PROJET 7

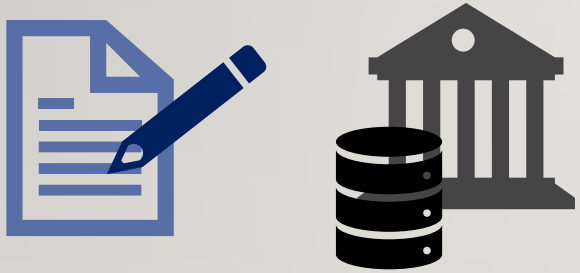
UN MODÈLE DE SCORING

HOME CREDIT

EVA BOOKJANS

OBJECTIVE – un Outil de Scoring Crédit

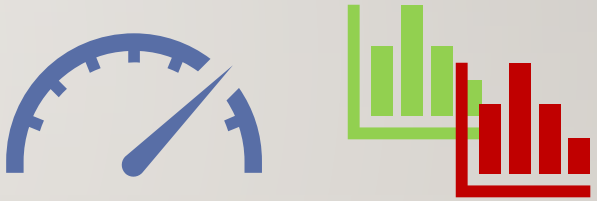
Données sur le Client



- Application
- Instituts financiers

Algorithme de Classification

Prédiction de Remboursement

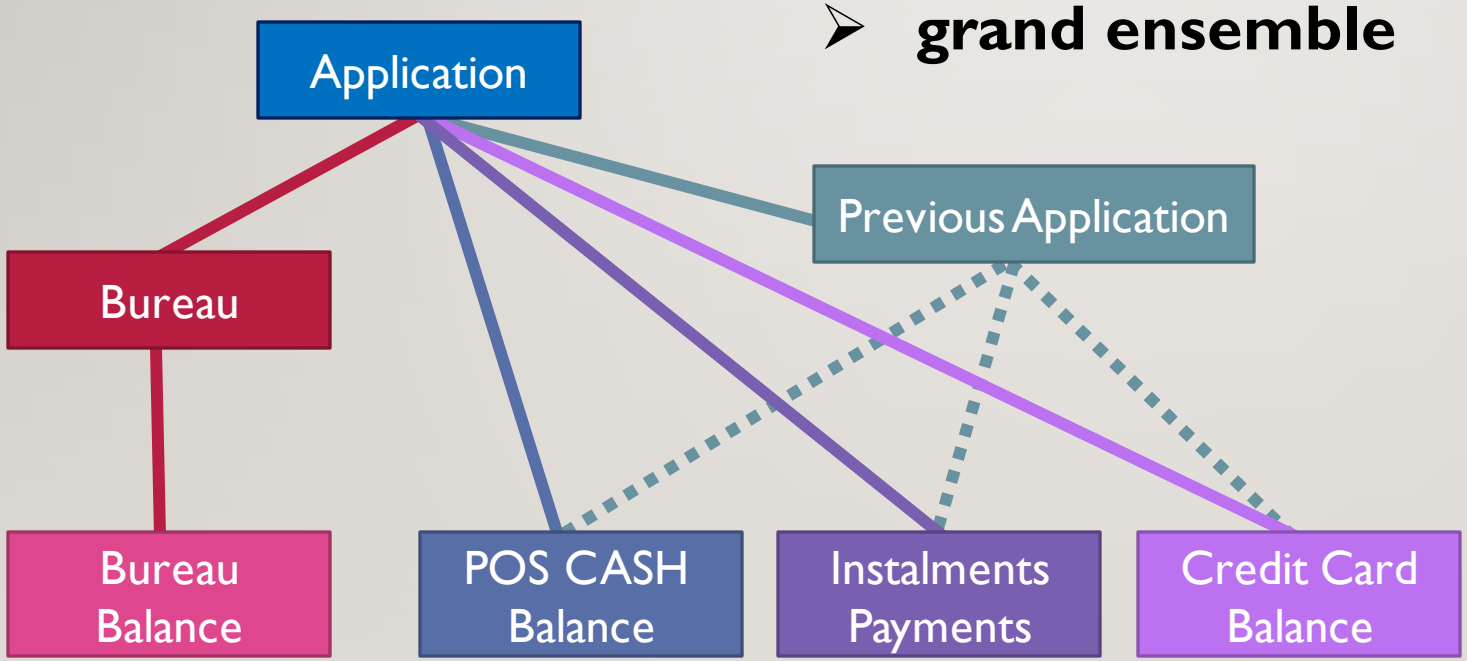


- Dashboard interactif
- Critères importants

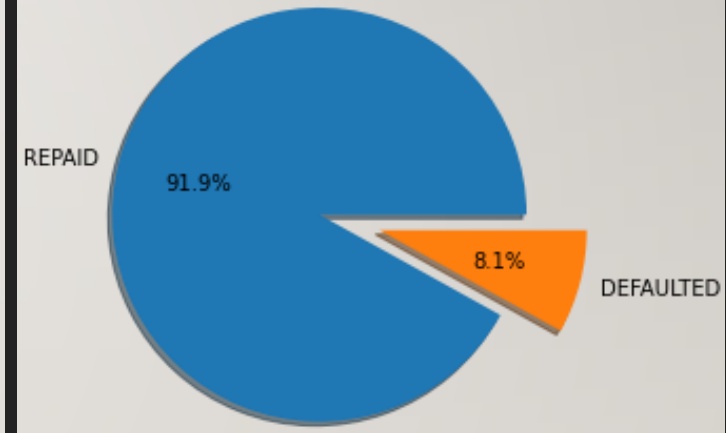
LE JEU DES DONNÉES*

Structure de Données

- 7 fichiers
- 307507 individus
- ➢ **grand ensemble**



La Classe Cible



➢ fortement **déséquilibré**

* Home Credit Default Risk | Kaggle

PRÉTRAITEMENT*

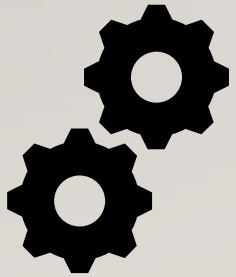
→ 773 Variables

Nettoyage

- remplacement des **valeurs aberrantes** par une valeur manquante (NaN)
- regroupement cohérente des valeurs catégorielles pour des valeurs peu fréquentes

Numérisation des Données

- **transformation des valeurs catégorielles en one-hot variables** (encodage binaire) en **variables ordinales** (e.g. niveau d'éducation, niveau de rendement,...)



Feature Engineering

- création des **nouvelles variables métier** potentiellement plus pertinentes
- **agrégations statistiques** des données venant des fichiers supplémentaires avec des différents fonctions (e.g. minimum, maximum, moyenne, déviations, ...)

* LightGBM with Simple Features | Kaggle ; Start Here:A Gentle Introduction | Kaggle

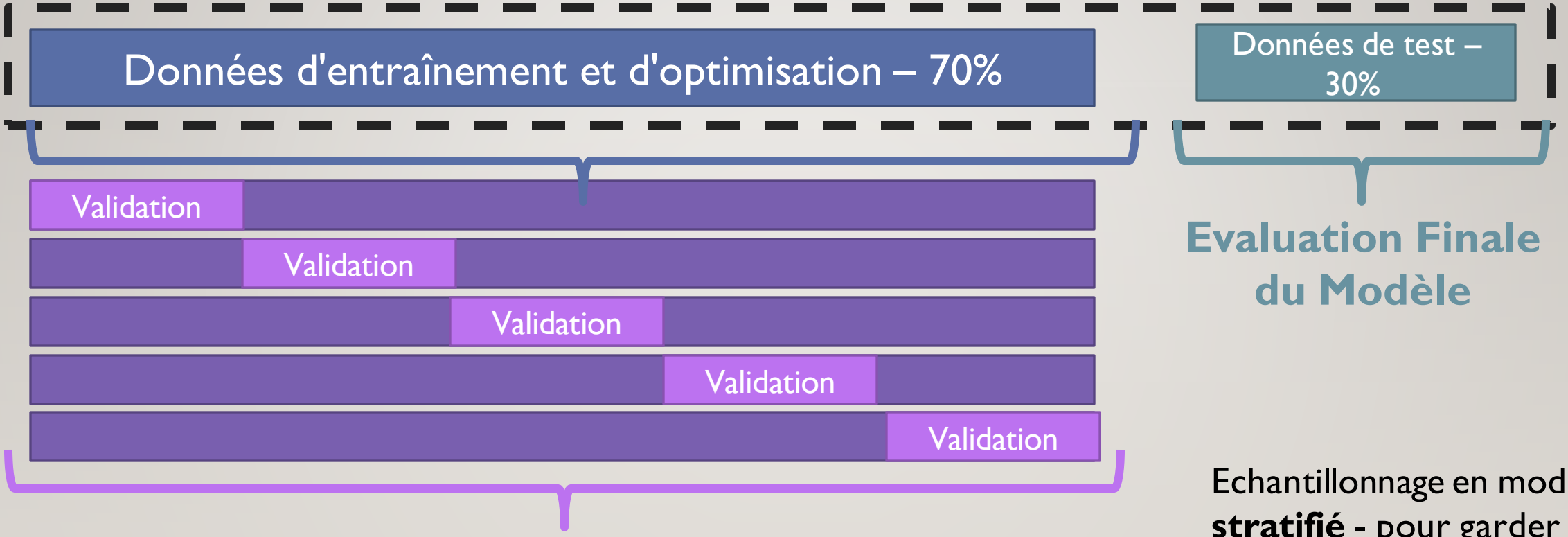
RÉDUCTION DE DIMENSION → 377 Variables

Peu Peuplées - plus de 70% des valeurs manquantes

Fortement Corrélés - plus de 90% de corrélation

Presque Constantes – moins de 10% de déviation
(si encodé entre 0 et 1)

MÉTHODE D'ÉVALUATION



Echantillonnage en mode **stratifié** - pour garder le même ratio entre remboursement et défaut

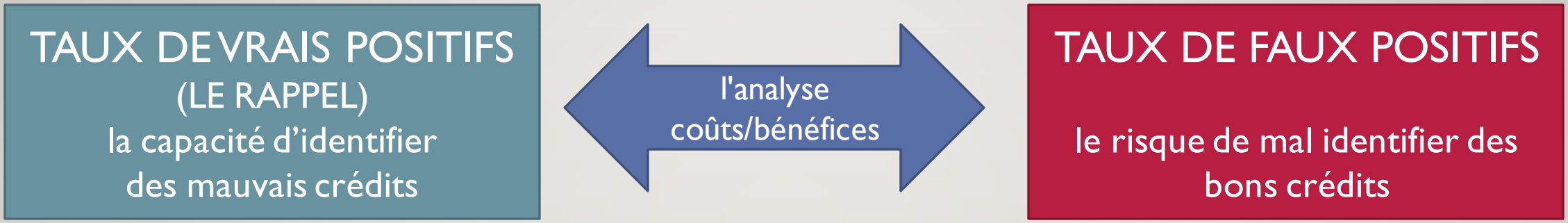
LA CLASSE CIBLE DÉSÉQUILIBRÉE

STRATÉGIES:

- **sous-échantillonnage** aléatoire de la classe majoritaire
- **SMOTE** (Synthetic Minority Oversampling Technique)
  = suréchantillonnage de la classe minoritaire
- **pondération** des individus selon la classe cible (class_weight = 'balanced')

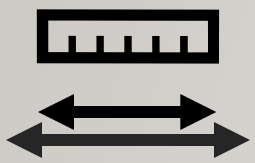
- **un métrique d'évaluation adapté** (e.g. ROC-AUC, précision moyenne, ...)
  - ➢ éviter les estimations de performances gonflées sur des jeux de données déséquilibrés

MÉTRIQUE D'ÉVALUATION - ROC-AUC



- Sélection des modèles optimaux
  - indépendamment de la distribution de classe
  - avant de spécifier le contexte de coût (mis en place d'un seuil de décision)

LES MODÈLES TESTÉS

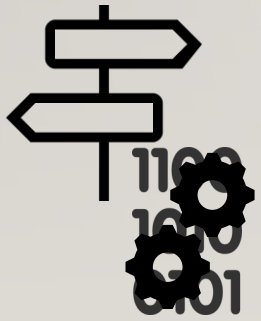


Mise à l'Echelle

- MinMaxScaler()
- Prétraitement customisé - prend en compte l'asymétrie de la distribution des valeurs (skew) pour éventuellement faire une mise à l'échelle logarithmique

Imputation des valeurs manquants

- médiane, moyenne



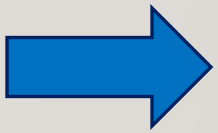
Modèle de Classification

- **LogisticRegression** (sklearn)
- **BalancedRandomForrest** (imblearn)
- **RUSBoostClassifier** (imblearn)
- **LightGBM – light gradient boosting machine** (lightgbm)

Pipeline

RÉSULTATS

Model	Mise à l'échelle	Imputation	Sous-échant.	AUC-ROC valid.	AUC-ROC test	Précision moyenne test	F1-score test	Temps (s)
LightGBM	--	--	non	0.7807	0.7780	0.1525	0.2935	20.5
LogisticRegr.	Cust.	médiane	non	0.7704	0.7694	0.1442	0.2754	45.2
RUSBoost	--	médiane	non	0.7701	0.7694	0.1452	0.2780	114.6
RandomForest	--	médiane	oui	0.7538	0.7516	0.1369	0.2627	31.7

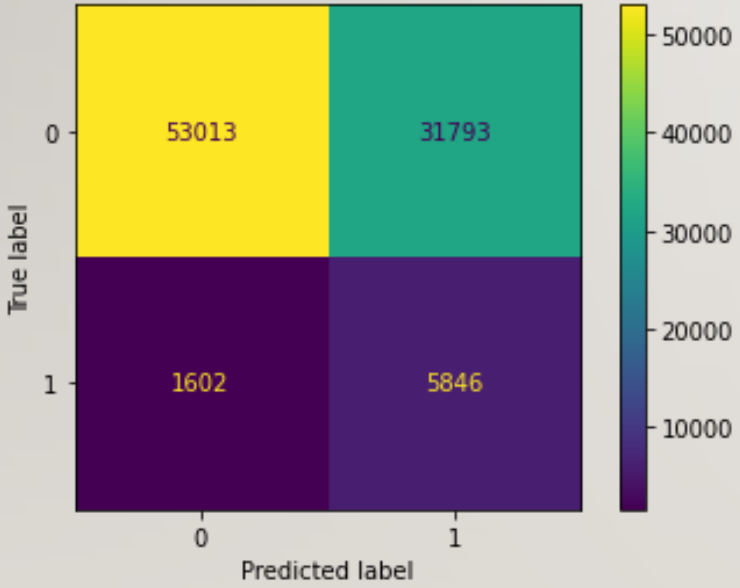


LightGBM

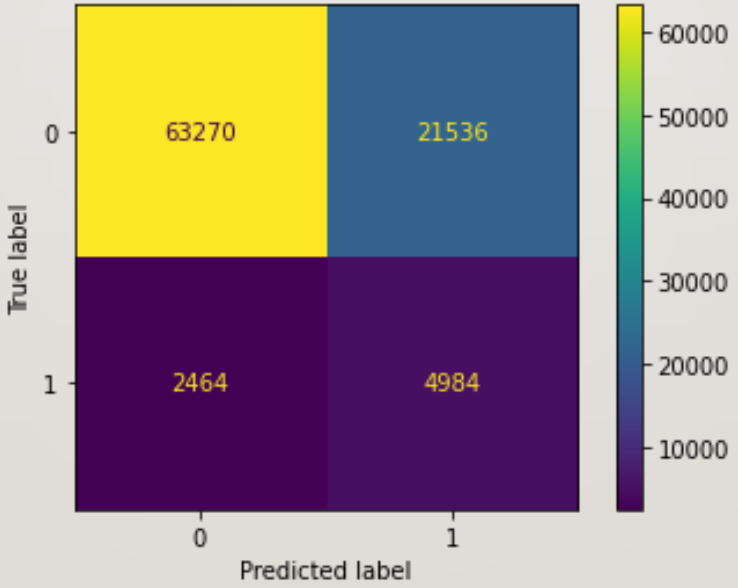
LE MODÈLE LIGHT-GBM

Matrix de Confusion avec différents seuils de décision

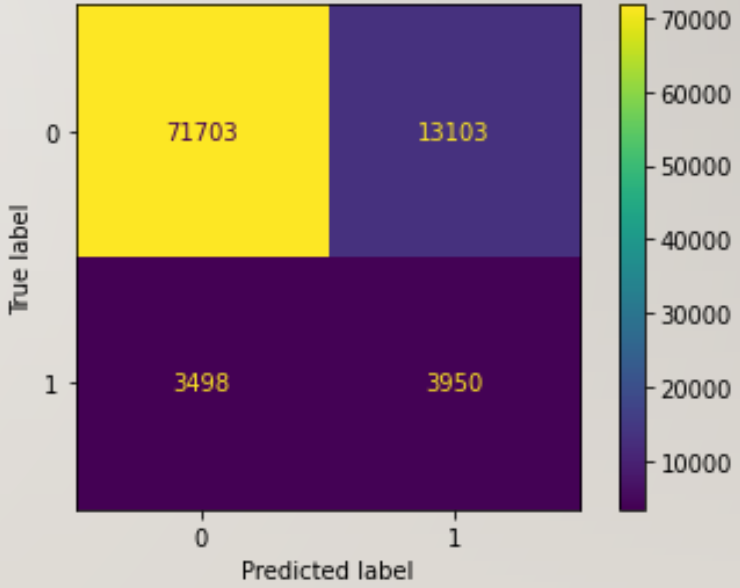
S = 0.4



S = 0.5

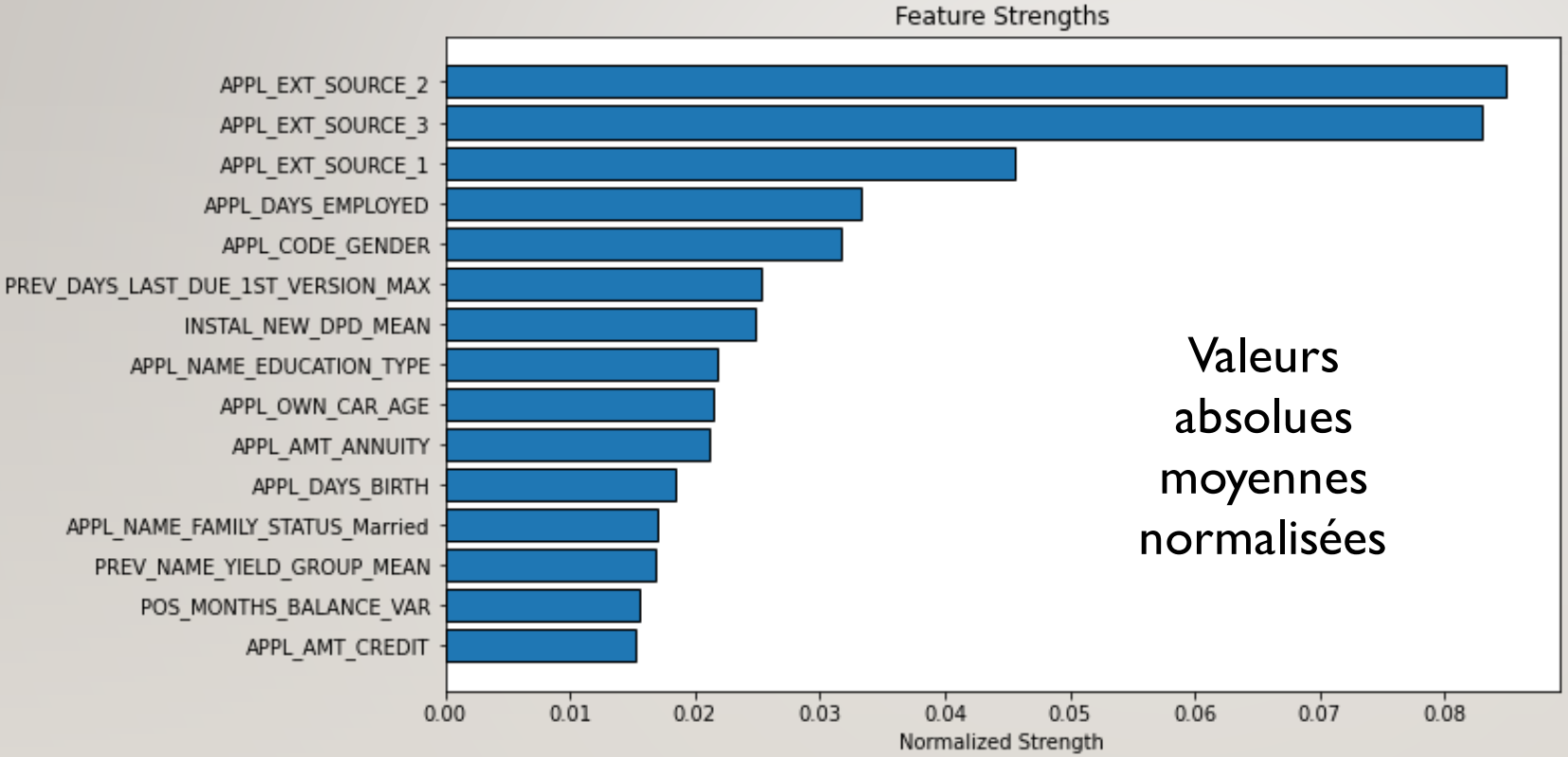


S = 0.6



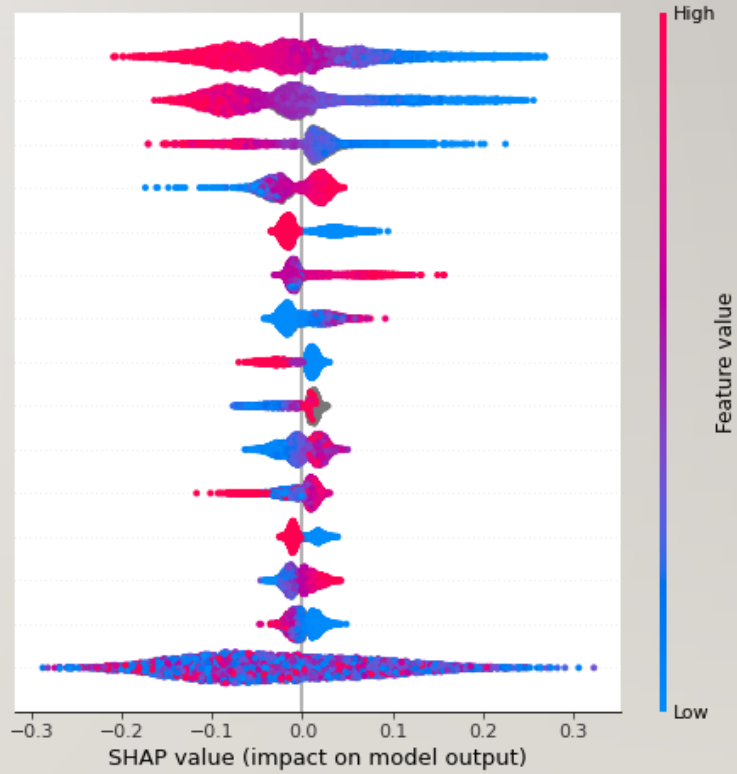
INTERPRÉTATION DU MODÈLE

Valeurs de Shapley (SHAP module)



Valeurs absolues moyennes normalisées

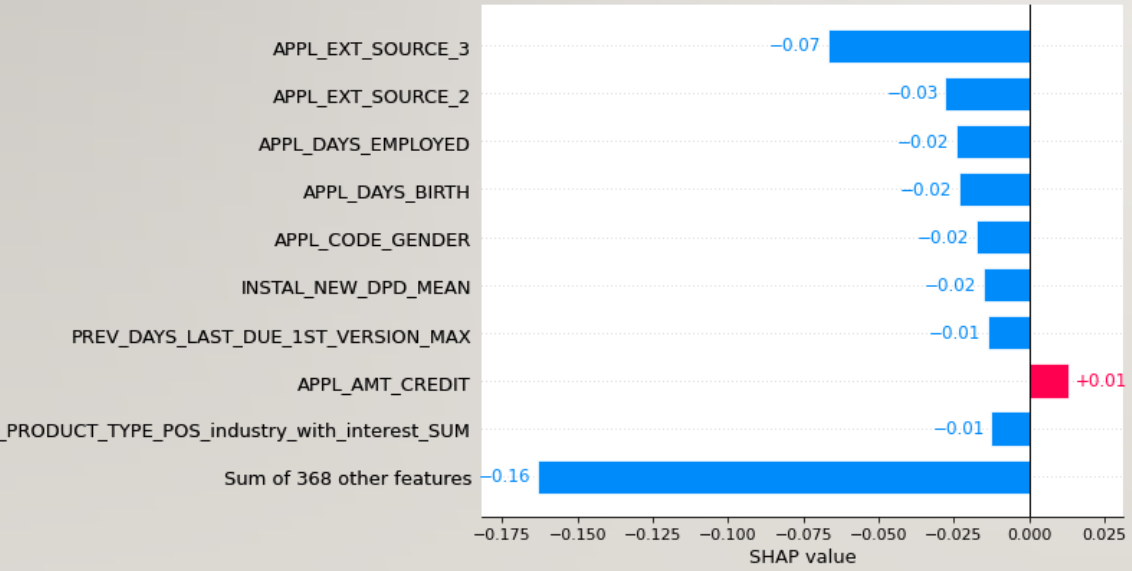
Remboursé **Défaut**



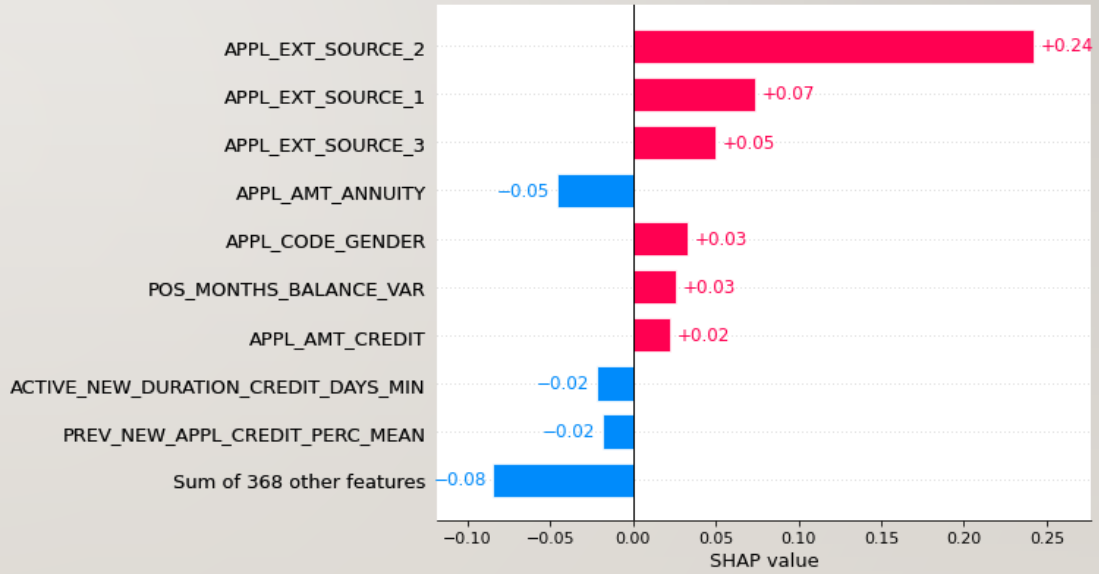
INTERPRÉTATION DU MODÈLE

Valeurs de Shapley (SHAP module)

Un individu qui a remboursé son crédit

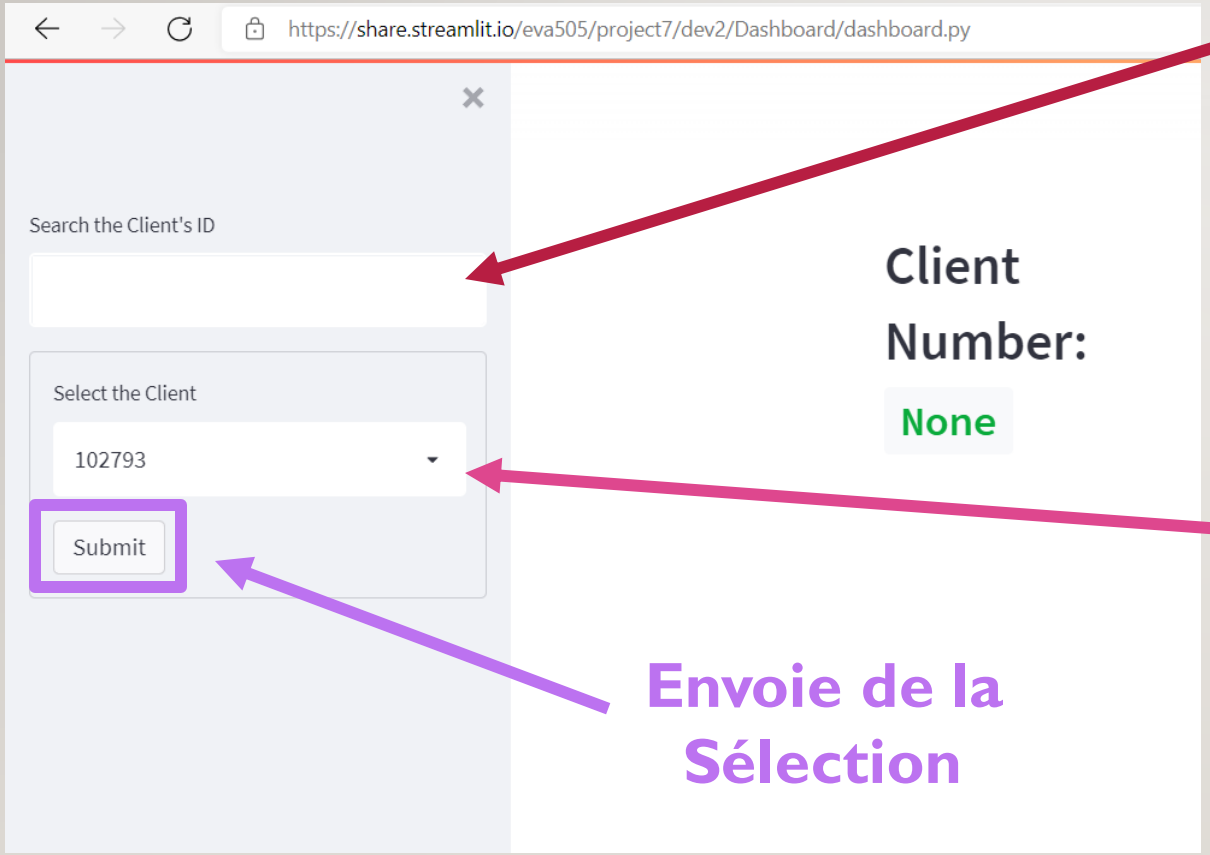


Un individu en défaut de crédit



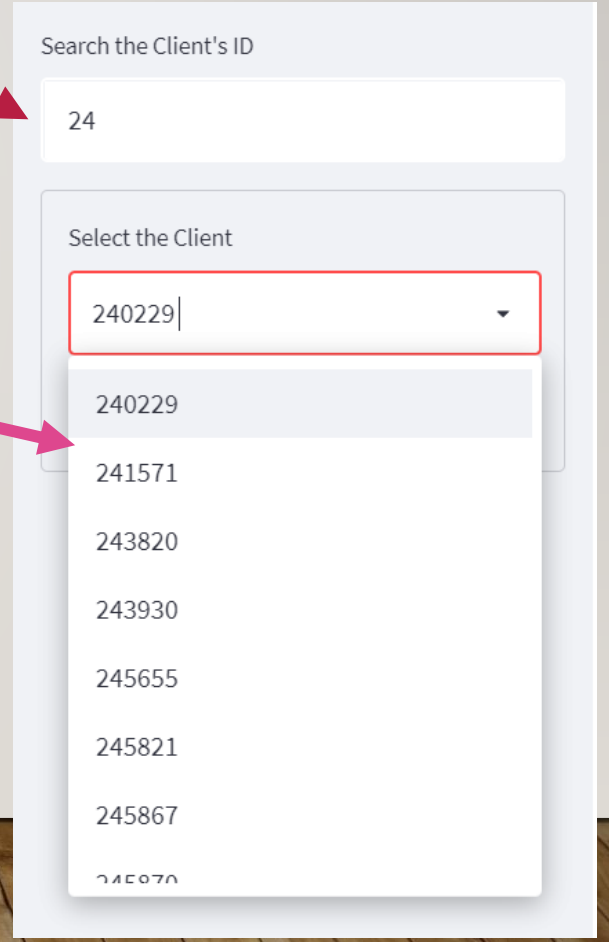
LE DASHBOARD – SÉLECTION CLIENT

Recherche du identifiant du client (correspondant à son début)



Sélection du identifiant du client

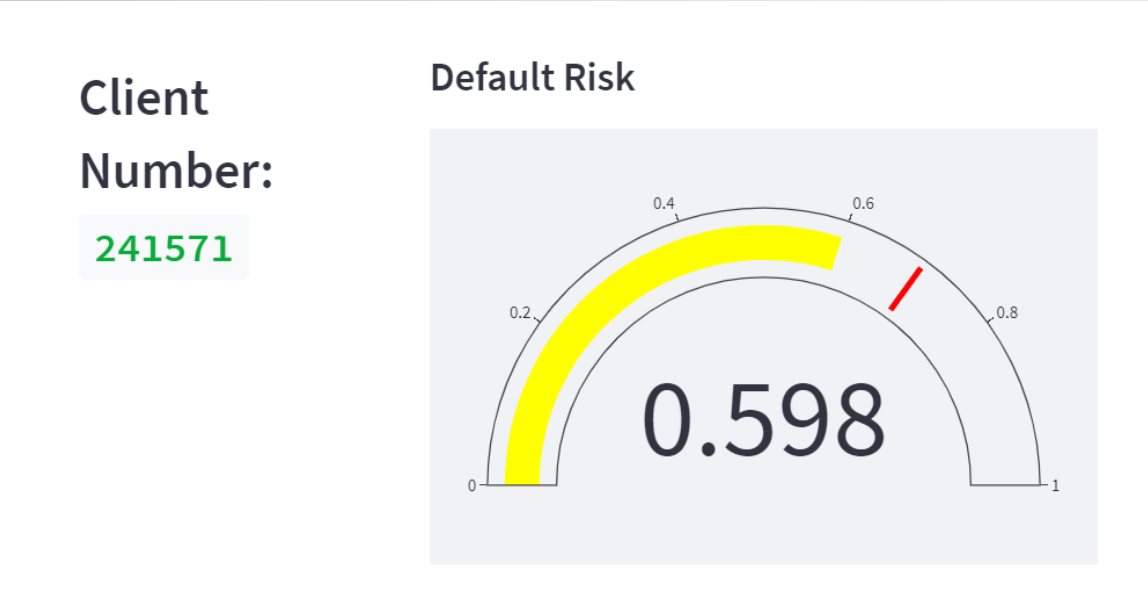
Envoie de la Sélection

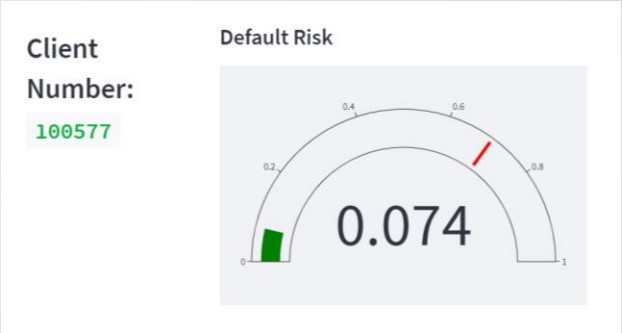


Streamlit - share.streamlit.io/eva505/project7/dev2/Dashboard/dashboard.py

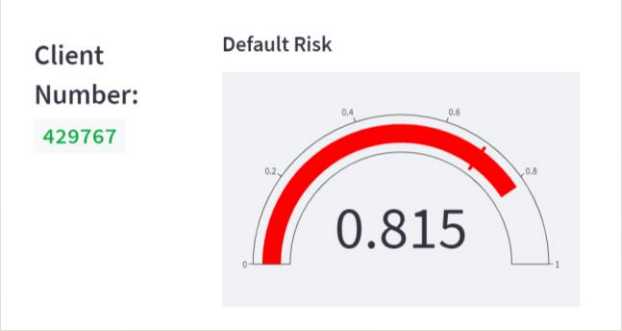
LE DASHBOARD – PRÉDICTION

Le Risque de Défaut de Crédit
comme prédit par le modèle





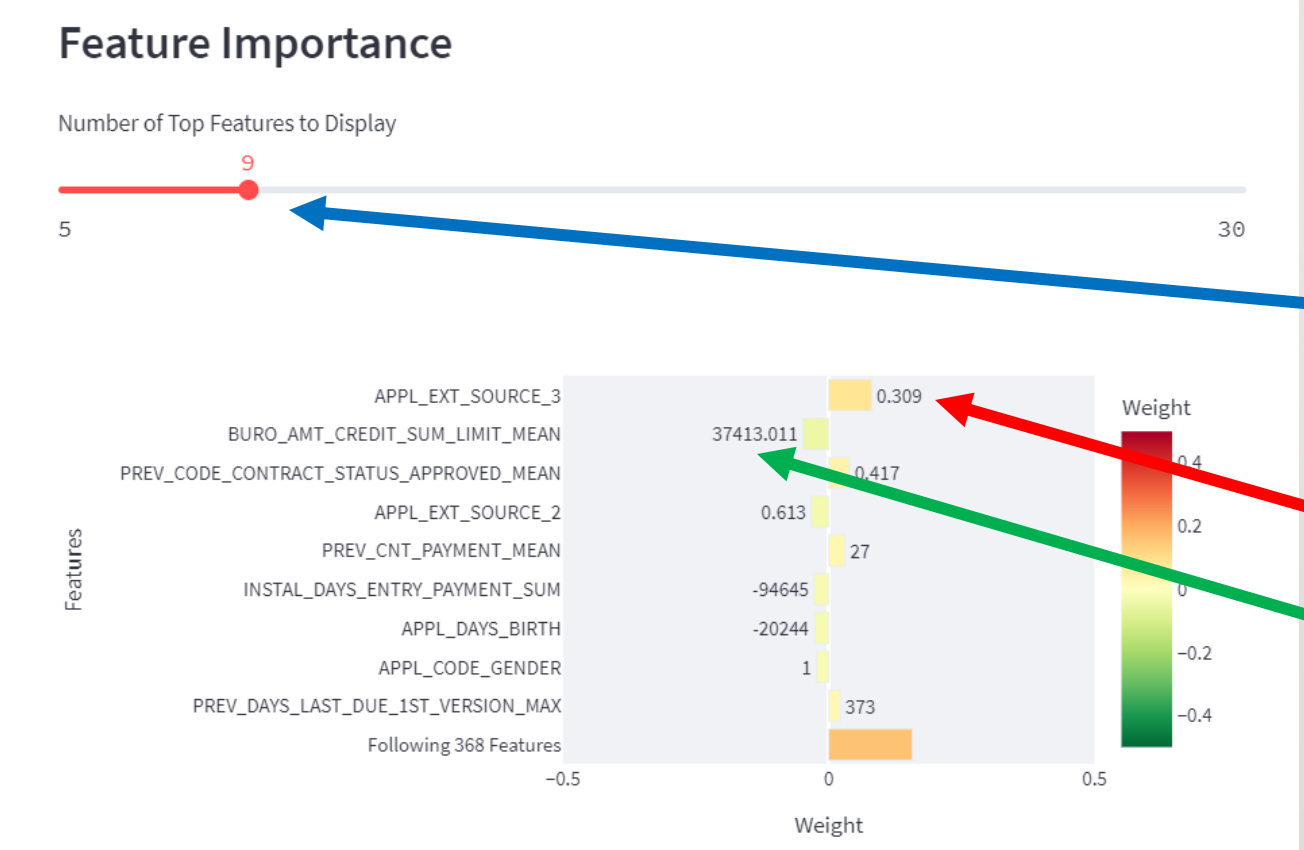
Très Faible Risque



Risque Elevée

Streamlit - share.streamlit.io/eva505/project7/dev2/Dashboard/dashboard.py

LE DASHBOARD – LES FACTEURS DE DÉCISION



Les Facteurs le plus importantes
dans la prédiction de modèle

Nombre des features à afficher

Défavorable

Favorable

Streamlit - share.streamlit.io/eva505/project7/dev2/Dashboard/dashboard.py

LE DASHBOARD – COMPARAISON DES CLIENTS

Choix des Filtres
pour faire la comparaison

Choix du Feature

Distribution de tous les clients
dans la base de données

Avec le filtre appliqué

Valeur du Client

Streamlit - share.streamlit.io/eva505/project7/dev2/Dashboard/dashboard.py

POINTS D'AMÉLIORATION

Préparation et Sélection des Données

- nettoyage plus soigneux, i.e. détection des valeurs aberrantes, features catégorielles,...
- réduction de nombre des features utilisés
- features métiers ou polynomiales

Le Modèle - LGBM

- re-évaluation du modèle, i.e. mise à l'échelle, hyperparamètres, ...

La fonction de coût

- analyse coûts-avantages, mauvais clients acceptés versus bons clients rejetés
- prendre en compte le montant du crédit

POINTS D'AMÉLIORATION

Evaluation du biais dans le modèle

- le genre est un feature le plus important dans le modèle, ce que problématique éthiquement

API et Dashboard

- retours des utilisateurs
- ajoute des fonctionnalités
- mise en place d'un propre base de données
- meilleur serveur