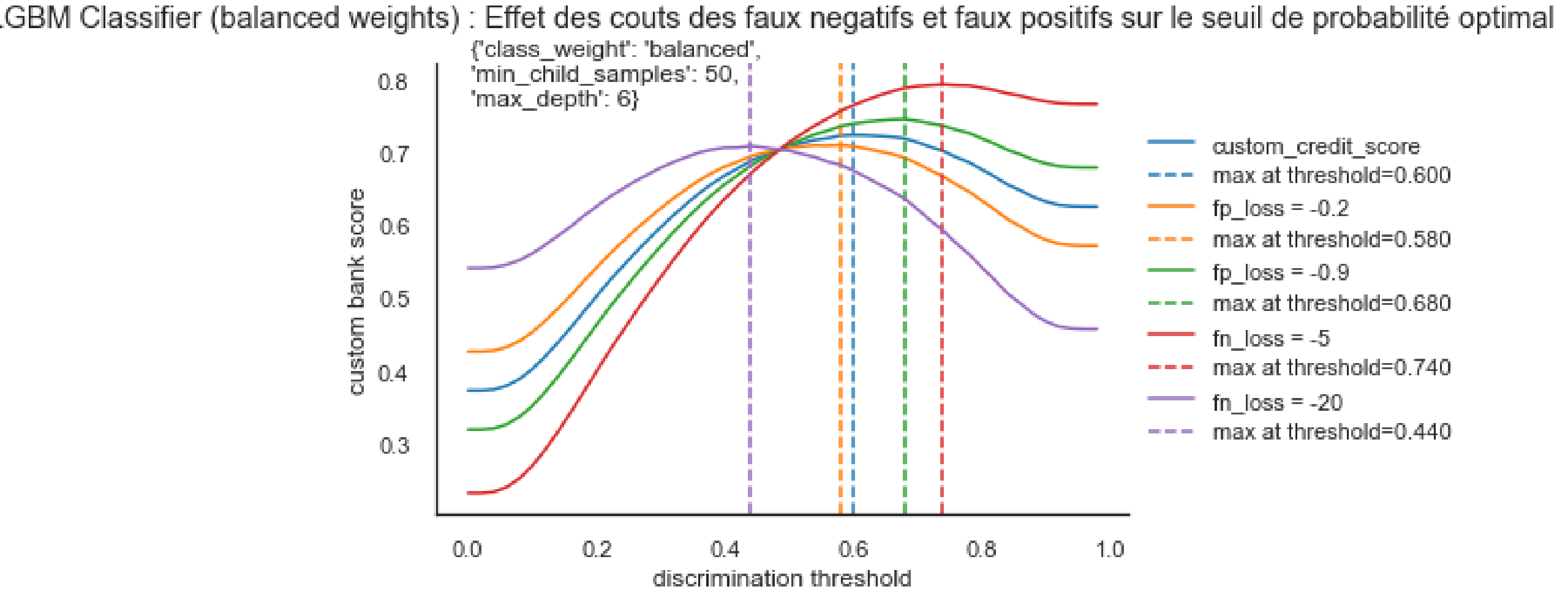
Classifiers

- Dummy
- Ridge
- LogisticRegression
- RandomForest
- LightGBM

Choix du modèle

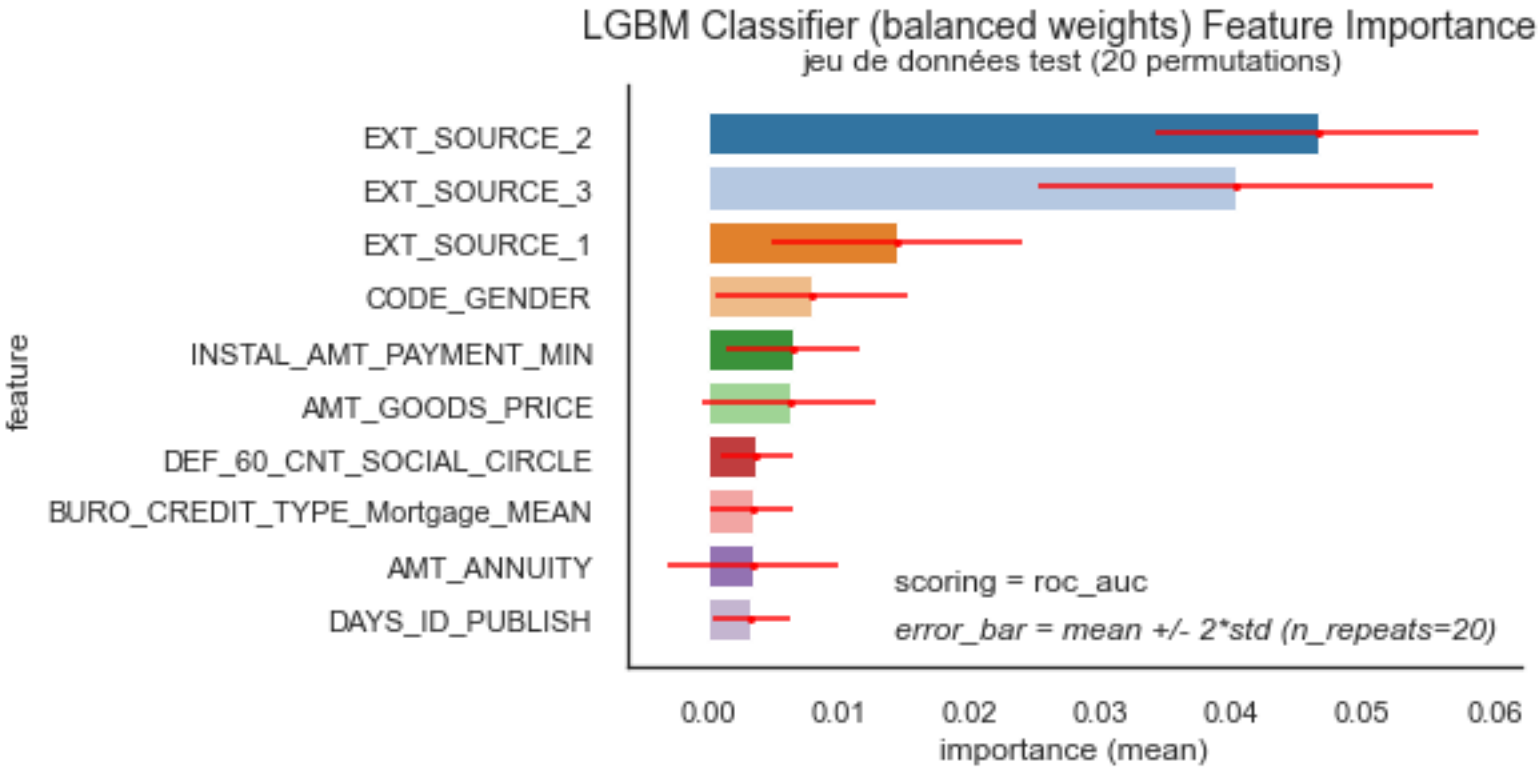


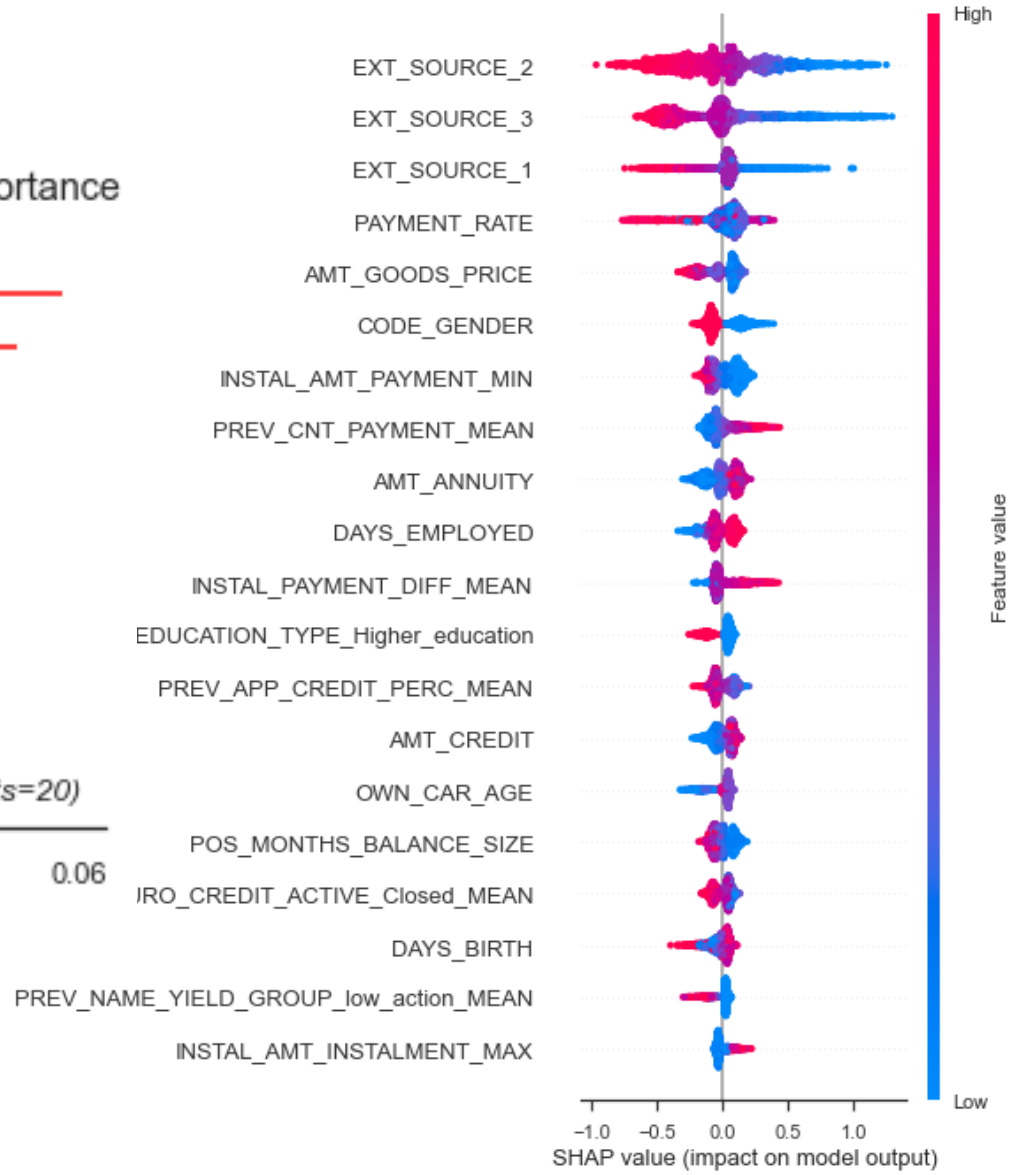
Optimisation du seuil (discrimination threshold)



Feature Importances - globale

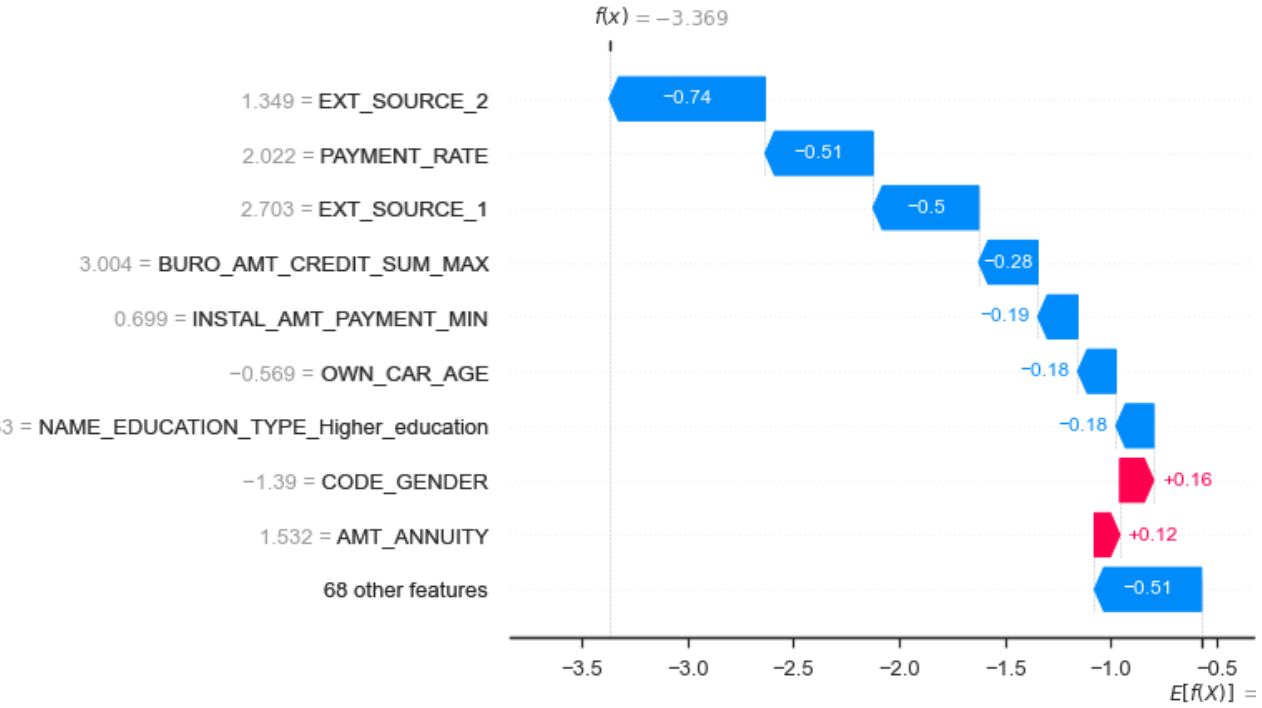
- Permutation importance

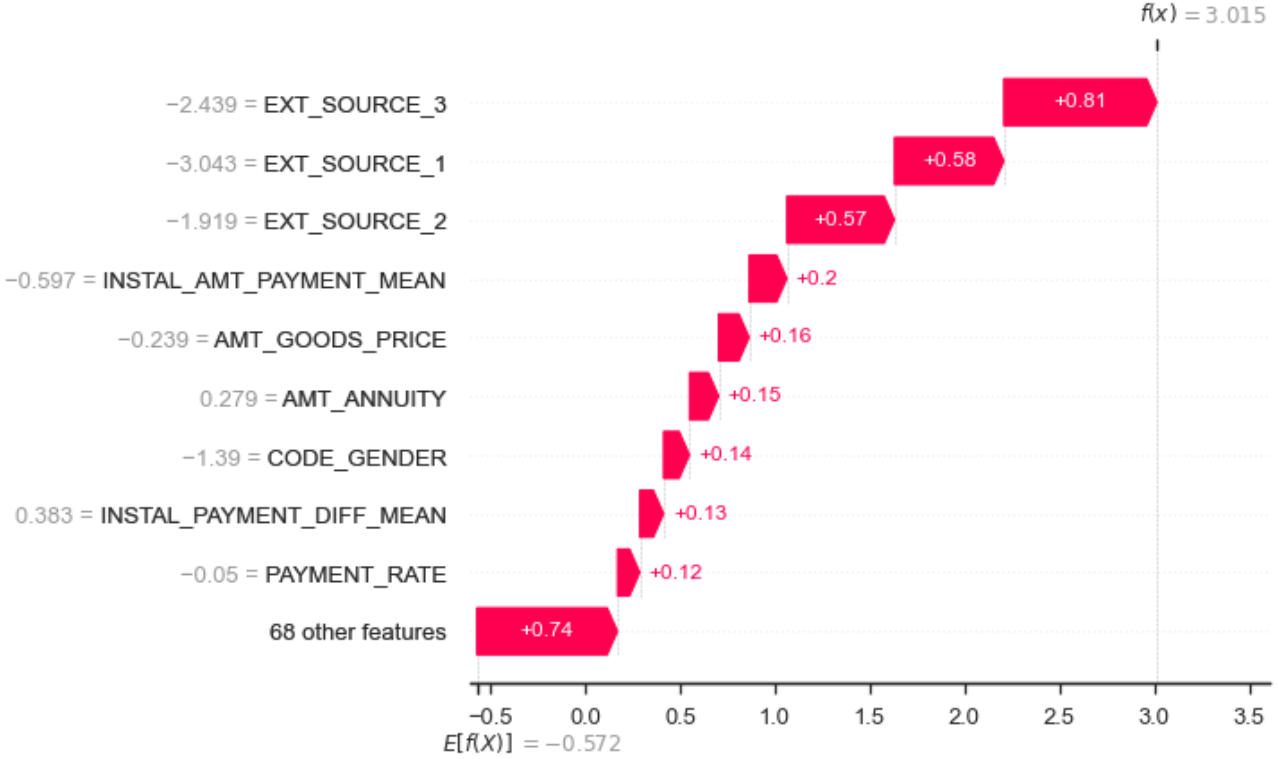




Interpretabilité du modèle

- Bon





04 Dashboard

API du dashboard

Fonction

- Gere l'accès aux prédictions
- Gere l'accès aux données (interpretabilité)
- Fournir un interface publique pour plusieurs clients

REST API

- Réponses json aux requetes GET

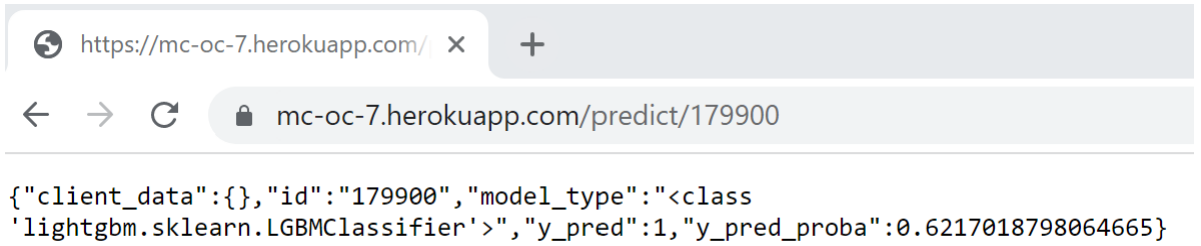
Code source dans dossier **api** sur dépôt : https://github.com/mrcreasey/oc-ds-p7-scoring-dashboard

Instructions pour développement

- Voir README.md dans dossier **api**

Deploiement sur heroku

- https://mc-oc-7.herokuapp.com/



Reste à Faire:

- Documentation (note: fastapi : autodocumentation)
- JWT authentication
- Cache des données en memoire

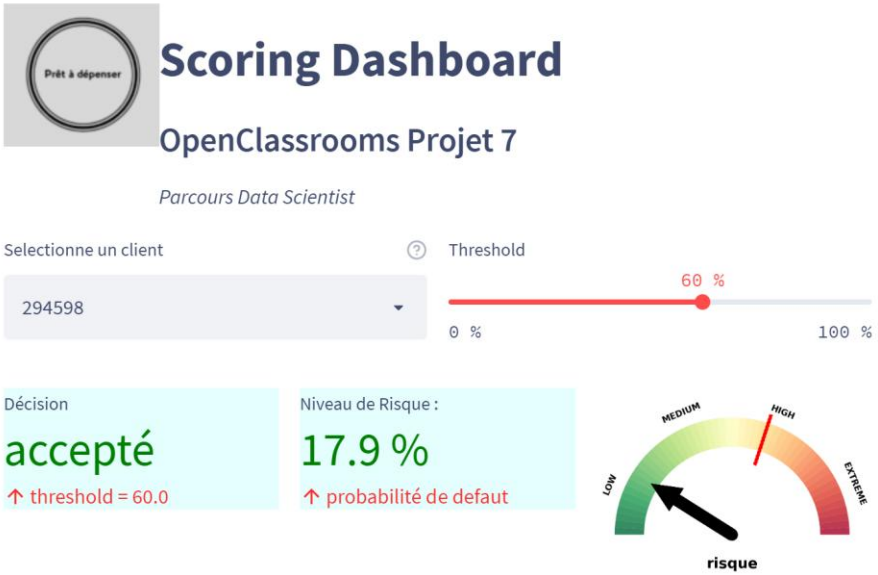
Visualisation du dashboard

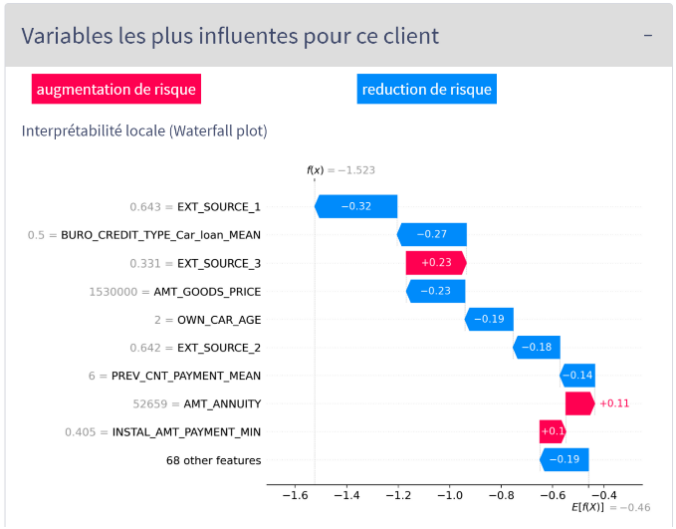
Fonction:

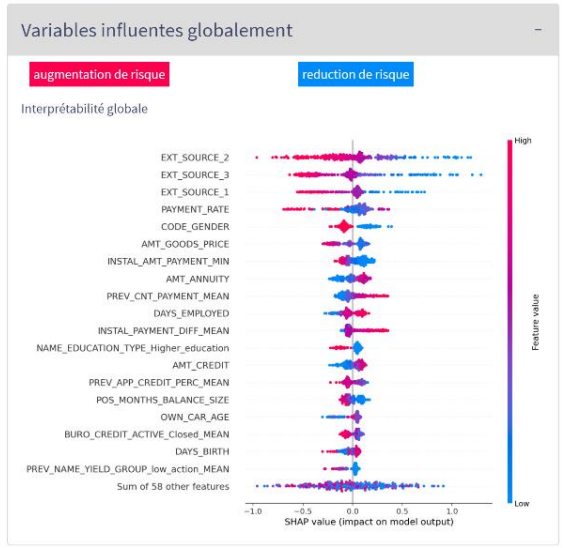
- Visualisation des prédictions et données fournies par l'api

lien vers dashboard deploiement:

- https://mrcreasey-oc-ds-p7-scoring-dashboard-dashboardmain-70agjx.streamlitapp.com/







05 Conclusion et améliorations à faire

Conclusions

Améliorations à faire

Modélisation

- Feature engineering avec experts du métier
- Revoire stratégie de traiter les valeurs manquantes
- Adapter la sélection de features à chaque modèle (Wrapper/Embedded)
- Optimiser la taille d'echantillons en analysant les learning curves des modeles
- Augmenter la recherche des meilleurs hyperparamètres des modèles

Deploiement

- Changer de Flask API vers fastapi (https://fastapi.tiangolo.com/)
- Ajout authentification pour accéder au dashboard
- Ajout encryptage des données client
- Stocker les données client séparément de l'API - par exemple dans un S3 bucket sur AWS
- Visualiser la position du client dans les distributions des features les plus importantes pour ce client

Questions

images: Mark Creasey

- mrcreasey@gmail.com

Code source: https://github.com/mrcreasey/oc-ds-p7-scoring-dashboard

- Merci !