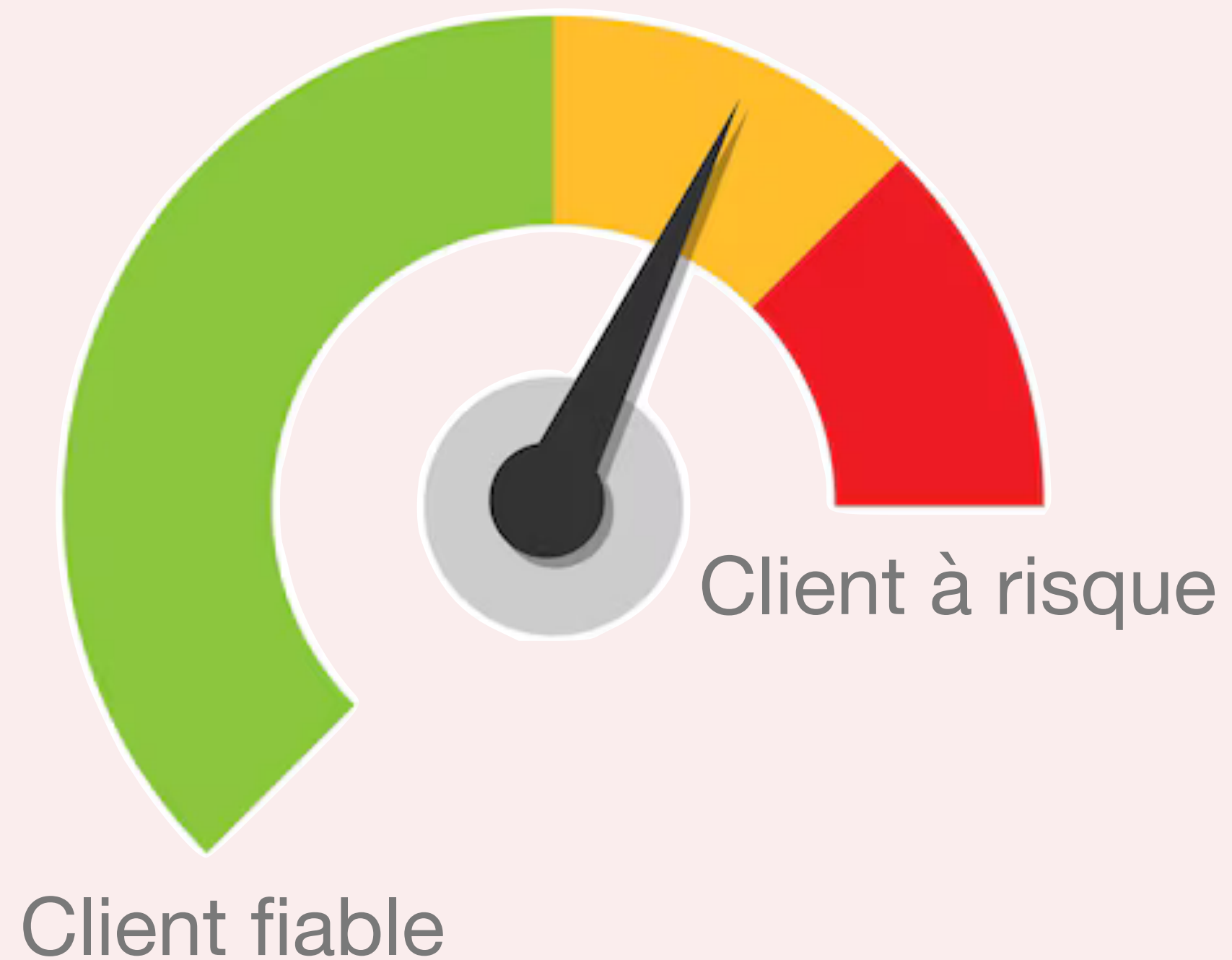


Prêt **à**
dépendre



■ Contexte et Objectif



Gestion du risque

Différencier les bons et moins bons clients en terme de **capacité de remboursement.**

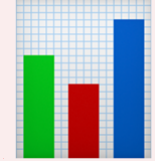


■ Contexte et Objectif



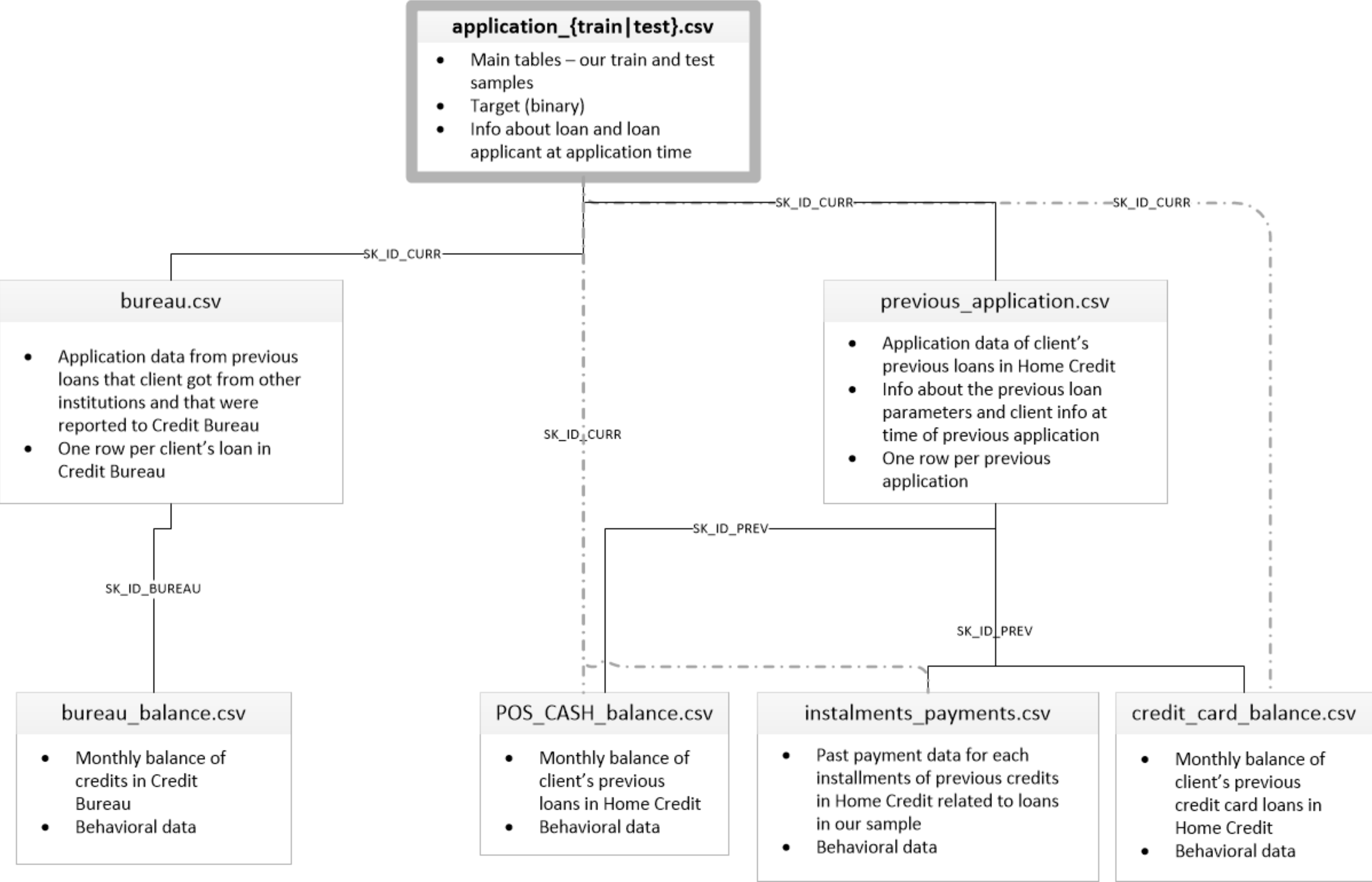
Établir une **classification** des clients basée sur leurs **caractéristiques démographiques**, leurs **comportements financiers** ainsi que sur leur **historique de crédit**.

Mise à disposition d'un **dashboard interactif** pour évaluer rapidement le **risque de défaut** des clients.

■ Plan

-  Review des données (analyse + préparation)
-  Modélisation : réduction du risque métier (metrics, scores)
-  Disponibilité du data product : API, Dashboard (MLOps)

■ Structure des données



■ Aperçu des données | feature engineering

CREDIT_INCOME_PERCENT	float64	rapport entre montant du crédit et les revenus du client
ANNUITY_INCOME_PERCENT	float64	rapport entre l'annuité du prêt et les revenus du client
CREDIT_TERM	float64	durée du paiement en mois
DAYS_EMPLOYED_PERCENT	float64	rapport entre l'expérience professionnelle et l'âge du client

307 511 observations

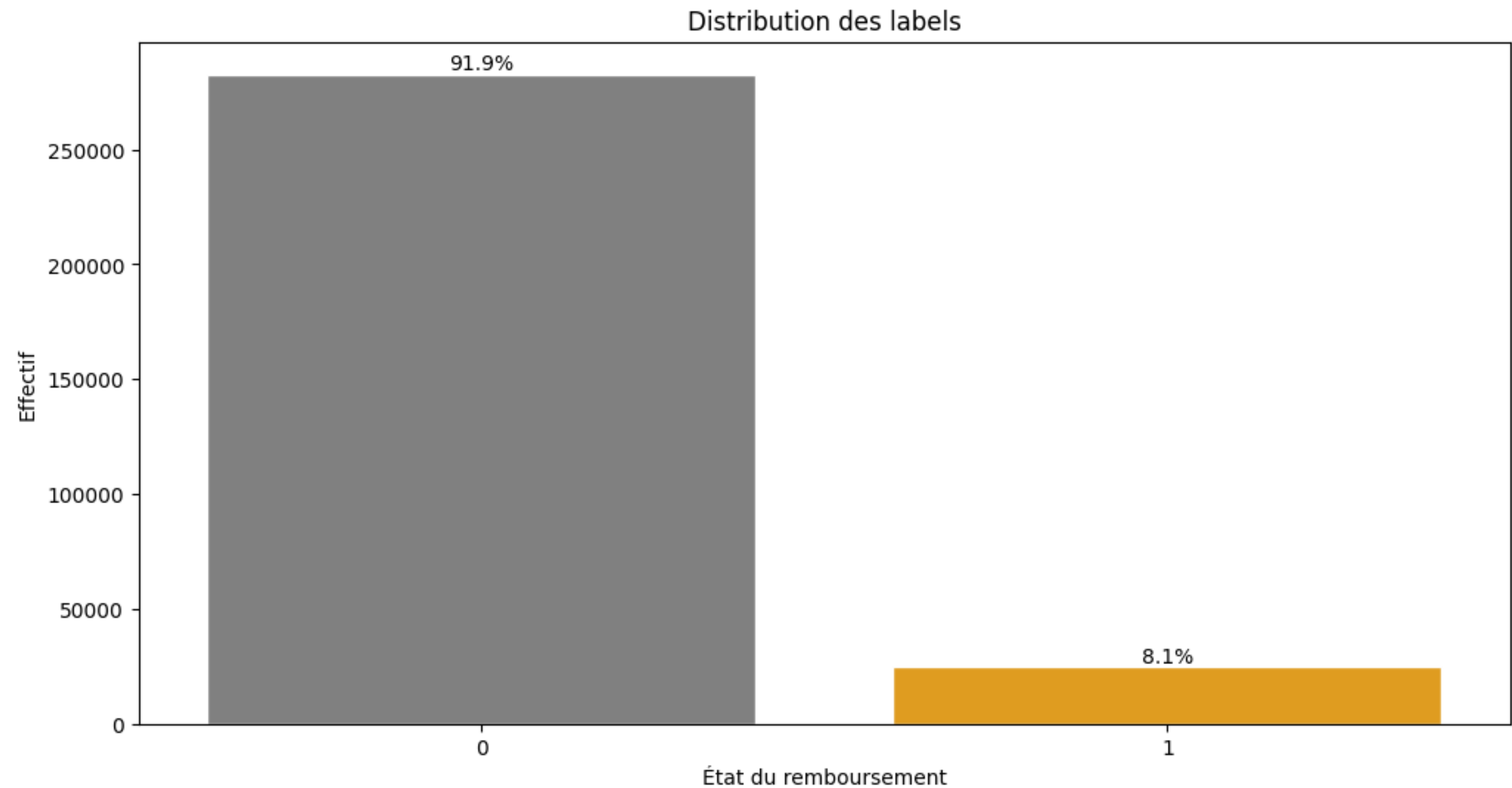
121 variables

105 quantitatives

16 qualitatives

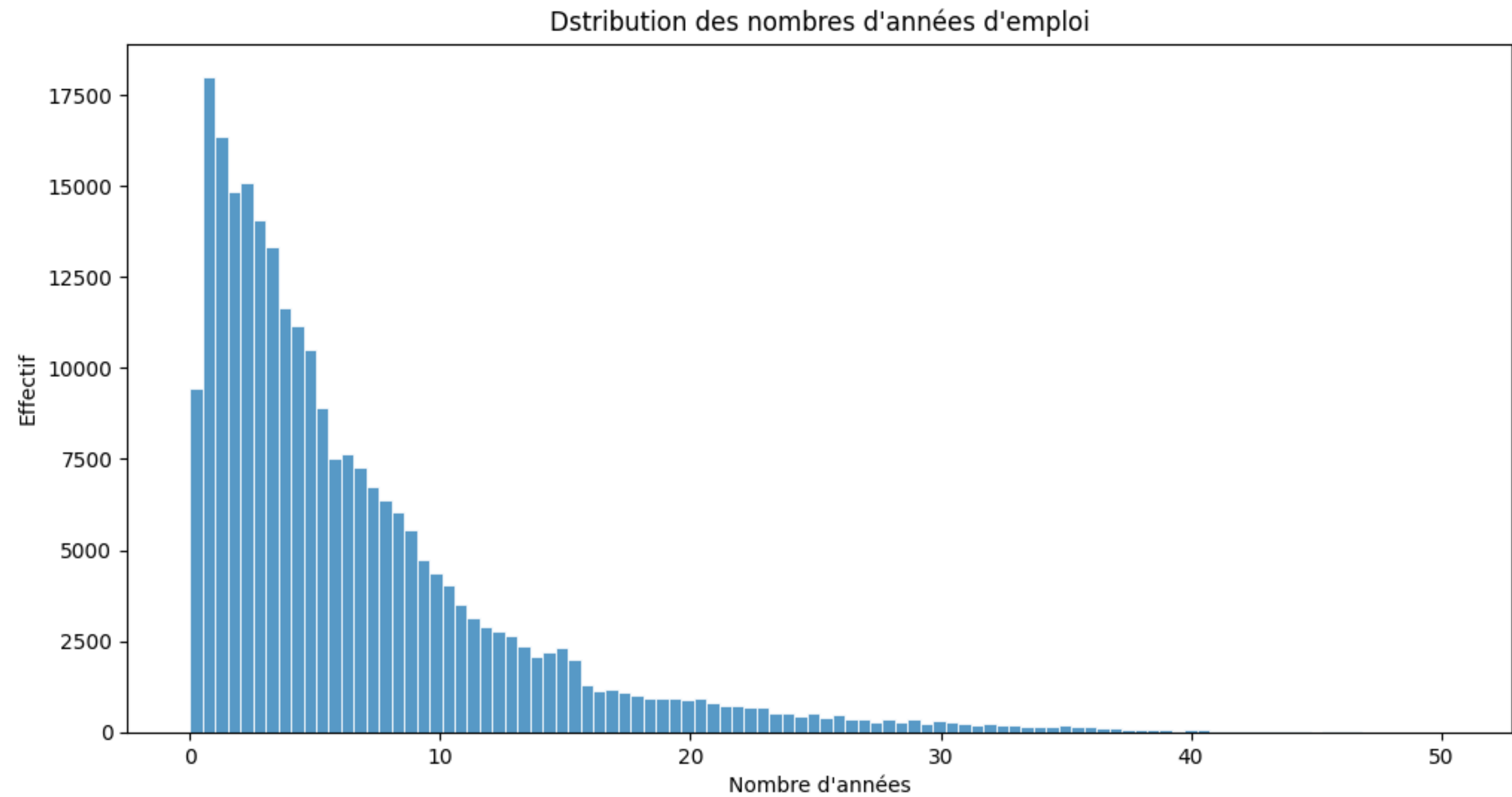
11 % nan

■ Exploration | labels



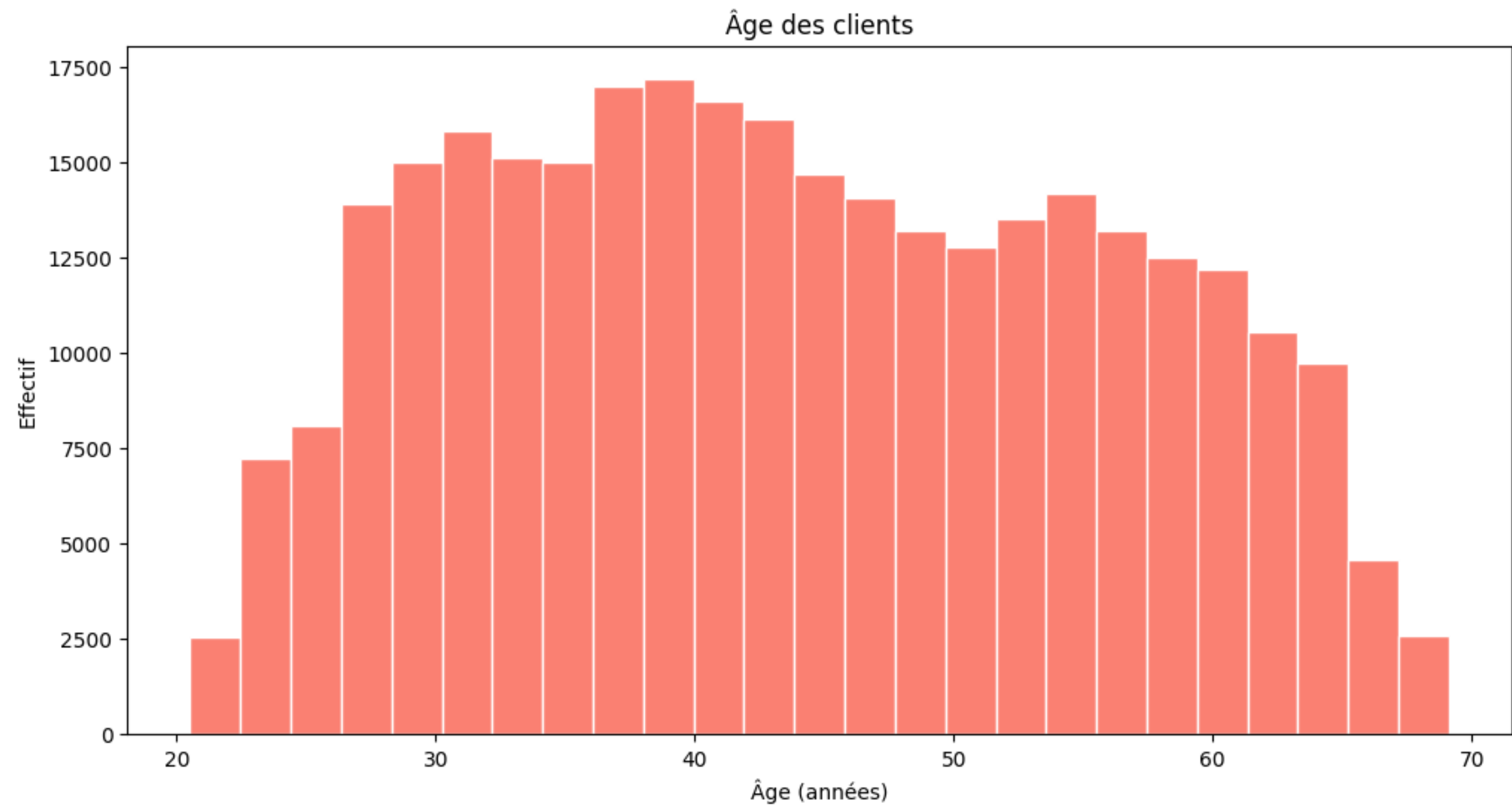
Déséquilibre évident entre les classes **positive** et **négative**.

■ Exploration | expérience professionnelle



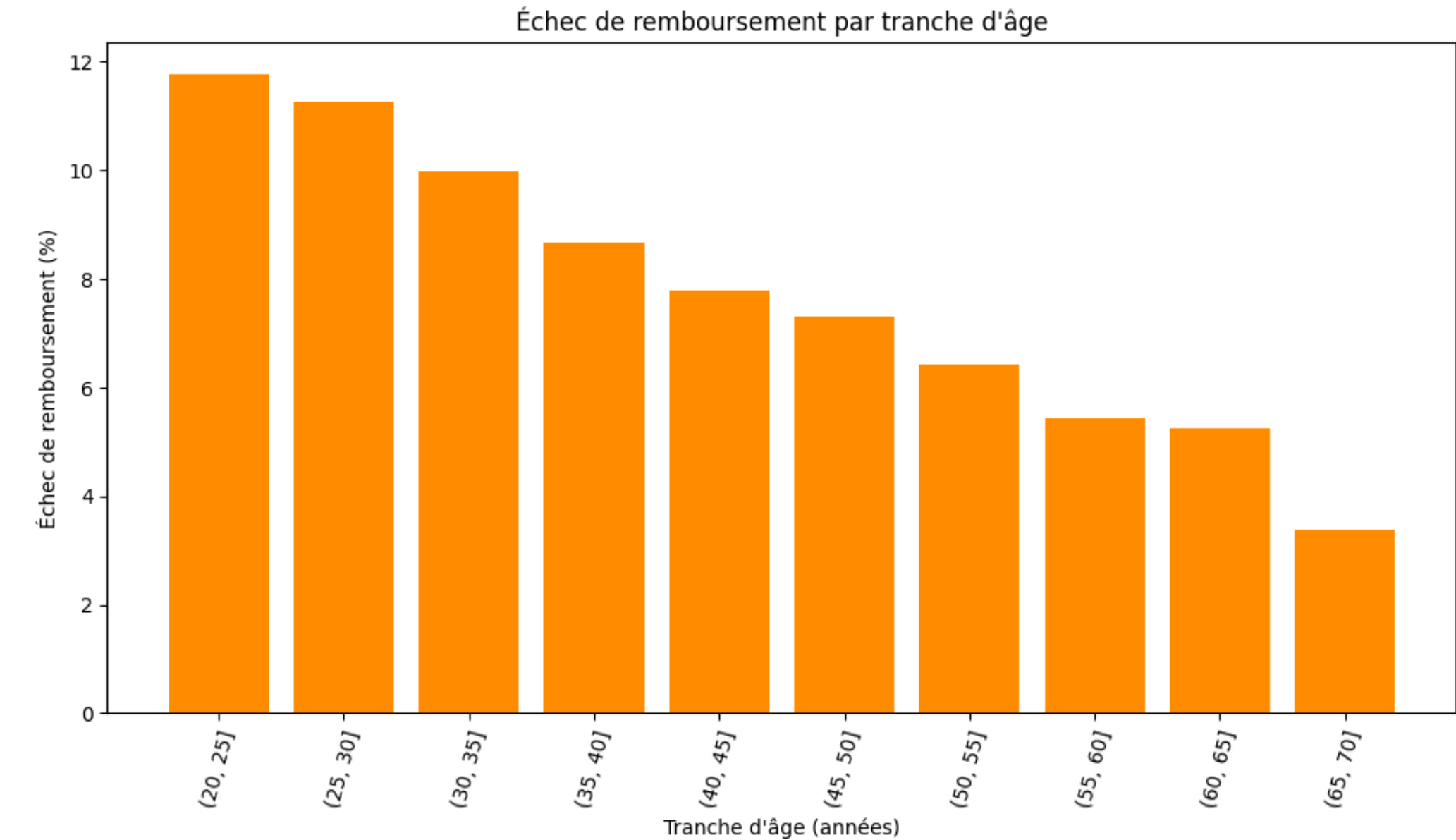
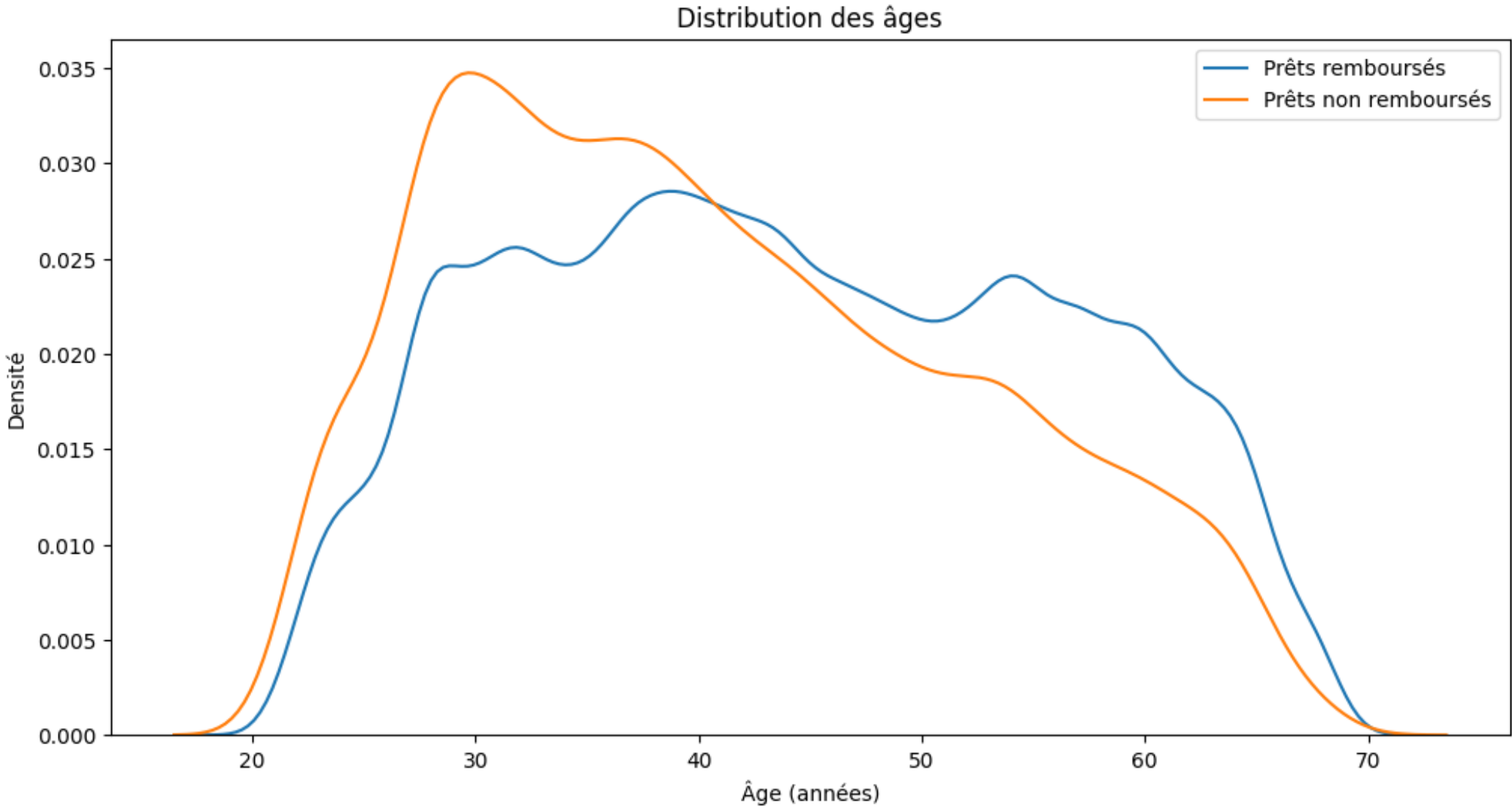
Une large majorité des clients ont **moins de 10 ans** d'expérience professionnelle.

■ Exploration | âge



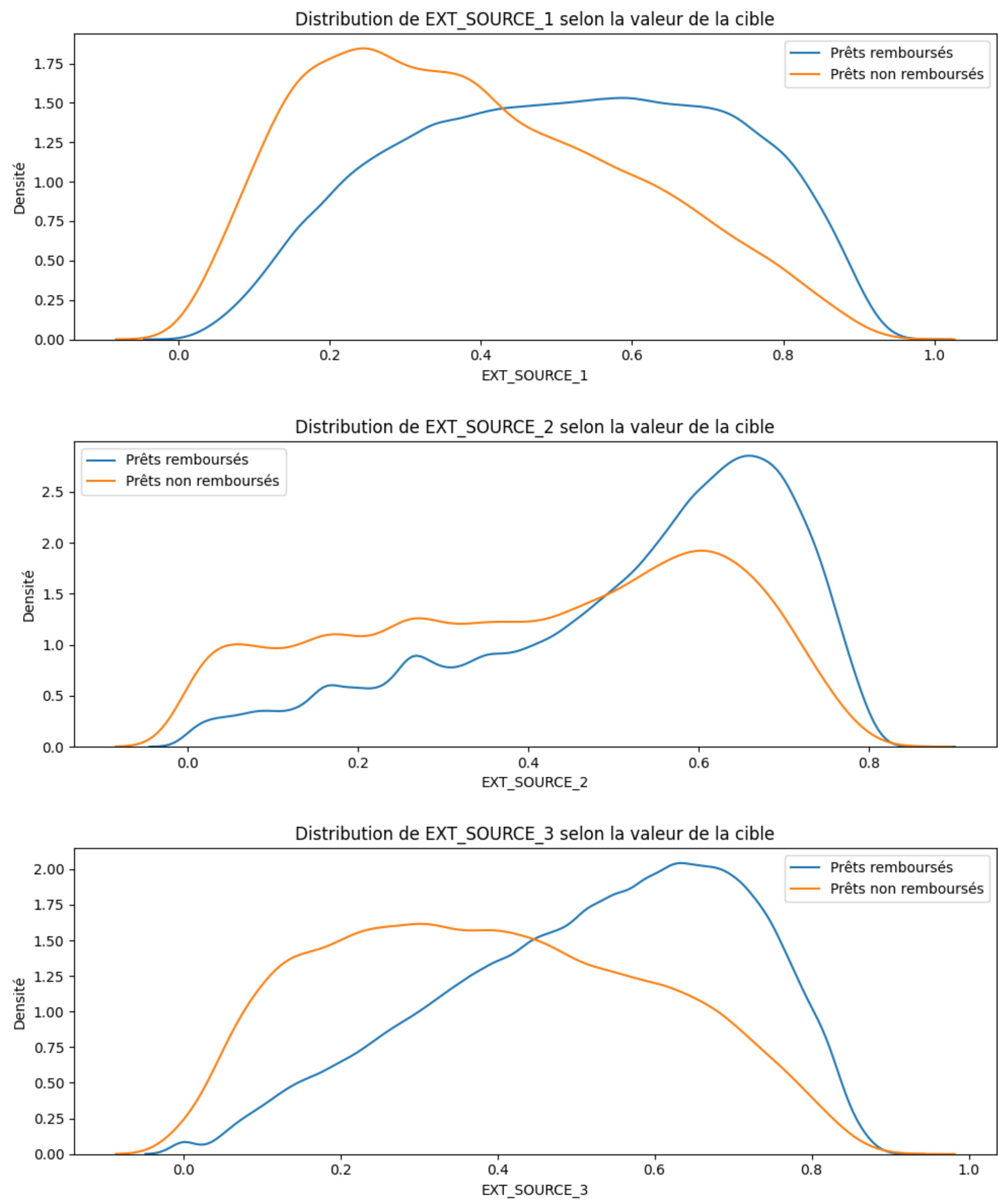
L'âge seul ne permet pas de tirer des conclusions sur les prêts.

Exploration | âge et remboursement



Les jeunes emprunteurs (**20-30 ans**) ont un taux d'**échec de remboursement** plus élevé, tandis que ce taux **diminue progressivement** avec l'âge, particulièrement **après 50 ans**.

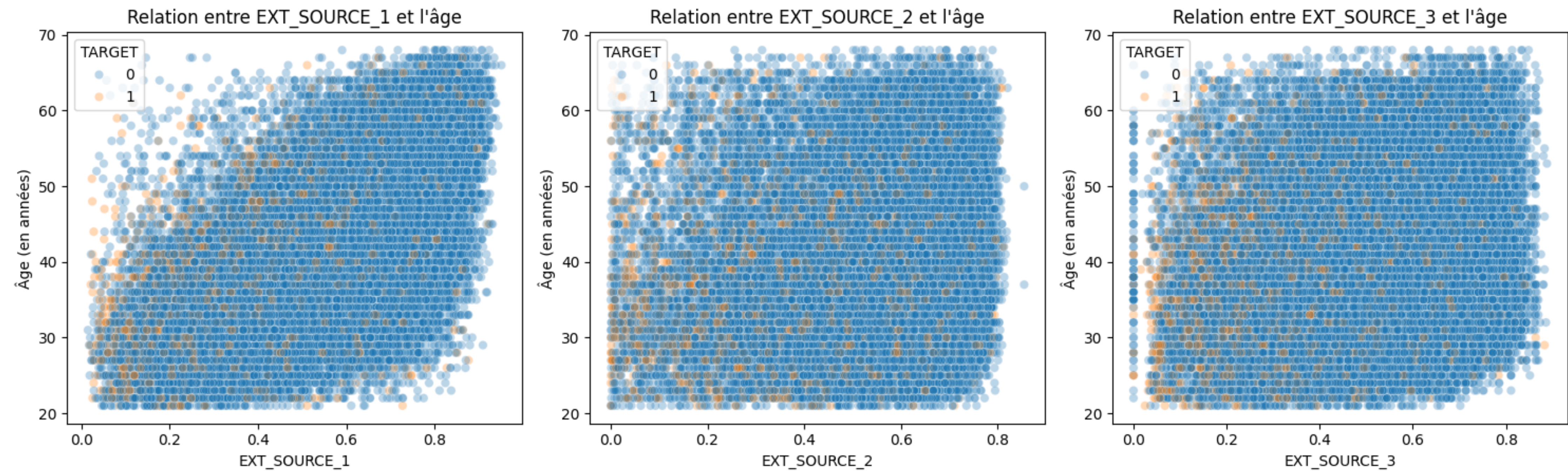
■ Exploration | sources externes et remboursement



Les emprunteurs ayant des scores plus faibles sur les variables **EXT_SOURCE_1**, **EXT_SOURCE_2**, et **EXT_SOURCE_3** ont une probabilité plus élevée de ne pas rembourser leurs prêts, tandis que les emprunteurs ayant des scores plus élevés tendent à mieux rembourser.

■ Exploration | sources externes et âge

Relation entre les sources externes et l'âge des clients



Existence d'un relation linéaire modérée entre l'âge et la première source externe.

■ Modélisation | entraînement des modèles (GridSearch)

Modèle linéaire

Régression Logistique

```
'C': [0.001, 0.01, 0.1, 1]
```

Bagging

Random Forest Classifier

```
'n_estimators': [50, 100, 200]  
'max_depth': [5, 10]  
'min_samples_split': [2, 5]
```

Boosting

XGBoost Classifier

```
'n_estimators': [50, 100, 200]  
'max_depth': [3, 5, 7]  
'learning_rate': [0.01, 0.1, 0.2]  
'subsample': [0.8, 1.0]
```

■ Modélisation | tracking MLFlow

Parameters (1)

Q Search parameters

Parameter	Value
C	1

Régression Logistique

Metrics (3)

Q Search metrics

Metric	Value
roc_auc_score	0.7491800957952321
lowest_cost	33066
best_threshold	0.54

Parameters (3)

Q Search parameters

Parameter	Value
min_samples_split	5
max_depth	10
n_estimators	200

Random Forest Classifier

Metrics (3)

Q Search metrics

Metric	Value
roc_auc_score	0.7352799202571751
lowest_cost	34601
best_threshold	0.49

Parameters (4)

Q Search parameters

Parameter	Value
subsample	1.0
learning_rate	0.2
max_depth	3
n_estimators	200

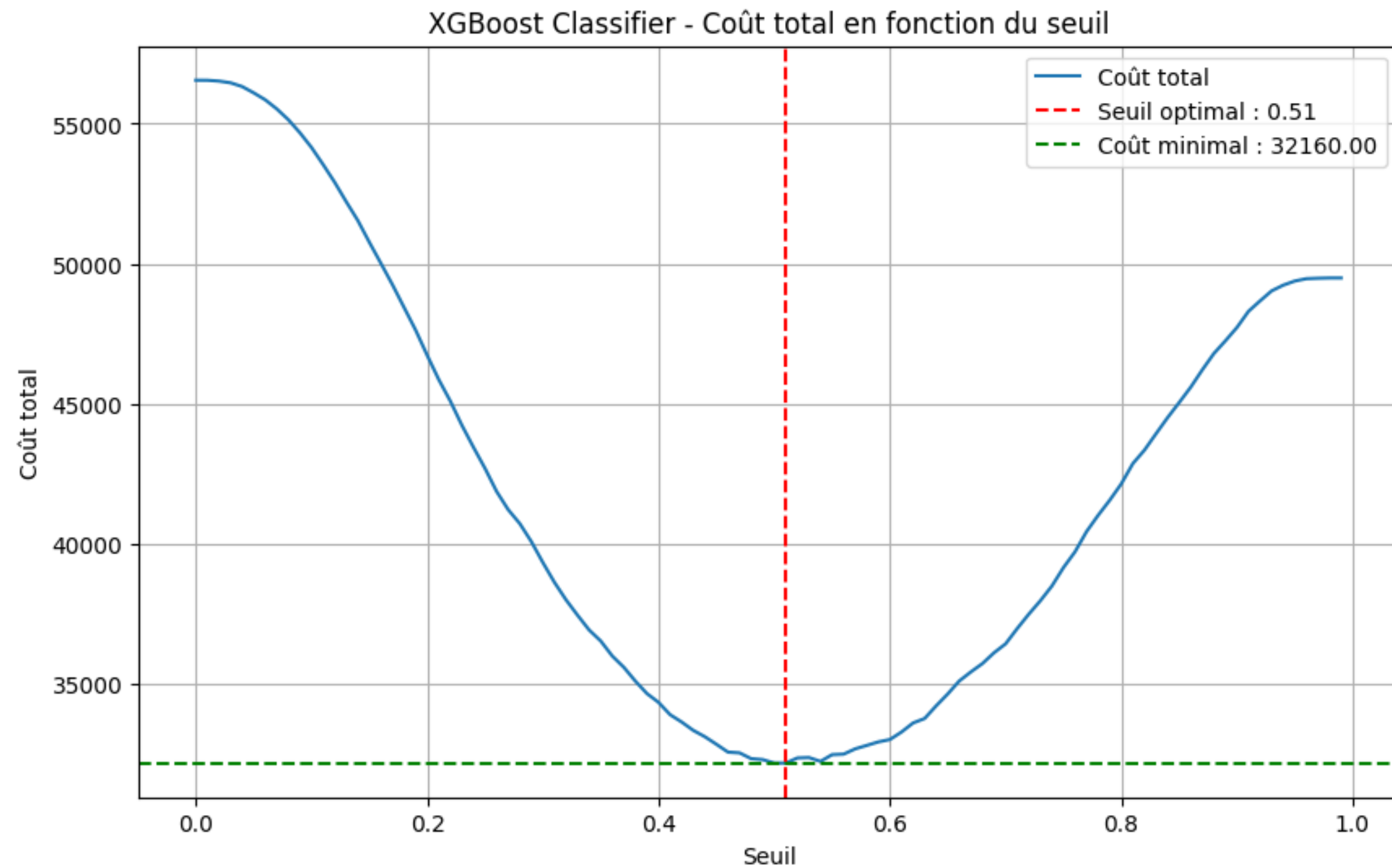
XGBoost Classifier

Metrics (3)

Q Search metrics

Metric	Value
roc_auc_score	0.7643944290038979
lowest_cost	32160
best_threshold	0.51

■ Modélisation | seuil optimal du meilleur modèle



Fonction de coût

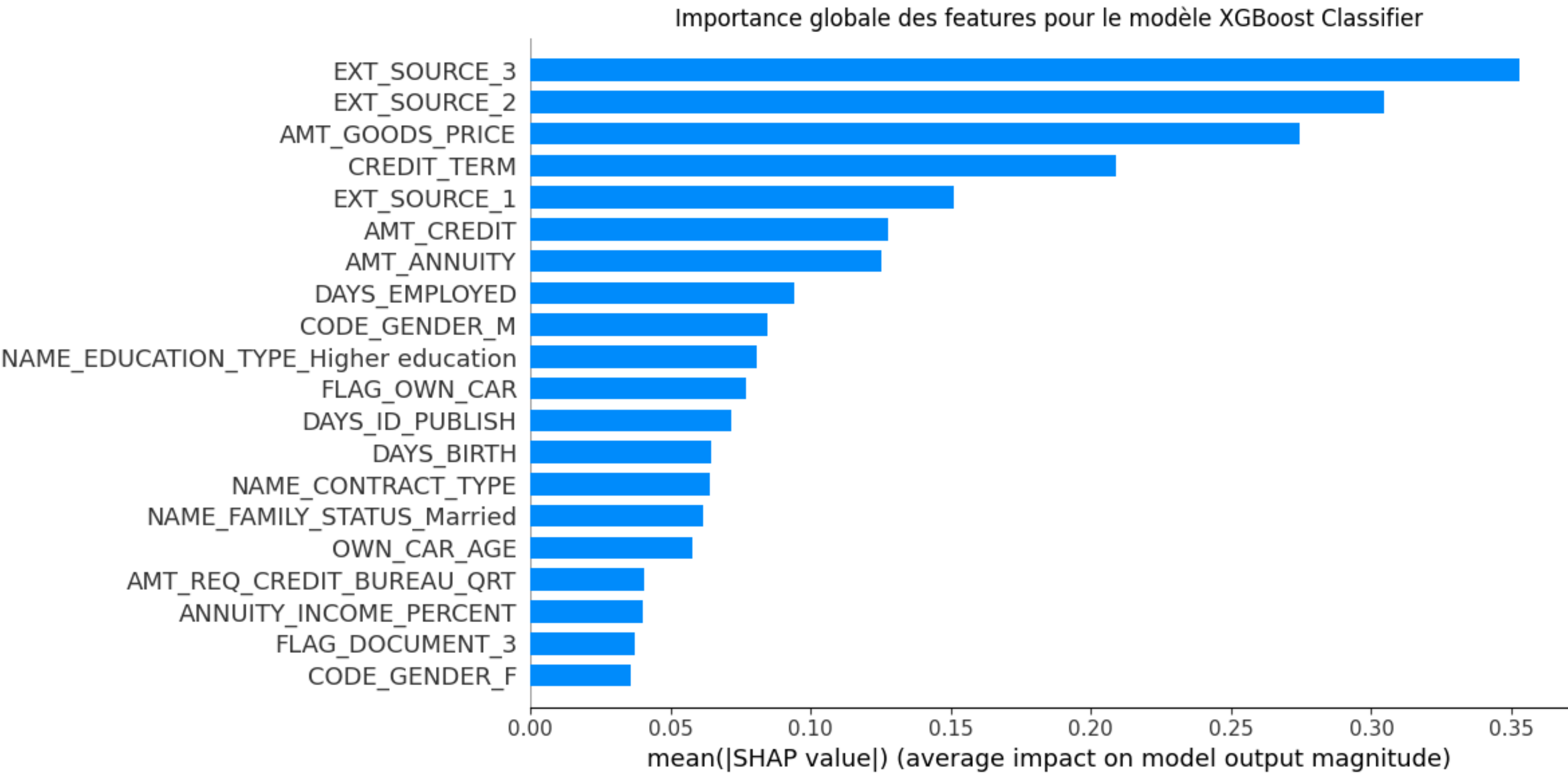
$$C = 1 * FP + 10 * FN$$

AUC : **0.76**

Seuil optimal : **0.51**

Coût minimal : **32160**

■ Modélisation | feature importances



Modélisation | data drift

Dataset Drift

Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5




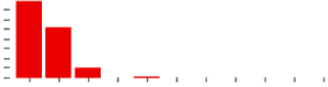

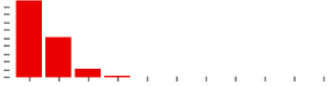



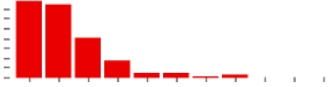
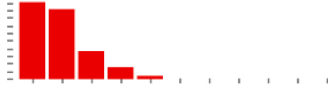
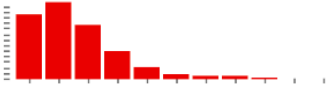

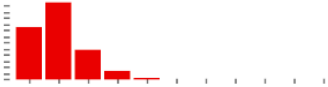




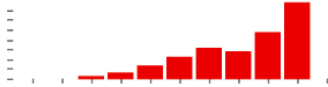
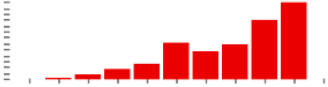


243
Columns

11
Drifted Columns

0.0453
Share of Drifted Columns

Data Drift Summary

Drift is detected for 4.527% of columns (11 out of 243).

<div><div></div><div>Search</div><div>×</div></div>						
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> CREDIT_TERM	num			Detected	Wasserstein distance (normed)	0.575103
> AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.359052
> CREDIT_INCOME_PERCENT	num			Detected	Wasserstein distance (normed)	0.293721
> AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	0.281765
> AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.210785
> AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.207334
> AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.161102
> AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.15426
> NAME_CONTRACT_TYPE	num			Detected	Jensen-Shannon distance	0.14755
> DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.138977
> FLAG_EMAIL	num			Detected	Jensen-Shannon distance	0.122121

■ Data product | API

POST

⌵

http://13.38.185.52:5000/scoring

Send

⌵

Params

Authorization

Headers (9)

Body

Scripts

Tests

Settings

Cookies

none

form-data

x-www-form-urlencoded

raw

binary

GraphQL

JSON

⌵

Beautify

1

{

2

|

"sk_id_curr": 100002

3

}

Body

Cookies

Headers (5)

Test Results

200 OK

• 39 ms • 12.19 KB •

🌐

|

📄 e.g. Save Response

⋮

Pretty

Raw

Preview

Visualize

JSON

⌵

↺↻

📄

🔍

1

{

2

>

"feature_importances_negative": [...

211

|

],

212

>

"feature_importances_positive": [...

633

|

],

634

|

"probability": 0.90053790807724,

635

|

"sk_id_curr": 100002,

636

|

"threshold": 0.51

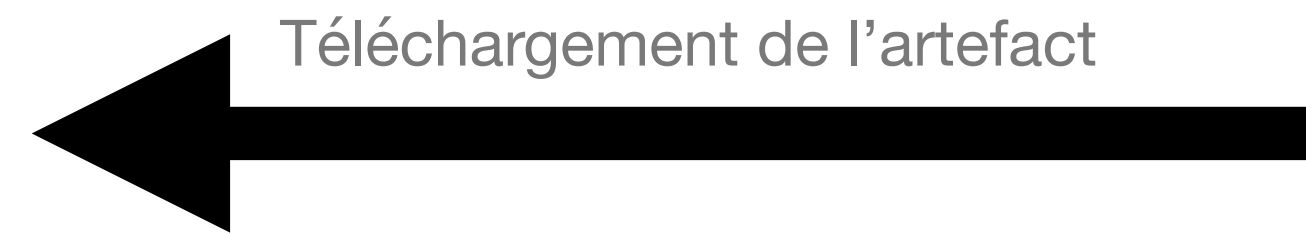
637

}

■ Data product | API



Instance API en ligne



Instance MLFlow en ligne

```
model_uri = 'runs:/8b9687c798834eb5b9e070154b21cf44/xgboost_classifier_best_model'
```

■ Data product | dashboard interactif

Prédiction du **score** du client via son numéro d'**identifiant unique**.

Informations sur les **caractéristiques les plus influentes** négativement ou positivement lors de la prédiction.



Évaluation du risque de crédit

Saisir le numéro de client

100002

Vérifier le risque

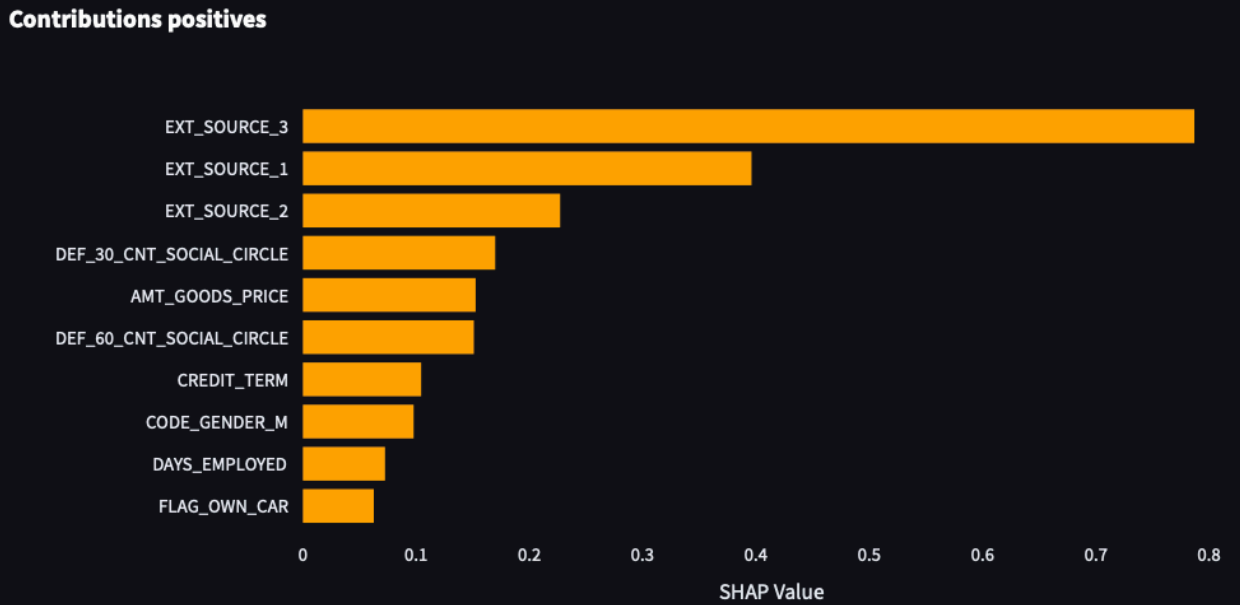
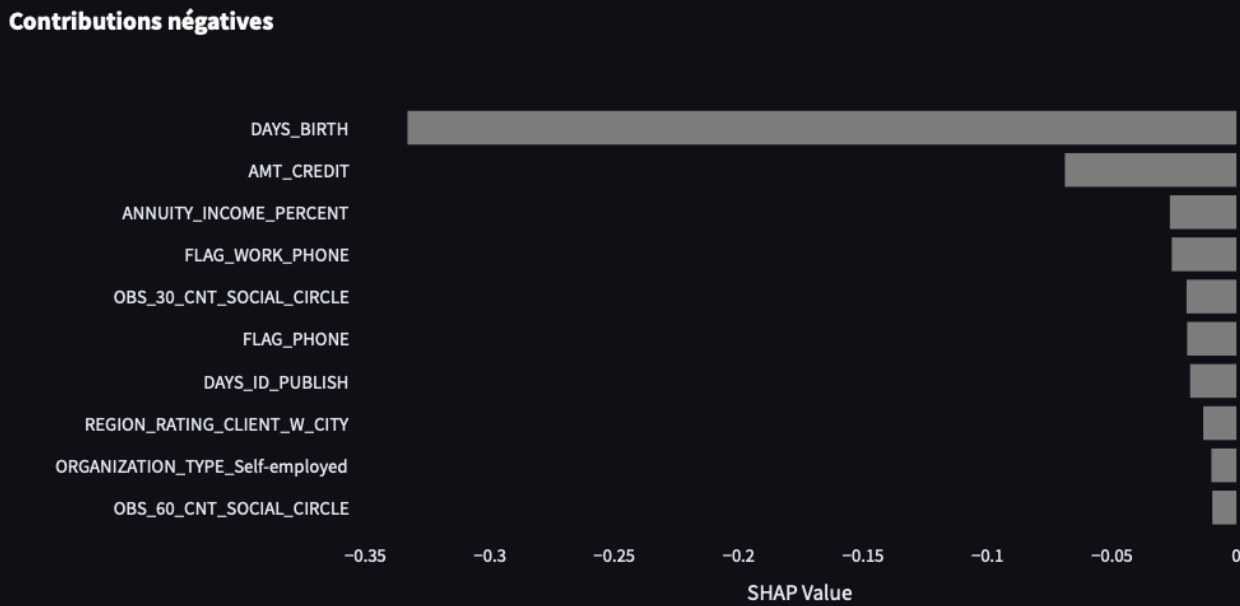
Résultat de l'évaluation pour le client 100002

Seuil de décision : 51.00

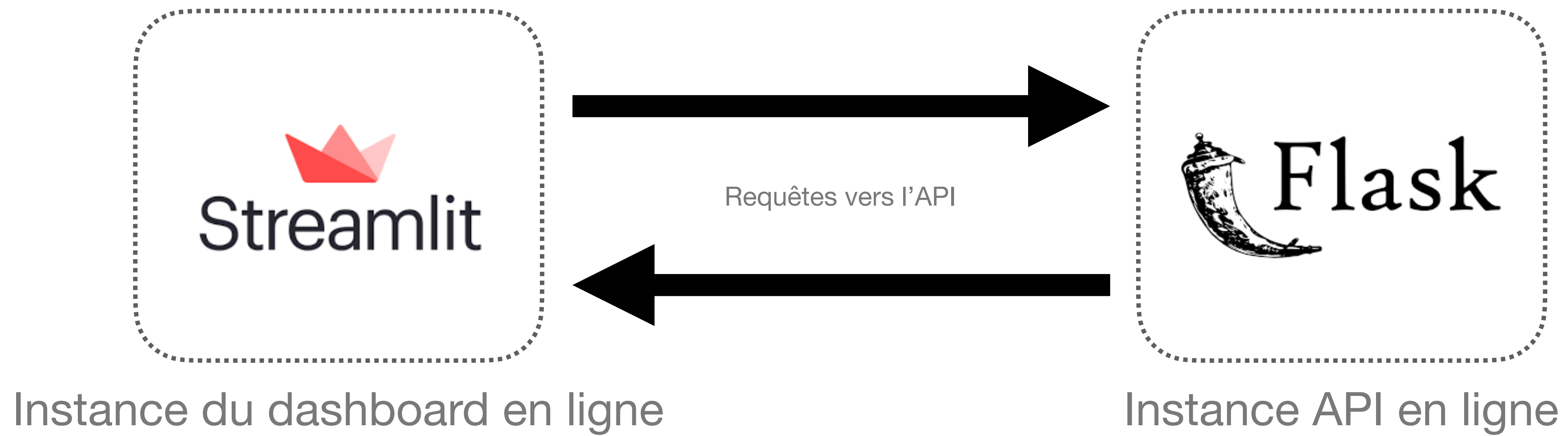
Score du client : 90.05

Client à risque

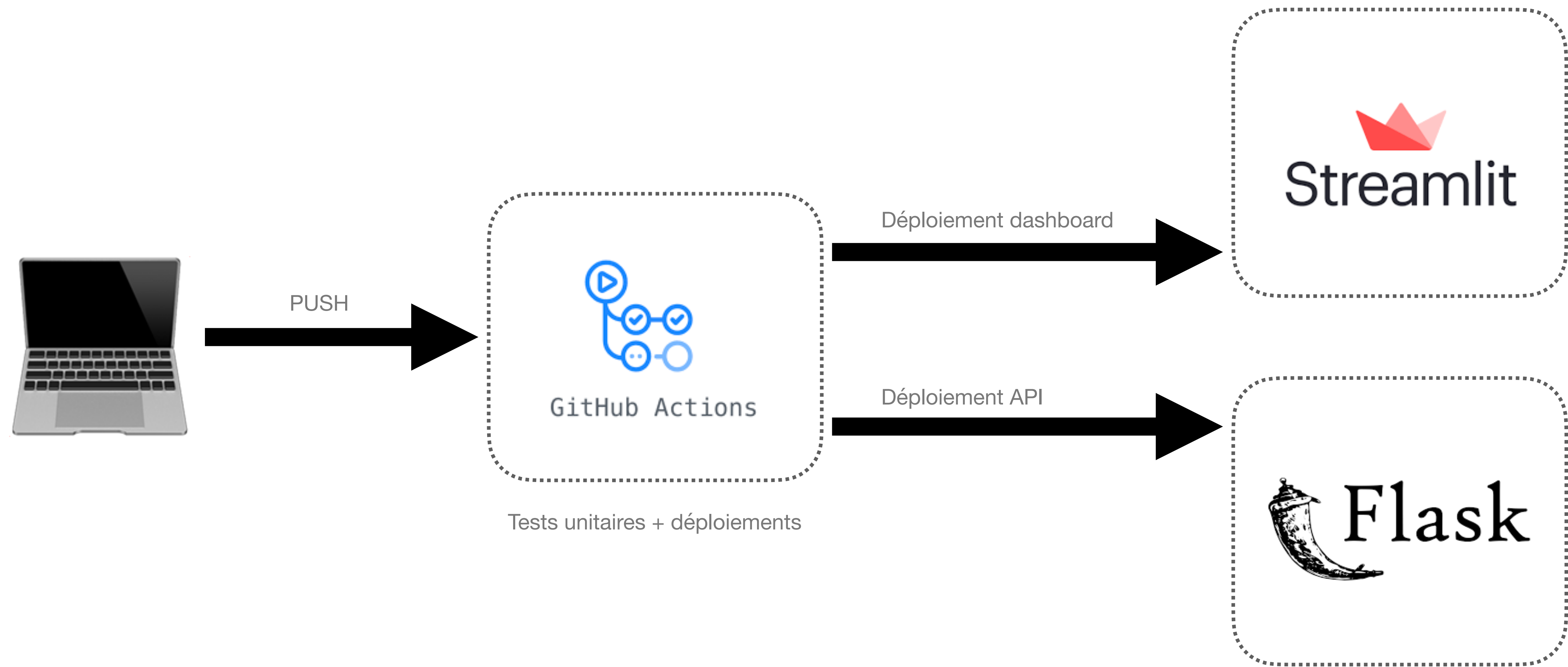
Contribution des features



■ Data product | dashboard interactif



■ Data product | CI - CD



Merci pour votre écoute