



Projet 7

Implémentation d'un modèle de scoring

Moustafa **ZMERLI**

03 Sept. 2022, Paris

Parcours

Data Scientist

Projet

❖ Il s'agit d'une société financière « **Prêt à dépenser** »:

➡ **Problématique:**

Souhaite mettre en œuvre un outil de « scoring crédit » pour calculer la probabilité qu'un client rembourse son crédit d'après diverses données







➡ **Mission:**

- Développer un algorithme de classification binaire:
 - crédit accepté
 - crédit rejeté.
- Développer un dashboard interactif pour assurer une transparence sur les décisions prises.

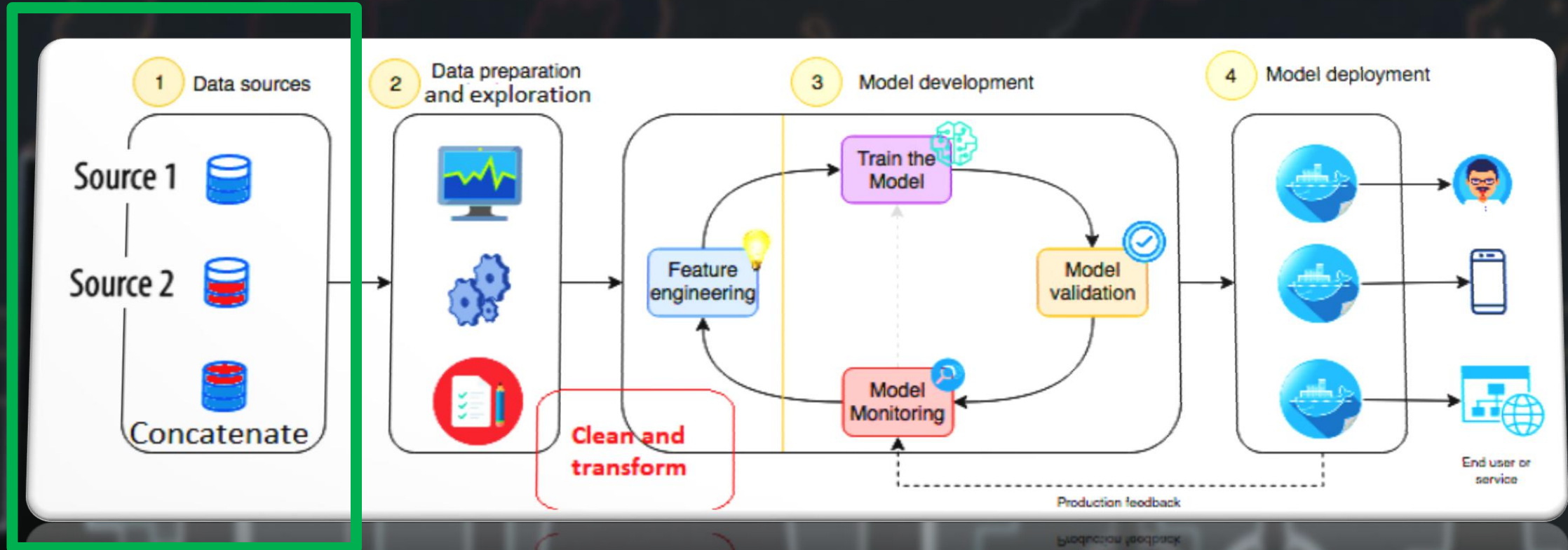
**HOME
CREDIT**

Prêt à dépenser

PLAN

- ☐ Mission/Projet 
- ☐ Jeu de données 
- ☐ Analyse et traitement 
- ☐ Modélisation 
- ☐ Dashboard 
- ☐ Conclusion 



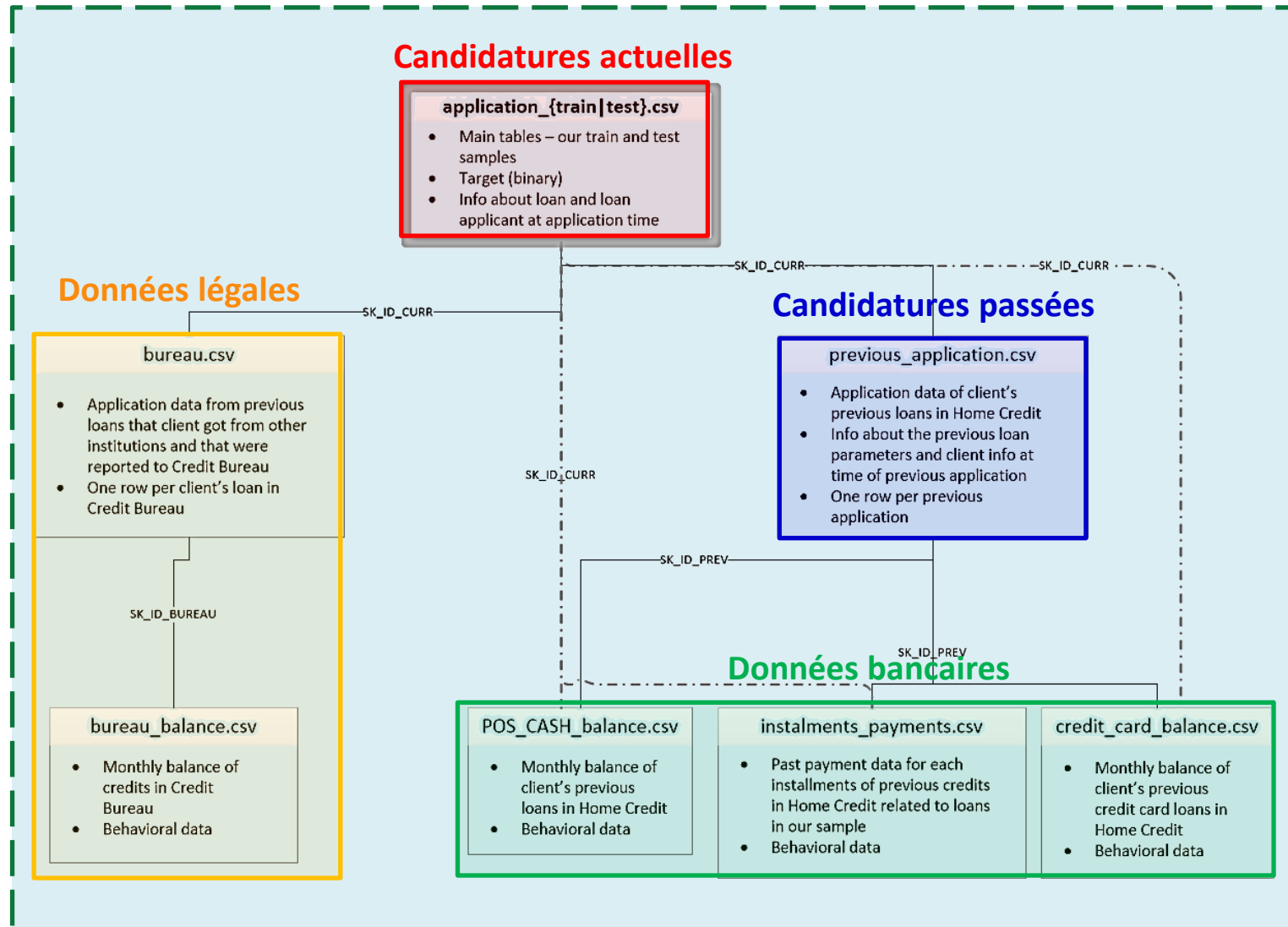


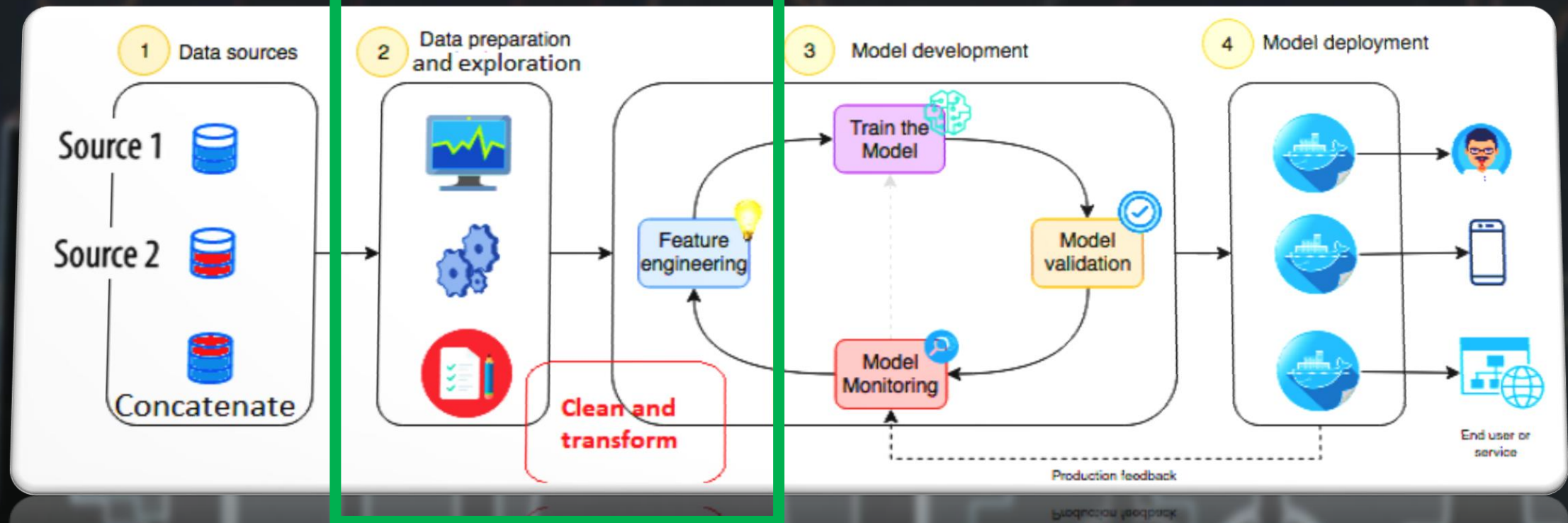
Description des données

7 sources de données

relatives aux clients et à la société:
(précédentes demandes de crédit,
balance de crédit, cash, etc.)

- **Base de données principale :**
- 307 000 clients
- 121 features : âge, sexe, emploi, logement, revenus, informations relatives au crédit, etc.
- Taux de remplissage = 75%
- **Labels cible :** défaut de crédit / pas de défaut de crédit

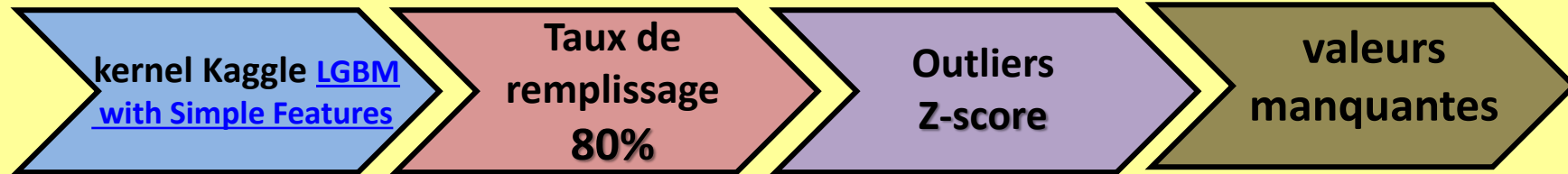




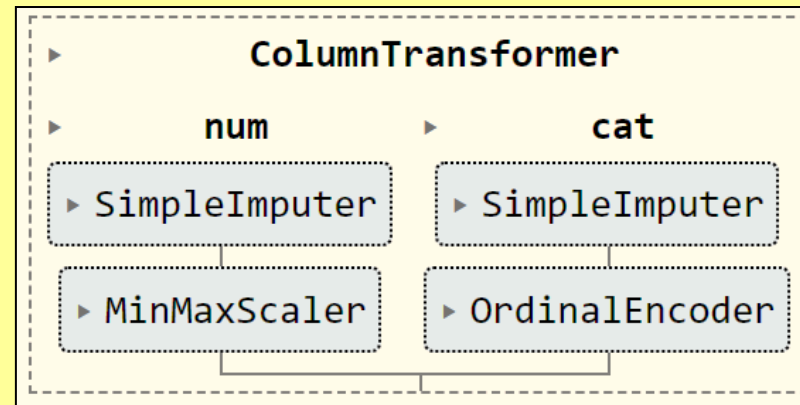
Traitement des données

----- ✂ PRETRAITEMENT ✂ -----

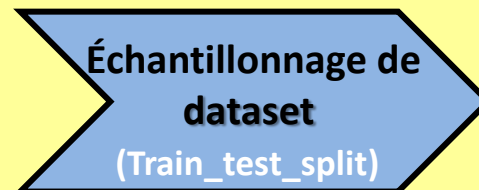
Données



Données numériques
Imputation: Valeur médiane
Standardiser: MinMaxScaler



Données catégorielles
Imputation: Valeur la plus fréquente
Encodage: OrdinalEncoder



80% données d'entraînement

20% données de test

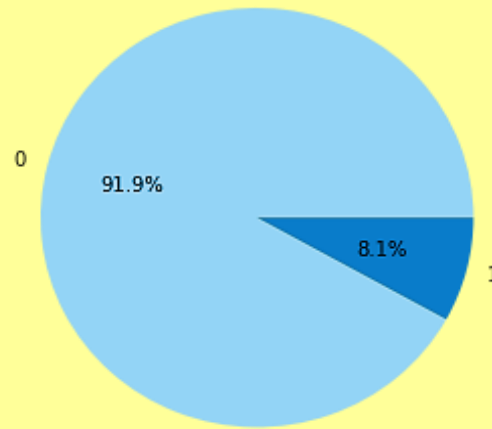
Données déséquilibrées

----- ✂ PRETRAITEMENT ✂ -----

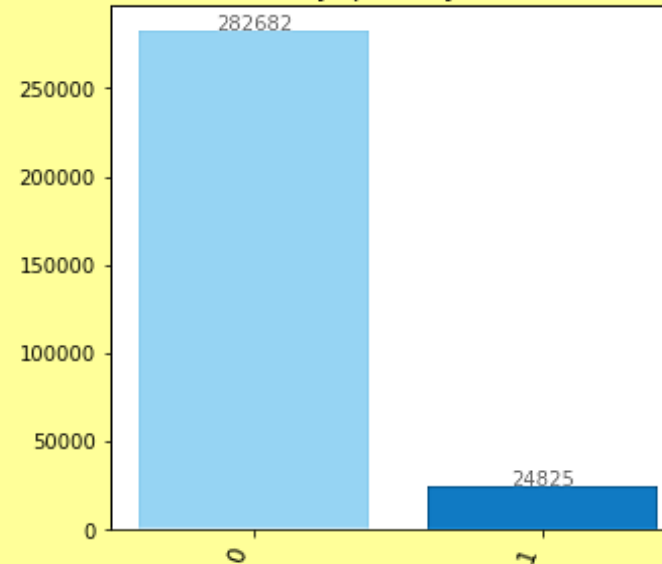
Target distribution

Target distribution

by percentage (%)



by quantity (#)

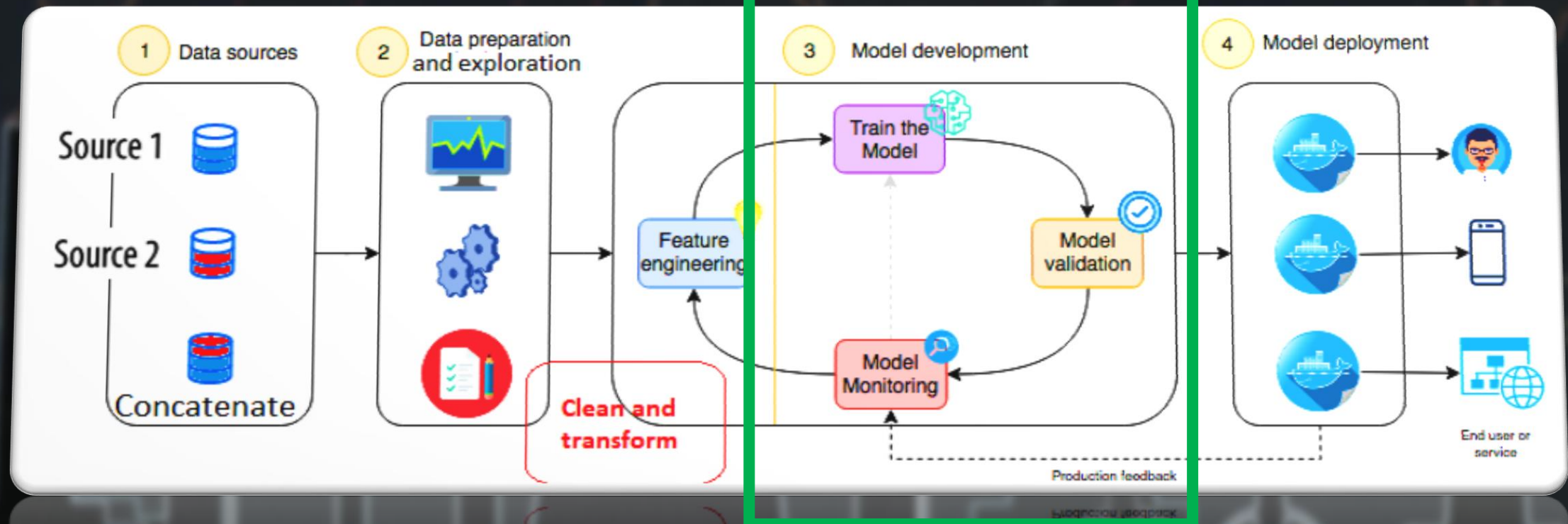


Class-weight

- Faire une pondération inversement proportionnellement à la fréquence des classes

**Oversampling
SMOTE**

- Augmenter les données dans la classe minoritaire

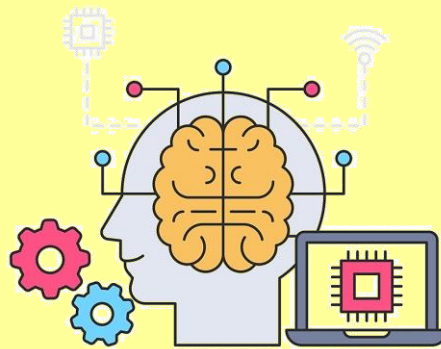


Modélisation

----- ✂ Variable prédite : Target ✂ -----

Choix des modèles

- Modèles:



Machine Learning

Random Forest

LightGBM

Gradient Boosting

- Métriques:

ROC-AUC

Matrice de
confusion

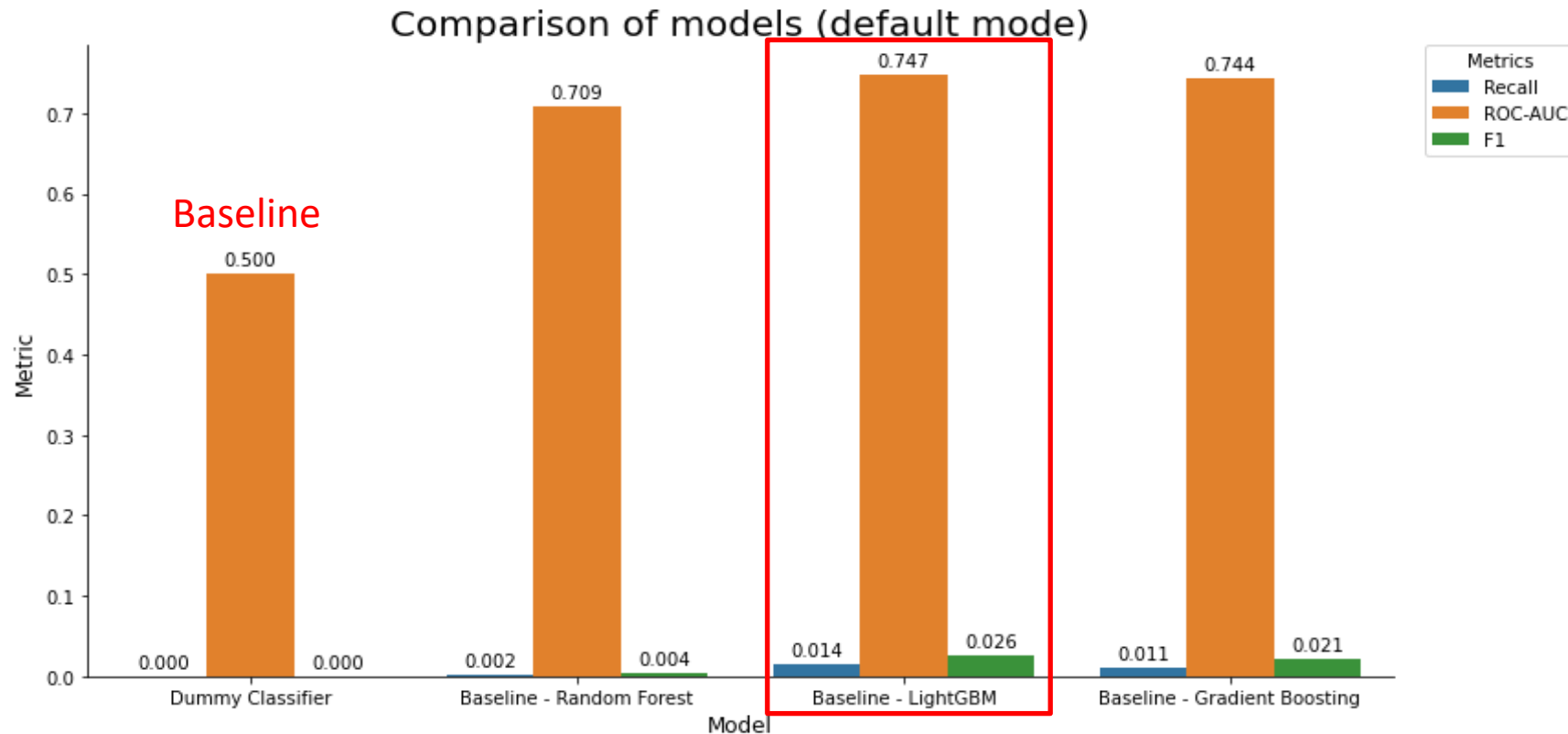
Fonction coût
(métier)

Modélisation



Variable prédite: Target

Modelés par défaut



Model	ROC-AUC	Recall	F1	Duration
Dummy Classifier	0.500	0.000	0.000	0.0
Random Forest	0.708	0.002	0.004	3.0
LightGBM	0.747	0.014	0.026	0.2
Gradient Boosting	0.744	0.011	0.021	6.2

Fonction de scoring

----- ✂ Pénaliser les FN ✂ -----

Loss Score

Erreur type I:

- Un faux positif (FP) constitue une perte d'opportunité pour la banque

Erreur type II:

- Un faux négatif (FN) constitue une perte pour créance irrécouvrable.

		Reality	
		Negative : 0	Positive : 1
Prediction	Negative : 0	True Negative : TN	False Negative : FN
	Positive : 1	False Positive : FP	True Positive : TP

Diagram illustrating the scoring function with arrows pointing to the confusion matrix cells:

- Green arrow from TN to score 1
- Red arrow from FN to score -10
- Red arrow from FP to score -1
- Green arrow from TP to score 1

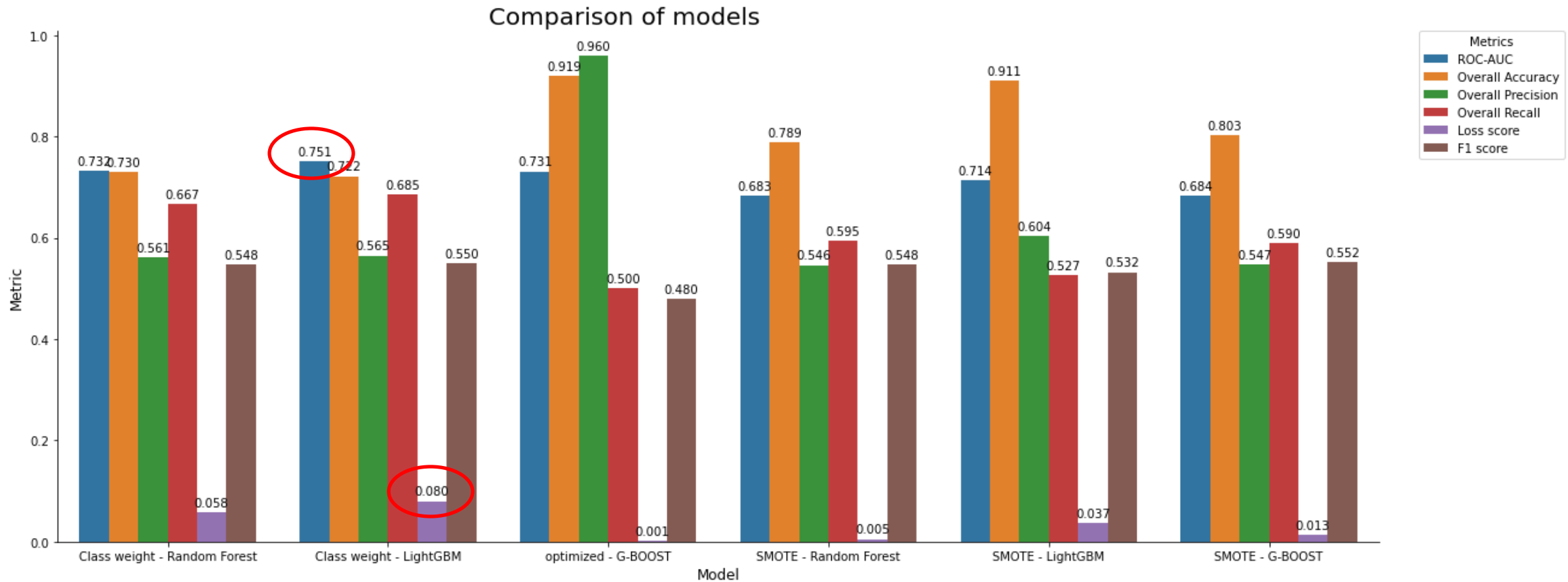
$$\text{Score} = (\text{gain_total} - \text{gain_minumun}) / (\text{gain_maximun} - \text{gain_minumun})$$

gain_total = TN*TN_rate + TP*TP_rate + FP*FP_rate + FN*FN_rate
gain_maximun = total_not_default*TN_rate + total_default*TP_rate
gain_minumun = total_not_default*TN_rate + total_default*FN_rate

Modélisation



Après tuning + SMOTE/Class weight



- Class-Weight donne des résultats meilleurs que SMOTE.
- Métriques (ROC-AUC et Loss Score) donnent le meilleur modèle « Class-Weight LightGBM »

Features importance



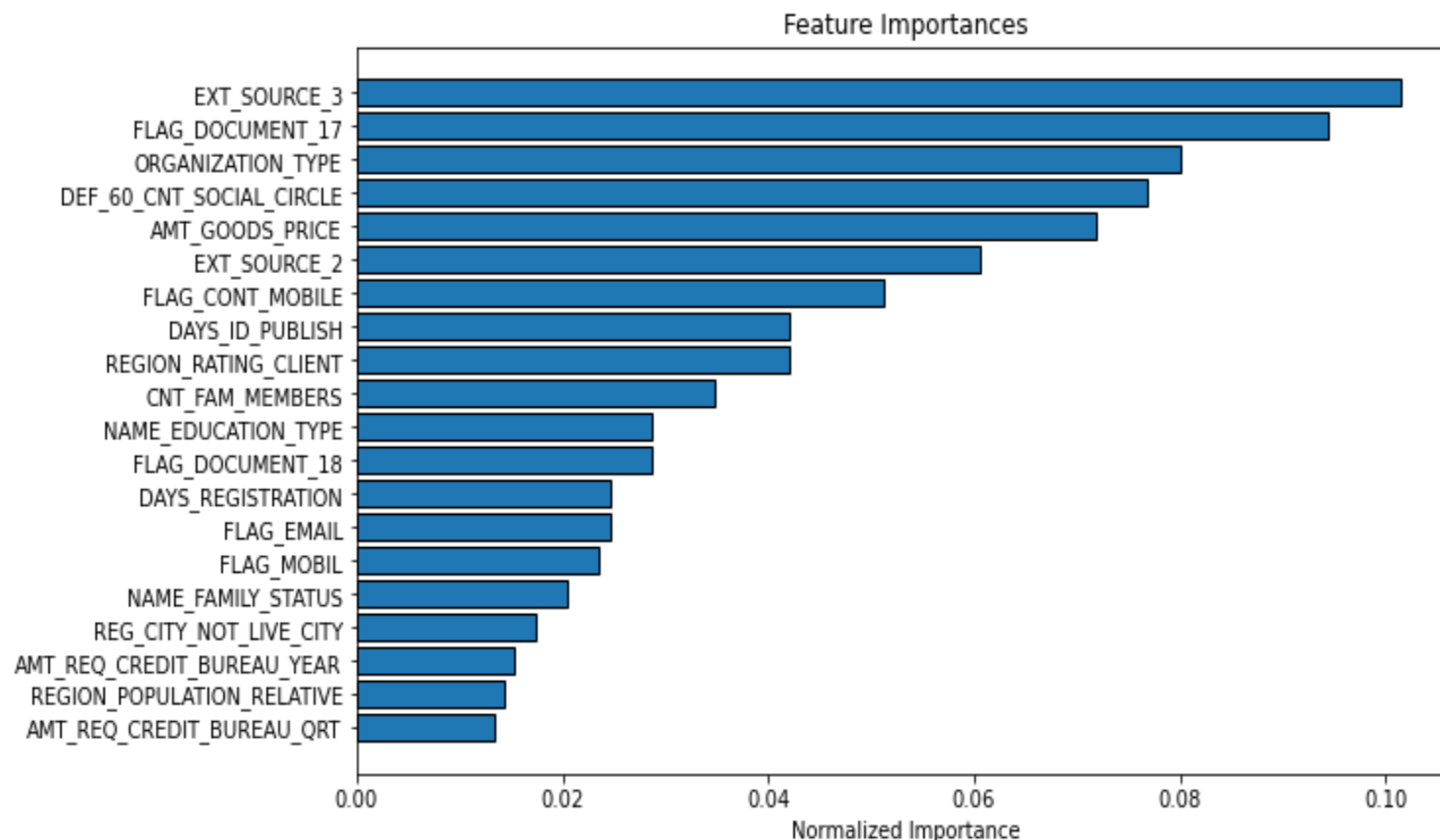
LGBM – Class-Weight

Les variables les plus importantes

- **source externe 3 et 2** : ce sont des scores normalisés provenant d'autres institutions.
- Les documents fournis par le client
- L'entreprise où il travaille
- L'environnement social
- Le montant du prêt demandé...



- ✓ En conclusion le modèle a pris en compte les différents types de variables :
Variables personnelles, Variables bancaires, Variables externes.



Features importance



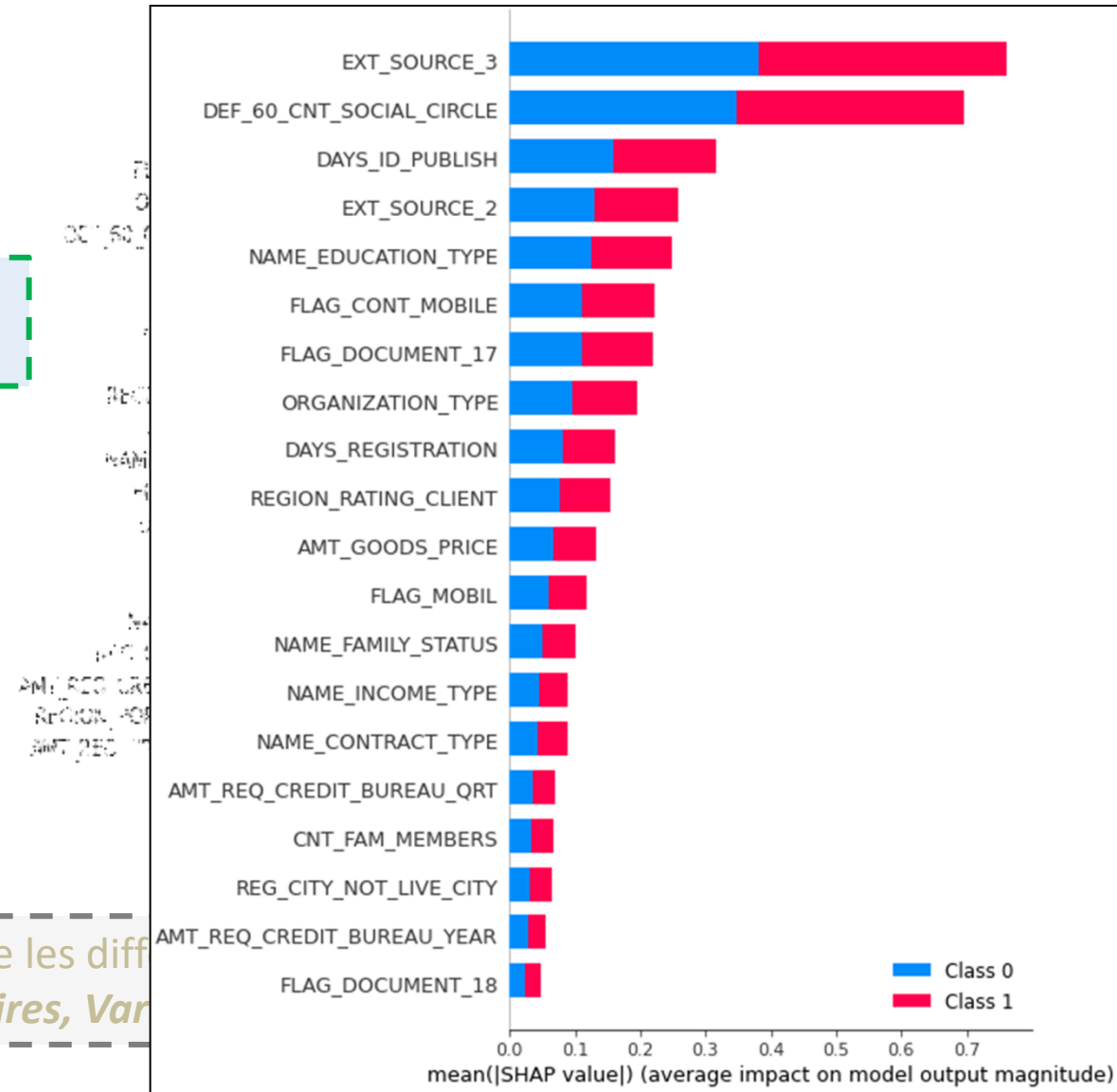
LGBM – Class-Weight

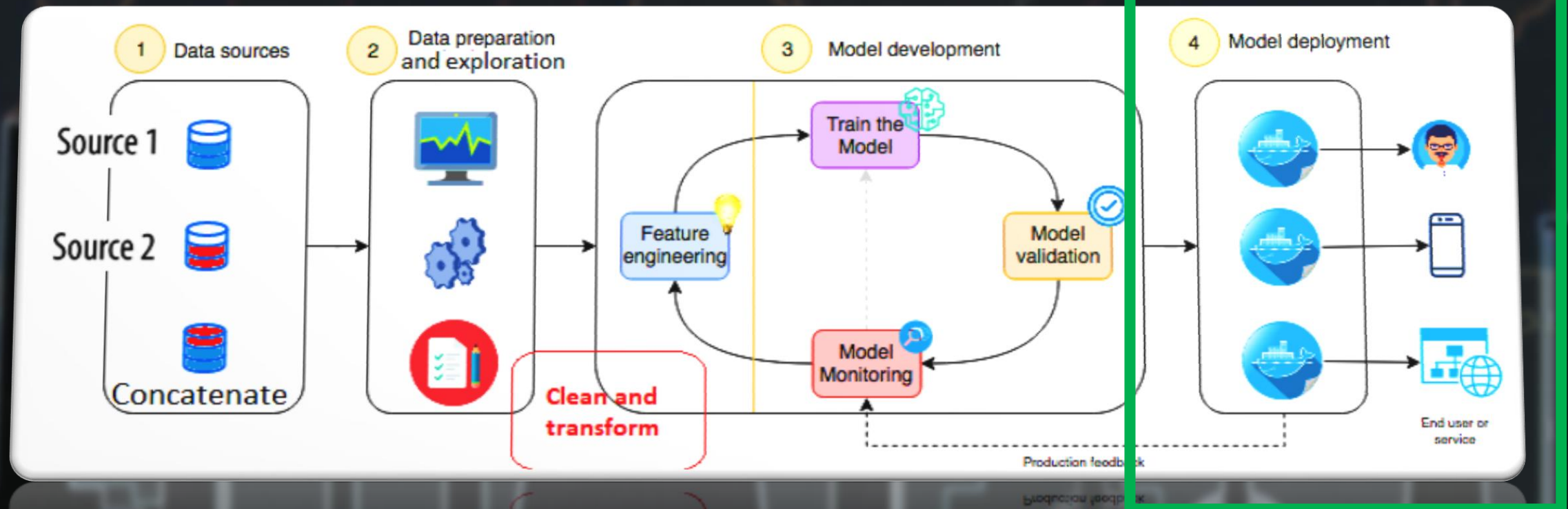
SHAP (SHapley Additive exPlanations)

- Influence de chacune des variables localement pour chaque prédiction.



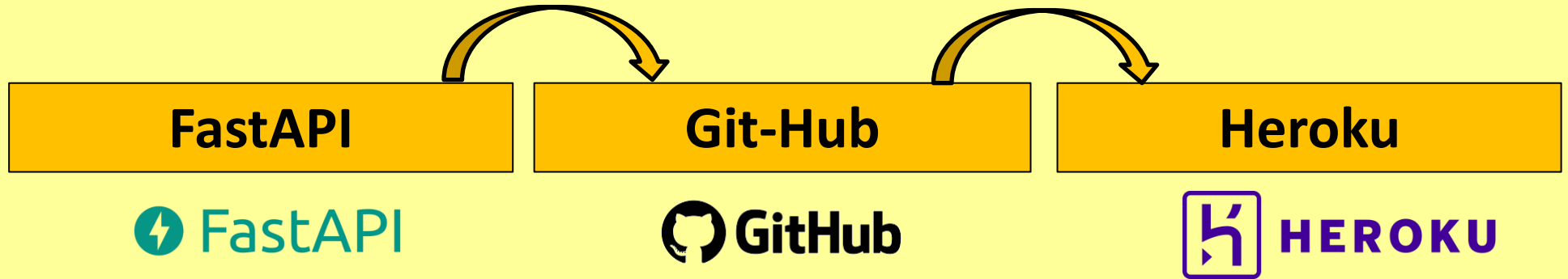
✓ En conclusion le modèle a pris en compte les diff
Variables personnelles, Variables bancaires, Var



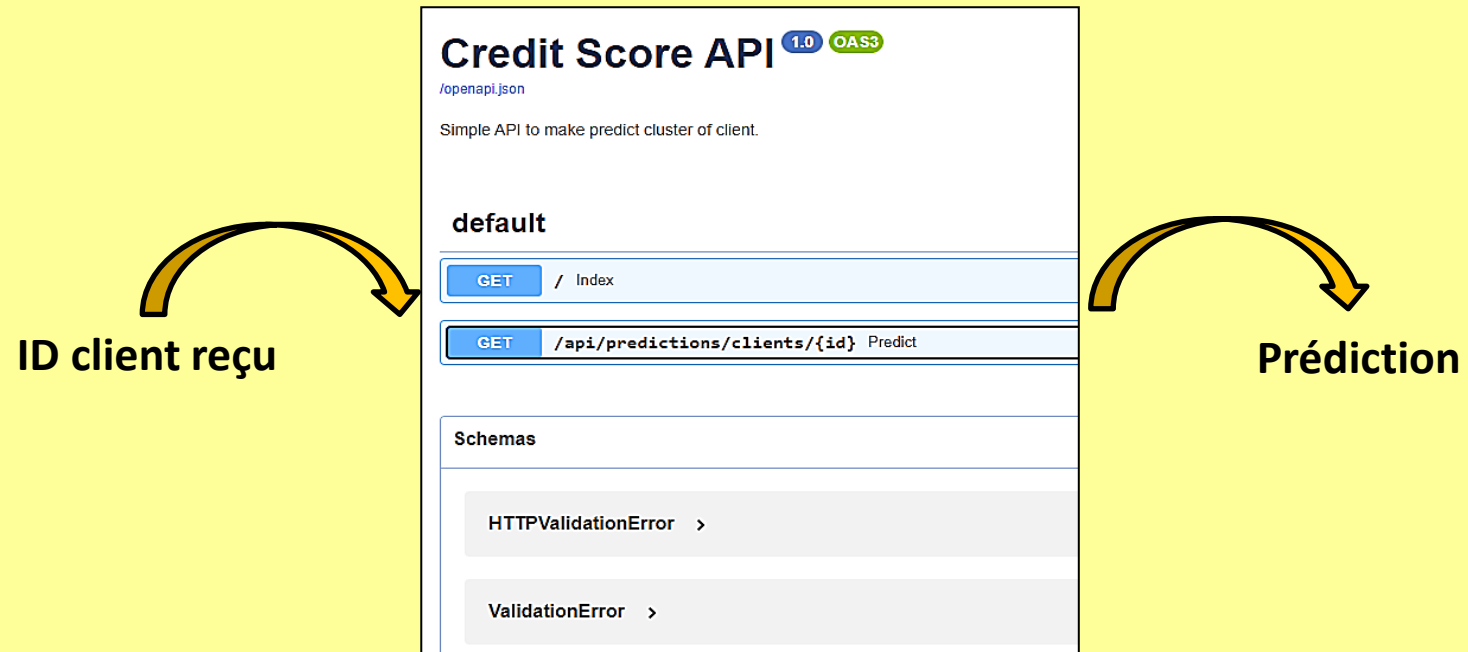


Dashboard

End Points - API



API permettant d'appeler la prédiction à partir de l'ID du client



Dashboard

Front



Page 1

Page 2



Navigation

- Home
- Client prediction

How to use it ? Select the Home page to see information related with the project. Select Client prediction to know whether a specific client will pay the loan based on his information.

Implement a scoring model

This project is part of [OpenClassRooms Data Scientist training](#) and has two main objectives:

- Building a scoring model that will give a prediction about the probability of a client paying the loan. The mission will be treated as a **binary classification problem**. So, 0 will be the class who repaid/pay the loan and 1 will be the class who did not repay/pay the loan.
- Build an interactive **dashboard** for customer relationship managers to interpret the predictions made by the model, and improve customer knowledge of customer relationship loaders.

How to use it ?

You can navigate through the Home page where you will find information related with the project. Also, you can go to the Client prediction to know whether a specific client will pay the loan based on his information

Other information

Data

The data used to develop this project are based on the [Kaggle's](#) competition: [Home Credit - Default Risk](#)

Repository

You can find more information about the project's code in its [Github' repository](#).

User Input Values

Select customer number

Identifiant client

289891

General informations:

Gender: Female

Age: 32

Education level: Secondary / secondary special

Marital status: Separated

Family members : 1 (including 0 children)

Work: Working

Work experiences: 2 years

Credit informations:

Credit amount: 312,768 \$

Annuity amount: 24,840 \$

Credit application result

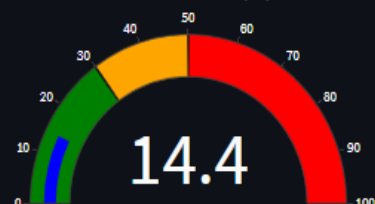
Check credit score:

- Later ✗
- Now ✓

You've selected client #289891.

Client #289891 has 14.40 % of risk to make default.

Risk of default (%)

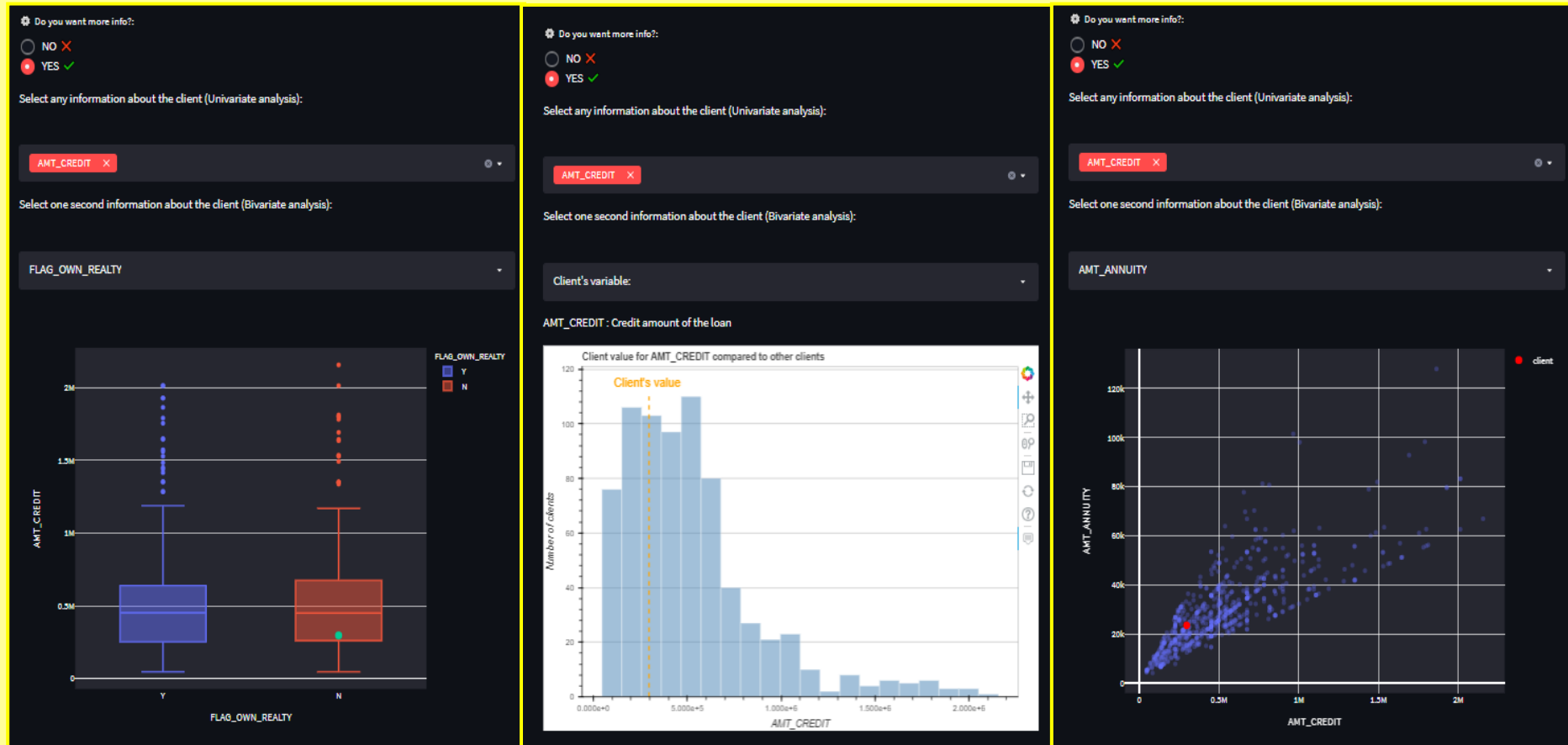


We recommend to accept client's application to loan 😊.

Below 30% of default risk, we recommend to accept client application. Above 50% of default risk, we recommend to reject client application. Between 30 and 50%, your expertise will be your best advice in your decision making. You can use the "client more informations" page to help in the evaluation.

Dashboard

Front



<https://mzmerli-credit-capacity-streamlit-streamlit-file-gzjerk.streamlitapp.com/>

Conclusion

➤ Un modèle plus performant

- | Feature engineering plus poussé
- | Traitement spécifique des valeurs manquantes
- | Métrique d'évaluation basé sur des hypothèse métier
 - | Loss-Score avec des coefficients plus adaptés au métier
- | Réduction de dimensionnalité (éviter le Curse of Dimensionality)

➤ Tableau de bord

- | Un onglet de simulation permettant aux clients de faire une estimation de leur prêt

**HOME
CREDIT**

