



OPENCLASSROOMS

# Projet 7

-

## Implémentez un modèle de scoring



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science  
*Openclassrooms – CentraleSupélec*

*août 2023 → avril 2024*



# Problématique

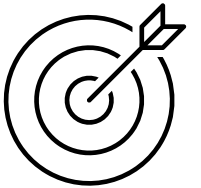
L'entreprise « Prêt à dépenser » propose des crédits à la consommation. Elle souhaite mettre en œuvre un «*scoring credit*» pour accorder ses crédits selon la probabilité qu'un client rembourse son crédit.



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

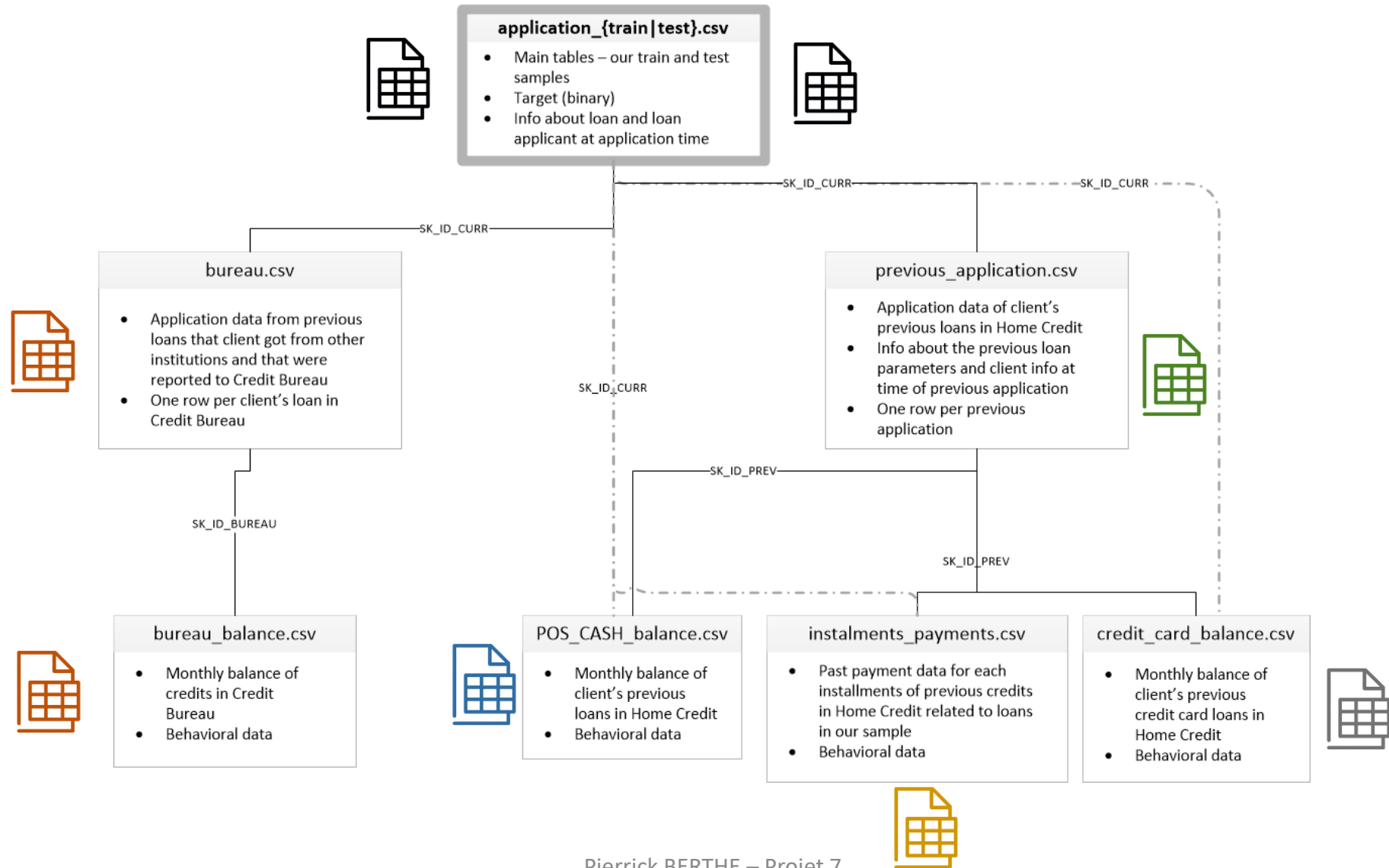
## Missions :

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)



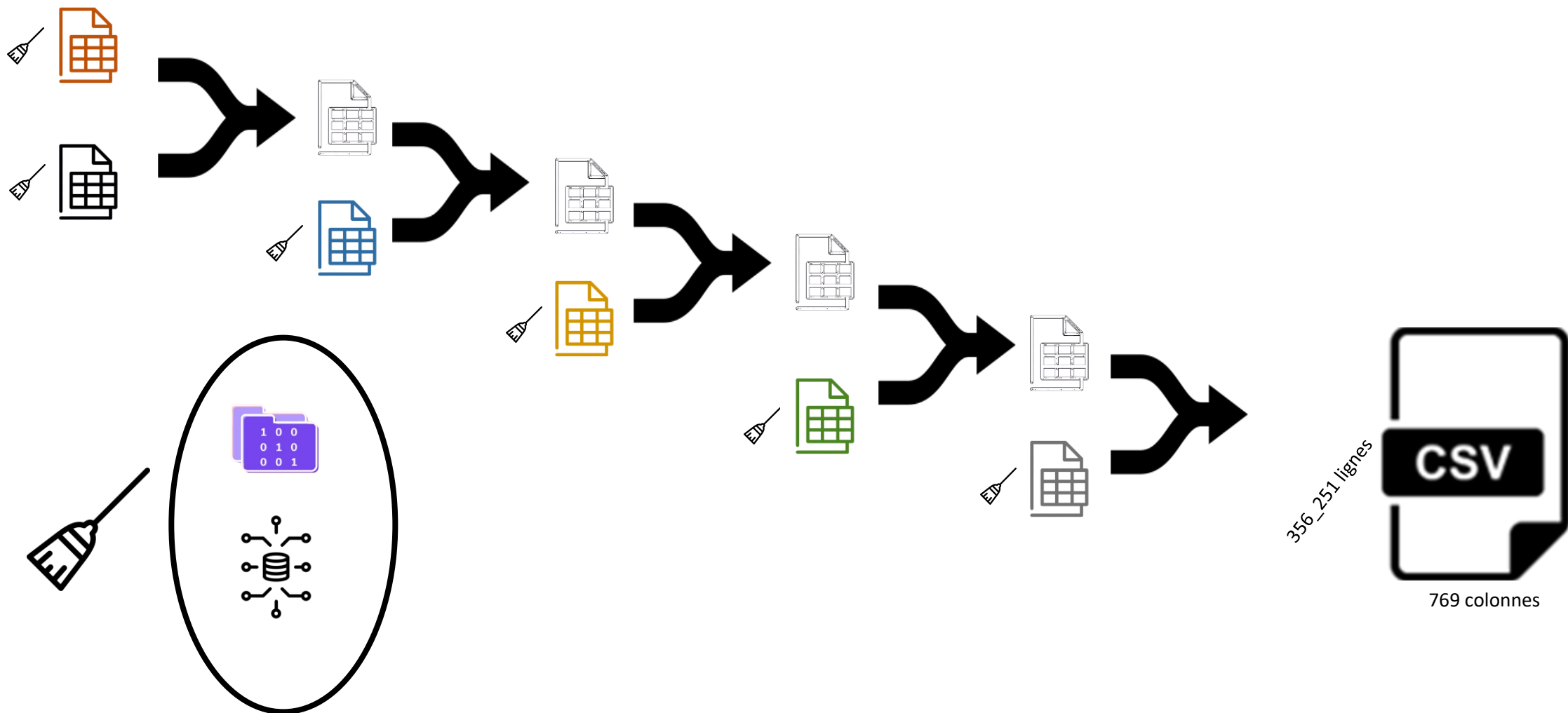


# Présentation du jeu de données



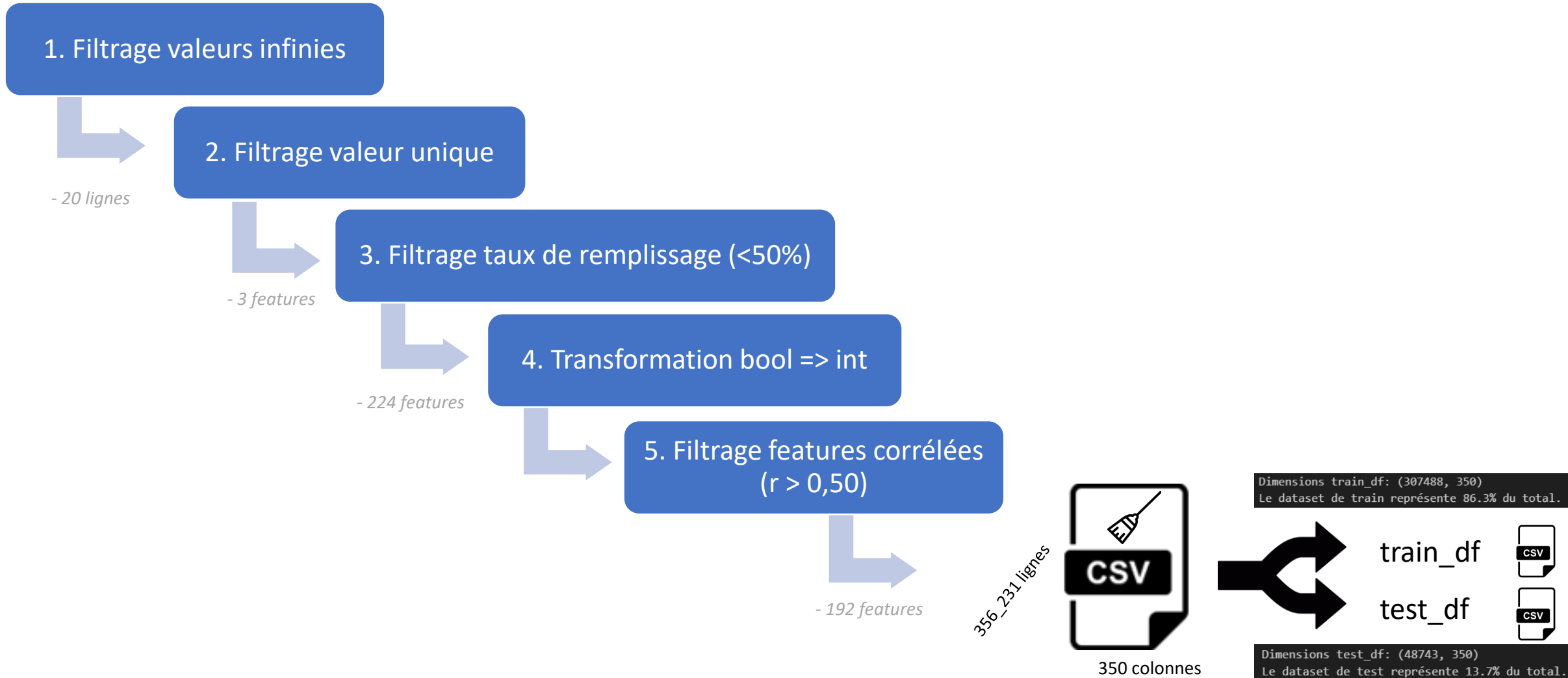


# Nettoyage et Feature engineering



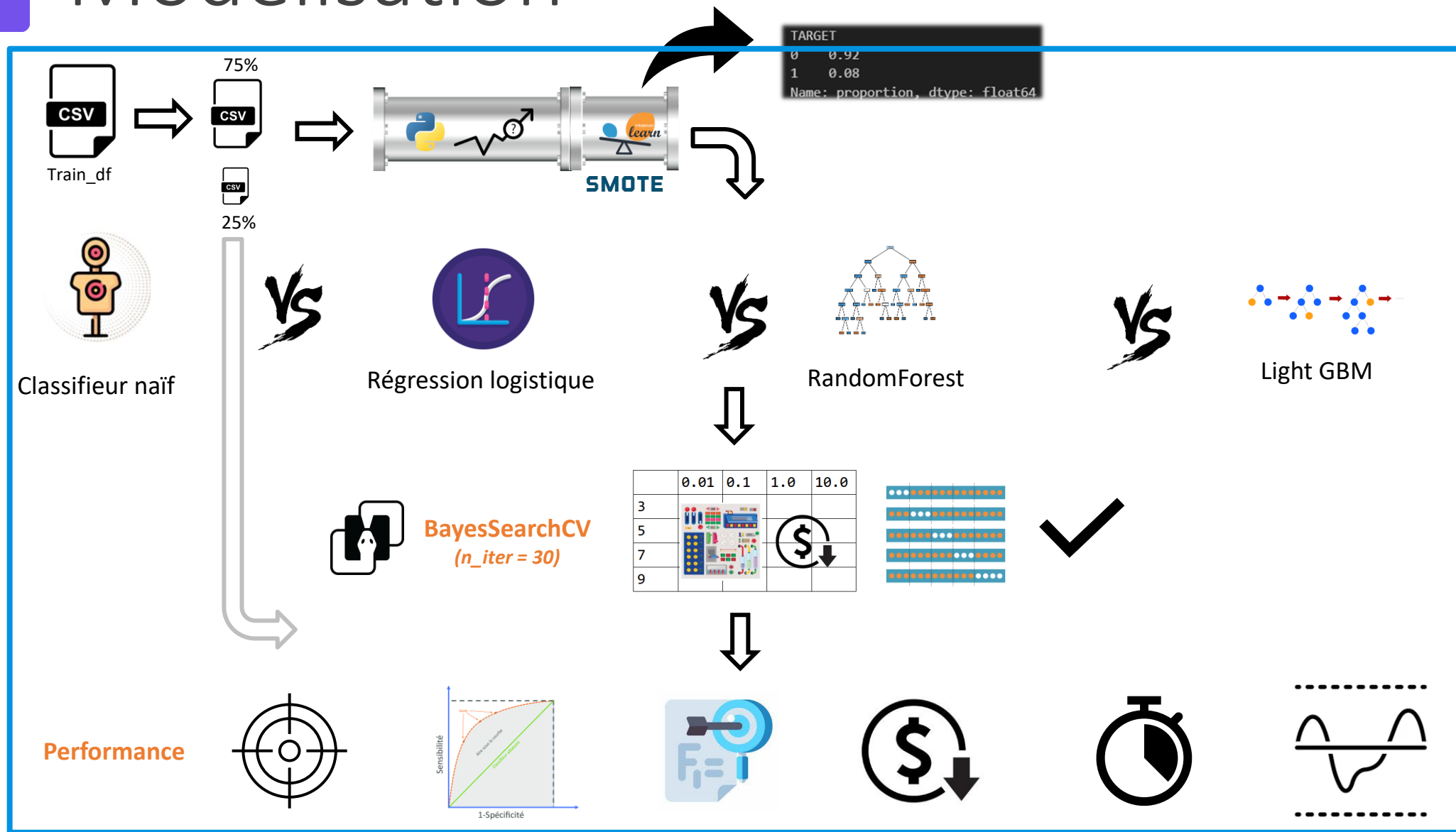


# Nettoyage et Feature engineering



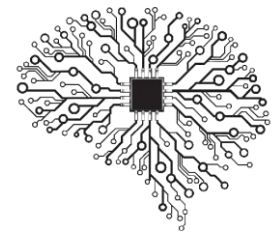
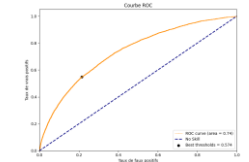
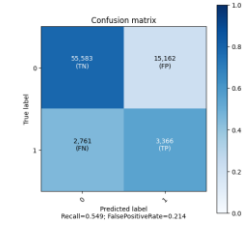
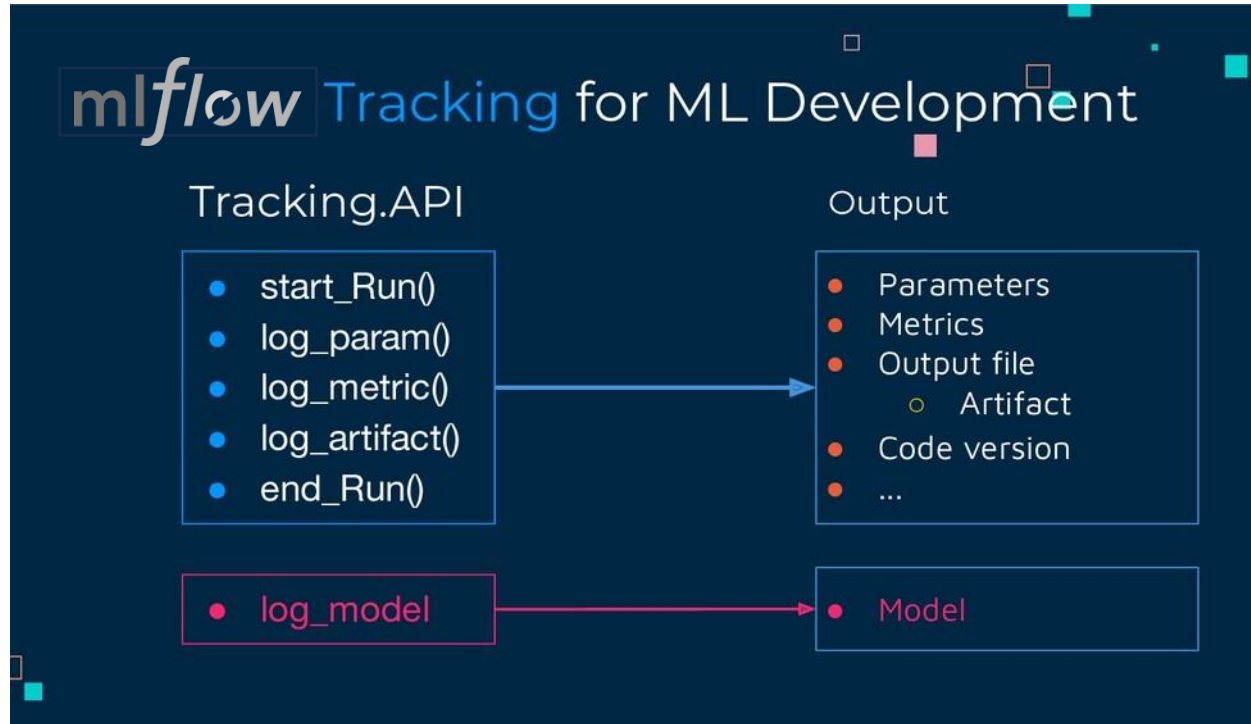


# Modélisation





# Modélisation





# Modélisation

mlflow 2.10.2 Experiments Models

Search Experiments

- ☐ Default
- ☒ Comparaison\_models\_n\_iter\_30
- ☐ Comparaison\_models

Comparaison\_models\_n\_iter\_30 Provide Feedback

Experiment ID: 806667750584134873 Artifact Location: file:///C:/Users/pierr/VSC\_Projects/Projet7\_OCR\_DataScientist/mlruns/806667750584134873

> Description Edit

metrics.rmse < 1 and params.model = "tree"

Time created State: Active Datasets

Sort: Created Columns

Table Chart Evaluation Experimental

		Run Name	Created	Dataset	Duration	Source	Models
<input type="checkbox"/>	<input type="checkbox"/>	light_gbm_20240319_144...	1 month ago	-	1.1min	colab_ke...	model_ligh.../1
<input type="checkbox"/>	<input type="checkbox"/>	random_forest_20240319...	1 month ago	-	1.5min	colab_ke...	sklearn
<input type="checkbox"/>	<input type="checkbox"/>	regression_logistic_20240...	1 month ago	-	51.3s	colab_ke...	sklearn
<input type="checkbox"/>	<input type="checkbox"/>	dummy_classifier_202403...	1 month ago	-	46.5s	colab_ke...	sklearn

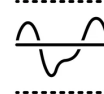
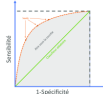


MLflow UI : <http://127.0.0.1:5000>

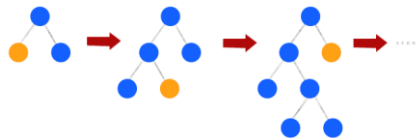




# Modélisation



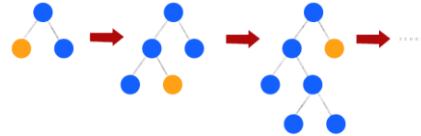
	accuracy	auc	f1_score	metier_score	temps_fit	temps_predict	seuil_predict
dummy_classifier	0.92	0.50	0.00	0.00	31.12	0.32	2.00
regression_logistic	0.77	0.74	0.27	0.54	43.91	0.37	0.57
random_forest	0.66	0.67	0.21	0.55	79.43	0.71	0.43
light_gbm	0.69	0.76	0.26	0.67	55.54	0.93	0.08



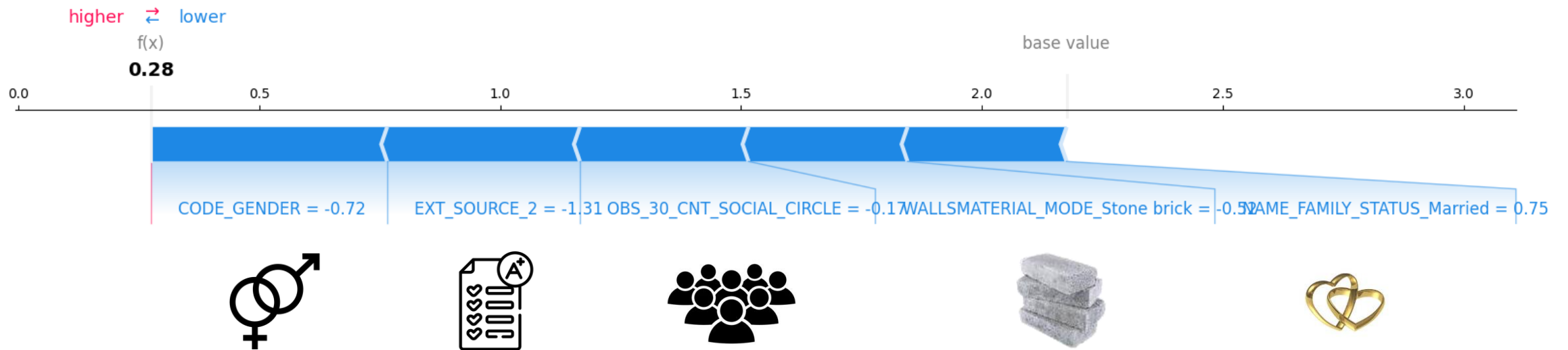
mlflow



# Modélisation

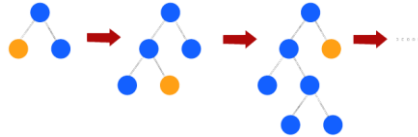


## Feature importance locale

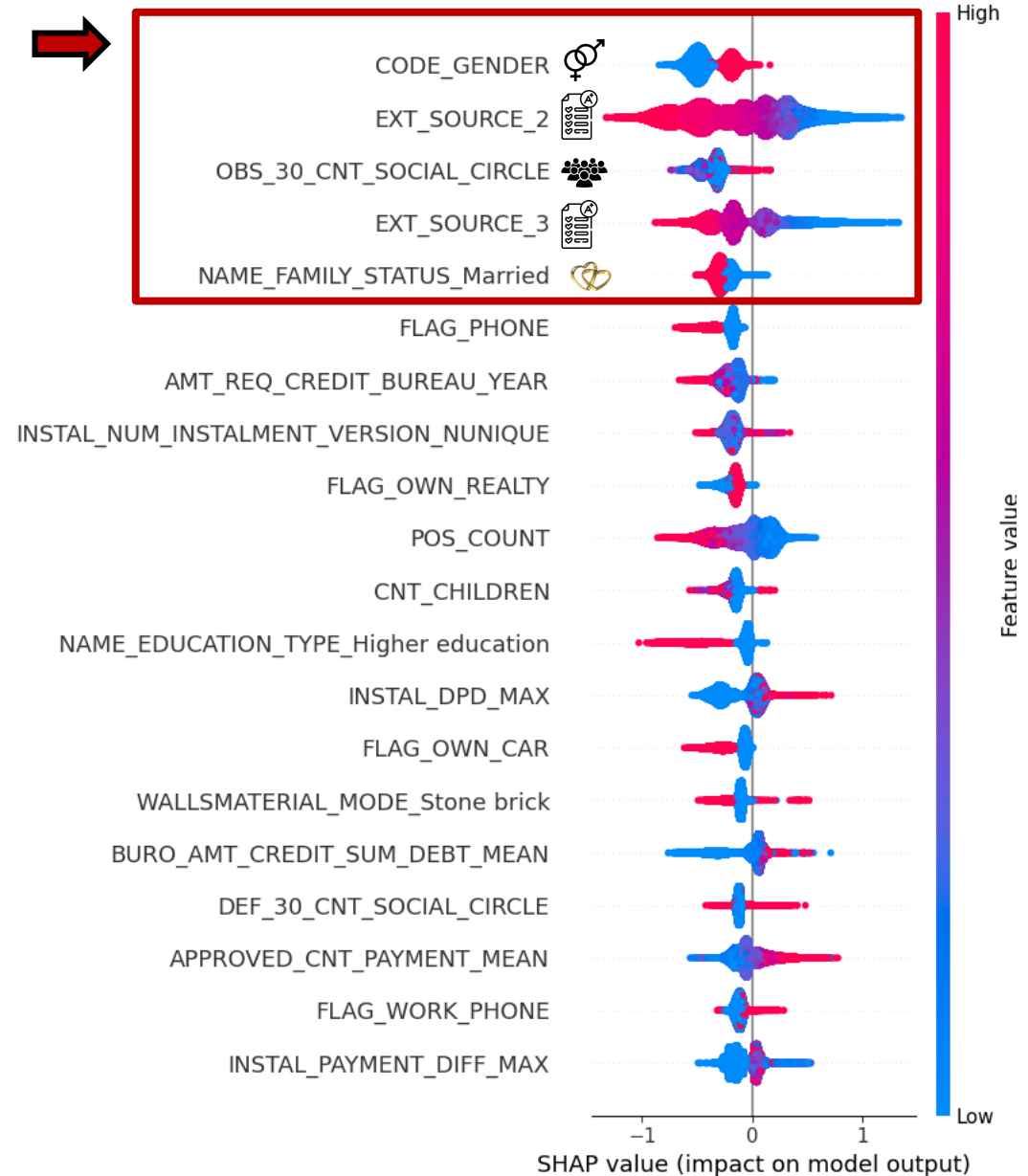




# Modélisation



Feature importance globale





# Pipeline de déploiement



```
✓ Berthe_Pierrick_4_dossier_code_022024
> .vscode
> data
> data_drift
> mlflow_model
> notebooks
> tests_unitaires
+ api.py
+ Fichier explicatif dossier Github.pdf
+ requirements.txt
```



The screenshot shows the GitHub repository page for 'pierrickBERTHE / Projet7\_OCR\_DataScientist'. The repository is private and has 1 branch and 0 tags. The commit history is displayed, with the most recent commit highlighted in a red box. The commit message is '17/04 : modif requirements avec statsmodels==0.14.1' and it was made 12 minutes ago. The commit hash is 07a9e6c. The commit is associated with the file 'Berthe\_Pierrick\_4\_dossier\_code\_022024', which is also highlighted in a red box. The file was added 2 weeks ago. Other files listed include .devcontainer, .github/workflows, .streamlit, data\_drift, script, .gitignore, dashboard.py, and requirements.txt.

File	Commit Message	Time
Berthe_Pierrick_4_dossier_code_022024	17/04 : modif requirements avec statsmodels==0.14.1	12 minutes ago
requirements.txt	17/04 : modif requirements avec statsmodels==0.14.1	12 minutes ago
dashboard.py	17/04 : version pour soutenance	53 minutes ago
script	04/04 : Refonte arborescence projet	2 weeks ago
data_drift	run api sans port et ajout notebook data drift	2 days ago
.streamlit	changement port streamlit port = 8501	2 weeks ago
.github/workflows	17/04 : version pour soutenance	53 minutes ago
.devcontainer	Added Dev Container Folder	2 weeks ago



Lien GitHub:

[https://github.com/pierrickBERTHE/Projet7\\_OCR\\_DataScientist](https://github.com/pierrickBERTHE/Projet7_OCR_DataScientist)



# Pipeline de déploiement



GitHub Actions

✓ **PROJET7\_OCR\_DATASCIENTIST**

> .devcontainer

✓ .github\workflows

ci\_deploiement\_API\_pythonanywhere.yml

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

pierrickBERTHE / Projet7\_OCR\_DataScientist

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

17/04 : modif requirements avec statsmodels==0.14.1 #77

Summary

Jobs

- build
- deploy

Run details

Usage

Workflow file

**build**  
succeeded 6 minutes ago in 1m 10s

Set up job 0s

Run actions/checkout@v4 8s

installation\_python\_version\_3.11.5 9s

print current directory 0s

creation\_virtual\_env 3s

installation\_requirements 42s

Print dossier source 0s

Print dossier Berthe\_Pierrick\_4\_dossier\_code\_022024 0s

tests\_unitaires\_API 3s

```
1 ▶ Run cd Berthe_Pierrick_4_dossier_code_022024
12 np.find_common_type is deprecated. Please use `np.result_type` or `np.promote_types`.
13 See https://numpy.org/devdocs/release/1.25.0-notes.html and the docs for more information. (Deprecated NumPy 1.25)
```

```
38
39 Ran 4 tests in 0.939s
```



# Pipeline de déploiement



GitHub Actions

pierrickBERTHE / Projet7\_OCR\_DataScientist

<> Code Issues Pull requests **Actions** Projects Wiki Security Insights Settings

← ci\_deploiement\_API\_pythonanywhere\_ssh-deploy

✓ 17/04 : modif requirements avec statsmodels==0.14.1 #77

Summary

Jobs

- ✓ build
- ✓ **deploy**

Run details

- Usage
- Workflow file

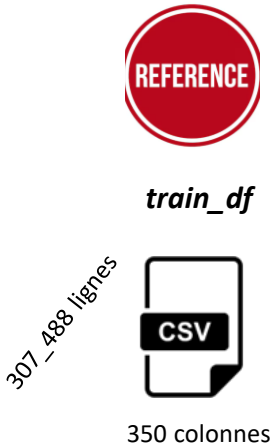
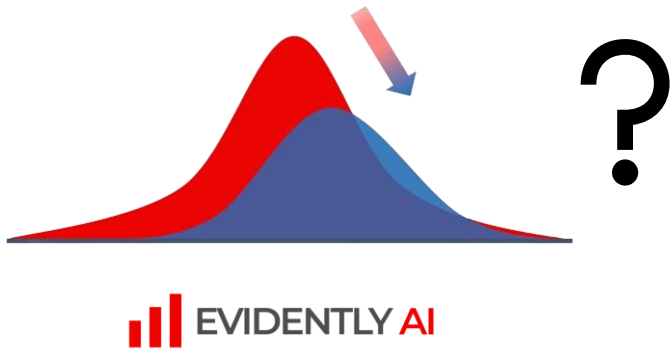
**deploy**

succeeded 4 minutes ago in 4m 5s

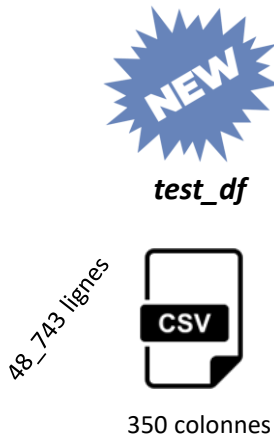
- > ✓ Set up job
- > ✓ testing pythonanywhere server ssh connection
- > ✓ Complete job



# Data drift



VS



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 Columns	99 Drifted Columns	0.284 Share of Drifted Columns
Data Drift Summary		





# Dashboard Streamlit

API cloud :

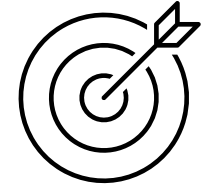
<http://pierrickberthe.eu.pythonanywhere.com/>

Test de l'API:

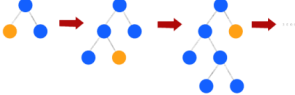



[Lien vers Dashboard](#)



# 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  de bout en bout (tracking experimentation => data drift) ✓

mlflow

EVIDENTLY AI



# Conclusion



## Limites :

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

OPENCLASSROOMS

Merci pour votre attention



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science  
*Openclassrooms – CentraleSupélec*

*août 2023 → avril 2024*