Projet 7
-
Implémentez un modèle de scoring

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 –> avril 2024

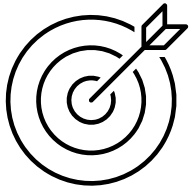
Problématique

L'entreprise « Prêt à dépenser » propose des crédits à la consommation. Elle souhaite mettre en œuvre un «*scoring credit*» pour accorder ses crédits selon la probabilité qu'un client rembourse son crédit.

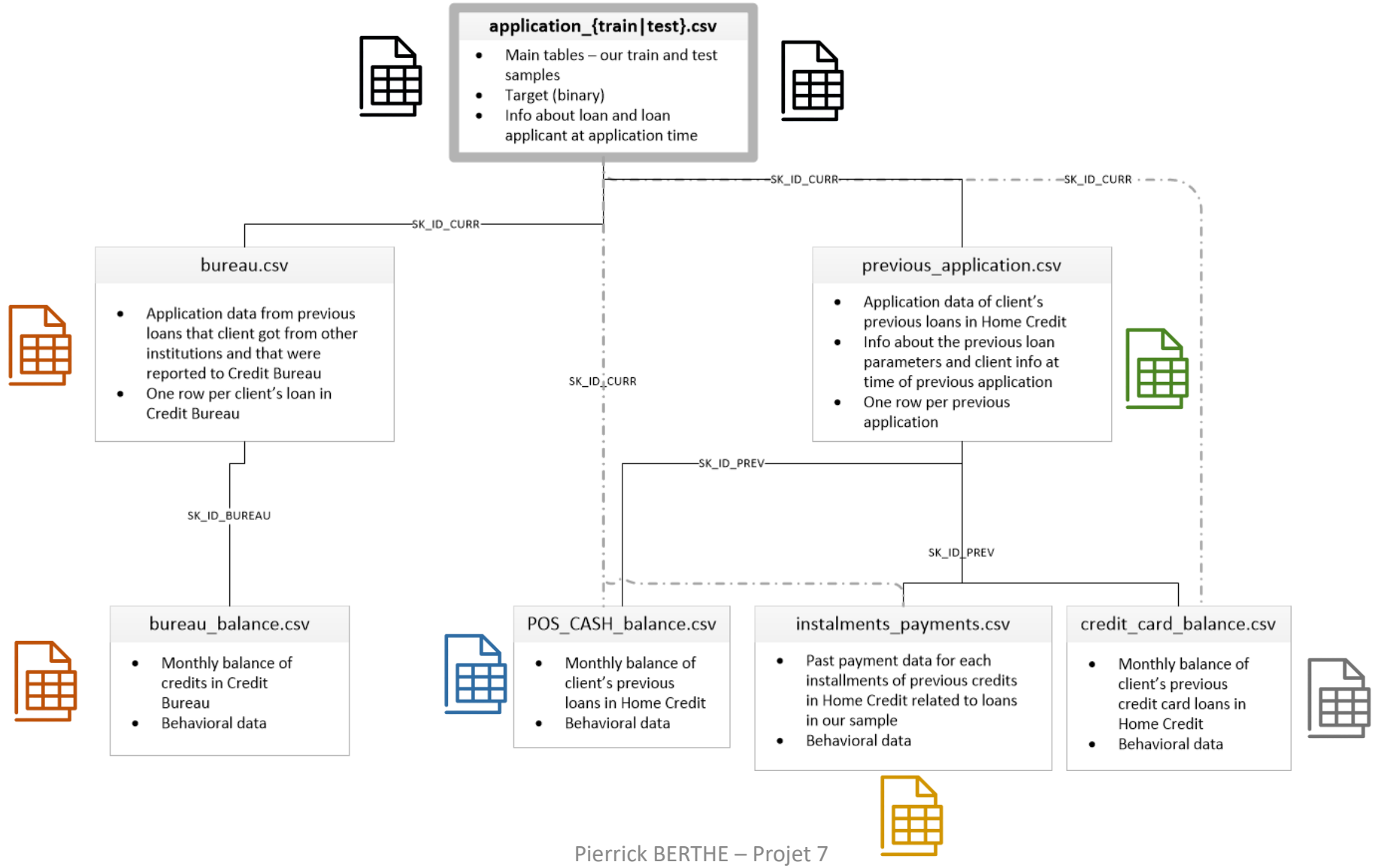
=> Automatiser la prise de décision d'accord de prêt grâce à un algorithme de classification

Missions :

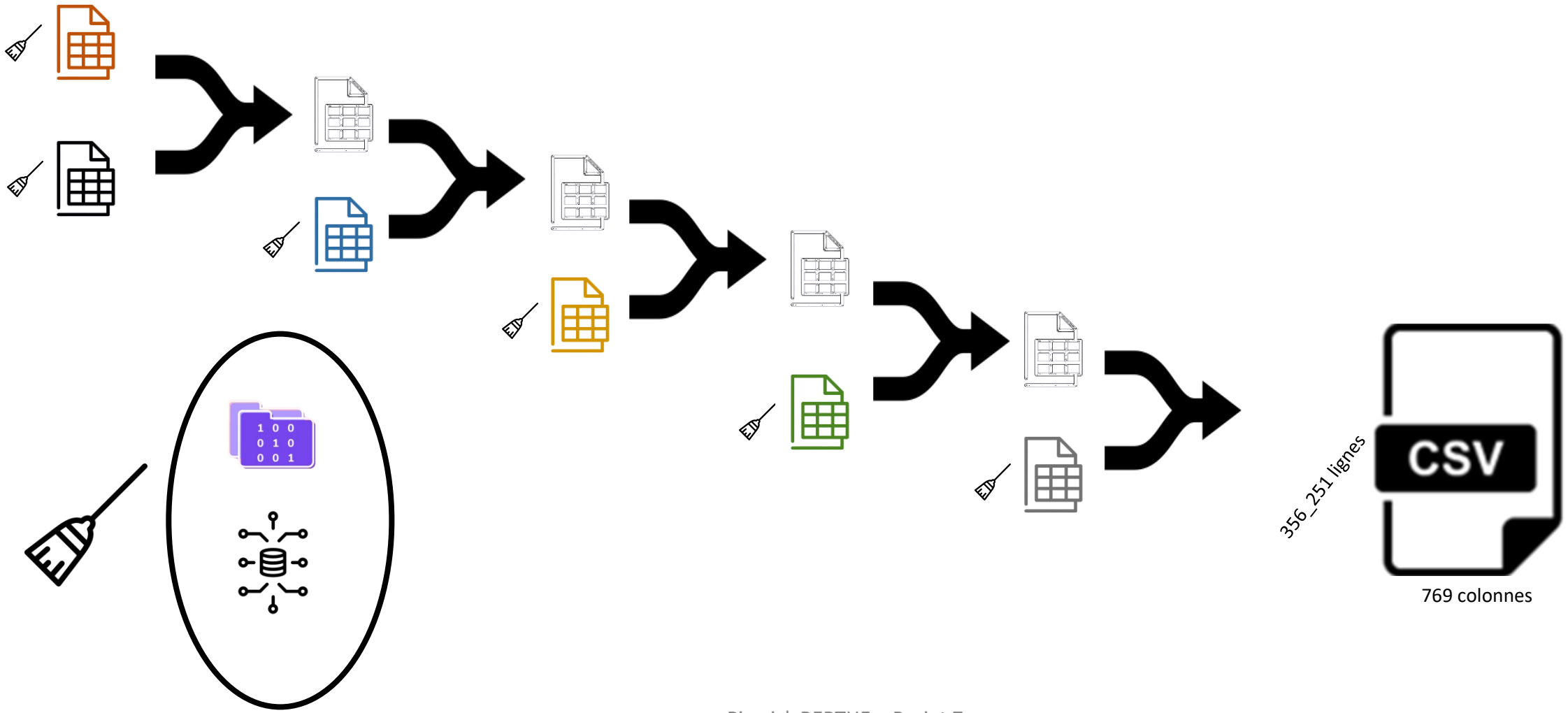
1. Construire le modèle de scoring
2. Analyser les features ayant le plus d'impact sur le scoring de manière générale et au niveau d'un client
3. Mettre en production le modèle de scoring dans une API
4. Mettre en œuvre une approche globale MLOps de bout en bout (tracking expérimentation => data drift)



Présentation du jeu de données



Nettoyage et Feature engineering



Nettoyage et Feature engineering

1. Filtrage valeurs infinies

- 20 lignes

2. Filtrage valeur unique

- 3 features

3. Filtrage taux de remplissage (<50%)

- 224 features

4. Transformation bool => int

5. Filtrage features corrélées (r > 0,50)

- 192 features

356_231 lignes

CSV

350 colonnes

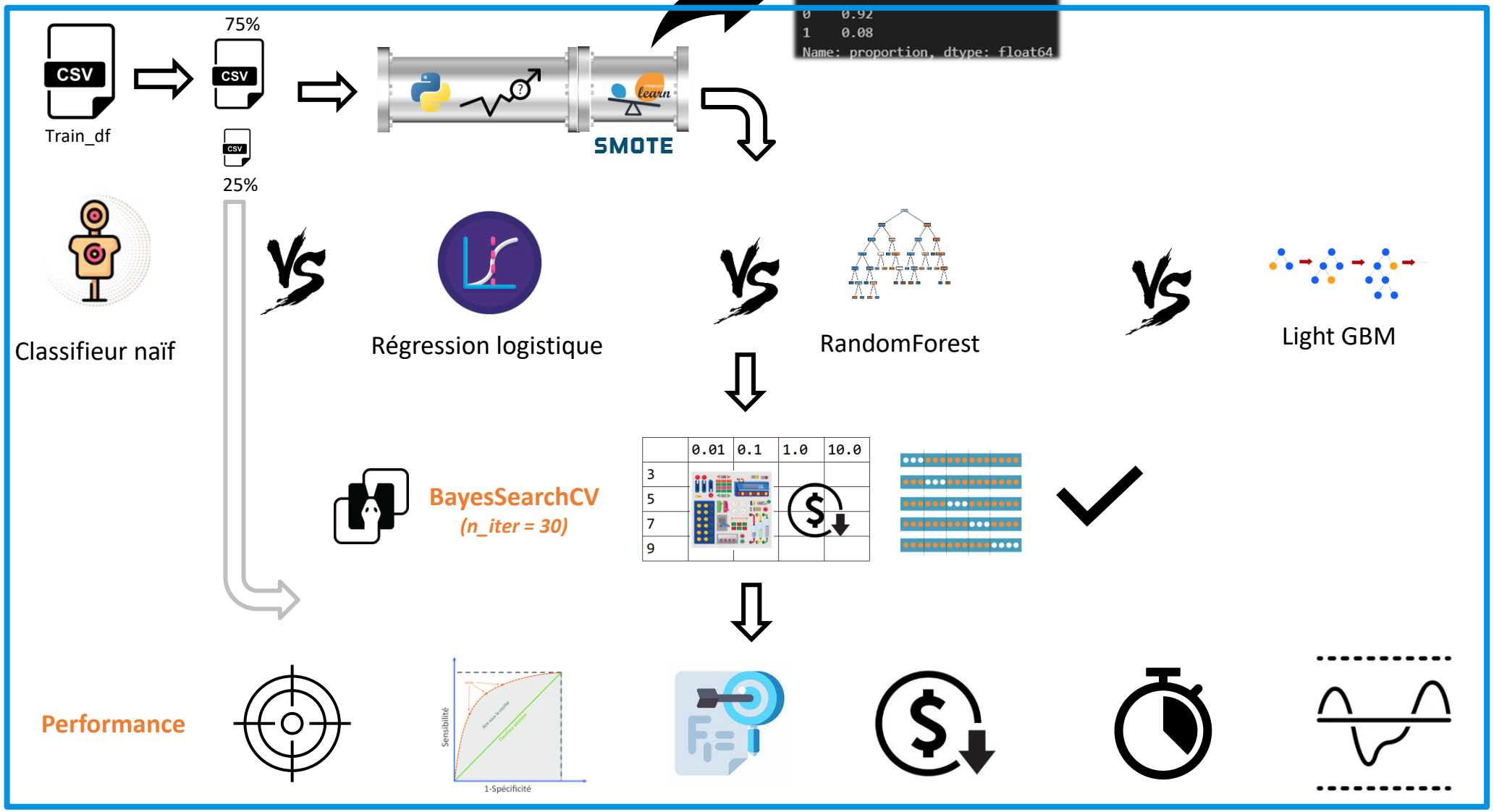
```
Dimensions train_df: (307488, 350)
Le dataset de train représente 86.3% du total.
```

train_df

test_df

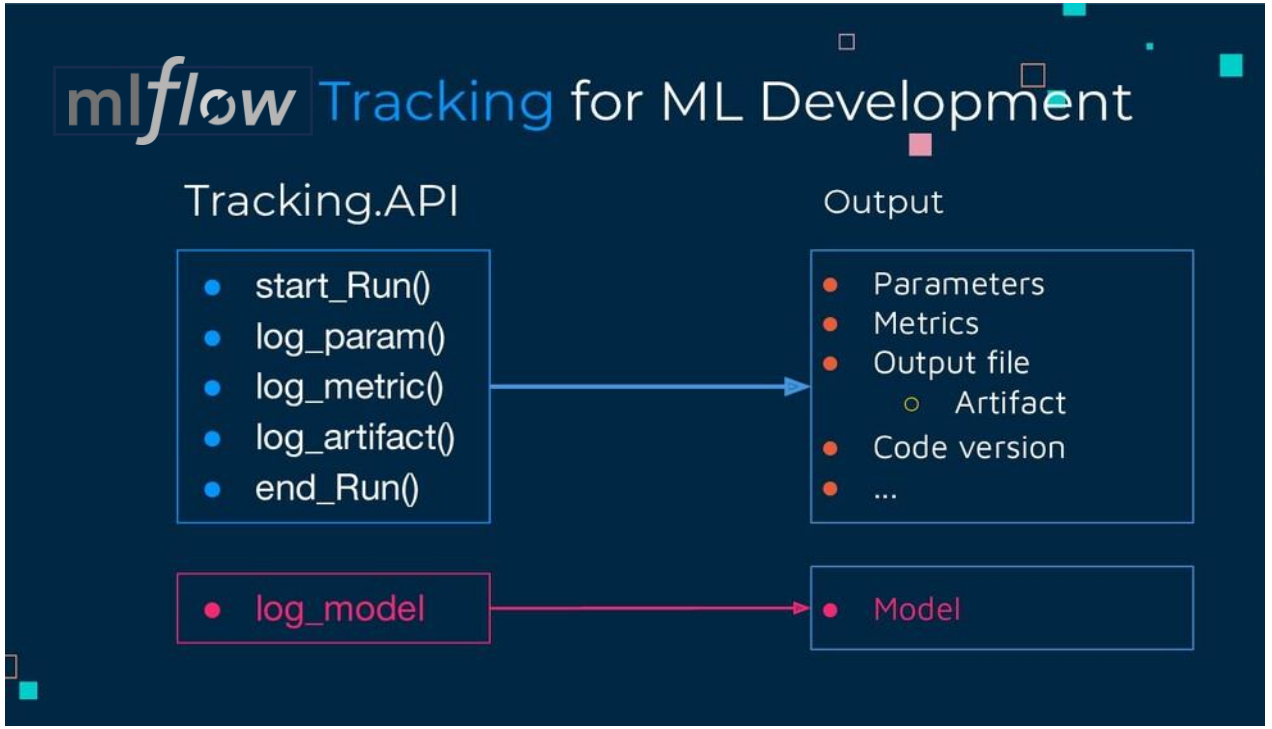
```
Dimensions test_df: (48743, 350)
Le dataset de test représente 13.7% du total.
```

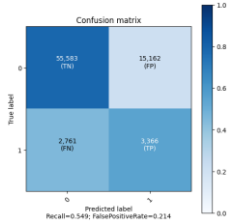
Modélisation

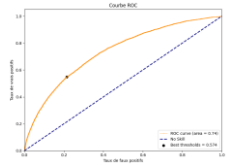


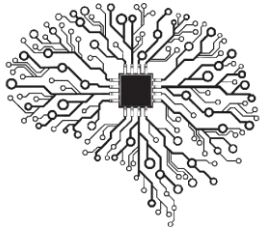
mlflow

Modélisation

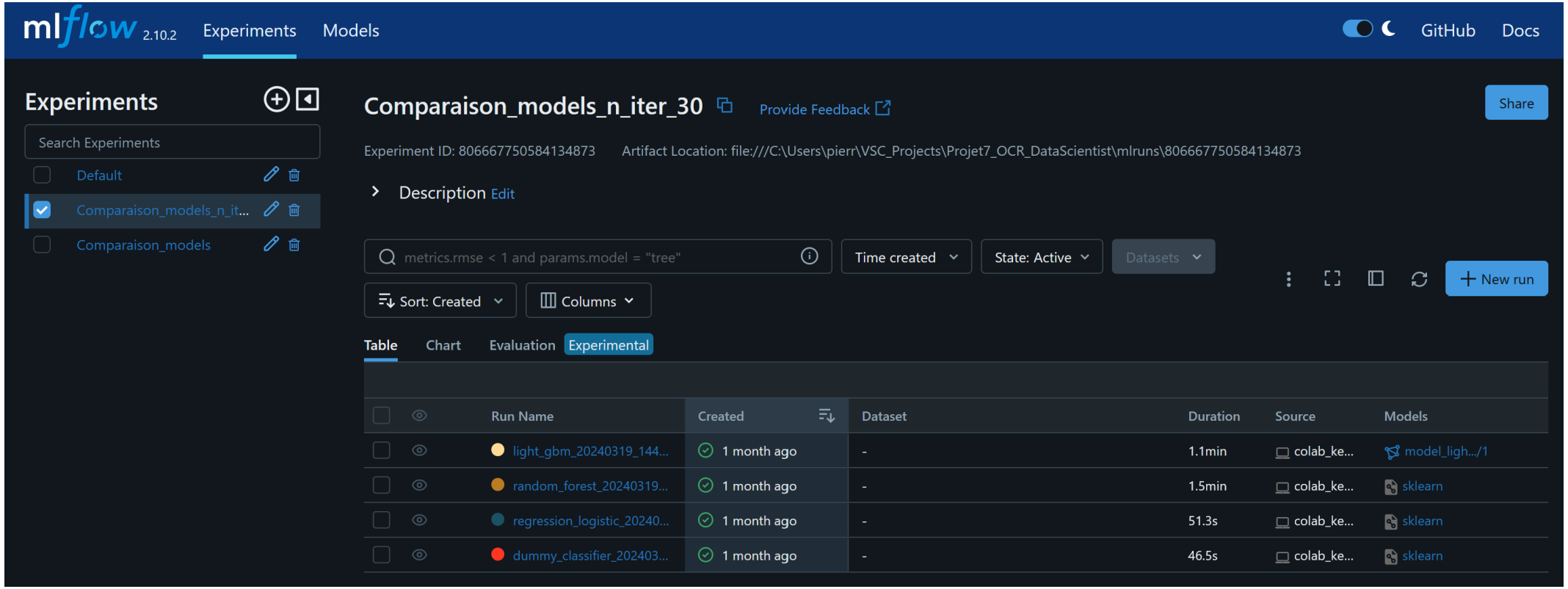








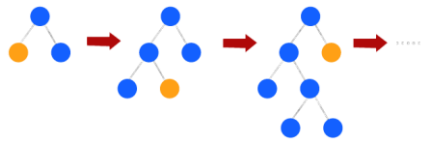
Modélisation



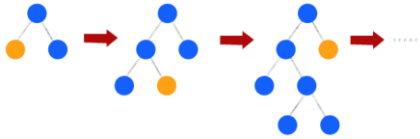
MLflow UI : http://127.0.0.1:5000

Modélisation

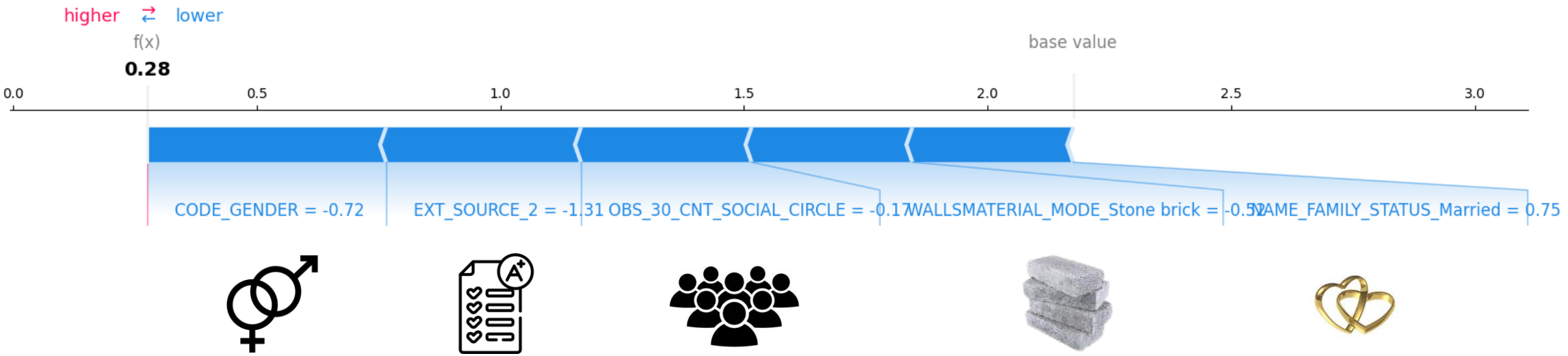
	accuracy	auc	f1_score	metier_score	temps_fit	temps_predict	seuil_predict
dummy_classifier	0.92	0.50	0.00	0.00	31.12	0.32	2.00
regression_logistic	0.77	0.74	0.27	0.54	43.91	0.37	0.57
random_forest	0.66	0.67	0.21	0.55	79.43	0.71	0.43
light_gbm	0.69	0.76	0.26	0.67	55.54	0.93	0.08



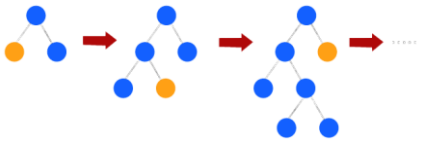
Modélisation



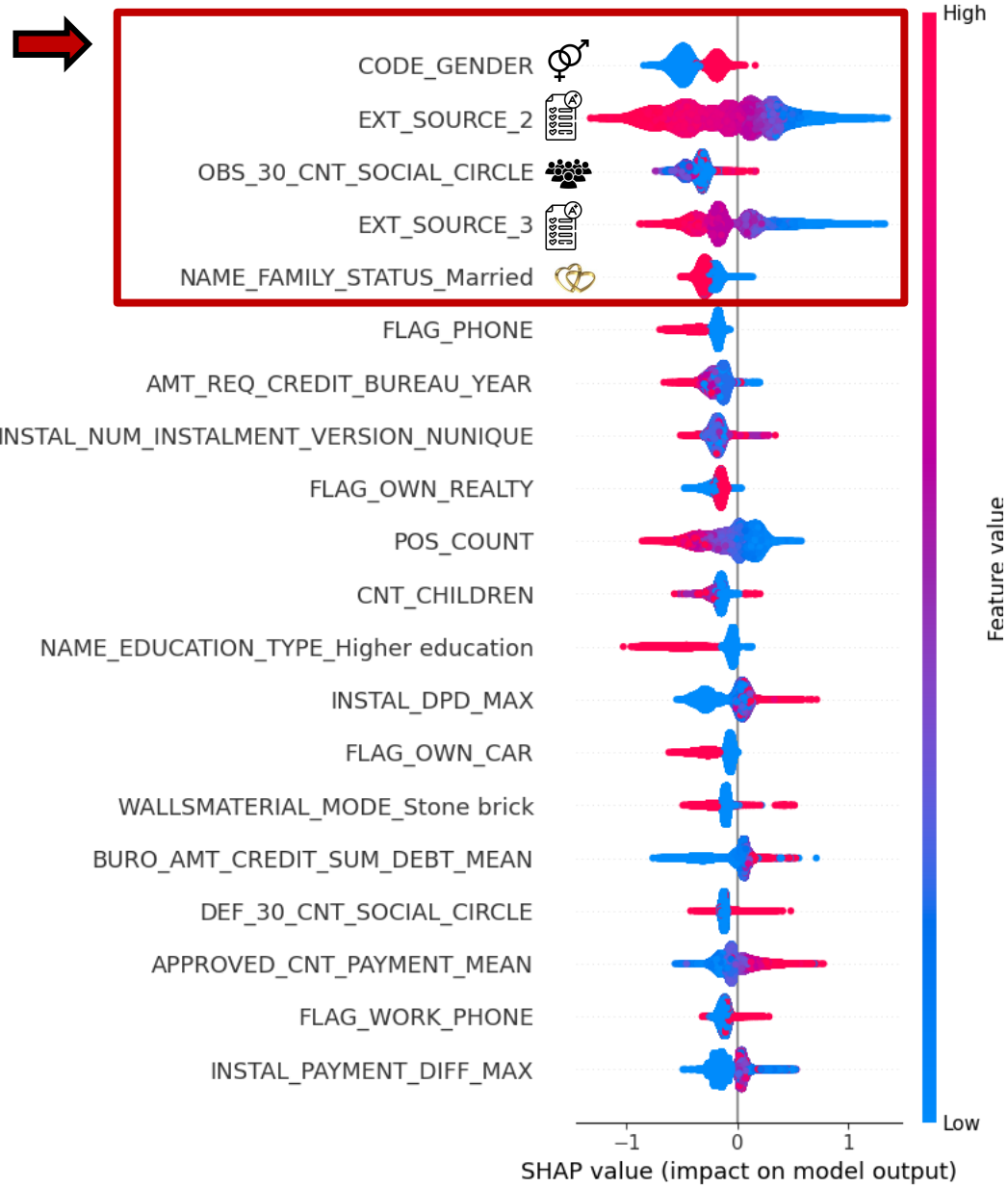
Feature importance locale



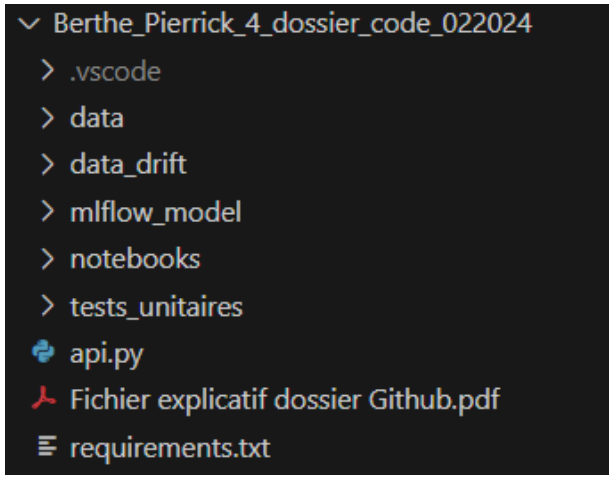
Modélisation



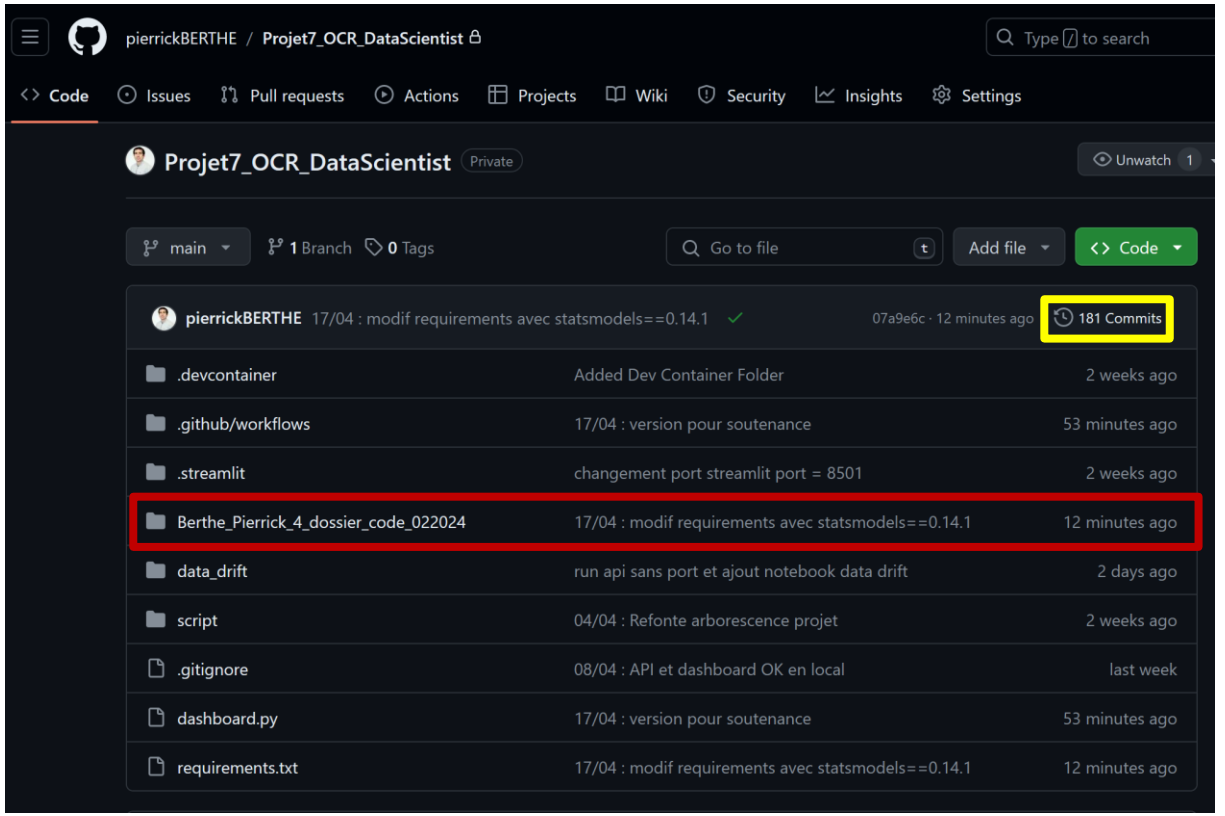
Feature importance globale



Pipeline de déploiement



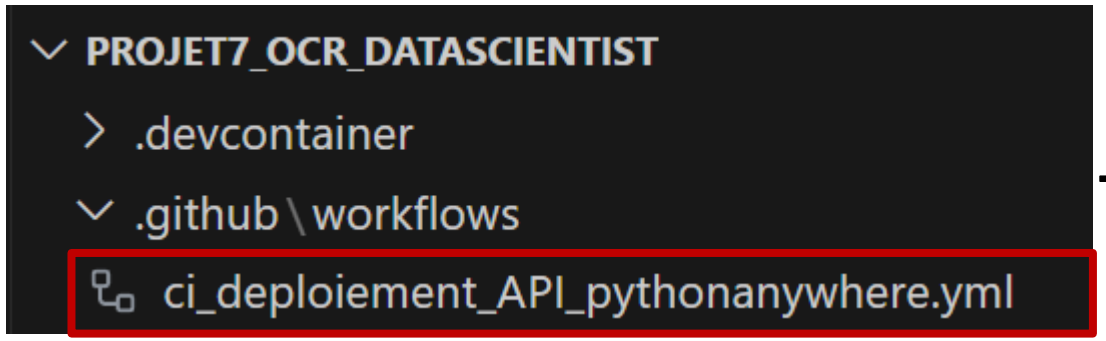
GitHub



Lien GitHub:
https://github.com/pierrickBERTHE/Projet7_OCR_DataScientist

Pipeline de déploiement

GitHub Actions



Push
- Dossier sur Github

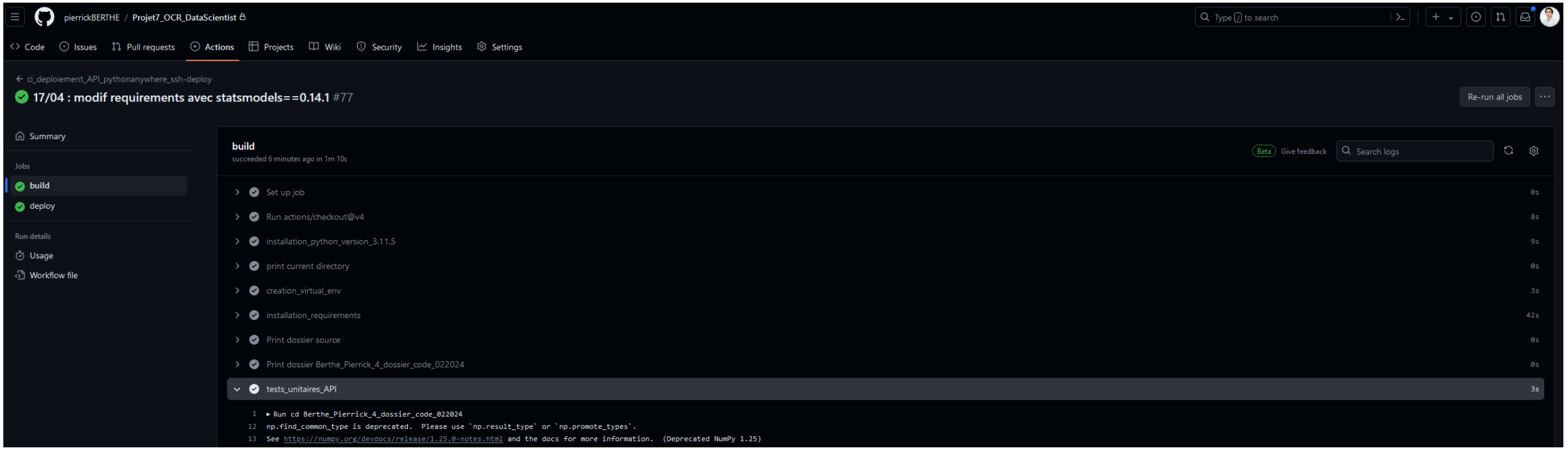
build
- Installation python 3.11.5
- Création env
- Installation requirements
- Tests unitaires

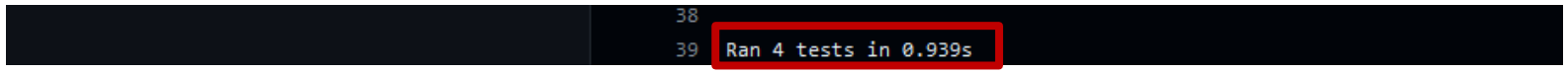
deploy
- Déploiement sur pythonanywhere avec ssh

pythonanywhere

Pipeline de déploiement

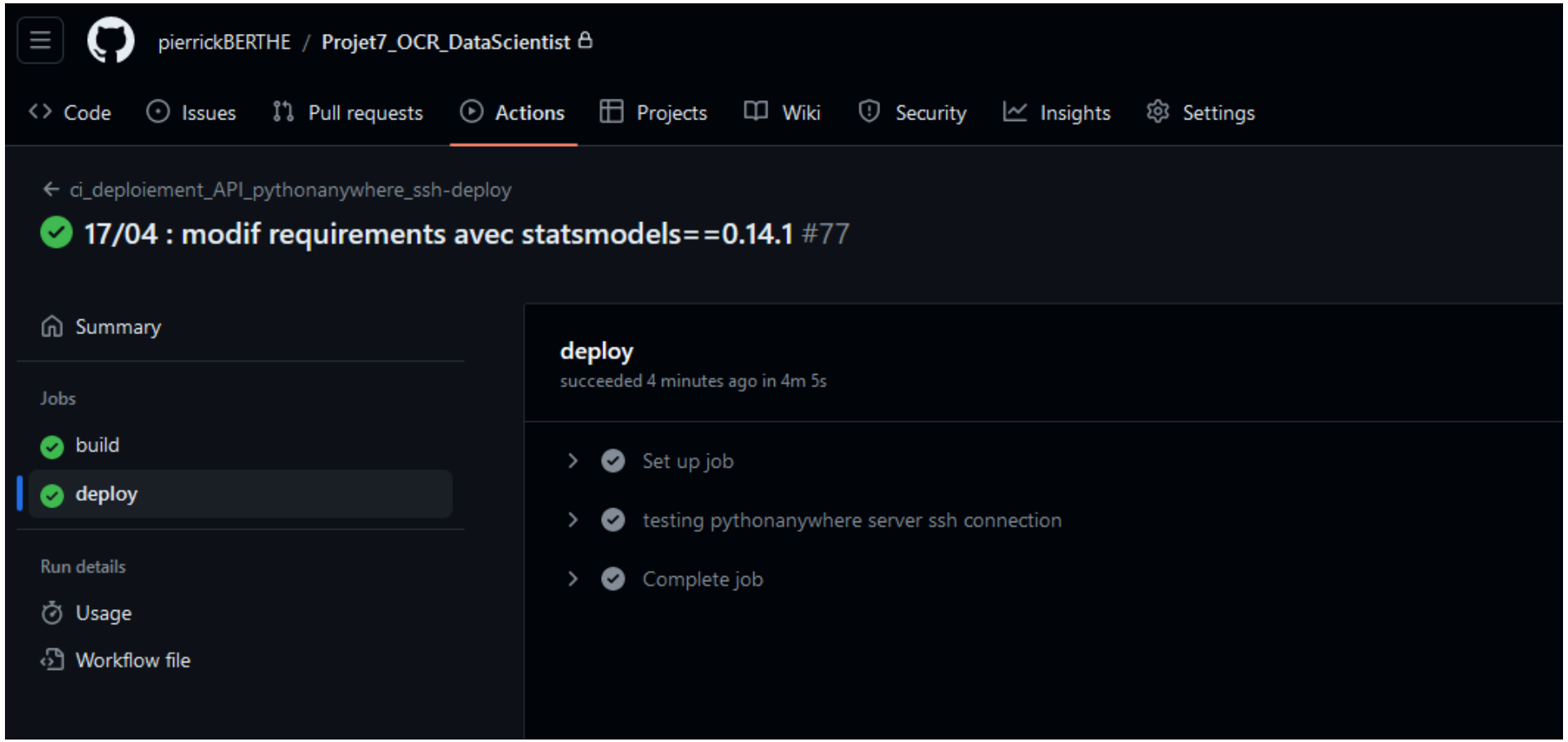
GitHub Actions



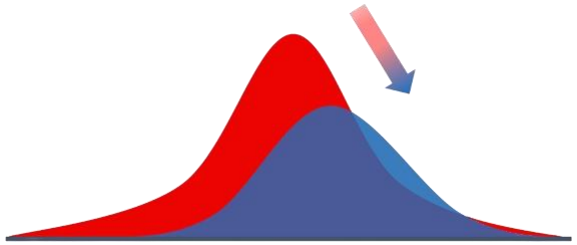


Pipeline de déploiement

GitHub Actions



Data drift



EVIDENTLY AI

?

train_df

307_488 lignes

350 colonnes

test_df

48_743 lignes

350 colonnes

Lien page HTML:

Berthe_Pierrick_5_Tableau_HTML_data_drift_evidently_022024.html

Dataset Drift

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

348	99	0.284
Columns	Drifted Columns	Share of Drifted Columns

Data Drift Summary

Dashboard Streamlit

API cloud :
http://pierrickberthe.eu.pythonanywhere.com/

Test de l'API:
Lien vers Dashboard

Conclusion

Missions :

1. Construire le modèle de scoring ✓
2. Analyser les features ayant le plus d'impact sur le scoring de manière générale et au niveau d'un client ✓
3. Mettre en production le modèle de scoring dans une API ✓
4. Mettre en œuvre une approche globale ML Ops de bout en bout (tracking expérimentation => data drift) ✓

mlflow

EVIDENTLY AI

Conclusion

Limites :

- Manque de connaissance métier pour comprendre toutes les features
- Dataset sur 10% des clients

Merci pour votre attention

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 –> avril 2024