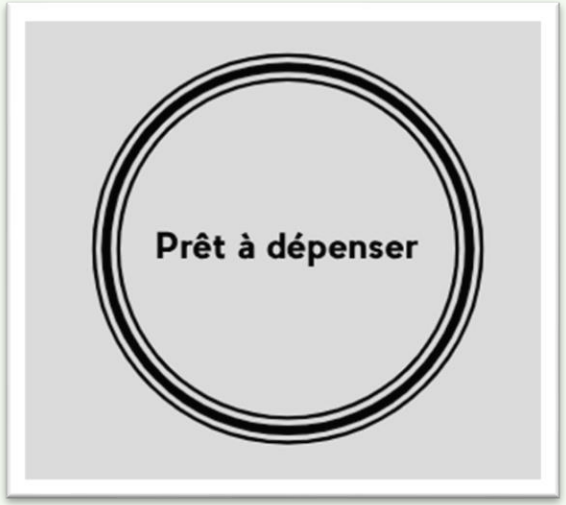
OPENCLASSROOMS




# P7 : Implémentez un modèle de scoring

11/08/2022

DUBART Maxime

# Implémentez un modèle de scoring

OBJETIFS

## Création d’un outil de « scoring crédit »

Probabilité de défaut de paiement et classification des demandes

## Développement d’un dashboard interactif

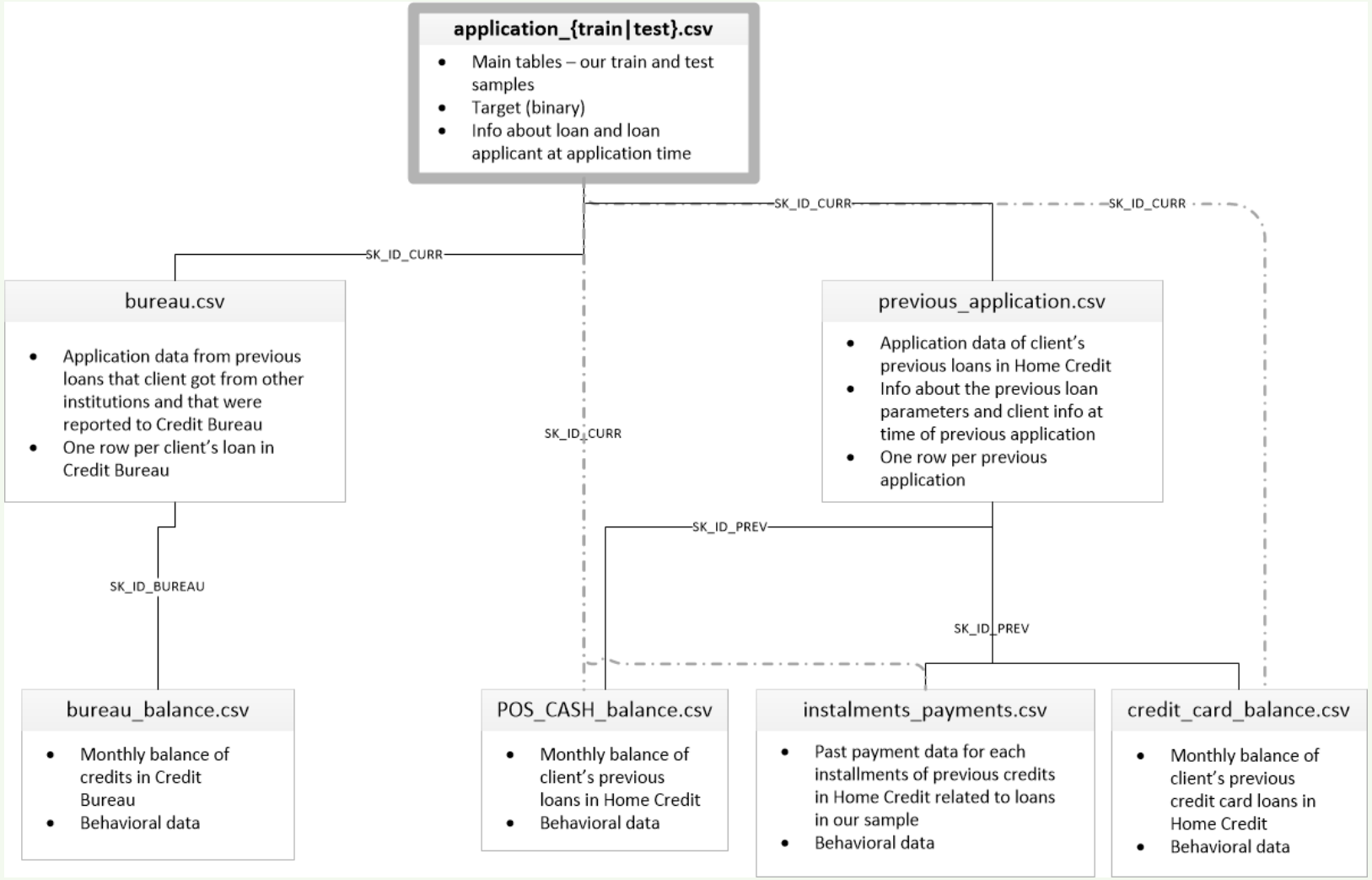
Permettre plus de transparence vis-à-vis des décisions d’octroi de crédit pour les clients

# Implémentez un modèle de scoring

DONNEES

## Fichiers de données

**7 fichiers**



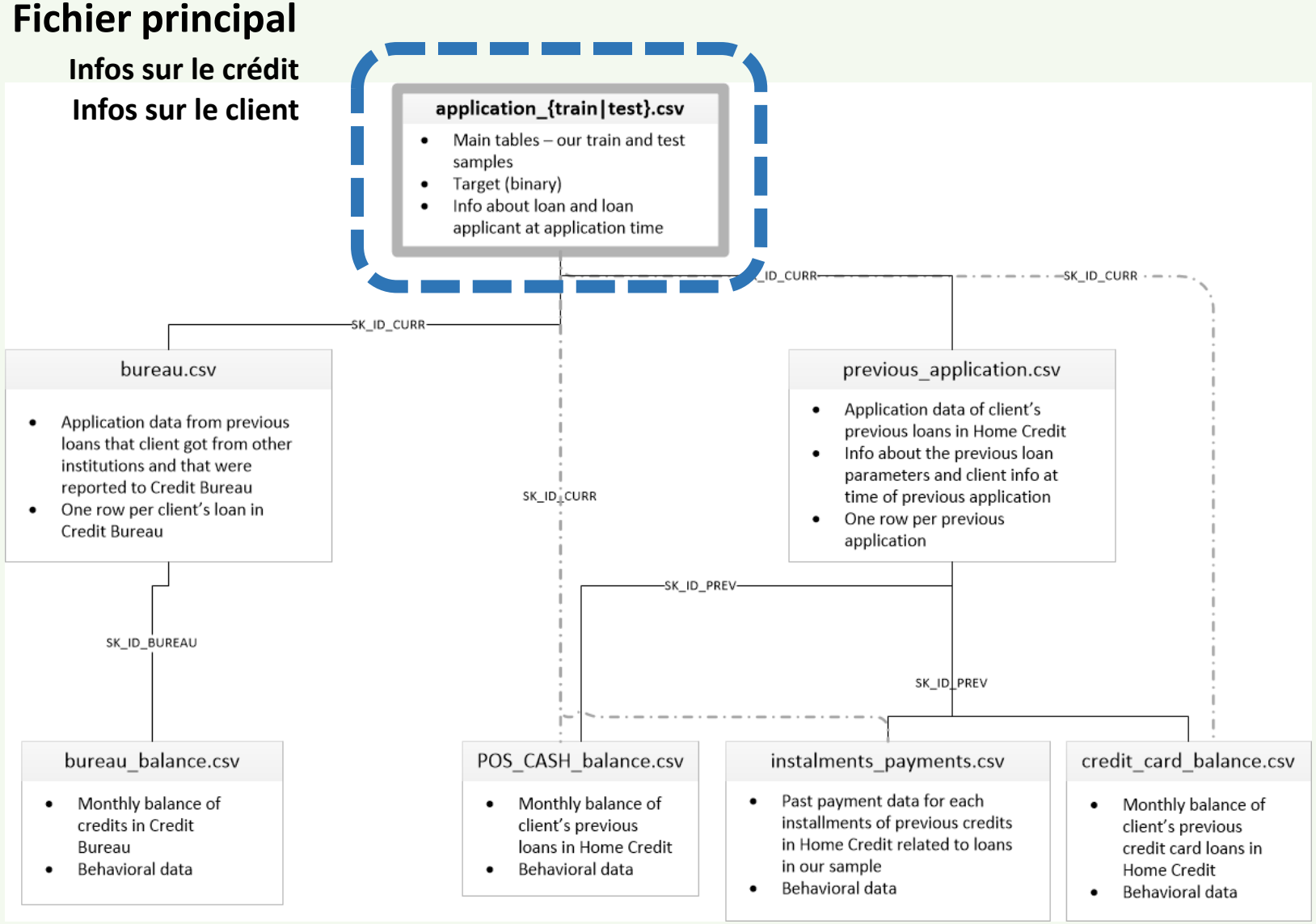
# Implémentez un modèle de scoring

DONNEES

## Fichiers de données

**7 fichiers**

**8% de positifs (i.e. défaut)**



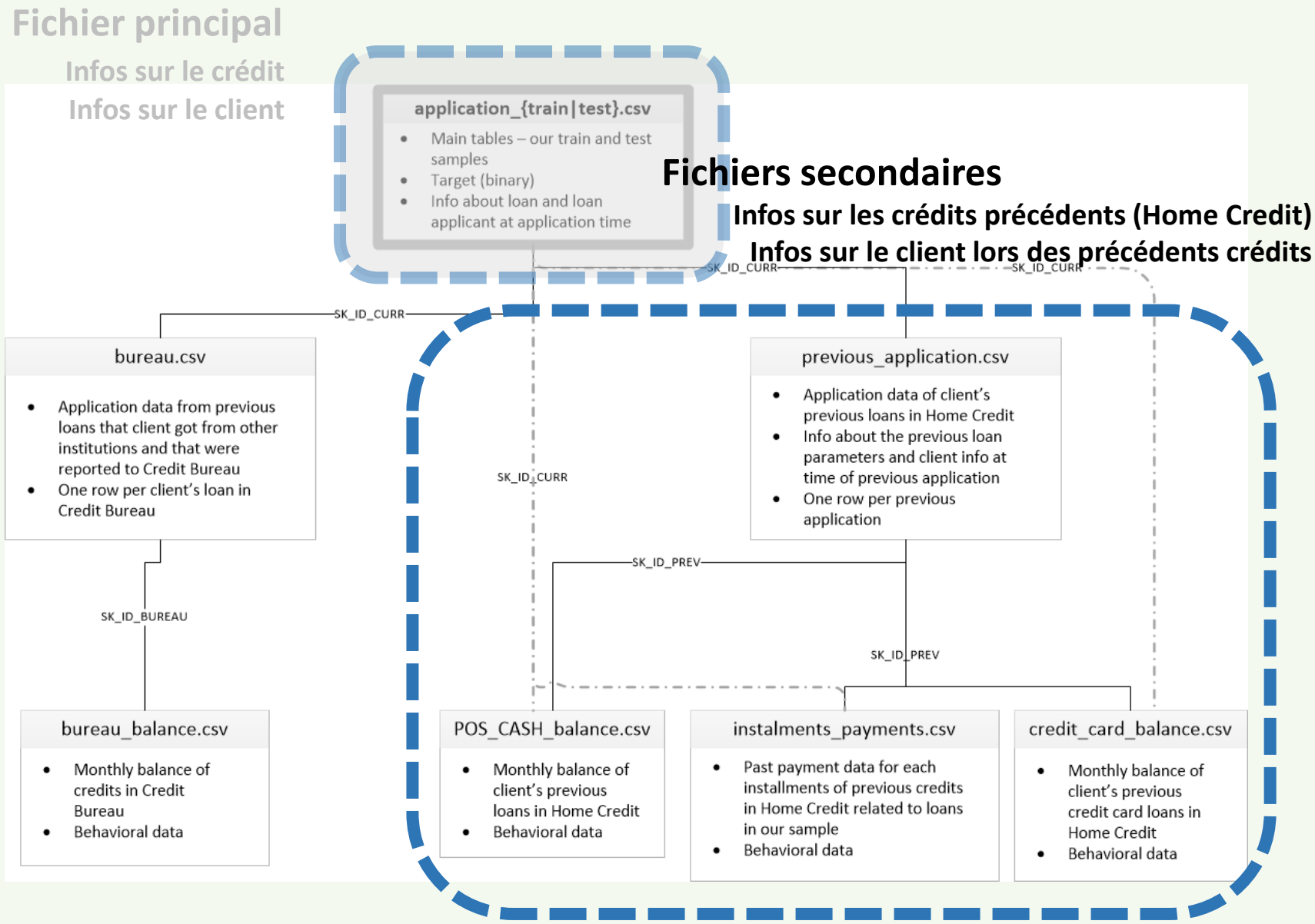
# Implémentez un modèle de scoring

DONNEES

## Fichiers de données

**7 fichiers**

**8% de positifs (i.e. défaut)**



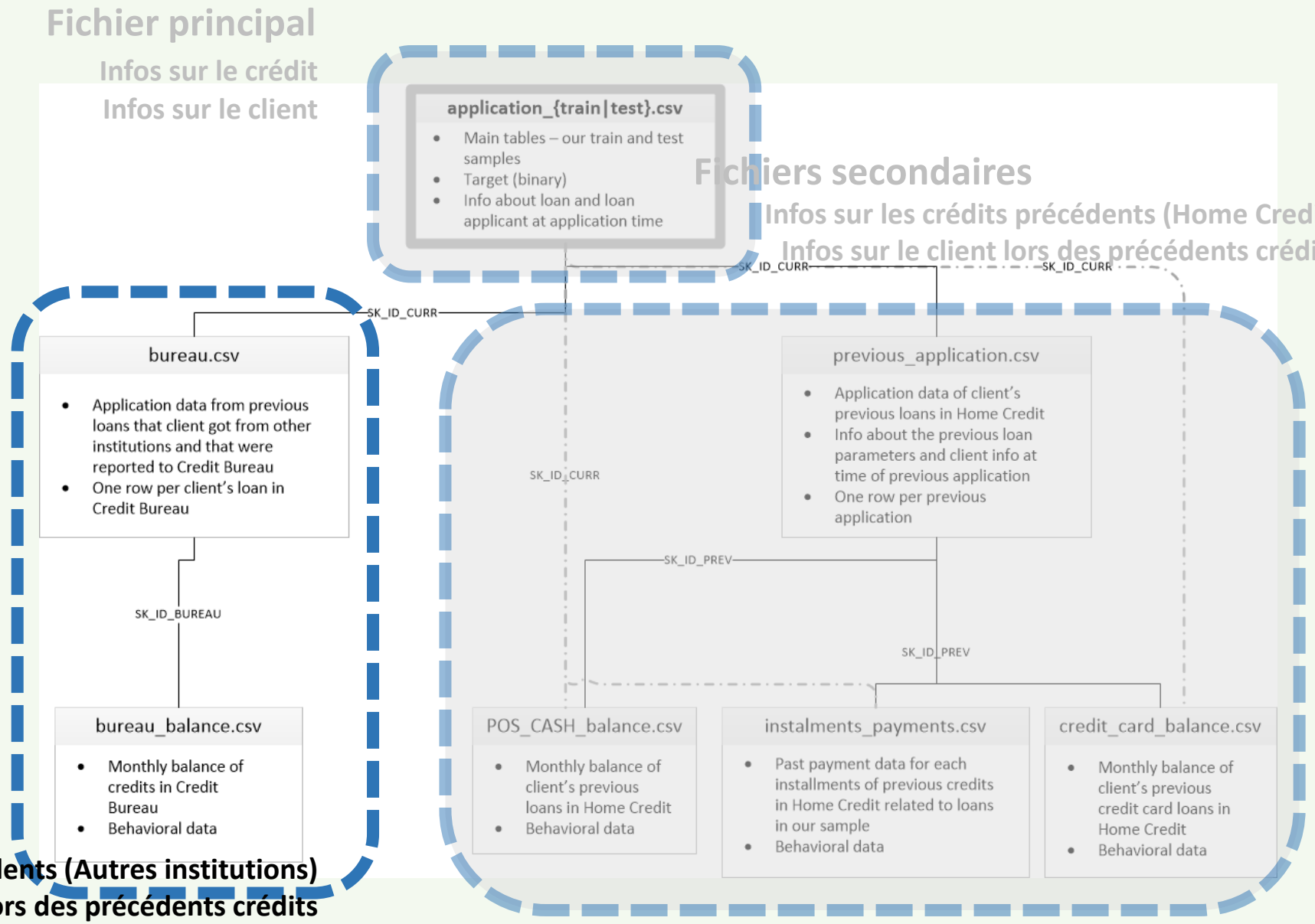
# Implémentez un modèle de scoring

DONNEES

## Fichiers de données

**7 fichiers**

**8% de positifs (i.e. défaut)**




**Fichiers secondaires**

**Infos sur les crédits précédents (Autres institutions)**

**Infos sur le client lors des précédents crédits**

# Implémentez un modèle de scoring

DONNEES

## Fichiers de données

**7 fichiers**

**8% de positifs (i.e. défaut)**

**Fichier principal**

**Infos sur le crédit**

**Infos sur le client**


**121 colonnes – différents types d'informations**

- Sur le logement du client (surface, localisation, etc.)
- Sur les demandes à propos du client au Bureau des Crédits
- Informations administratives (e.g., quels documents ont été fournis)
- Informations sur le travail du client (localisation, type, salaire, etc.)
- Informations sur les hobbies du client
- Informations sur la situation du client (e.g., âge, statuts marital, # enfants etc.)
- Informations sur la demande de crédit (montant, mensualités, type de crédit)
- Informations provenant de sources externes

# Implémentez un modèle de scoring



## (i) Réduire la dimensionalité, suppression de :

- features avec > 40% de valeurs manquantes – excepté EXT_SOURCE_1
- features 'flag' *a priori* inutile pour la classification (e.g. enregistrement de tel document)
- features relatives à l'enregistrement dans la BDD (e.g. heure de la demande)

→ **Aboutit à la suppression de env. 70% des *features* (36 au lieu de 121)**

# Implémentez un modèle de scoring



## (i) Réduire la dimensionalité, suppression de :

- features avec > 40% de valeurs manquantes – excepté EXT_SOURCE_1
- features 'flag' *a priori* inutile pour la classification (e.g. enregistrement de tel document)
- features relatives à l'enregistrement dans la BDD (e.g. heure de la demande)

→ **Aboutit à la suppression de env. 70% des *features* (36 au lieu de 121)**

## (ii) Adapté le jeu de données aux analyses :

- suppression des valeurs aberrantes (remplacées par NA)
- encodage des *features* catégorielles (one-hot encoding)
- création de variables composées (e.g., % de jours employés, valeurs du crédit relative au salaire, revenu moyen par membre du foyer, etc.)

→ **Au final, 143 *features* pour env. 300k clients**

# Implémentez un modèle de scoring

MODELES

## Modèles de classification

### Régression logistique (Lasso/Ridge)

**Avantages** : Coefficients facilement interprétables, considération du déséquilibre de classe via poids, temps fit/prédiction, sélection variables possible, peu paramétrique

**Inconvénients** : Contrainte linéarité/additivité, imputation des valeurs manquantes (ici, par la moyenne)

# Implémentez un modèle de scoring

MODELES

## Modèles de classification

### Régression logistique (Lasso/Ridge)

**Avantages** : Coefficients facilement interprétables, considération du déséquilibre de classe via poids, temps fit/prédiction, sélection variables possible, peu paramétrique
**Inconvénients** : Contrainte linéarité/additivité, imputation des valeurs manquantes (ici, par la moyenne)

### XGBoost

**Avantages** : Pas de contrainte de linéarité, considération des valeurs manquantes
**Inconvénients** : Interprétation plus délicate (i.e. nécessite analyses posterieurs), temps fit/prediction plus important, hautement paramétrique

# Implémentez un modèle de scoring

MODELES

## Classes non balancées

**Peut-être problématique pour les algo. de ML**

Pas vraiment pour la reg. log.

# Implémentez un modèle de scoring



## Classes non balancées


**Peut-être problématique pour les algo. de ML**

Pas vraiment pour la reg. log.

**Stratégies possibles**

- Sur-échantillonner la classe minoritaire
- Sous-échantillonner la classe majoritaire (1:2)
- Création d'échantillons synthétiques (e.g., SMOTE)

# Implémentez un modèle de scoring

MODELES

## Classes non balancées


**Peut-être problématique pour les algo. de ML**

Pas vraiment pour la reg. log.

**Stratégies possibles**

- Sur-échantillonner la classe minoritaire
- **Sous-échantillonner la classe majoritaire (1:2)**
- Création d'échantillons synthétiques (e.g., SMOTE)

*Forte augmentation de la taille du jeu de données*

*Via k-nn, problème des variables binaires*

# Implémentez un modèle de scoring

MODELES

## Classes non balancées


**Peut-être problématique pour les algo. de ML**

Pas vraiment pour la reg. log.

**Stratégies possibles**

*Forte augmentation de la taille du jeu de données*

- Sur-échantillonner la classe minoritaire
- **Sous-échantillonner la classe majoritaire (1:2)**
- Création d'échantillons synthétiques (e.g., SMOTE)

*Via k-nn, problème des variables binaires*

**Entrainement sur jeux de données re-balancés ou non**

# Implémentez un modèle de scoring

MODELES

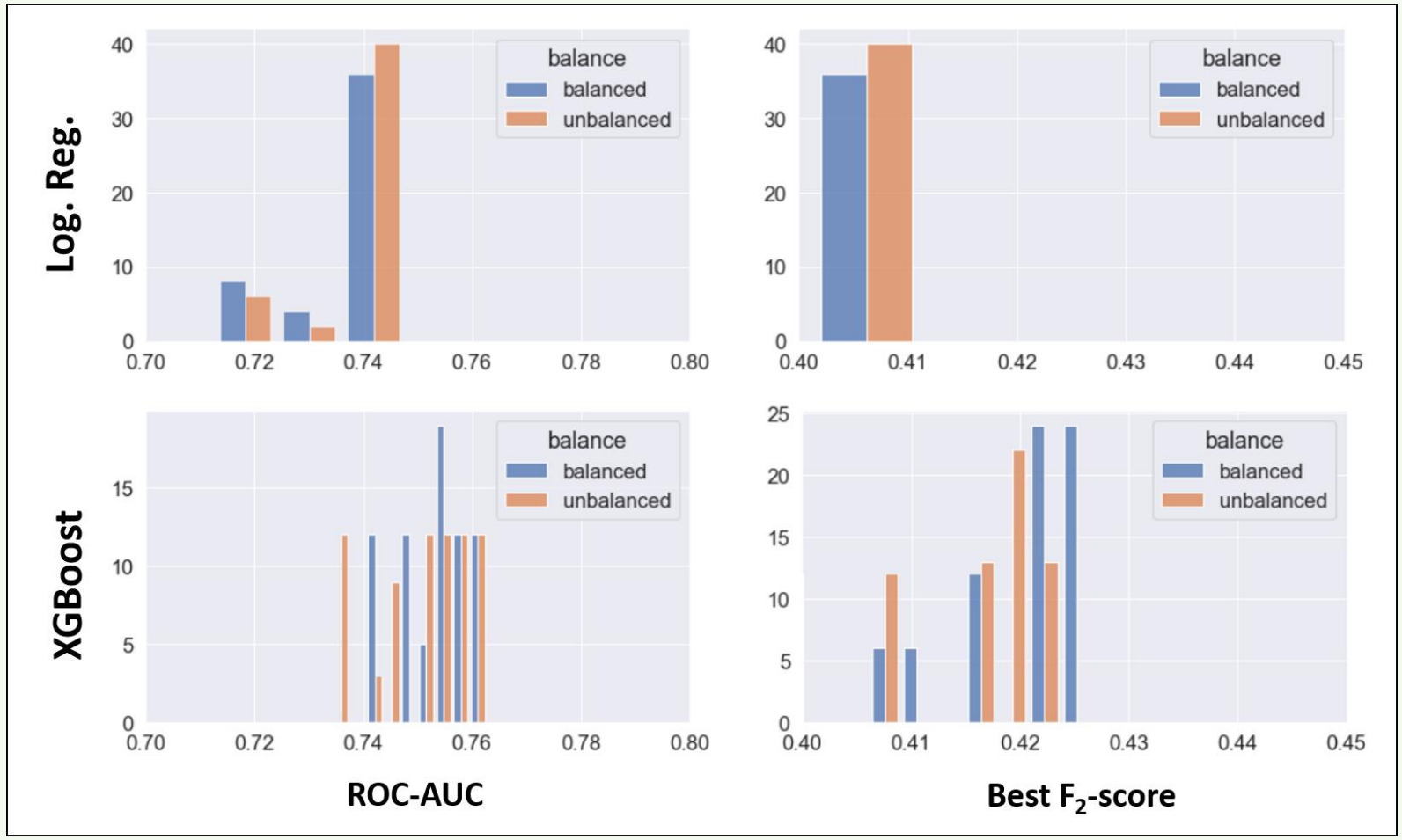
## Méthodologie

- **Datasets original ou rebalancé (downsampling)**
- **Choix des hyper-paramétres (recherche sur grille / cross-validation)**
  Split train / validation : 70/30
  5-folds cross-validation
- **Métriques : ROC-AUC et $F_\beta$-Score**

# Implémentez un modèle de scoring

MODELES

## Scores de validation




**XGBoost légèrement meilleur que la régression logistique**

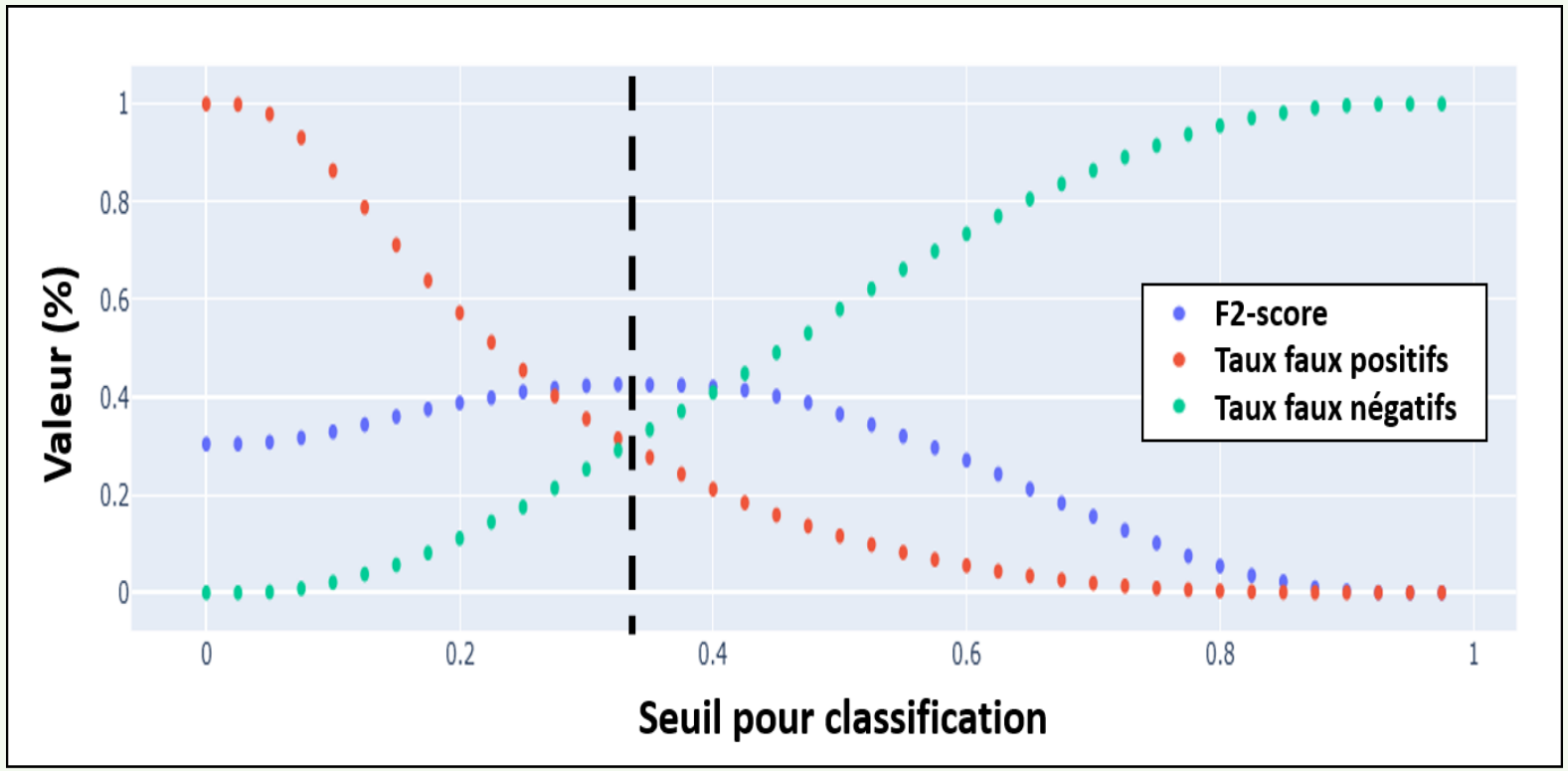
**Peu d'effet du rééquilibrage des classes**

**Scores proches de ceux obtenus sur les meilleurs résultats Kaggle (avec l'ensemble des fichiers)**

# Implémentez un modèle de scoring

MODELES

## Sélection du seuil



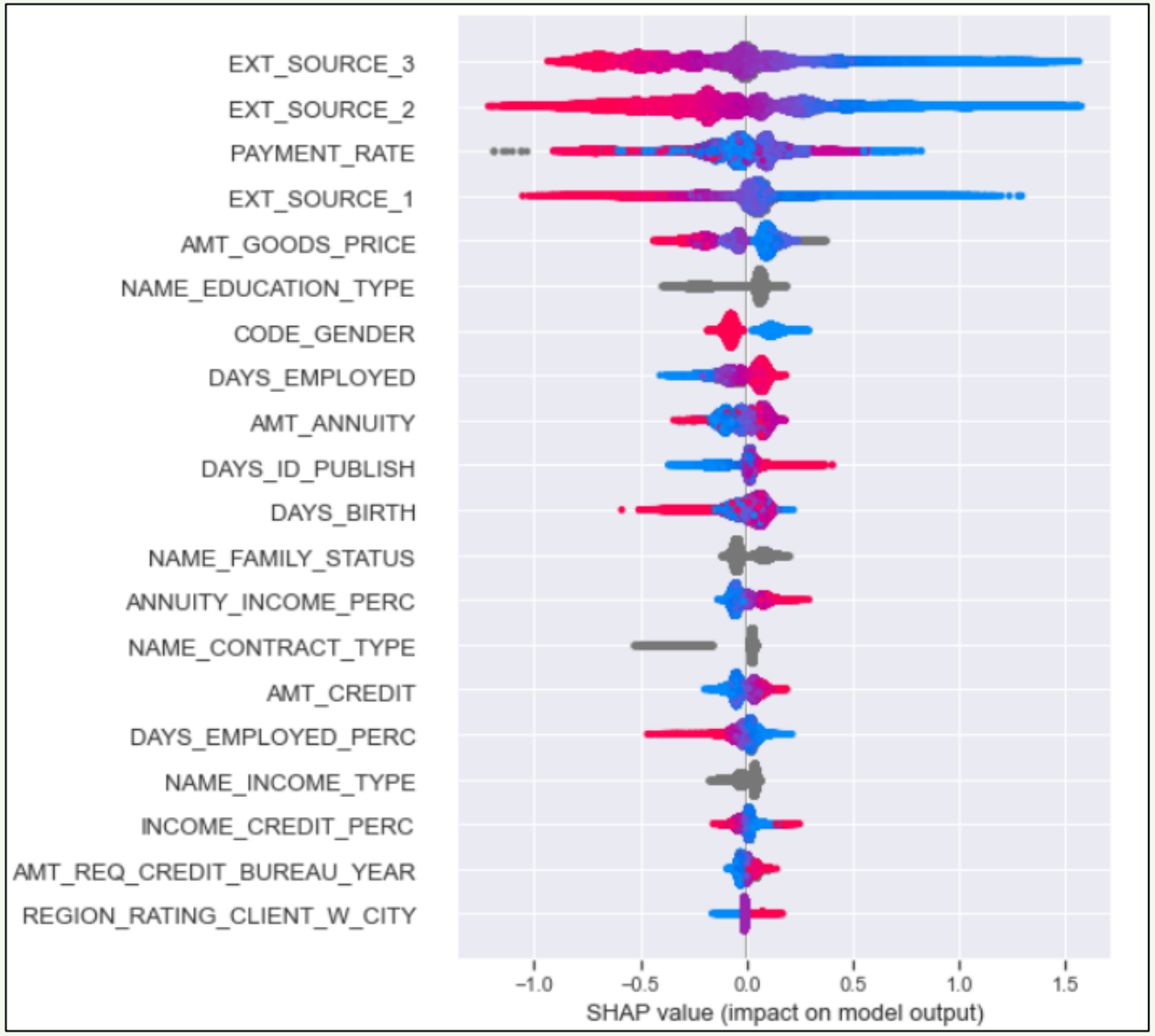

**Seuil de 0.3/0.35 maximise le $F_2$-score**

**A ce seuil : env. 30% de FP et FN**

# Implémentez un modèle de scoring

MODELES

## Global features importances (SHAPs)




**Sources extérieures d'info important beaucoup**

**Le taux de remboursement (i.e., annuité / total)**

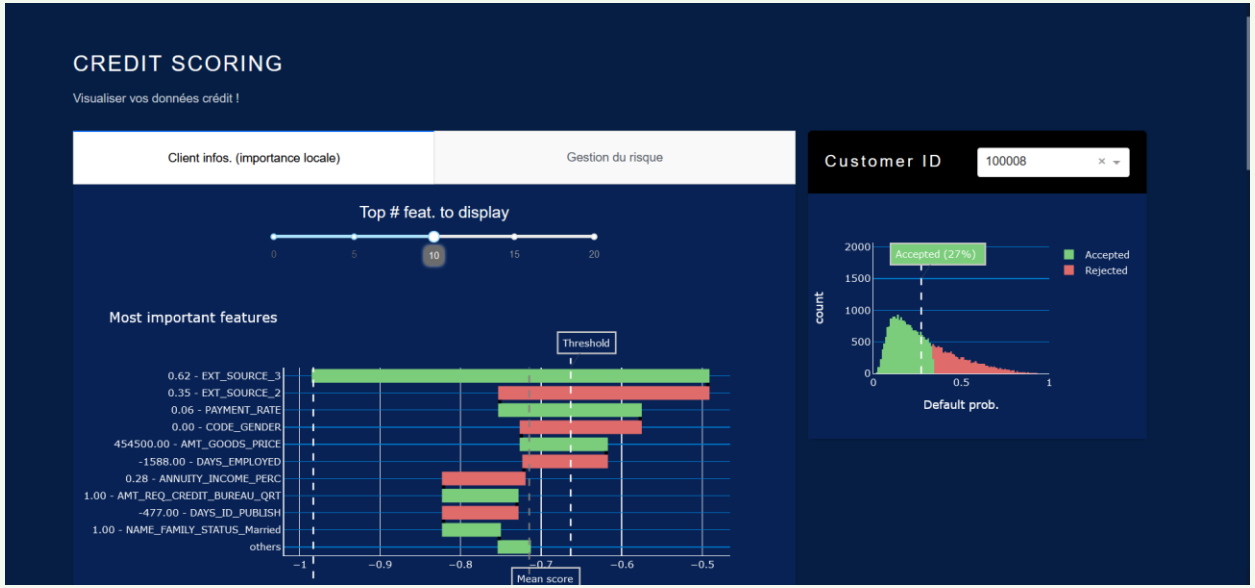
**Le niveau d'éducation, le genre, la durée depuis laquelle le client travaille (inversée)**

# Déployer le modèle & dashboard

## API & Dashboard

main.py — API\app | main.py — Dashboard\app | style.css

```python
# -*- coding: utf-8 -*-

from flask import Flask, jsonify, request
import pandas as pd
import shap
import pickle
import sklearn
import xgboost
import requests
import numpy as np

app = Flask(__name__)

# Read Scaler
with open('standard_scaler.pkl', 'rb') as inp:
    ss = pickle.load(inp)

# Read model
with open('xgb_model.pkl', 'rb') as inp:
    xgb = pickle.load(inp)

# Read explainer
with open('shap_xgb_explainer.pkl', 'rb') as inp:
    xgb_explainer = pickle.load(inp)

# Read data (X)
df = pd.read_csv('./us_data_subset.csv')
df_unscaled = df.copy()

# Get targets & ids as separate lists
target = df.pop('TARGET')
sk_id_curr = df.pop('SK_ID_CURR')

# Scaled df
df_ss = pd.DataFrame(ss.transform(df), columns = df.columns)
probs = xgb.predict_proba(df_ss)[:,1]

# Get fn, fp, tp, tn
def fp_fn_tp_tn(pp_, y_):
    thresh = np.arange(0,1,0.025)

    output = []

    for t in thresh:
        yp_ = (pp_ > t).astype('int')
        cm = sklearn.metrics.confusion_matrix(y_, yp_)
        tn_ = cm[0,0]
        fp_ = cm[0,1]
        fn_ = cm[1,0]
        tp_ = cm[1,1]
        prec_ = tp_ / (tp_ + fp_)
        recall_ = tp_ / (tp_ + fn_)
```



# Déployer le modèle & dashboard



## Outils utilisés

### API

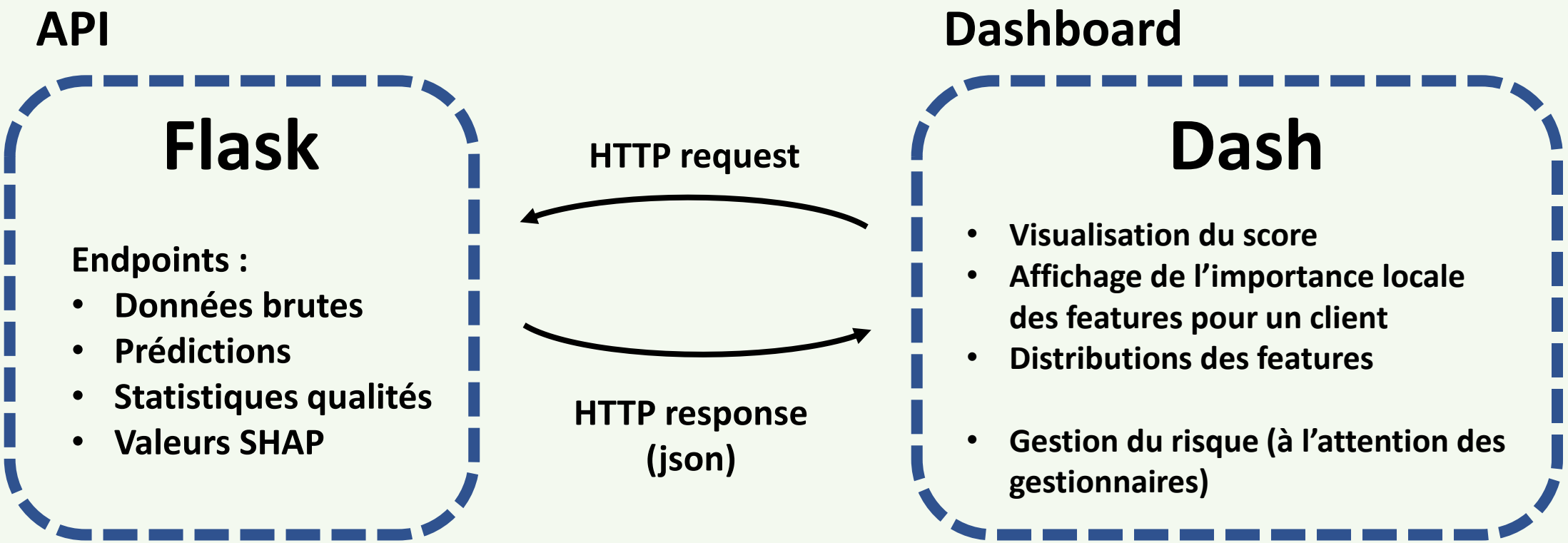
**Flask**

**Endpoints :**

- **Données brutes**
- **Prédictions**
- **Statistiques qualités**
- **Valeurs SHAP**

# Déployer le modèle & dashboard

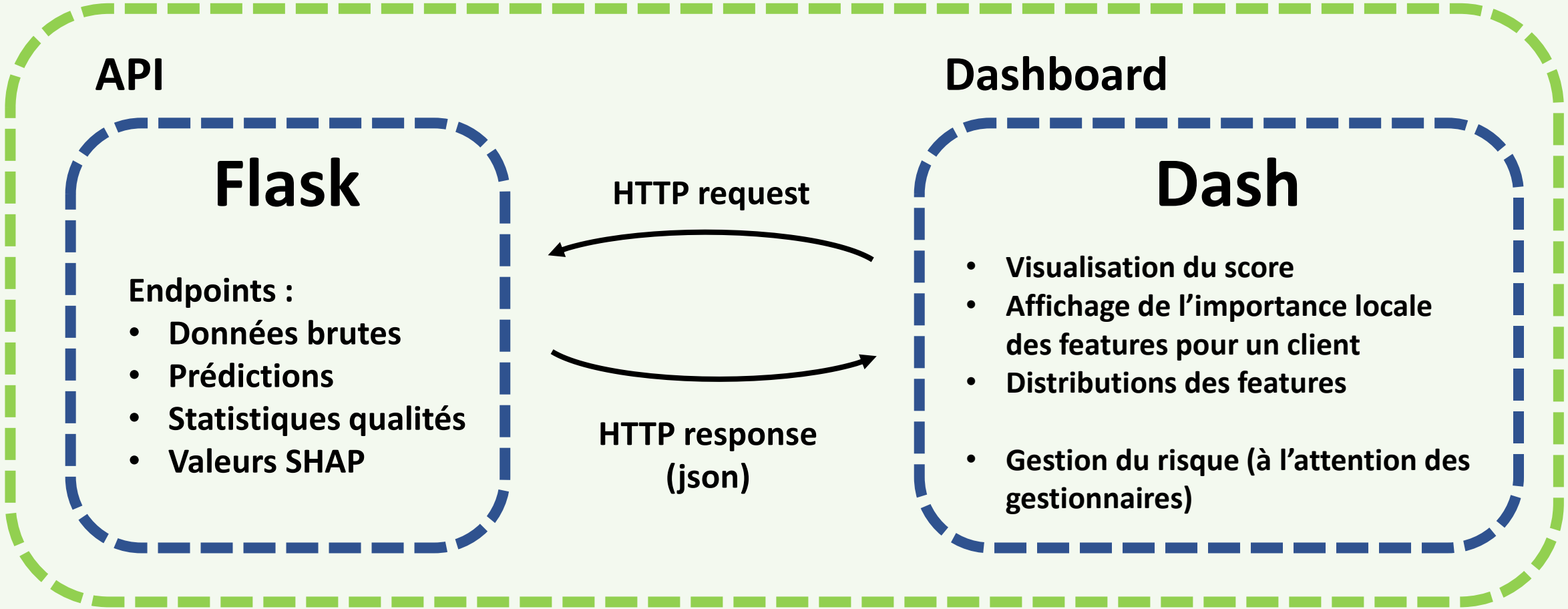
## Outils utilisés



# Déployer le modèle & dashboard

## Outils utilisés

**Versionnées sur GitHub & déployées sur Heroku**



# Déployer le modèle & dashboard

API & Dashboard

**https://credits-ocr-dashboard.herokuapp.com/**

**Merci**

# Implémentez un modèle de scoring

MODELES

## Sélection du seuil