

Projet 7

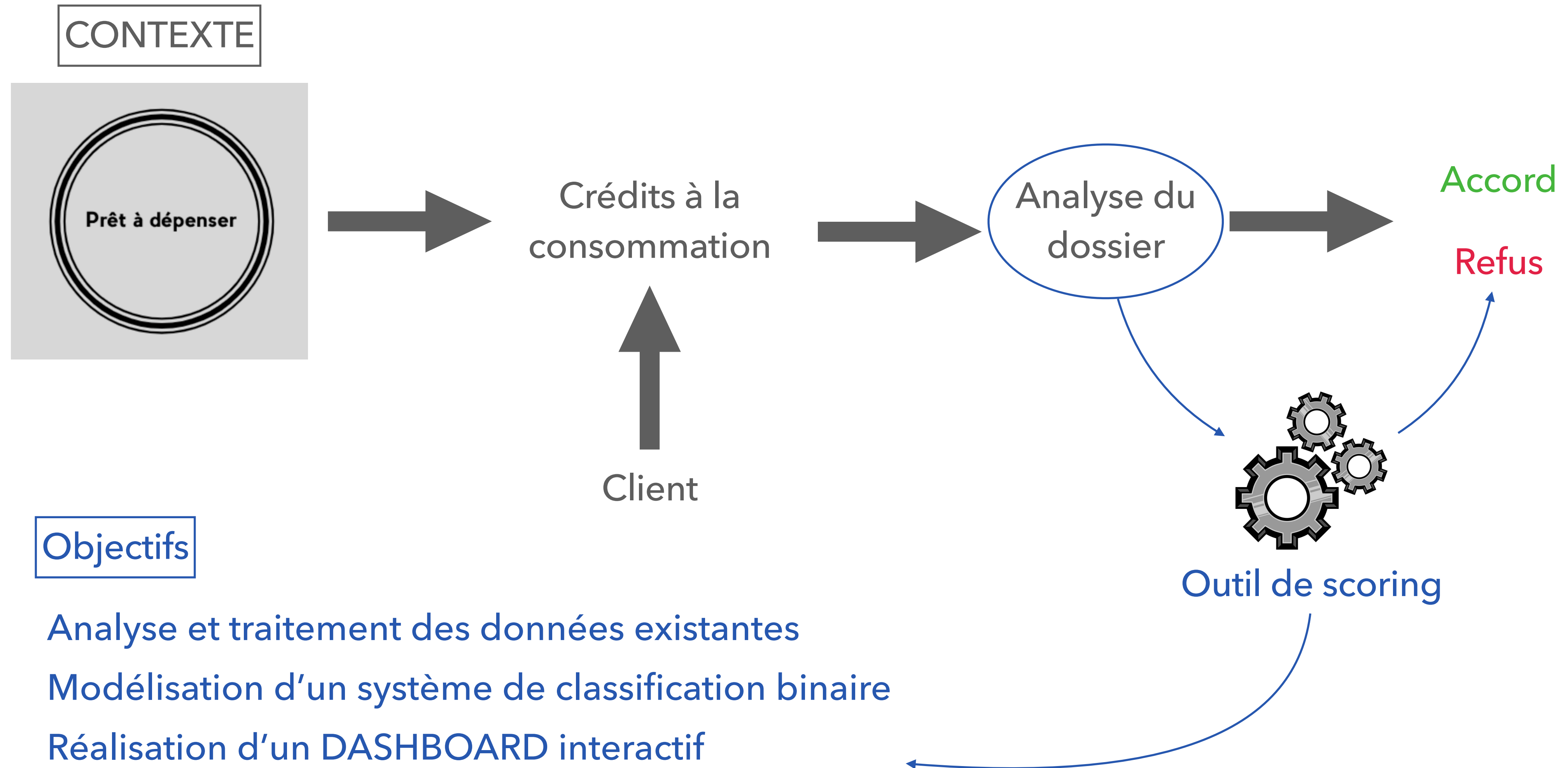
Implémentez un modèle de scoring

Développement d'un « outil de scoring » pour la société

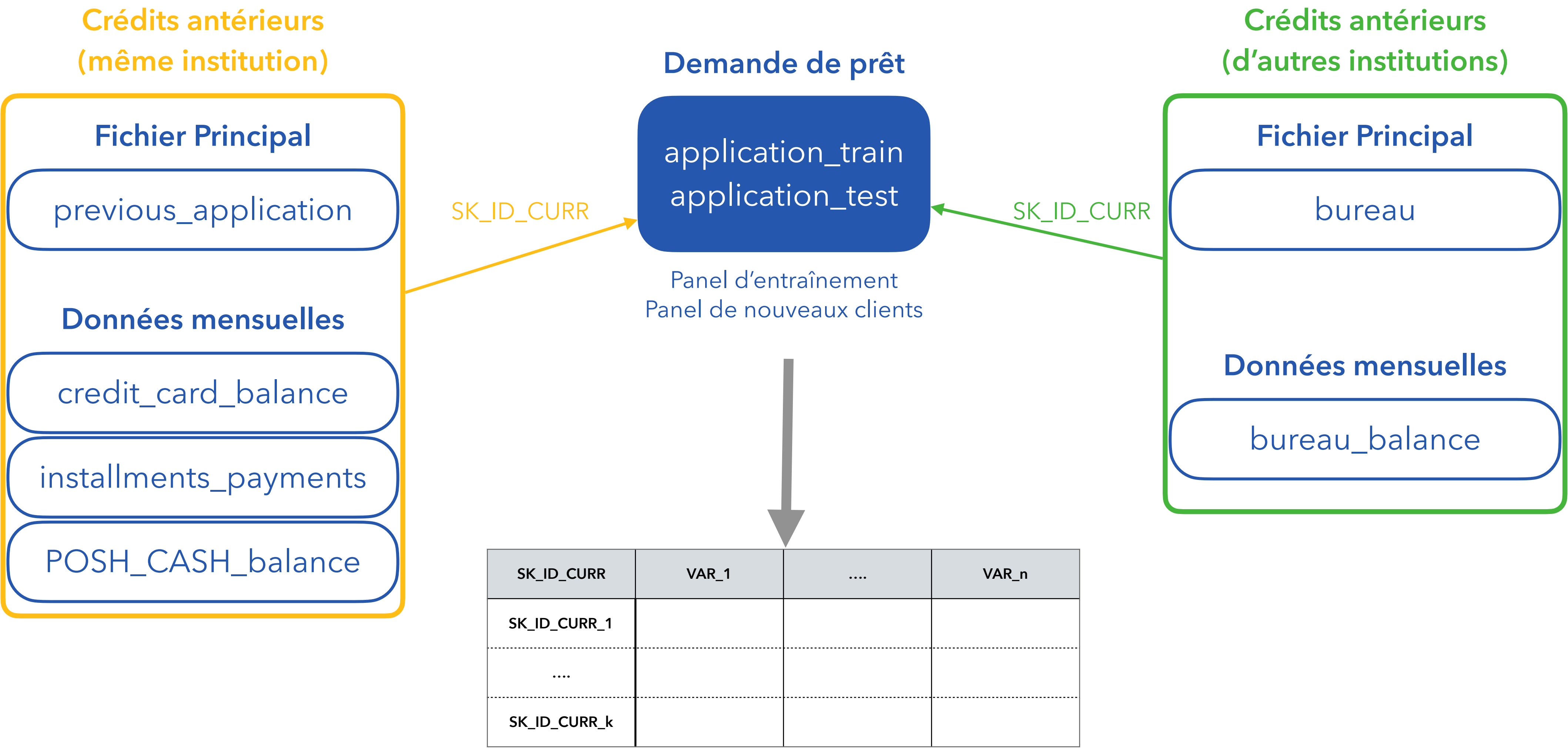
Prêt à dépenser

- 1 - Contexte et objectifs du projet
- 2 - Présentation et préparation du jeu de données
- 3 - Modélisation
- 4 - Analyse de DataDrift
- 5 - Construction et le déploiement du DASHBOARD
- 6 - Démonstration de l'API
- 7 - Conclusion

1- Contexte et objectifs

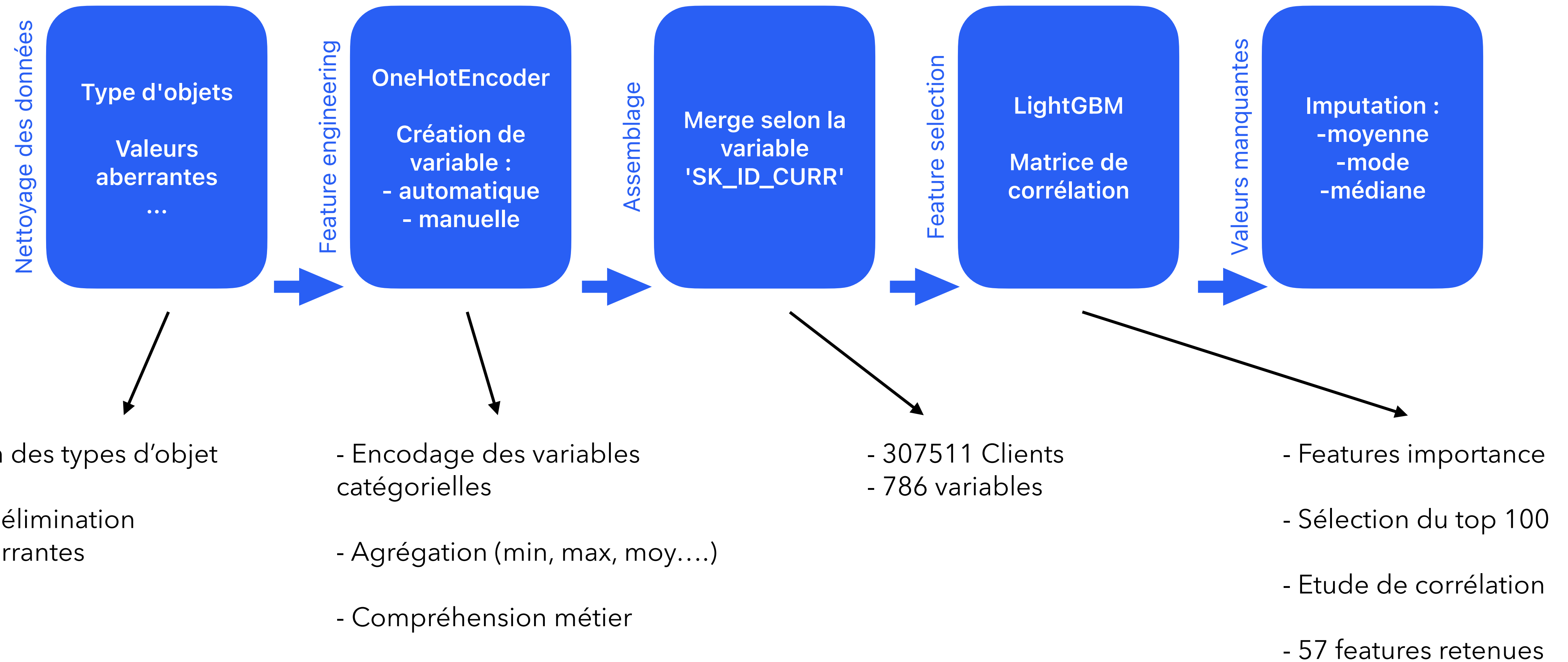


2 - Présentation et préparation



2 - Présentation et préparation

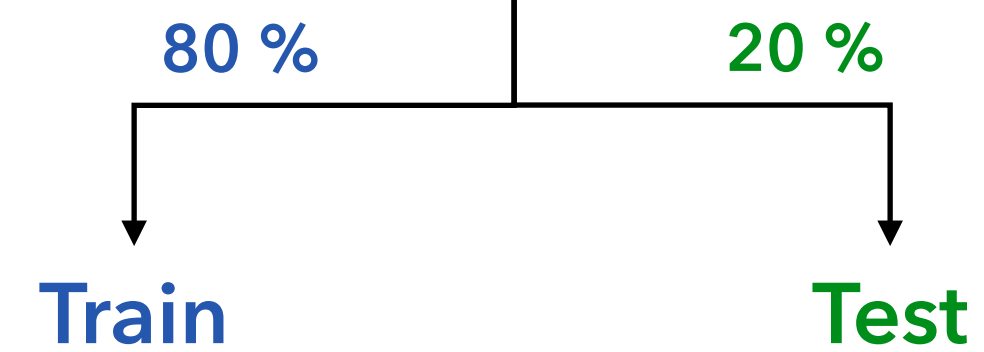
Utilisation de Kernels Kaggle



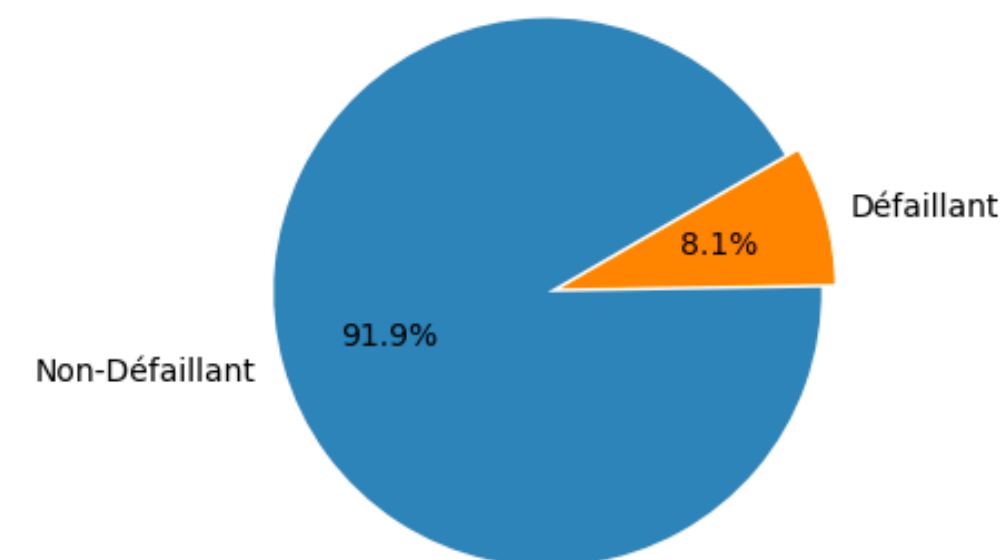
3 - Modélisation

Dataset avec TARGET

Utilisation pour l'entrainement
des modèles



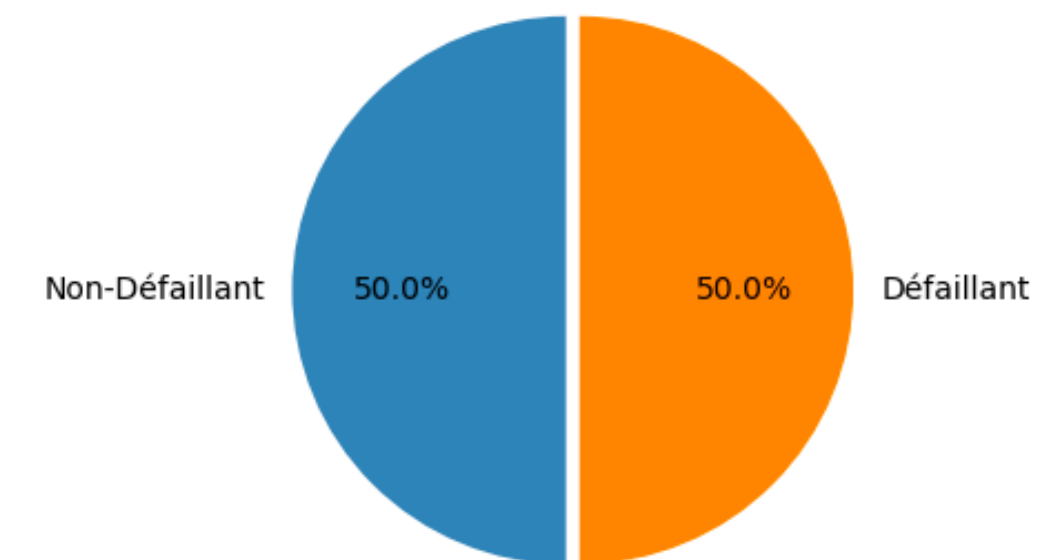
Répartition de la cible avant rééquilibrage



Dataset sans TARGET

Simulation de nouveaux clients
DASHBOARD

Répartition de la cible après rééquilibrage



PREPROCESSING

Ré-équilibrage

Standardisation





SMOTE

Sur-échantillonnage

3 - Modélisation

Métriques d'évaluation

Classification Binaire

		Classes Prédites	
		0	1
Classes Réelles	0	 TN (True Negative)	 FP (False Positive)
	1	 FN (False Negative)	 TP (True Positive)

FP :
Perte des intérêts
de remboursement

FN :
Perte de la somme prêtée

Evaluation de la performance du modèle

Evaluation de la qualité de la classification

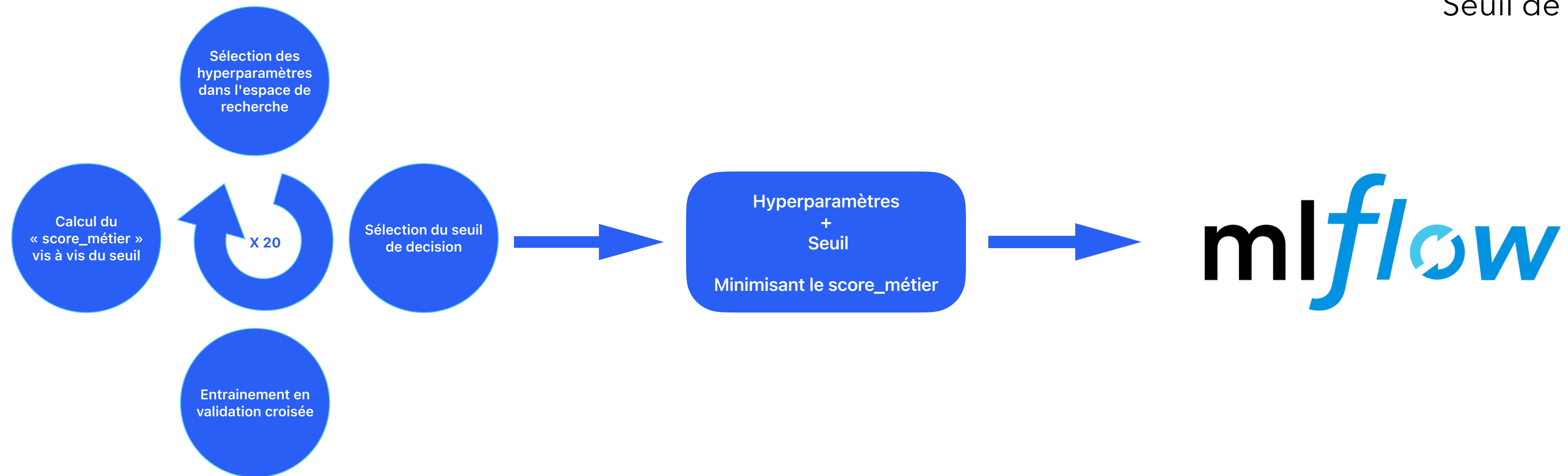
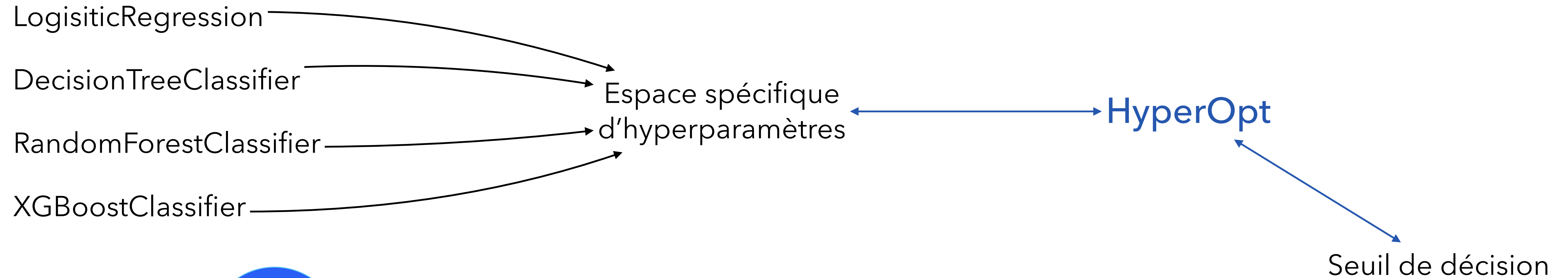
Dépendance de la problématique

Minimiser les pertes financières

$$\text{« Score métier »} = \frac{10 * FN + FP}{TN + TP + FP + 10 * FN}$$

3 - Modélisation

Entrainement et optimisation des modèles



3 - Modélisation

Tracking via MLFlow et Résultat

TableChartEvaluationExperimental

<input type="checkbox"/>		Run Name	Created		Dataset	Duration	Source	Models
<input type="checkbox"/>		DecisionTreeClassifier	13 days ago		-	17.6s	ipykerne...	sklearn
<input type="checkbox"/>		XGBoost	13 days ago		-	30.7s	ipykerne...	sklearn
<input type="checkbox"/>		RandomForest	13 days ago		-	47.1s	ipykerne...	sklearn
<input type="checkbox"/>		LogisticRegression	13 days ago		-	3.2s	ipykerne...	sklearn

Comparaison_des_modèles >

RandomForest

Run ID: b4a341fb4d464ee694a798ad3c767628

Date: 2023-11-30 11:27:08

Duration: 47.1s

Status: FINISHED

> Description [Edit](#)

> Datasets

▼ Parameters (2)

Name	Value
max_depth	28
n_estimators	100

▼ Metrics (13)

Name	Value
accuracy_test	0.6903515332834704
accuracy_val	0.9312786044352075
auc_test	0.6599826569562295
auc_val	0.9780669771802296
f1_test	0.2454235676361043
f1_val	0.9263826532044149
precision_test	0.1527647610121837
precision_val	0.940523403788174
recall_test	0.6237663645518631
recall_val	0.9203519865572973
score_metier_test	35856
score_metier_val	26065
threshold	0.2519995429386416

▼ Artifacts

▼ best_model

MLmodel

conda.yaml

model.pkl

python_env.yaml

requirements.txt

▼ Artifacts

▼ best_model

MLmodel

conda.yaml

model.pkl

python_env.yaml

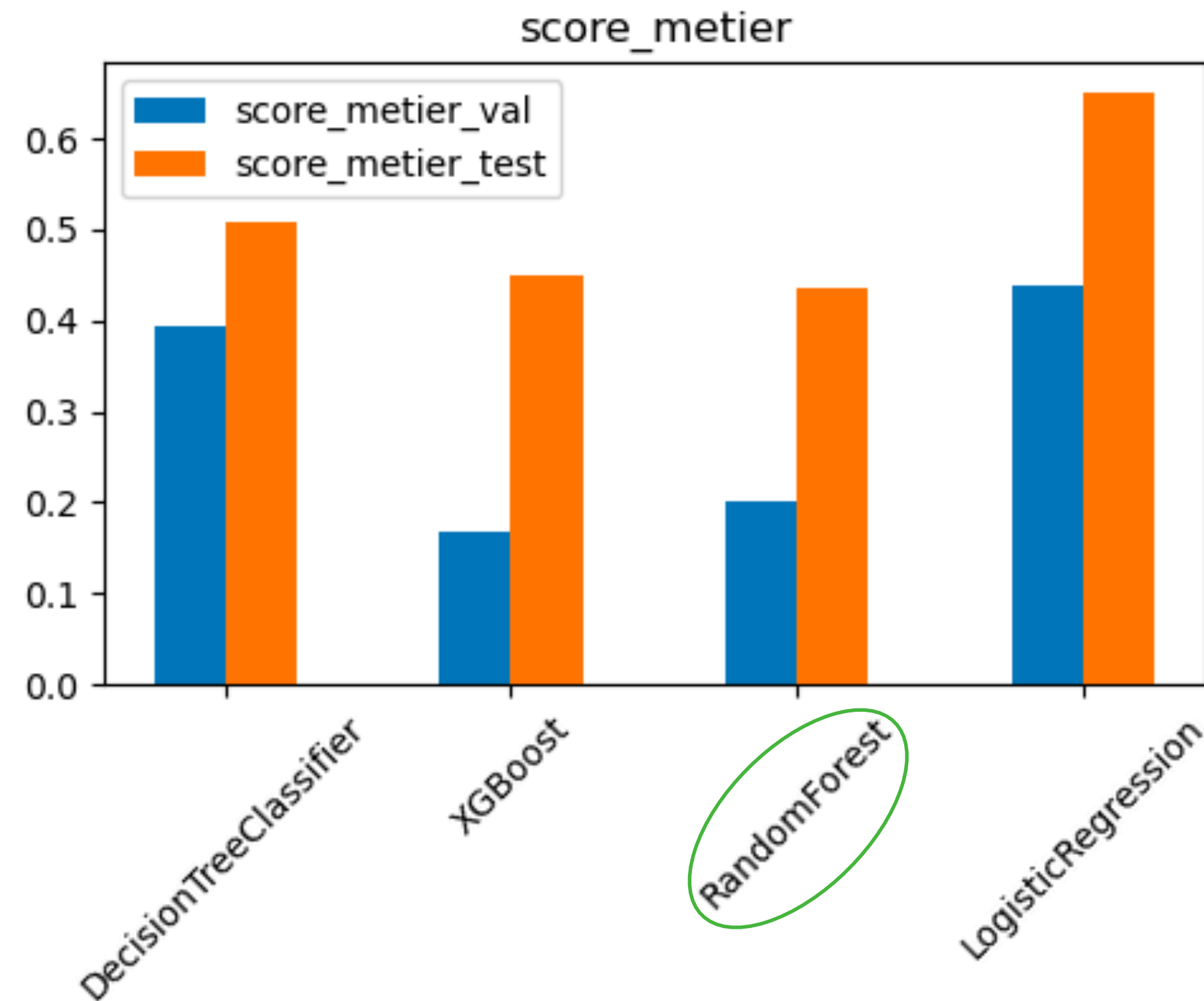
requirements.txt

Full Path:file:///Users/maxi
Size: 114B

mlflow==2.6.0
cloudpickle==2.2.1
configparser==5.3.0
numpy==1.24.3
psutil==5.9.0
scikit-learn==1.2.2
scipy==1.11.3

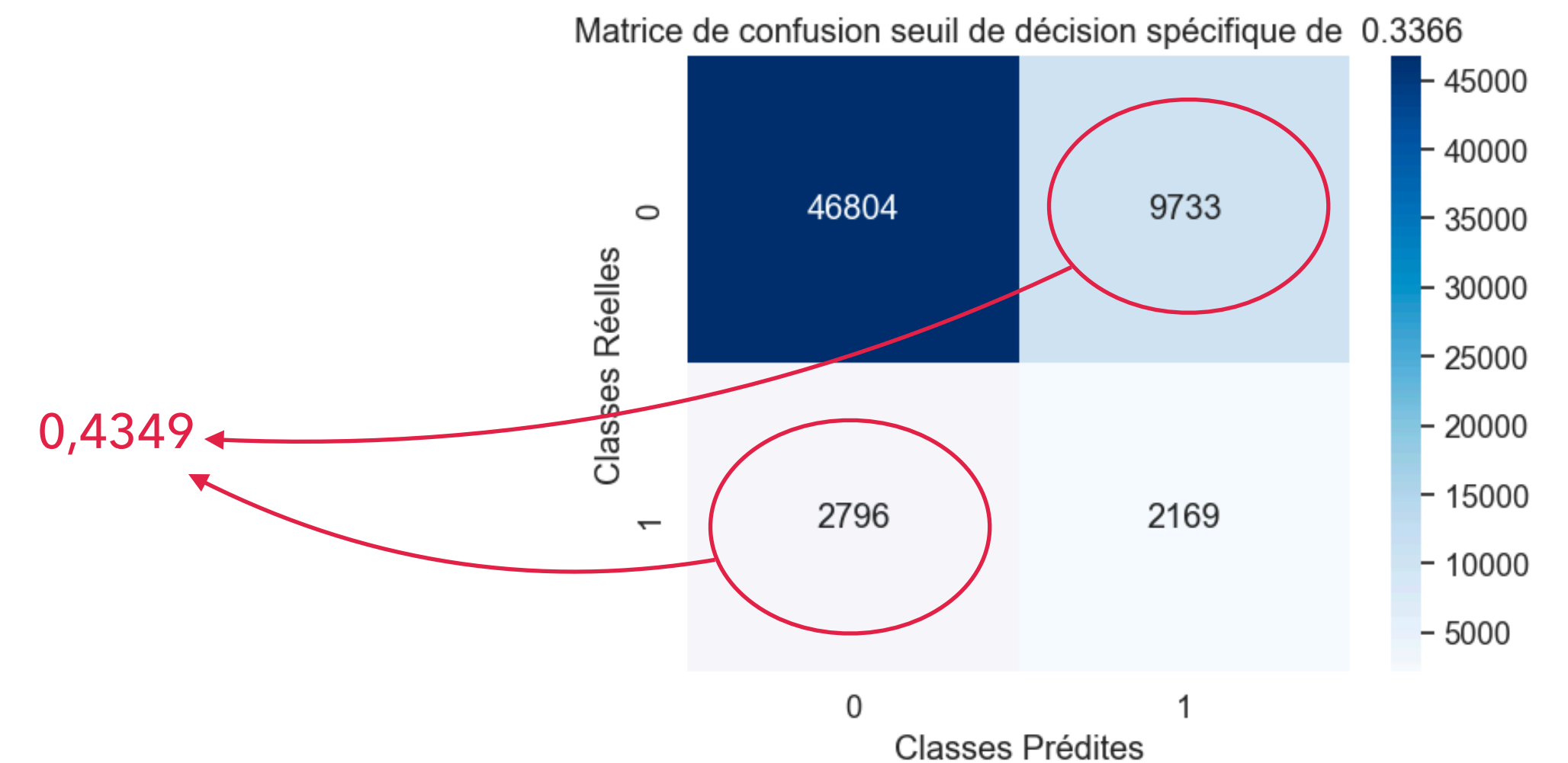
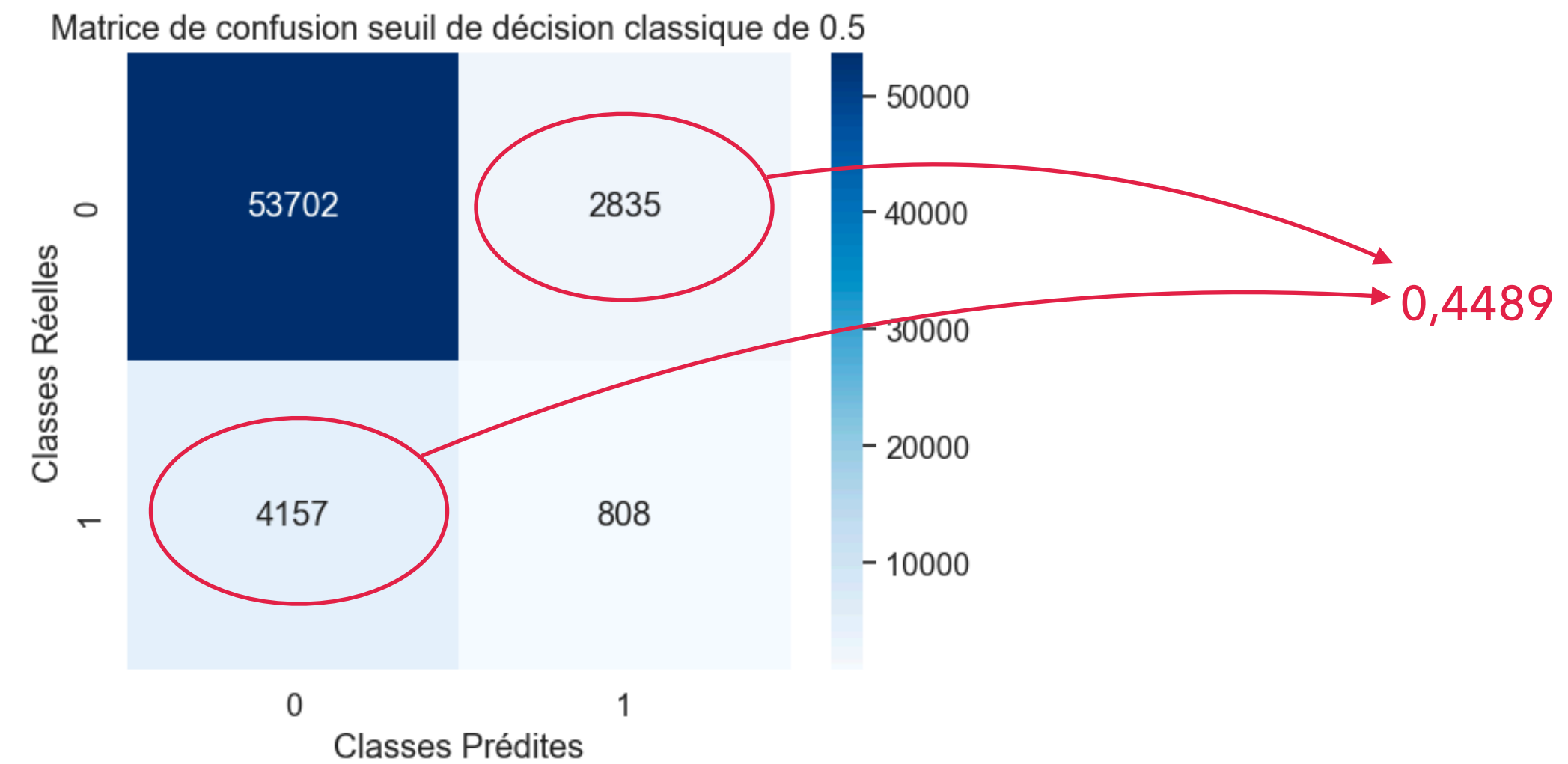
3 - Modélisation

Tracking via MLFlow et Résultat



RandomForestClassifier + StandardScaler

Estimateur



4 - Analyse de DATADRIFT

Evidently


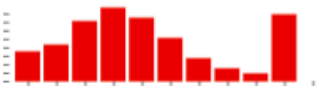








Dataset Drift

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

57 Columns	9 Drifted Columns	0.158 Share of Drifted Columns
---------------	----------------------	-----------------------------------

Data Drift Summary

Drift is detected for 15.789% of columns (9 out of 57).

<div><div></div><div>Search</div><div></div></div>						
	Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test
>	BURO_STATUS_0_MEAN_MEAN	num			Detected	Wasserstein distance (normed)
>	PAYMENT_RATE	num			Detected	Wasserstein distance (normed)
>	EXT_SOURCE_1	num			Detected	Wasserstein distance (normed)
>	INCOME_CREDIT_PERC	num			Detected	Wasserstein distance (normed)
>	AMT_ANNUITY	num			Detected	Wasserstein distance (normed)
						Drift Score

Pas de d rive significative sur l'ensemble

9 d rives sur 57 variables dont deux assez cons quentes

5 - Construction et déploiement d'un DASHBOARD



*Estimateur
sérialisé en pickle*

Backend
FastAPI

https://github.com/maxsch38/P7_OCR_API_Backend

Format de transfert : JSON

DataBase

Frontend
Streamlit

https://github.com/maxsch38/P7_OCR_API_Frontend

*Autres fichiers
(Image...)*



5 - Construction et déploiement d'un DASHBOARD



Code Blame 25 lines (19 loc) · 392 Bytes Code 55% faster with GitHub Copilot

```
1 name: Run Tests
2
3 on:
4   push:
5     branches:
6       - main
7
8 jobs:
9   test:
10    runs-on: ubuntu-latest
11
12    steps:
13      - name: Checkout code
14        uses: actions/checkout@v3
15
16      - name: Set up Python
17        uses: actions/setup-python@v3
18        with:
19          python-version: 3.11.6
20
21      - name: Install dependencies
22        run: pip install -r requirements.txt
23
24      - name: Run tests
25        run: pytest
```

test
succeeded 3 days ago in 36s

Search logs

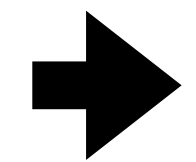
- > ✓ Set up job 1s
- > ✓ Checkout code 1s
- > ✓ Set up Python 9s
- > ✓ Install dependencies 20s
- ✓ Run tests 2s
 - 1 ▶ Run pytest
 - 7 ===== test session starts =====
 - 8 platform linux -- Python 3.11.6, pytest-7.4.3, pluggy-1.3.0
 - 9 rootdir: /home/runner/work/API_backend_P7/API_backend_P7
 - 10 plugins: anyio-3.7.1
 - 11 collected 4 items
 - 12
 - 13 Tests/test_backend.py [100%]
 - 14
 - 15 ===== 4 passed in 1.18s =====
- > ✓ Post Set up Python 0s
- > ✓ Post Checkout code 0s
- > ✓ Complete job 0s

Démonstration de l'API avec modèle un modèle léger

Conclusion

Développement d'un « outil de scoring »

- ✓ Entraînement et mise en place d'un estimateur pour la prise de décision automatique d'accord ou de refus de prêt.
- ✓ Création d'un DashBoard interactif pour le conseiller financier
- ✓ Mise en place d'un pipeline de déploiement continu



Axes d'amélioration :

- Solution technique pour déploiement du véritable estimateur (RandomForestClassifier)
- Solliciter l'avis d'expert financier pour la création de variable et d'un score spécifique plus adapté
- Obtenir un retour de la société de crédit sur l'interface de l'API pour l'amélioration des fonctionnalités

Question ?