SOUTENANCE PROJET 7

Implémentez un modèle de scoring

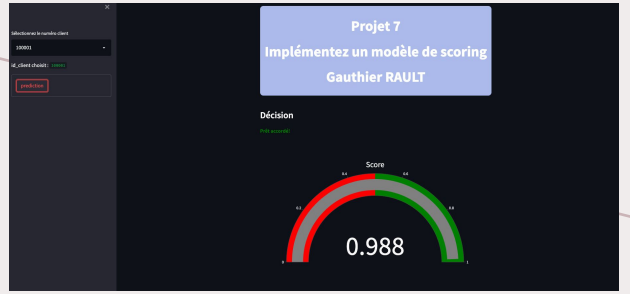
HOME CREDIT

GAUTHIER RAULT
PARCOURS DATA SCIENTIST
CHEZ OPENCLASSROOMS

EVALUATEUR
MONSIEUR Nassim LAOUITI

"le principal avantage des données est qu'elles vous disent quelque chose sur le monde que vous ne saviez pas auparavant"
Par Hilary MASON

ORDRE DU JOUR



Ouverture de la problématique - 5 min

Explication de l’approche de modélisation - 10 min

Présentation du dashboard - 5 min

Discussion - 5-10 min

OUVERTURE DE LA PROBLÉMATIQUE

L'entreprise "Prêt à dépenser" propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt

<u>Objectifs principaux:</u>

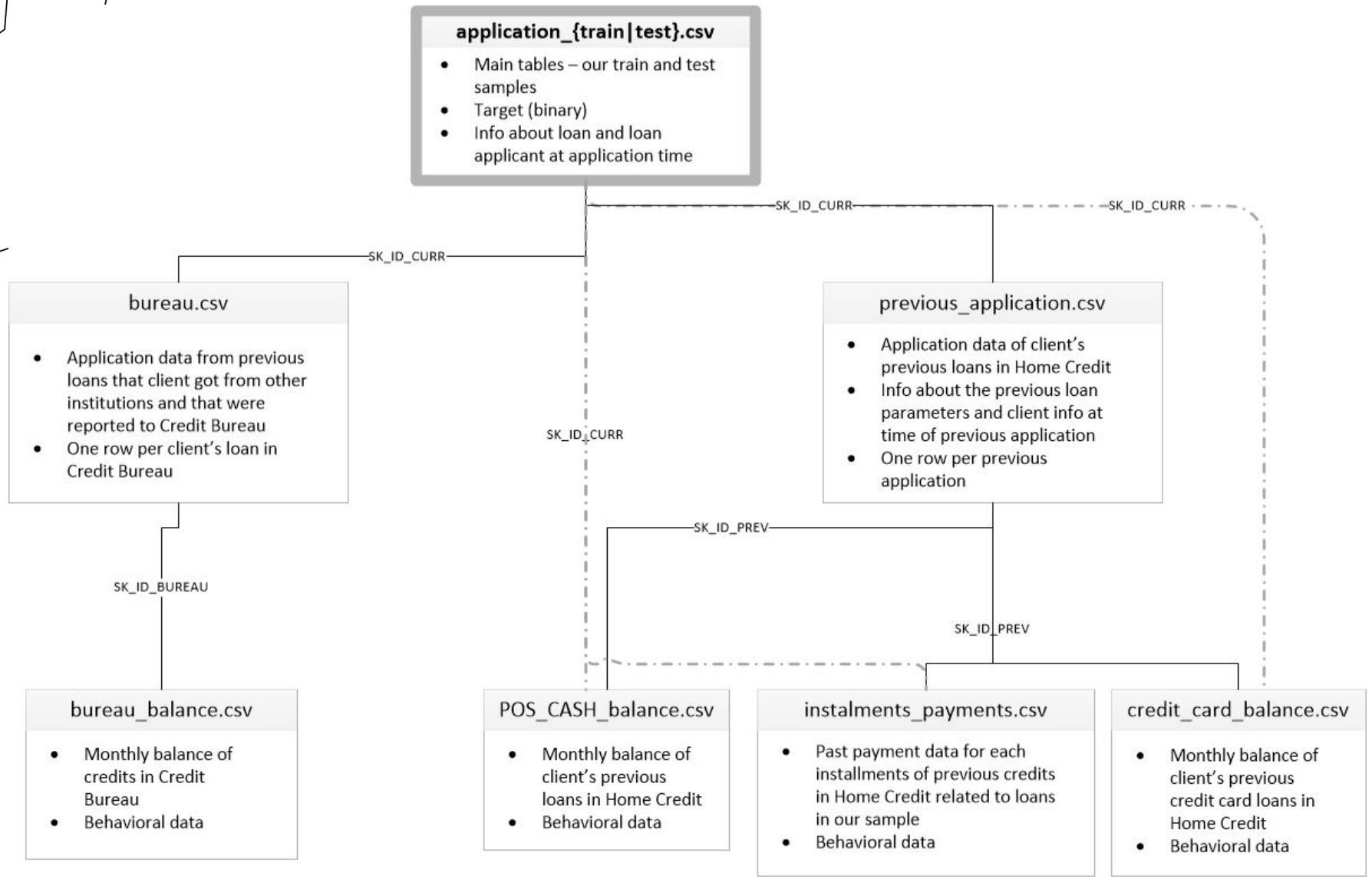
- Traiter les données mises à disposition
- Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique
- Construire un dashboard interactif à destination des gestionnaires de la relation client

<u>Moyens pour y parvenir:</u>

- Mise à disposition de 9 datasets

Datasets au format .csv

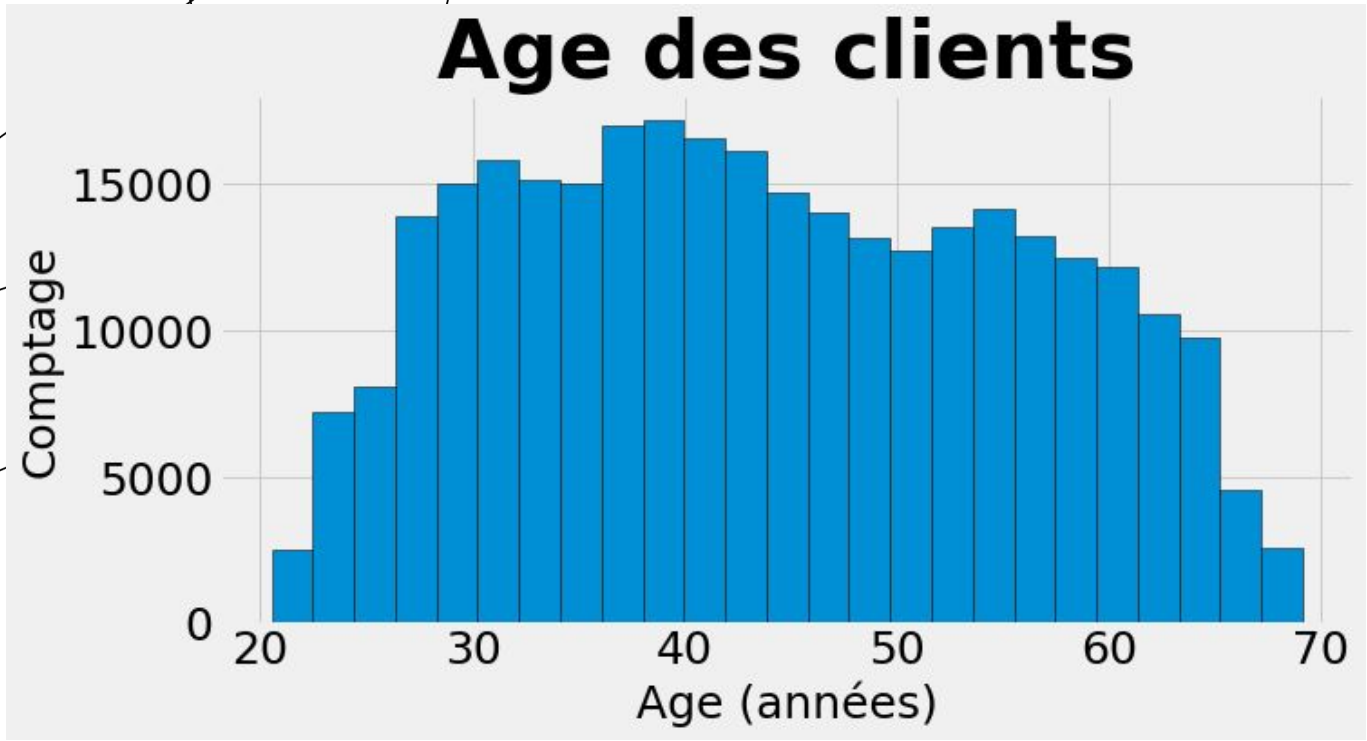
9 datasets reliés par la clé SK_ID_CURR

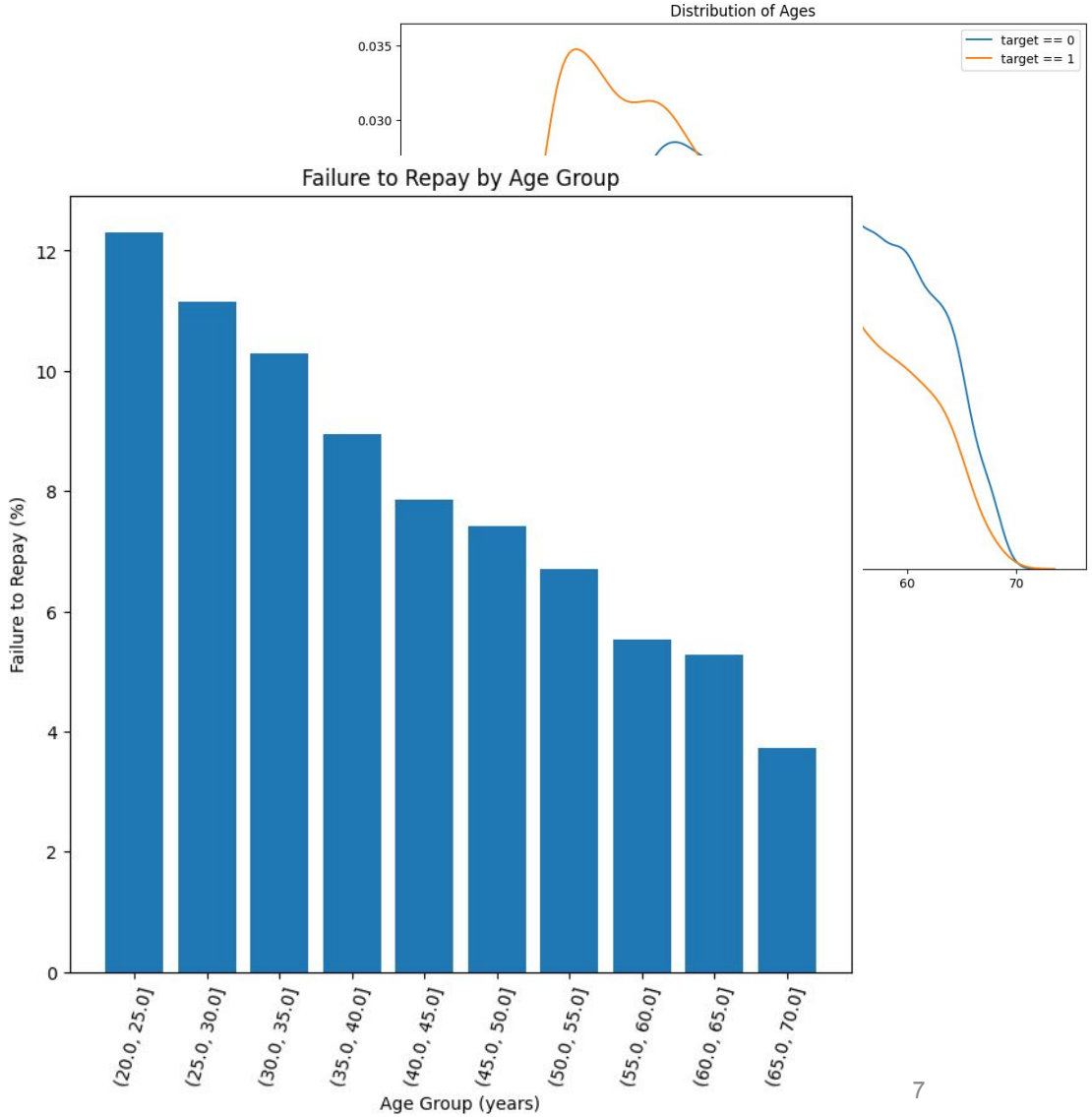


Découverte des datasets

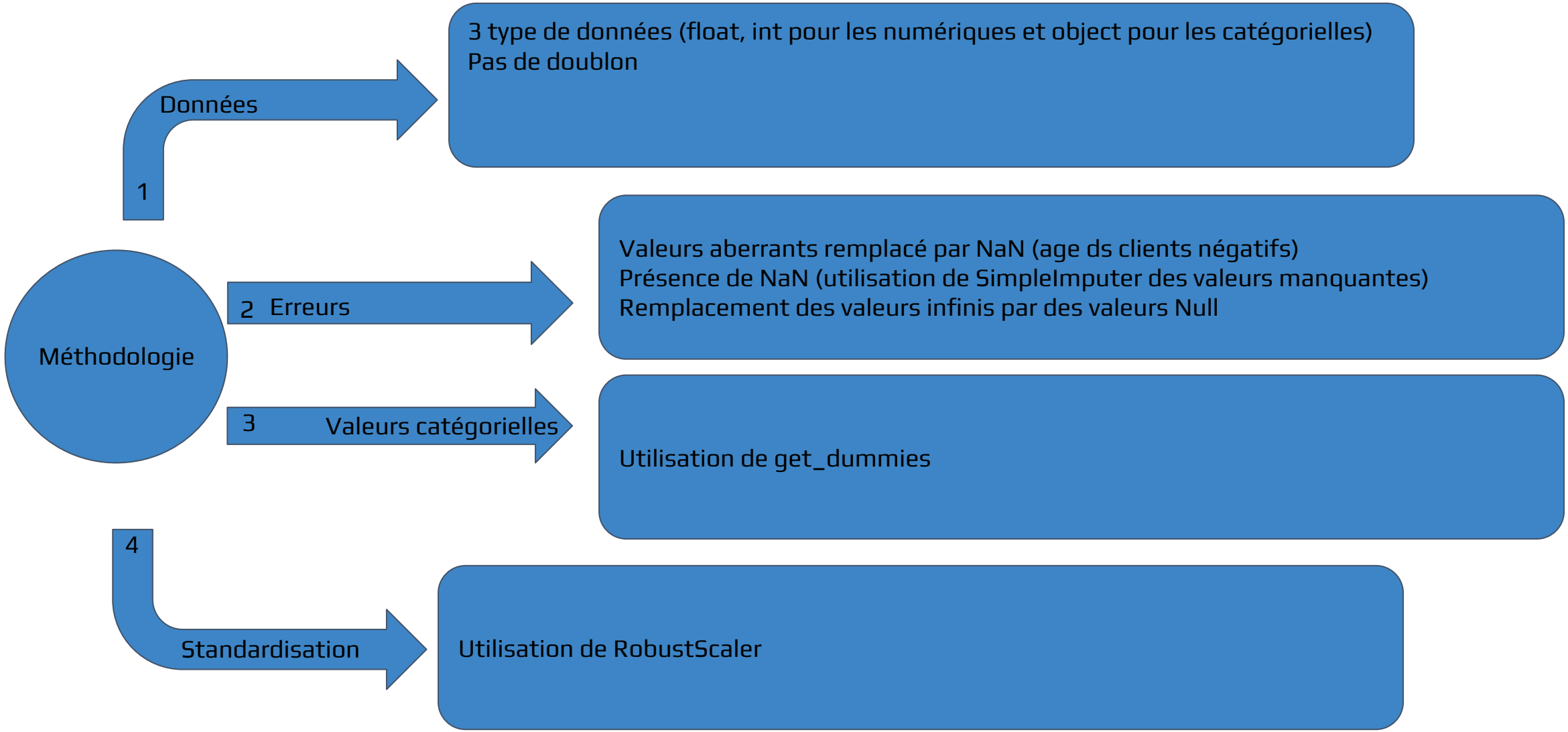
	Rows	Columns	%NaN	%Duplicate	object_dtype	float_dtype	int_dtype	bool_dtype	MB_Memory
./Datas/application_test.csv	48744	121	23.81	0.0	16	65	40	0	44.998
./Datas/POS_CASH_balance.csv	10001358	8	0.07	0.0	1	2	5	0	610.435
./Datas/credit_card_balance.csv	3840312	23	6.65	0.0	1	15	7	0	673.883
/Datas/installments_payments.csv	13605401	8	0.01	0.0	0	5	3	0	830.408
./Datas/application_train.csv	307511	122	24.40	0.0	16	65	41	0	286.227
./Datas/bureau.csv	1716428	17	13.50	0.0	3	8	6	0	222.620
./Datas/previous_application.csv	1670214	37	17.98	0.0	16	15	6	0	471.481
./Datas/bureau_balance.csv	27299925	3	0.00	0.0	1	0	2	0	624.846

Découverte des données





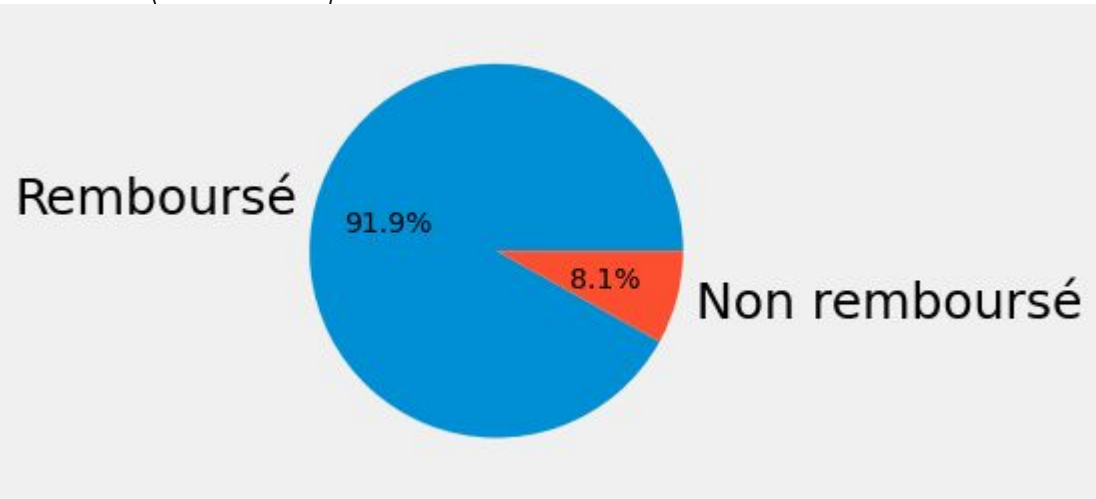
Préparation des données



Feature engineering

- **CREDIT_INCOME_PERCENT**: Pourcentage du montant du crédit par rapport au revenu d'un client
- **ANNUITY_INCOME_PERCENT**: Pourcentage de la rente de prêt par rapport au revenu d'un client
- **CREDIT_TERM**: Durée du paiement en mois
- **DAYS_EMPLOYED_PERCENT**: Pourcentage des jours employés par rapport à l'âge du client
- **INCOME_PER_PERSON**: Pourcentage des revenus des clients par rapport aux membres de la famille.

Répartition de la variable Target et rééquilibrage SMOTE



- Problématique de déséquilibre des classes avec près de 92% des données de la classe négative.
- Utilisation de la méthode SMOTE, qui réalise pour cela un KNN et crée un point à une distance aléatoire d'un point sélectionné et de ses plus proches voisins afin de rééquilibrer le dataset en vue d'un entraînement.

Sélection du modèle de Machine Learning

Sélectionner le meilleur algorithme adapté à notre problématique

Optimiser ses hyperparamètres

Sélection du modèle de Machine Learning

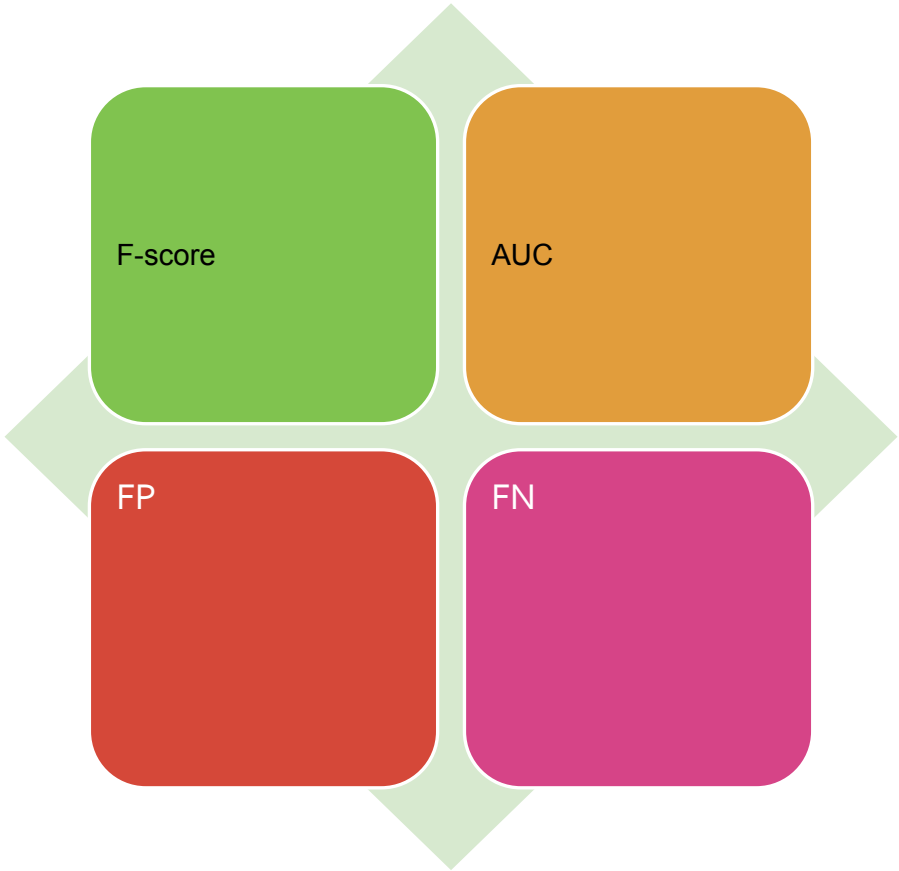
2 phases de sélection

1. Test de 6 algorithmes

2. Test des 2 meilleurs algorithmes avec fine-tuning

Sélection du modèle de Machine Learning

Critères d'évaluation



1ère phase - Sélection de l'algorithme

	Métriques			
Algorithmes	**F-score**	**AUC**	**Faux Positif (%)**	**Faux Négatif (%)**
DummyClassifier (Baseline)	0,5	0,5	46,13	4
RidgeClassifier	0,69	0,67	28,11	2,73
LogisticRegression	0,69	0,67	28,25	2,76
SGDClassifier	0,65	0,66	32,95	2,55
DecisionTreeClassifier	0,84	0,54	9,03	6,63
Random Forest	0,92	0,5	0,04	8,01
XGBoost	0,92	0,51	0,38	7,74

Random Forest et XGBoost

Fine tunning des 2 Meilleurs Algorithmes
Utilisation d'Hyperopt

	Métriques			
Algorithmes	**F-score**	**AUC**	**Faux Positif (%)**	**Faux Négatif (%)**
Random Forest	0,85	0,68	8,49	6,02
XGBoost	0,92	0,75	0,43	7,71

XGBoost

Définition du Seuil dans la problématique

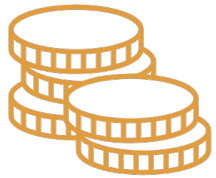
2 coûts à optimiser

- Prêter le moins possible à des « mauvais clients »
  - FN (Faux Négatif: Faux Bon Client)
- Prêter le plus possible aux « bon clients »
  - FP (Faux Positif: Faux Mauvais Client)

Définition du Seuil dans la problématique

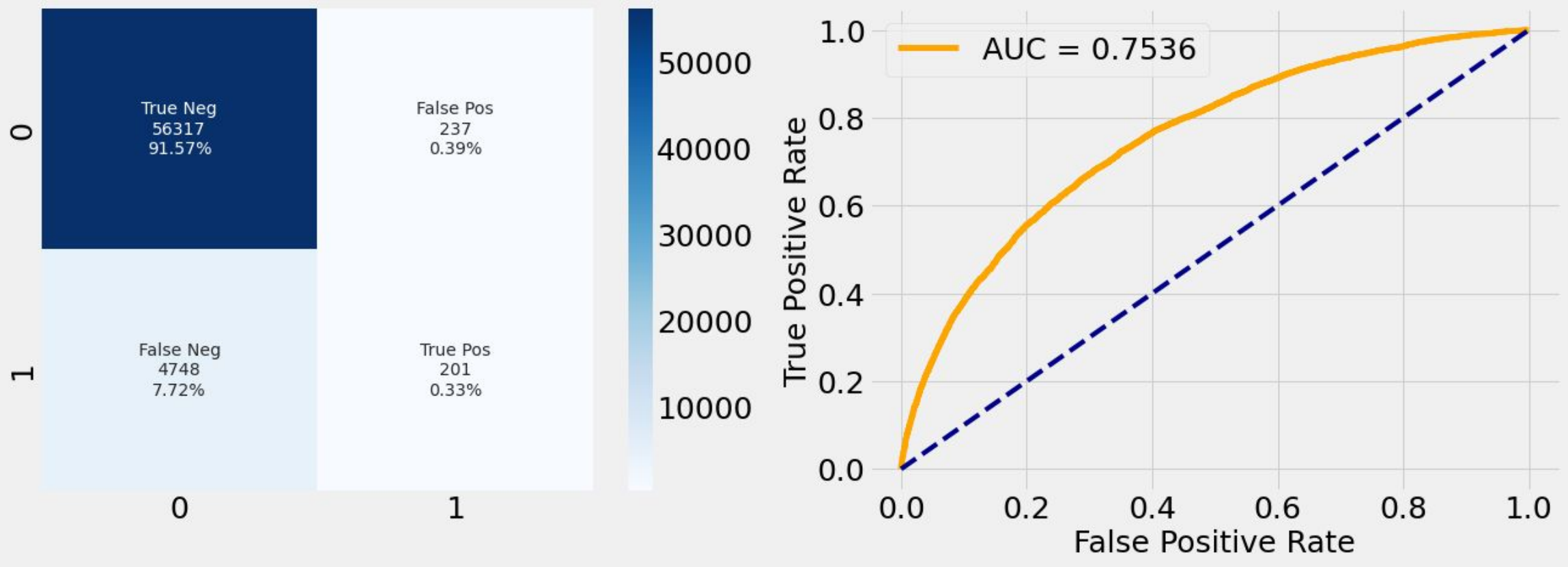
Rapport de coût entre FN et FP

1 FN coûte 10 fois plus chère qu'1 FP



1xFN=10xFP

Optimisation métier - Réentrainement d'XGBoost en ajoutant la fonction métier

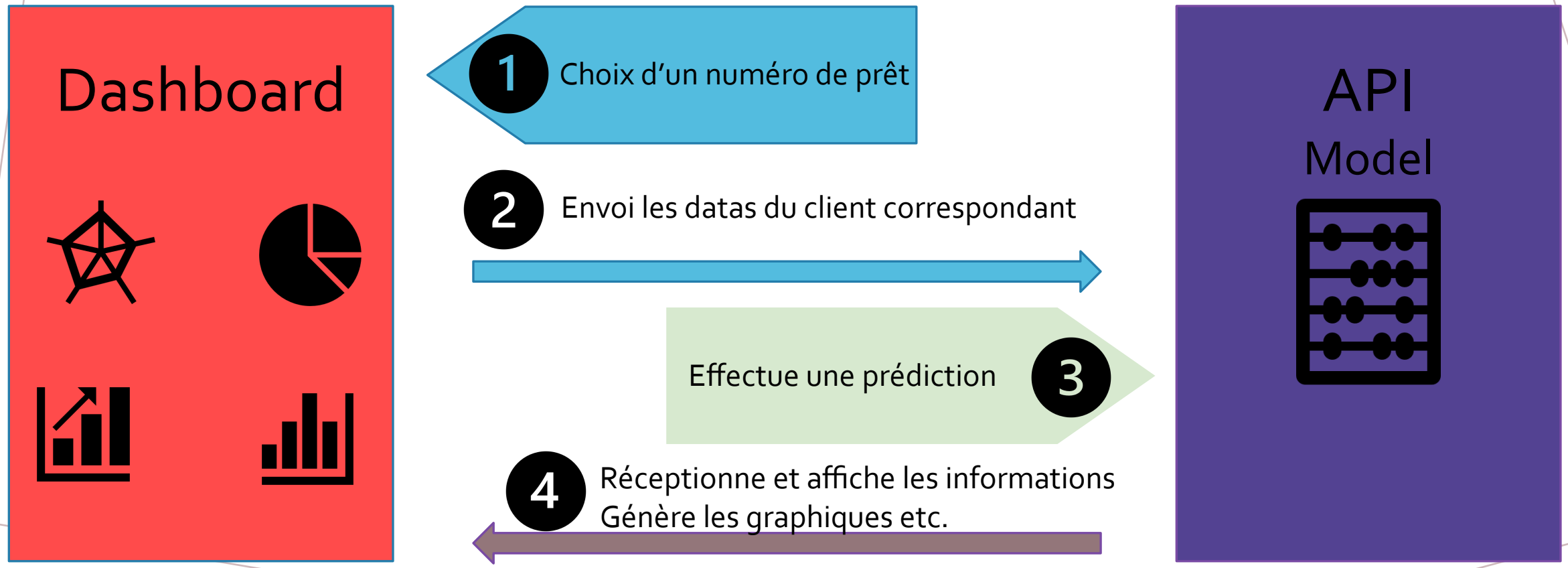


Dashboard & API

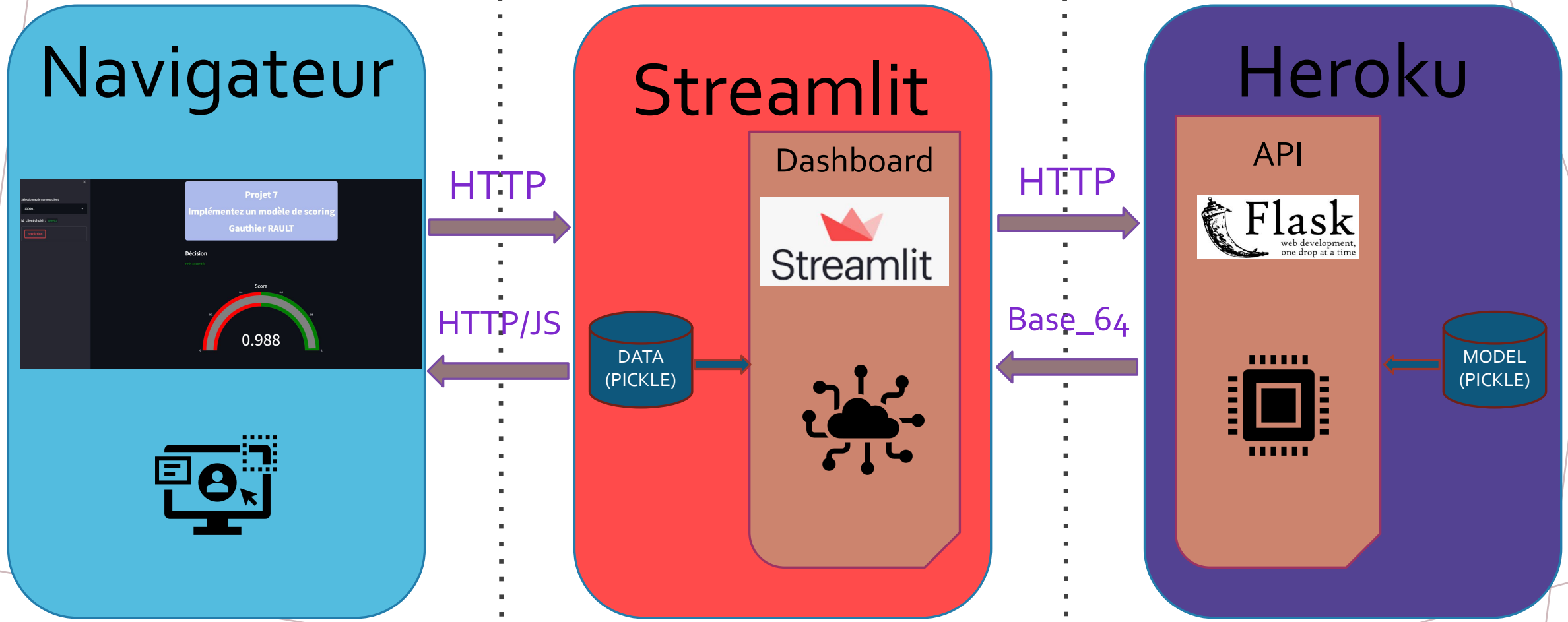
Schéma fonctionnel

Architecture

Dashboard & API
Schéma fonctionnel



Dashboard & API
Architecture

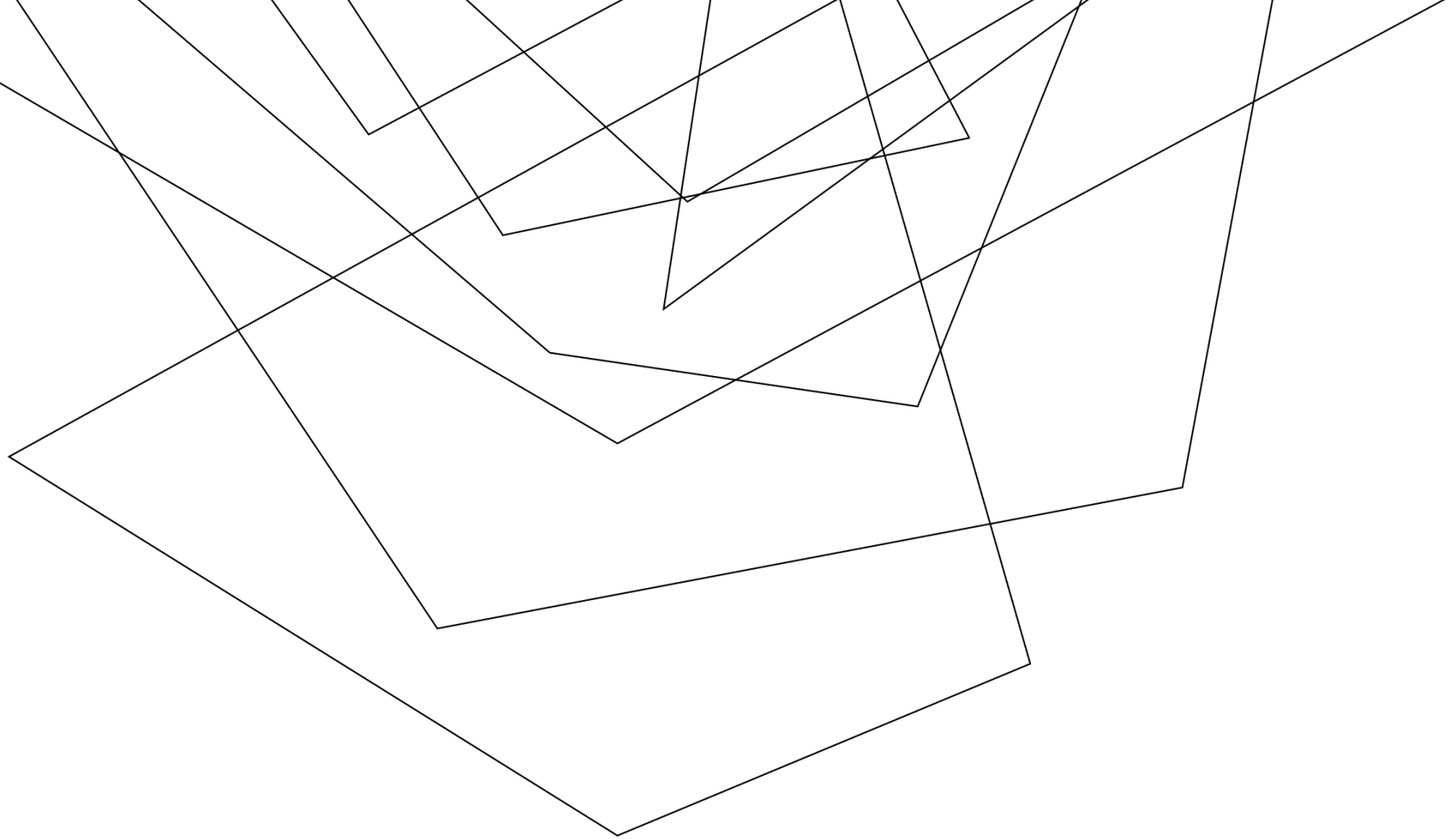


Dashboard
Démonstration

https://gauthierrault-p7-main-39hfff.streamlit.app/

AXES D'AMÉLIORATIONS DE L'ÉTUDE

1. Évaluer les algorithmes à partir d'une métrique personnalisée
2. Améliorer l'optimisation des algorithmes
3. Tester plus d'hyperparamètres
4. Tester plus de valeur par hyperparamètres
5. Dashboard
6. Permettre de modifier les données clients et effectuer une nouvelle prédiction



MERCI POUR VOTRE ATTENTION