Data scientist
Projet 7

Sommaire

I. Contexte & Problématique

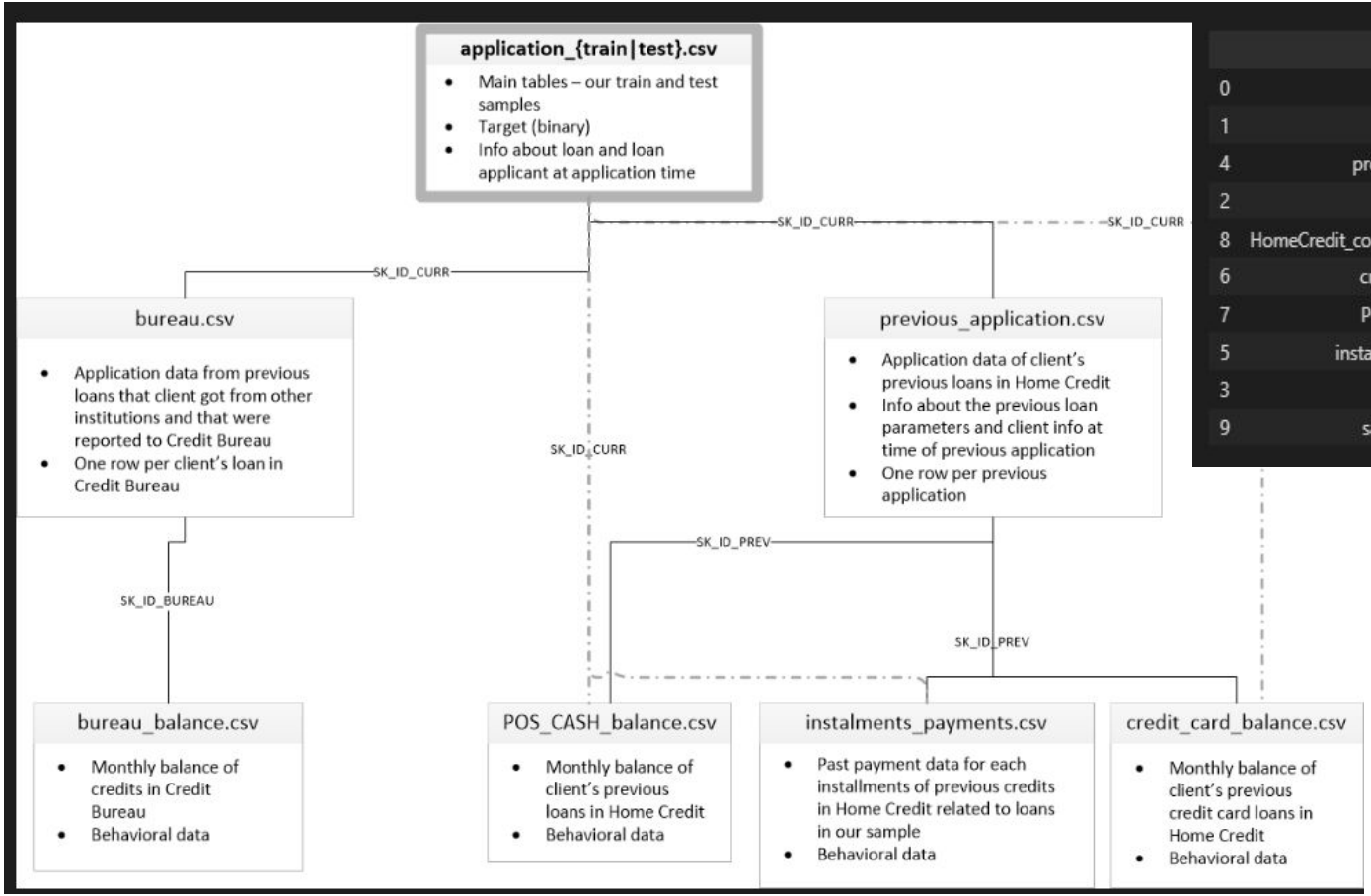
– 1. Titre & contexte

– 2 – Contexte métier

– 3 – Problématique data

II. Données & Enjeux Data

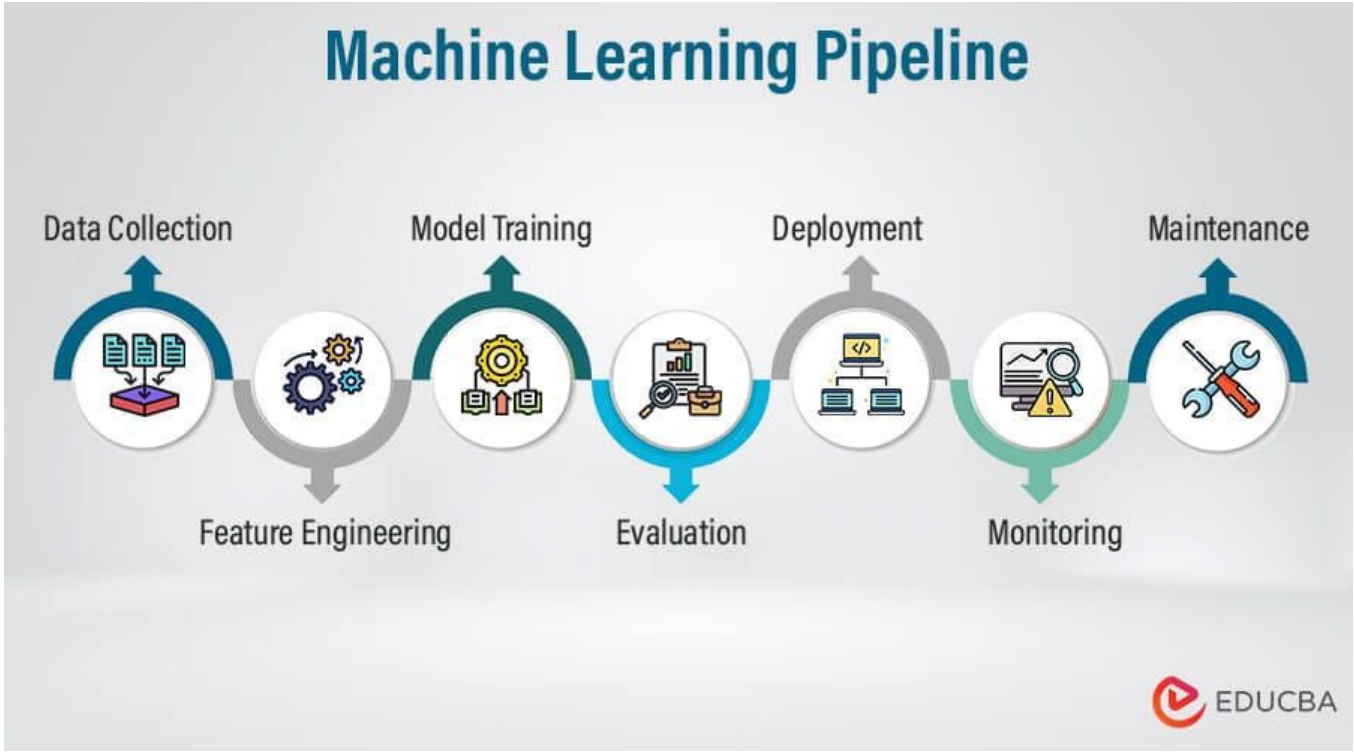
4 – Présentation du jeu de données



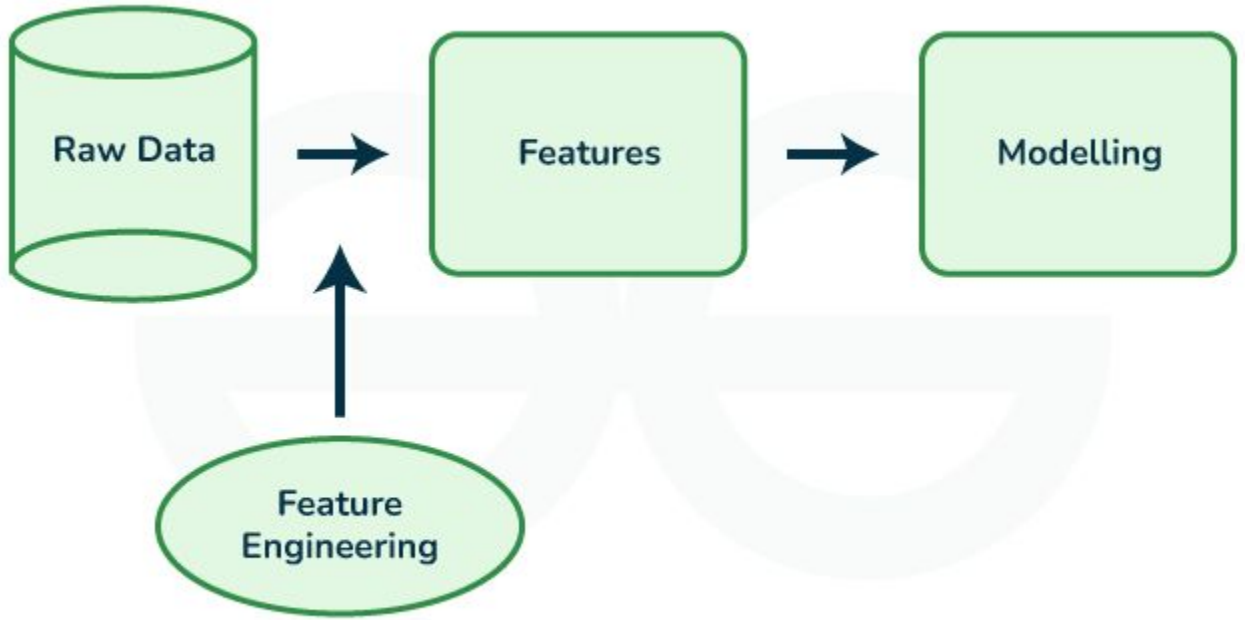
	Dataset	Lignes	Colonnes	%NA moyen
0	application_train	307511	122	24.40
1	application_test	48744	121	23.81
4	previous_application	1670214	37	17.98
2	bureau	1716428	17	13.50
8	HomeCredit_columns_description	219	5	12.15
6	credit_card_balance	3840312	23	6.65
7	POS_CASH_balance	10001358	8	0.07
5	installments_payments	13605401	8	0.01
3	bureau_balance	27299925	3	0.00
9	sample_submission	48744	2	0.00

5 – Déséquilibre des classes

6 – Pipeline global du projet

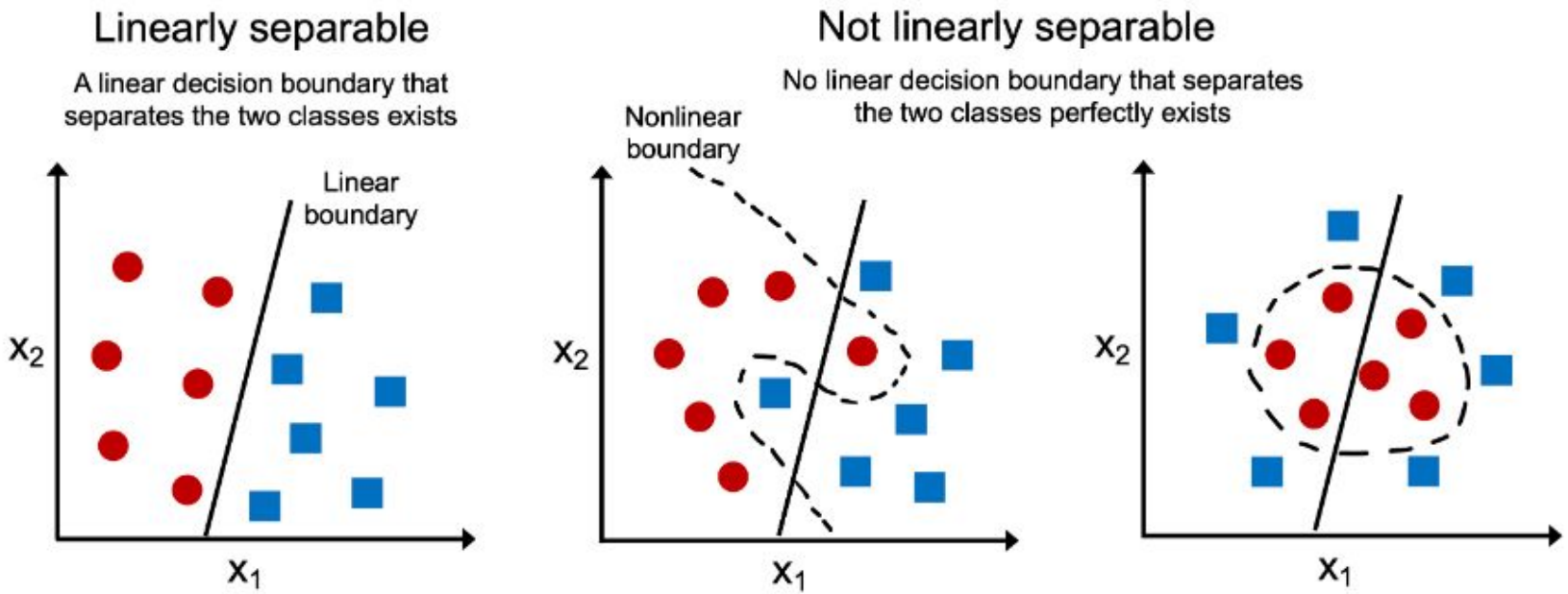


7 – Préprocessing & feature engineering

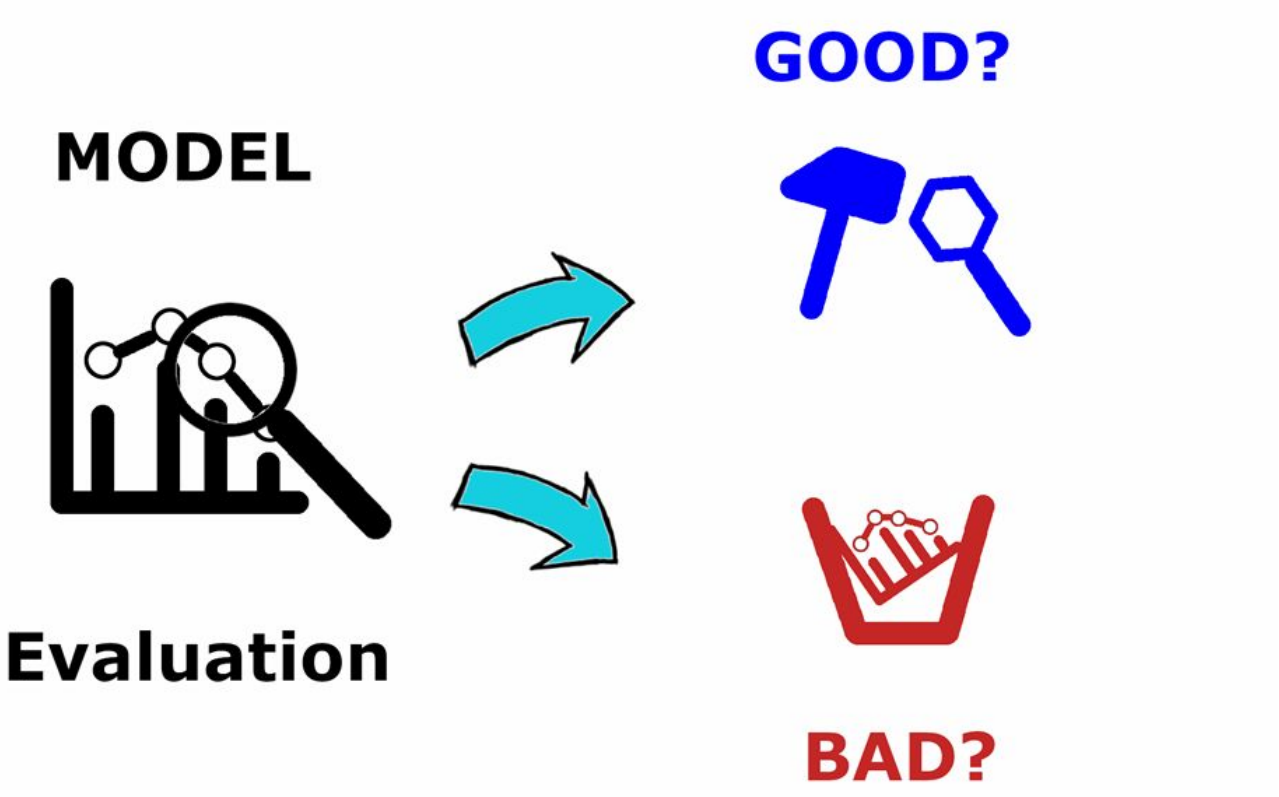


III. Modélisation & Performance

8 – Démarche de modélisation



9 – Choix des métriques



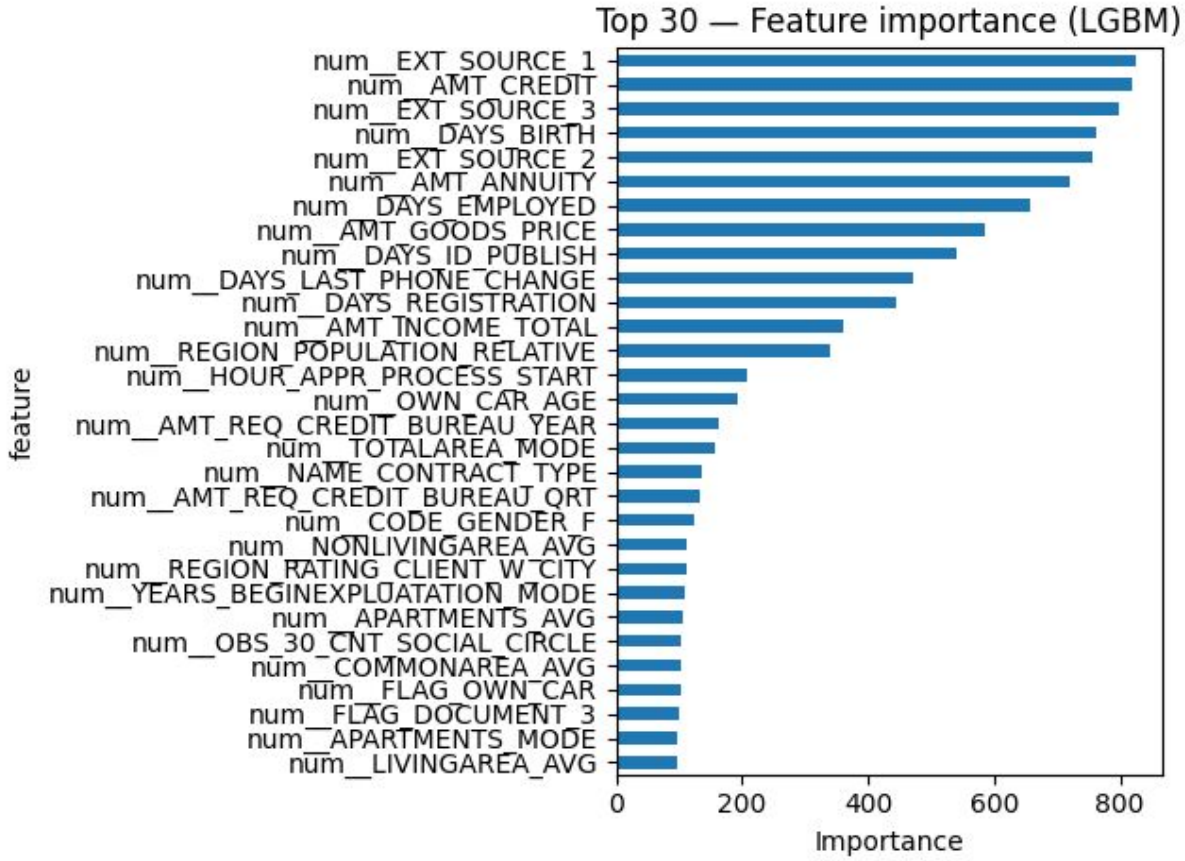
10 – Tracking des expériences avec MLflow

Résumé des performances :

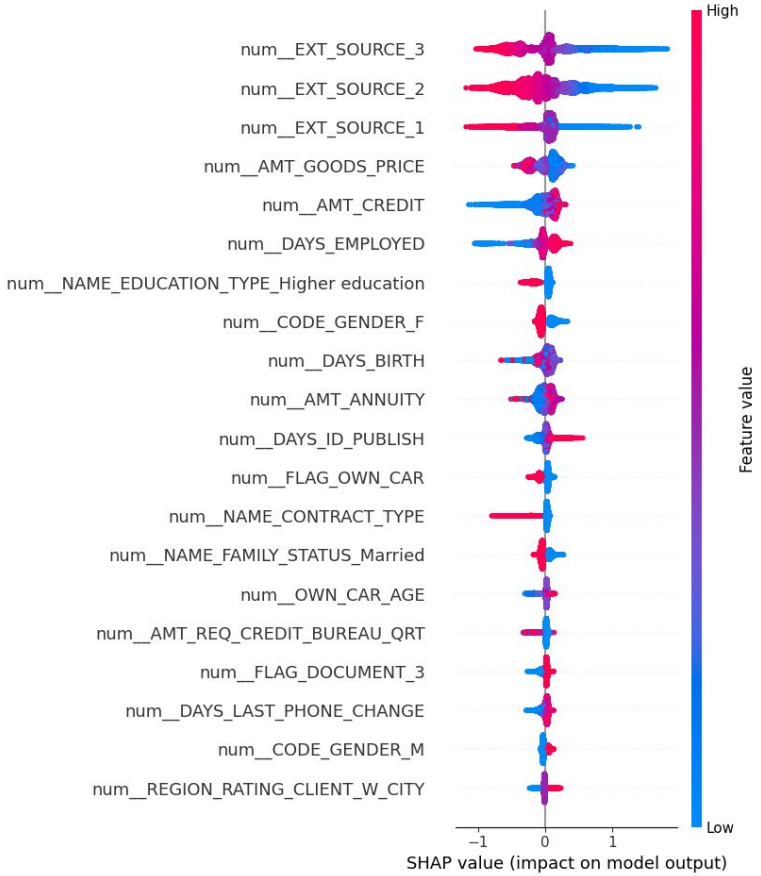
	auc	f1	business_cost	best_threshold	cm
log_reg	0.748872	0.260041	33250	0.534646	[[41828, 14710], [1854, 3111]]
xgb	0.760156	0.275017	32279	0.504949	[[41029, 15509], [1677, 3288]]
lgbm_model	0.761191	0.273358	32157	0.514848	[[41331, 15207], [1695, 3270]]
gb_model	0.757187	0.033991	32520	0.089192	[[42148, 14390], [1813, 3152]]

11 – Sélection du meilleur modèle

12 – Feature importance globale



13 – Feature importance locale (SHAP)



IV. MLOps, Déploiement & Monitoring

14 – Déploiement & MLOps

15 – Analyse du data drift

📊 Drift Report — Equivalent Evidently

Ce rapport couvre : **Feature Drift** (KS), **Prediction Drift** (KS sur scores), et **Target Drift** (KS sur labels). Aligné sur les features réellement utilisées par le modèle (post-preprocess).

Features testées	Features en drift	Alpha
239	**0** (0.0%)	**0.05**

📈 Prediction Drift (score distribution)

- **p-value** : 0.382463
- **KS statistic** : 0.004088
- **Drift** : ❌ (alpha=0.05)
- **Reason** : ks_test
- **Mean (train/test)** : 0.3990 / 0.3984
- **Std (train/test)** : 0.2073 / 0.2064

🎯 Target Drift (label distribution)

- **p-value** : 1.000000
- **KS statistic** : 0.000001
- **Drift** : ❌ (alpha=0.05)
- **Reason** : ks_test
- **Rate (train/test)** : 0.0807 / 0.0807
- **N (train/test)** : 246008 / 61503

16 – Démo scoring via API

💳 Home Credit — Simulation de décision de crédit

Cette application illustre un **système de scoring crédit** basé sur :

- un **modèle de machine learning**
- une **API FastAPI**
- une **règle métier indépendante** (seuil = 0,65)

Le modèle prédit une **probabilité de défaut**, la décision finale est ensuite appliquée.

Choisissez un client de démonstration :

Client_2 — Profil intermédiaire

Données envoyées à l'API


```
{
  "AMT_ANNUITY" : 550
  "AMT_CREDIT" : 12000
  "AMT_GOODS_PRICE" : 12000
  "AMT_INCOME_TOTAL" : 21600
  "AMT_REQ_CREDIT_BUREAU_QRT" : 1
  "AMT_REQ_CREDIT_BUREAU_YEAR" : 2
  "CODE_GENDER_F" : 0
  "DAYS_BIRTH" : -16425
  "DAYS_EMPLOYED" : -5475
  "DAYS_ID_PUBLISH" : -3500
  "DAYS_LAST_PHONE_CHANGE" : -900
  "DAYS_REGISTRATION" : -6000
  "EXT_SOURCE_1" : 0.45
  "EXT_SOURCE_2" : 0.5
  "EXT_SOURCE_3" : 0.48
  "HOUR_APPR_PROCESS_START" : 14
  "NAME_CONTRACT_TYPE" : 1
  "OWN_CAR_AGE" : 10
  "REGION_POPULATION_RELATIVE" : 0.025
  "TOTALAREA_MODE" : 0.18
}
```


Calculer la décision

🎯 Résultat du scoring

Probabilité de défaut

0.216

Seuil métier appliqué : 0.65

✔ Décision : **Crédit accordé**

ℹ Le modèle prédit une probabilité. La décision finale est prise via une règle métier indépendante.

V. Limites, Perspectives & Conclusion

17 – Limites du projet

18 – Perspectives d'amélioration

19 – Apports du projet

20 – Conclusion & message clé

Questions - Réponses