PROJET_07 : IMPLÉMENTEZ UN MODÈLE DE SCORING

Jérôme LE GAL
Etudiant OpenClassRooms – parcours Data Scientist
Le 08/12/2024

CONTEXTE ET OBJECTIFS

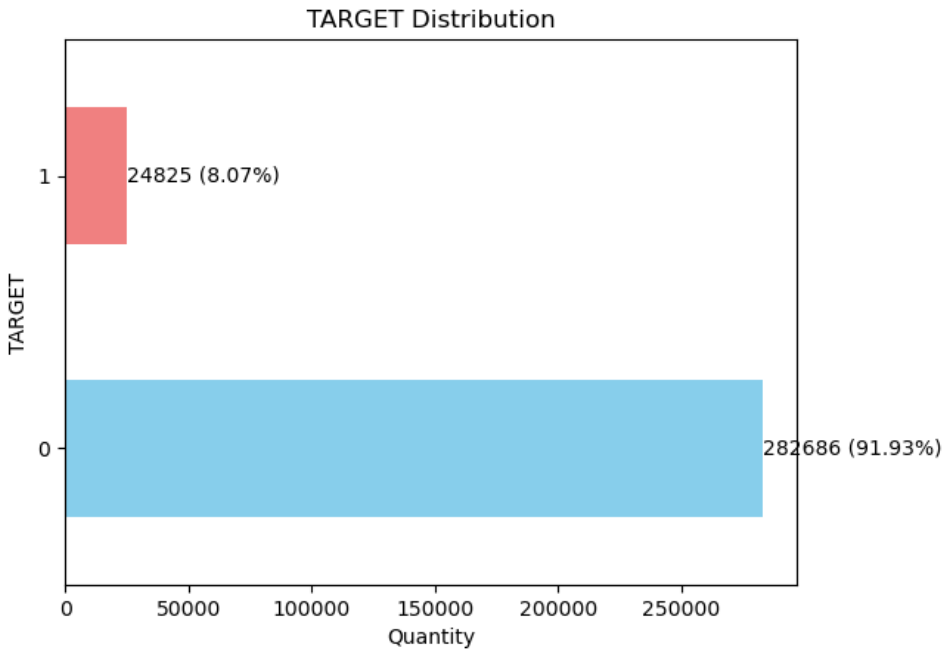
- Mission pour : « Prêt à dépenser »
- Demandes en crédit accordée ou refusée

- Modèle de scoring
- API et interface graphique déployées sur le Cloud
- Approche MLOps

DESCRIPTION DU JEU DE DONNÉES

- Fichiers csv :
  - Train / Test
  - Informations clients
  - Informations de prêts
  - Description
  - …

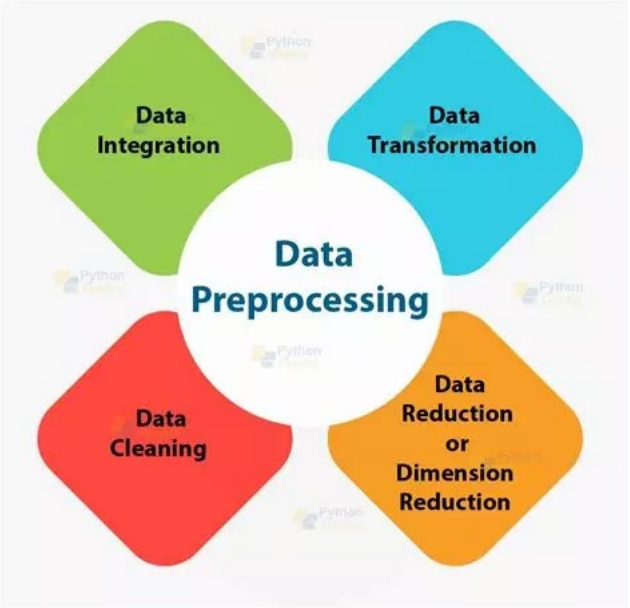
ANALYSE EXPLORATOIRE



- Distribution ‘target’ :
- Données manquantes
  - Max : ~70%
- Corrélations : Target vs Features
  - *Coef Pearson < 0.18*

PREPROCESSING

- Preprocessing :
  - Agrégation (min, max, sum, mean)
  - Encodage : OneHotEncoder, LabelEncoder
  - Transformations : StandardScaler
  - PCA



- Datasets finaux (avant PCA) :
  - Train shape : (307507, 695)
  - Test shape : (48744, 694)

MODÉLISATIONS :

STRATÉGIE

1. Evaluation des contraintes
2. Méthode d'évaluation des modèles
3. Choix modèles à optimiser
4. Méthode de « *tracking* »
5. Entrainements (GPU) et choix modèle final
6. Analyse « *Feature Importance* »

CONTRAINTES ET SCORING

<u>Contraintes :</u>

- Déséquilibre de la Target
- Perte de coûts : FN = 10 x FP

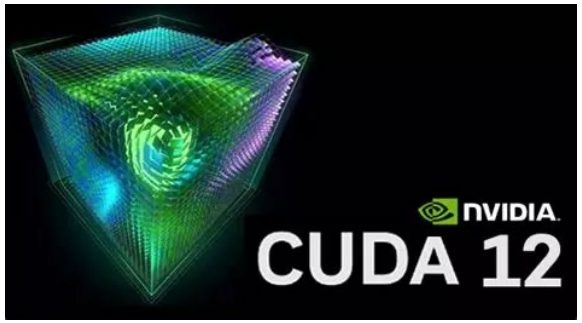
<u>Evaluation :</u>

- ✓ Score métier (make_scorer)
- ✓ Matrice de confusion
- ✓ AUC-ROC
- ✓ Temps de calcul

MODÈLES

- Baseline – LogisticRegression :
  - Imputation, standardisation
  - *Basic*
  - *SMOTE*
  - *Class_weight*
- RandomForestClassifier + Smote (*n_estimators, max_depth, max_features, sampling_strategy*)
- LightGBM + Smote (*num_leaves, max_depth, min_data_in_leaf, sampling_strategy*)
- XGBoost + Smote (*max_depth, gamma, sampling_strategy*)

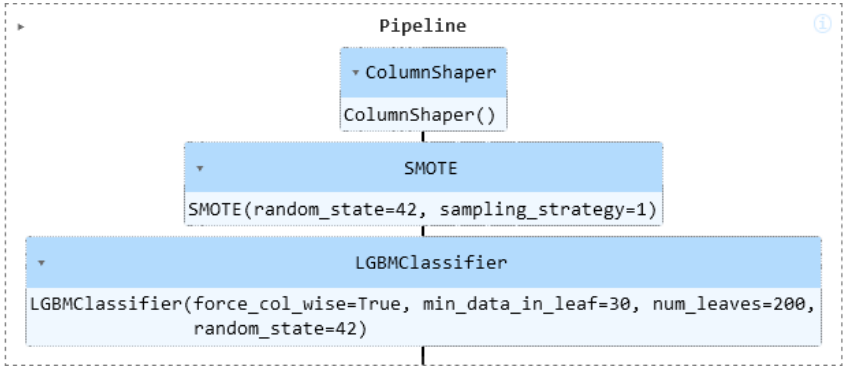
❖ Optimisation : GridSearchCV sur GPU



SYNTHÈSE DES RÉSULTATS

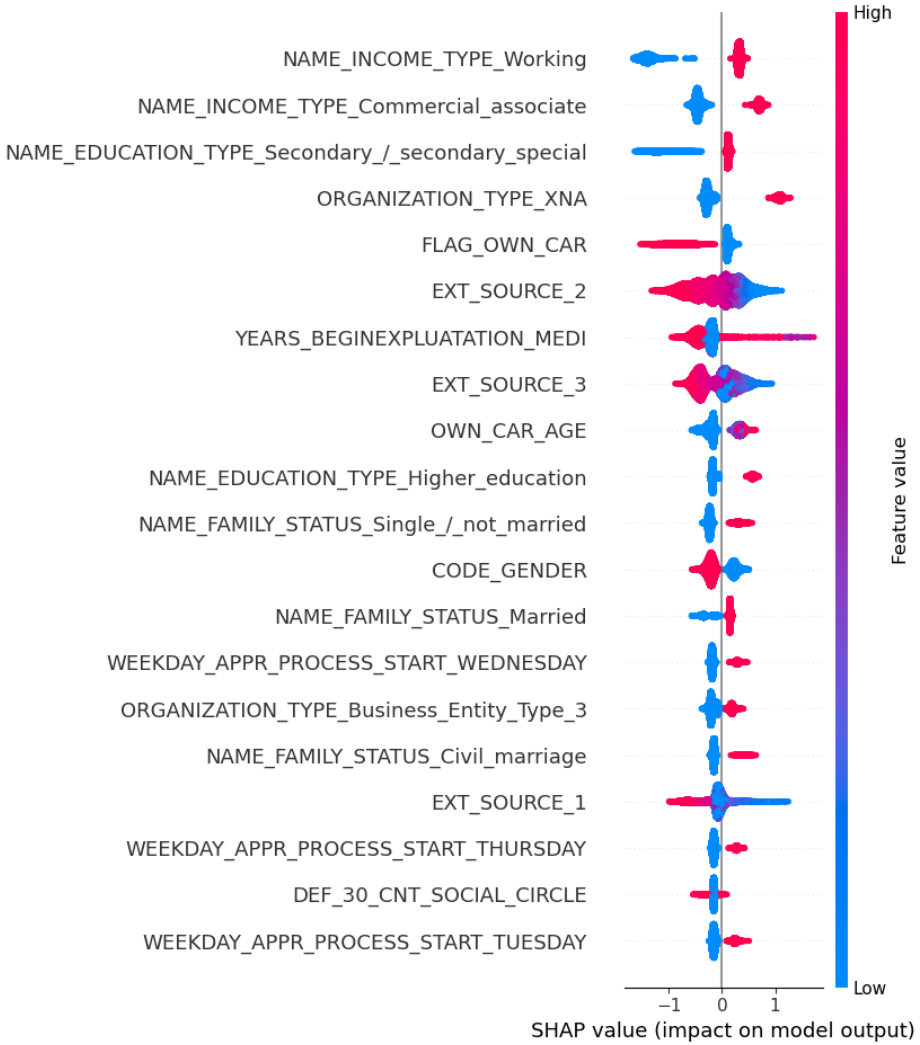
Modèle	CV Accuracy Mean	CV Accuracy StdDev	Confusion Matrix	Business Score	AUC	Training Time (s)
Baseline	0.919070	0.002403	TP=70520, FN=101 FP=6144, TN=112	0.944560	0.764050	11
SMOTE Baseline	0.481572	0.001224	TP=31402, FN=39219 FP=758, TN=5498	0.801910	0.764573	16
Augmented Baseline	0.481572	0.001224	TP=31402, FN=39219 FP=758, TN=5498	0.801910	0.764573	12
Random Forest	0.919265	0.002846	TP=70598, FN=23 FP=6234, TN=22	0.944140	0.692612	33
LightGBM	0.917747	0.002933	TP=70278, FN=343 FP=5943, TN=313	0.945180	0.772800	42
XGBoost	0.916641	0.002602	TP=70111, FN=510 FP=5851, TN=405	0.945190	0.764124	25

Meilleur compromis : **LightGBM**



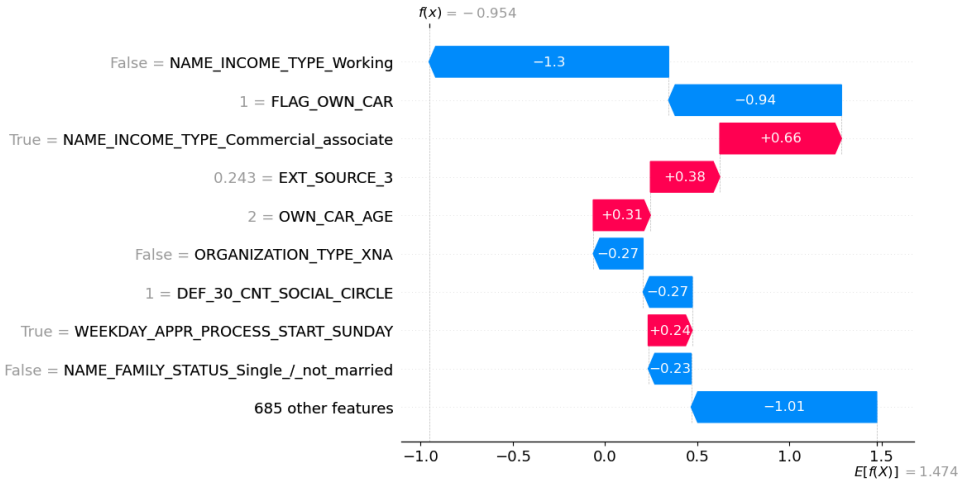
FEATURE IMPORTANCE GLOBALE

- Métier
- Education
- Vie maritale
- …

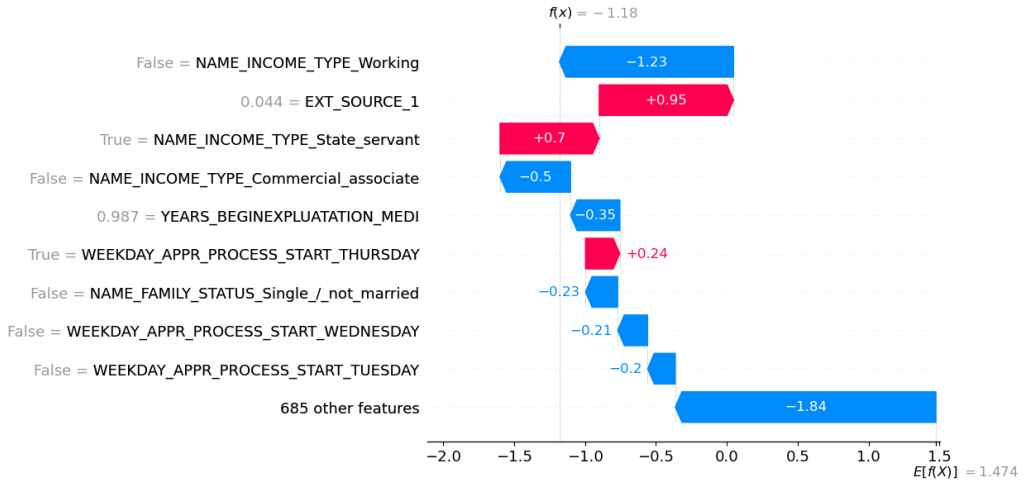


FEATURE IMPORTANCE LOCALE

Exemple - Classe : 0

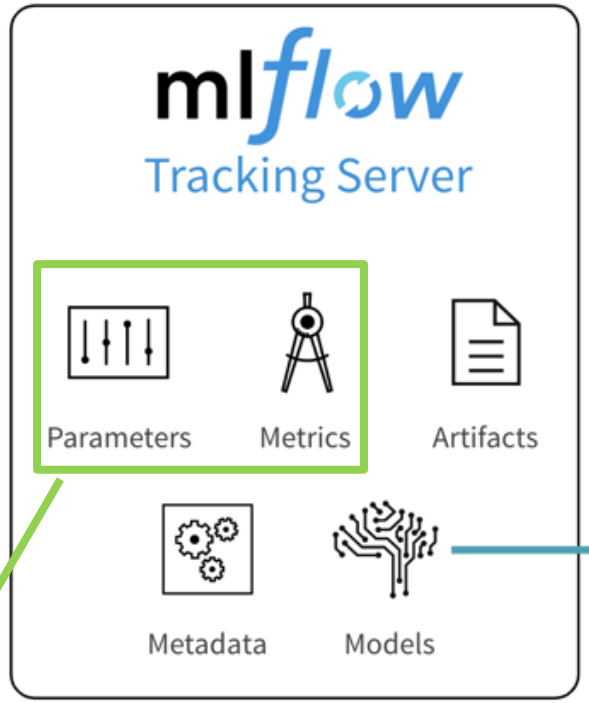


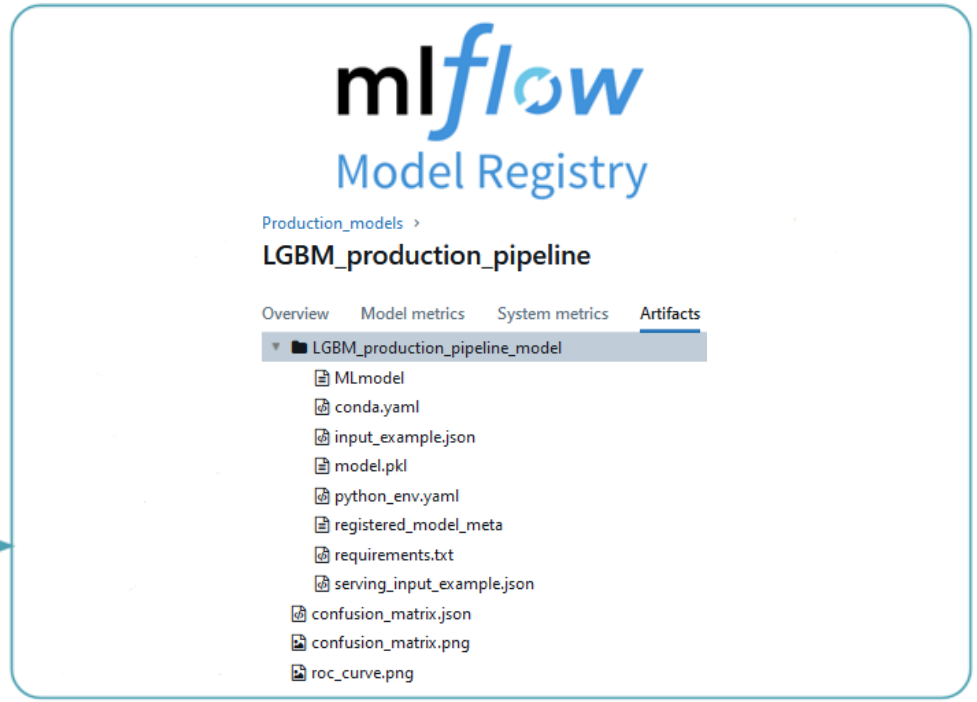
Exemple - Classe : 1



TRACKING

Lien Mlflow





Parameters (4)

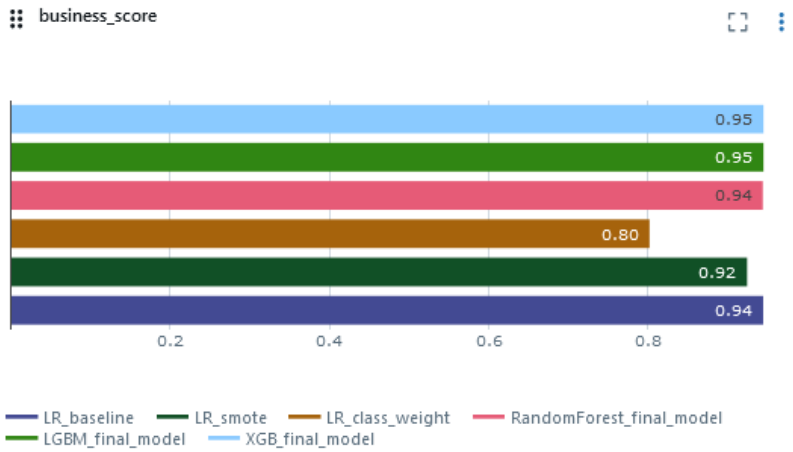
Search parameters

Parameter	Value
num_leaves	200
max_depth	-1
min_data_in_leaf	30
sampling_strategy	1

Metrics (6)

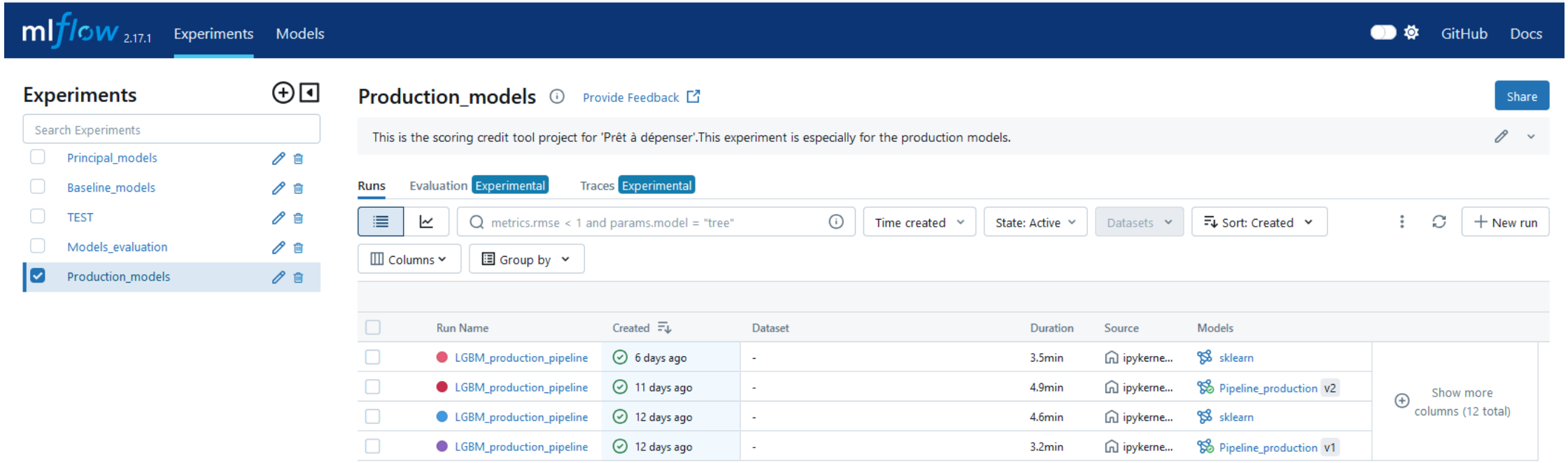
Search metrics

Metric	Value
mean_cv_score	0.918622902484499
std_cv_score	0.0015317995344439248
cv_time	429.7950897216797
training_time	51.25253462791443
business_score	0.94518
auc	0.7728000329267455



MLFLOW UI

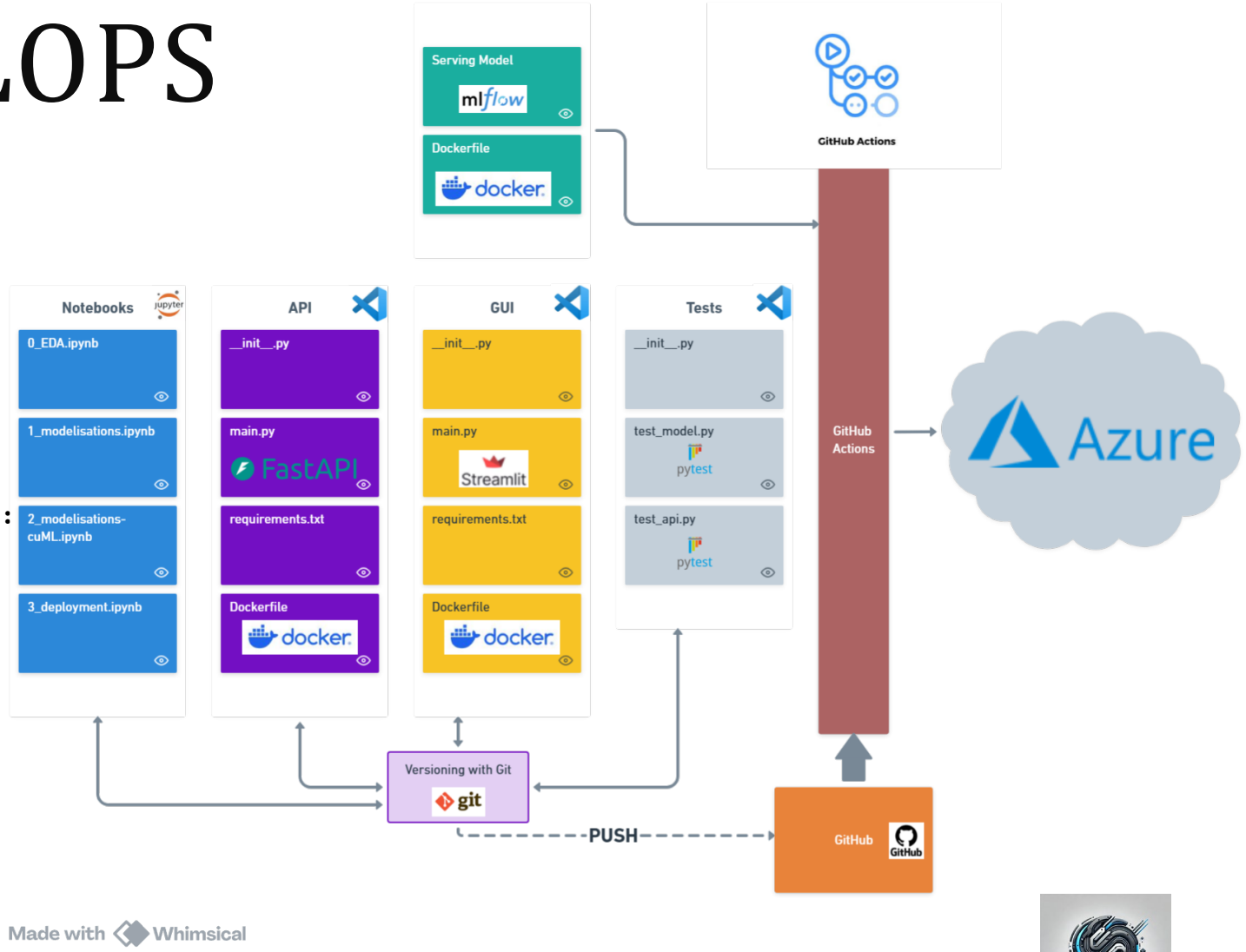
Lien Mlflow



DÉPLOIEMENT SUR LE CLOUD :

STRATÉGIE MLOPS

1. Serving du modèle (Model Registry – Mlflow)
2. Création API + interface graphique
3. Versioning avec Git / GitHub
4. Workflow : déploiement automatique en « containers » :
   - Modèle
   - API (FastAPI)
   - GUI (Streamlit)
   - Tests unitaires (Pytest)
5. Disponibilité sur le Cloud

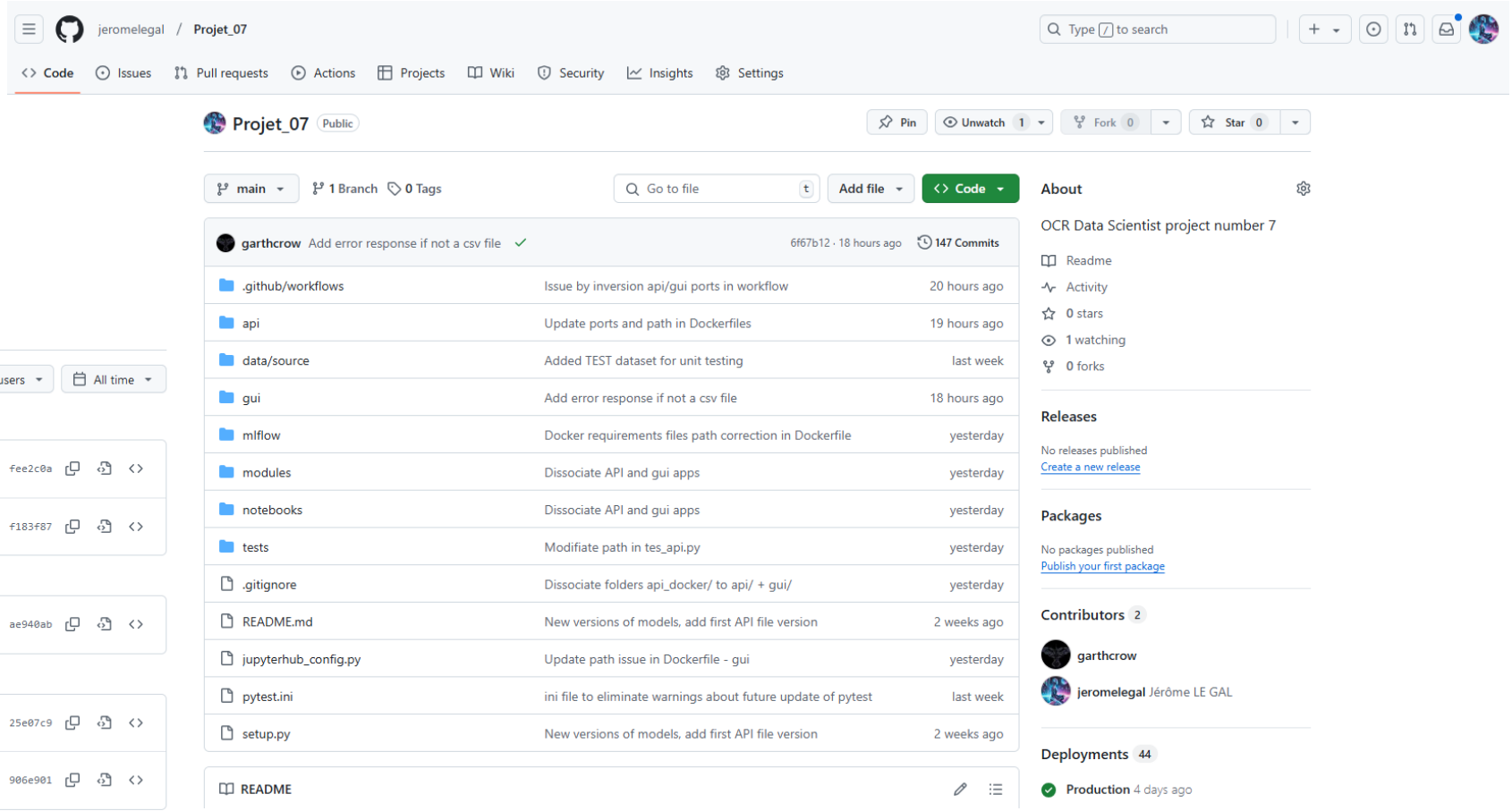


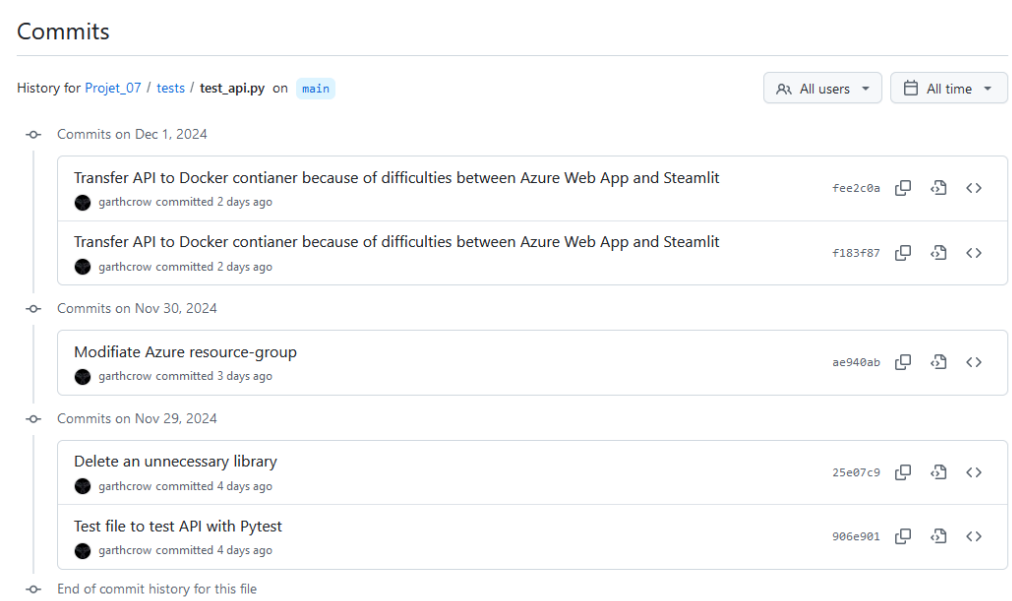
Made with Whimsical

Lien :
https://github.com/jeromelegal/Projet_07

GITHUB

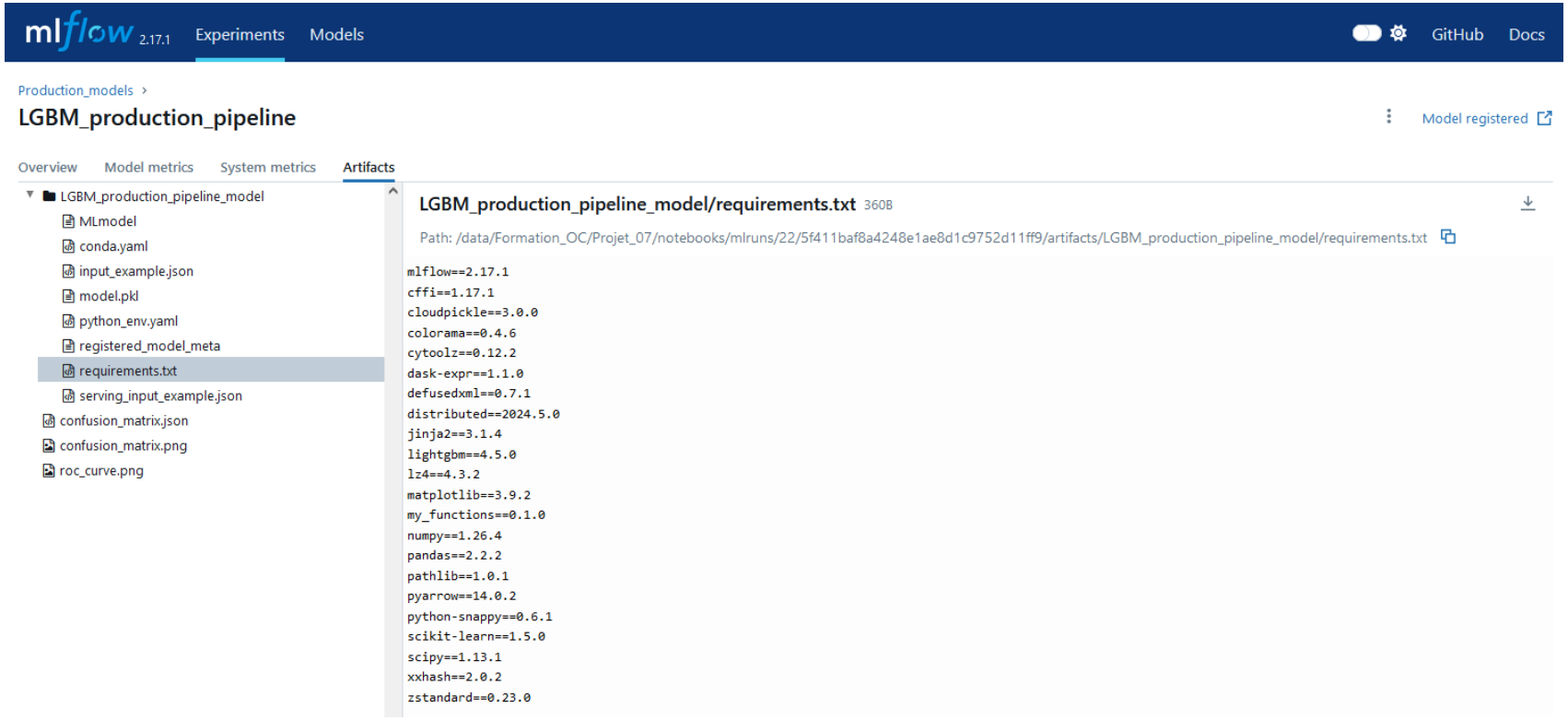
- Mise à disposition du code sur GitHub
- Commits





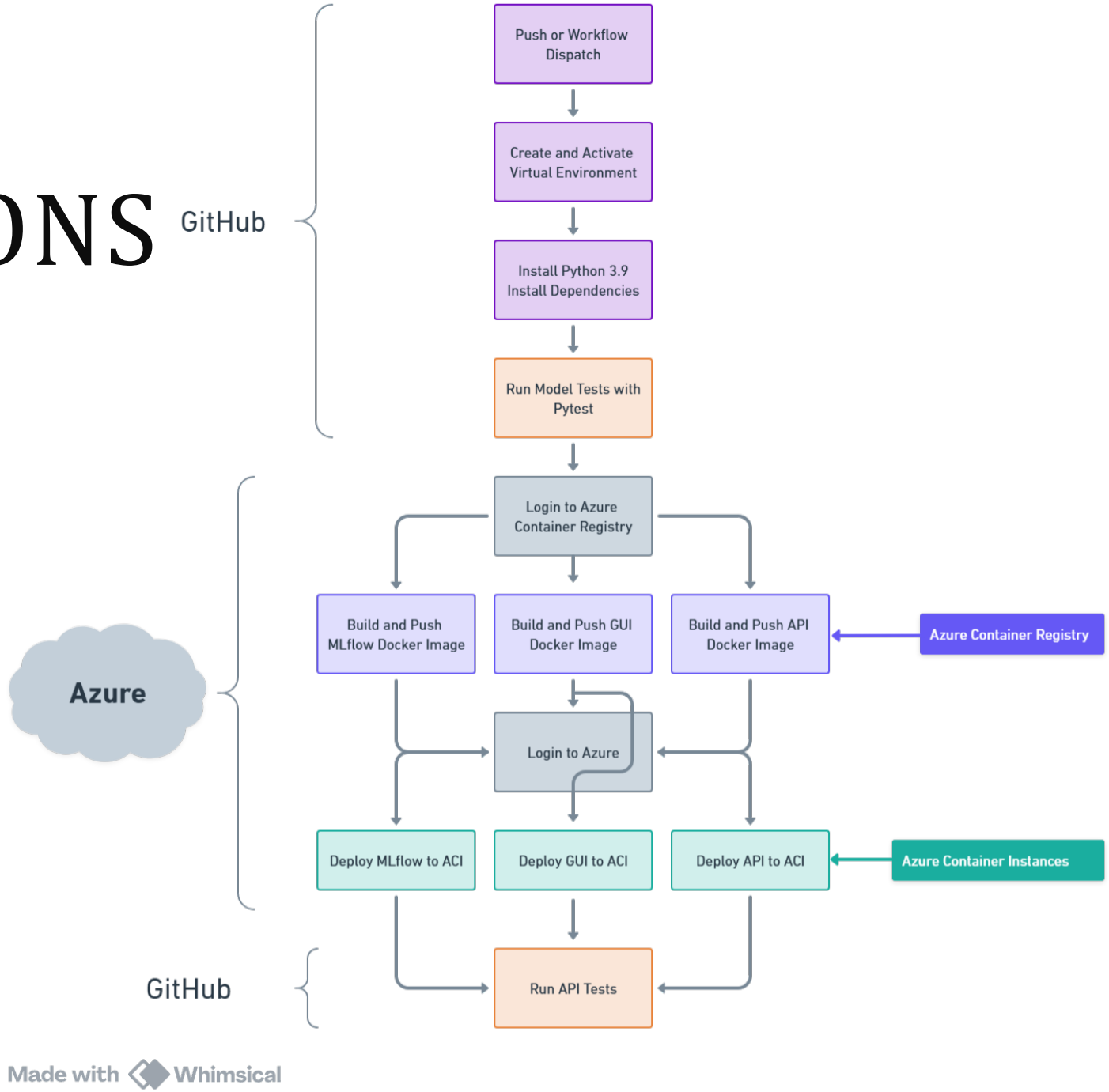
SUIVI DES REQUIREMENTS

Le suivi des packages requis est réalisé dans Mlflow.
Exemple ci-dessous du modèle en production.



WORKFLOW GITHUB ACTIONS

- Déclenchement sur « push »



TESTS

- Test du modèle avant déploiement

Run model tests with Pytest

```
▼ Run pytest -v test_model.py
  pytest -v test_model.py
  shell: /usr/bin/bash -e {0}
  env:
    pythonLocation: /opt/hostedtoolcache/Python/3.9.20/x64
    PKG_CONFIG_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib/pkgconfig
    Python_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
    Python2_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
    Python3_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
    LD_LIBRARY_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib
============================= test session starts ==============================
platform linux -- Python 3.9.20, pytest-8.3.4, pluggy-1.5.0 -- /opt/hostedtoolcache/Python/3.9.20/x64/bin/python
cachedir: .pytest_cache
rootdir: /home/runner/work/Projet_07/Projet_07
configfile: pytest.ini
plugins: anyio-4.6.2.post1
collecting ... collected 1 item

test_model.py::test_model_prediction PASSED                              [100%]

```

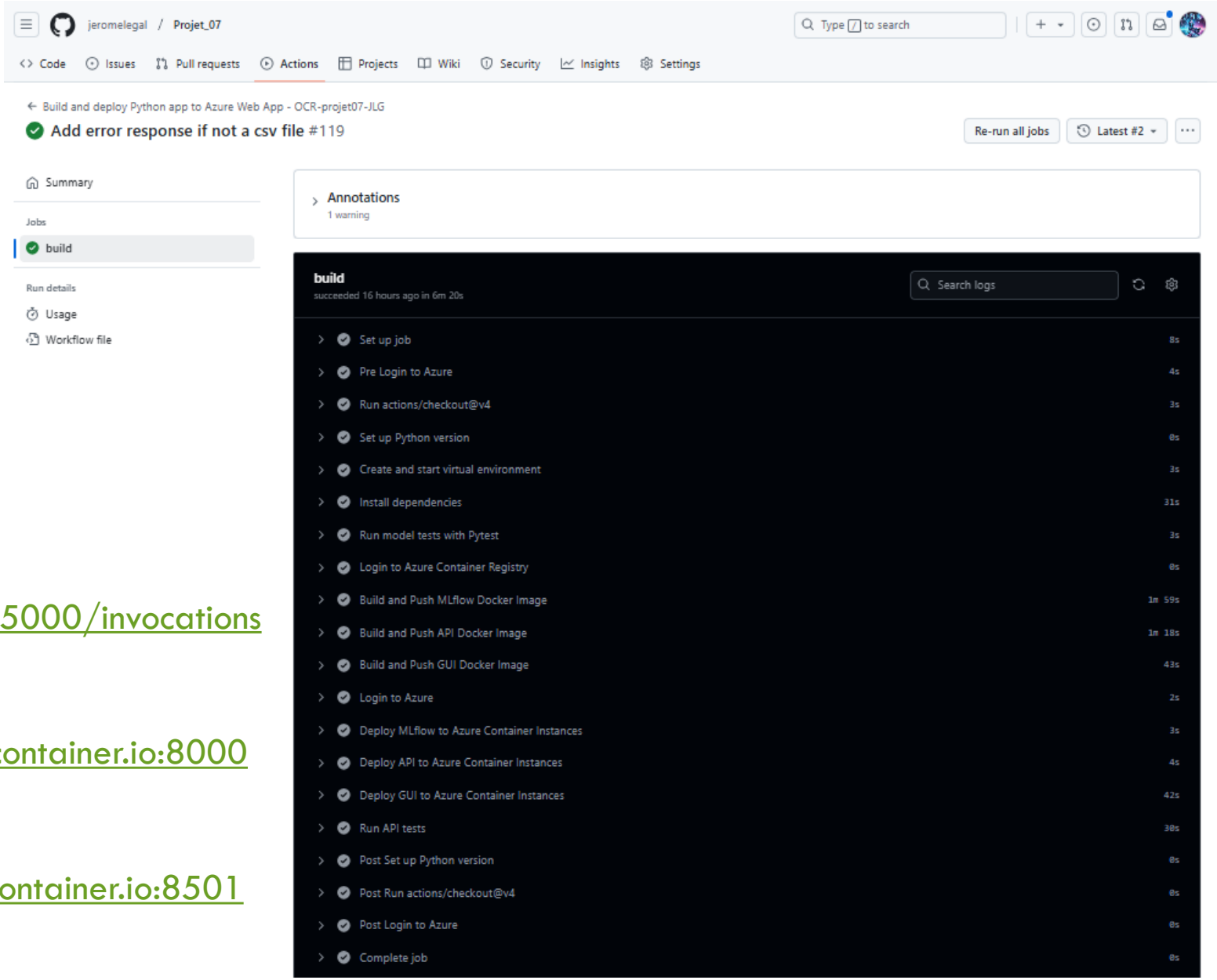
Run API tests

```
▼ Run export PYTHONPATH=$PYTHONPATH:$(pwd)/..
  export PYTHONPATH=$PYTHONPATH:$(pwd)/..
  pytest -v test_api.py
  shell: /usr/bin/bash -e {0}
  env:
    pythonLocation: /opt/hostedtoolcache/Python/3.9.20/x64
    PKG_CONFIG_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib/pkgconfig
    Python_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
    Python2_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
    Python3_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
    LD_LIBRARY_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib
    DOCKER_CONFIG: /home/runner/work/_temp/docker_login_1733505987747
============================= test session starts ==============================
platform linux -- Python 3.9.20, pytest-8.3.4, pluggy-1.5.0 -- /opt/hostedtoolcache/Python/3.9.20/x64/bin/python
cachedir: .pytest_cache
rootdir: /home/runner/work/Projet_07/Projet_07
configfile: pytest.ini
plugins: anyio-4.7.0
collecting ... collected 5 items

test_api.py::test_gui_container_started PASSED                           [ 20%]
test_api.py::test_api_container_started PASSED                           [ 40%]
test_api.py::test_model_container_started PASSED                         [ 60%]
test_api.py::test_format_data_for_api PASSED                             [ 80%]
test_api.py::test_predict PASSED                                         [100%]

============================== 5 passed in 24.62s ==============================
```

- Test de l'API et de l'Interface Graphique

DEPLOIEMENT

- Processus de déploiement GitHub Actions



Liens :

- Modèle: http://mlflowjlg-container.germanywestcentral.azurecontainer.io:5000/invocations

- API: http://api-container.germanywestcentral.azurecontainer.io:8000

- GUI: http://gui-container.germanywestcentral.azurecontainer.io:8501

EXEMPLE SUR LE CLOUD

- Accès API : http://api-container.germanywestcentral.azurecontainer.io:8501

➢ Exemples :

- *M.Jacques_0.csv*
- *M.Robert_1.csv*
- *Multi.csv*

DATA DRIFT

Datasets bruts

data_drift_report.html

Drift is detected for 7.438% of columns (9 out of 121).

Search

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.359052
AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	0.281765
AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.210785
AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.207334
AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.161102
AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.15426
NAME_CONTRACT_TYPE	cat			Detected	Jensen-Shannon distance	0.14755
DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.138977
FLAG_EMAIL	num			Detected	Jensen-Shannon distance	0.122121

Datasets preprocessés

data_drift_preprocessed_report.html

Drift is detected for 6.704% of columns (12 out of 179).

Search

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
PAYMENT_RATE	num			Detected	Wasserstein distance (normed)	0.574683
AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.33941
EXT_SOURCE_1	num			Detected	Wasserstein distance (normed)	0.249278
CODE_GENDER	num			Detected	Jensen-Shannon distance	0.234671
AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.209606
AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.207334
INCOME_CREDIT_PERC	num			Detected	Wasserstein distance (normed)	0.179298
AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.160558
NAME_CONTRACT_TYPE_Cash loans	num			Detected	Jensen-Shannon distance	0.14755
NAME_CONTRACT_TYPE_Revolving loans	num			Detected	Jensen-Shannon distance	0.14755
AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.143037
DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.138981

CONCLUSION

- Avec la création d'un score métier, le modèle est optimisé sur la cible la plus importante (limiter l'accord de prêt aux personnes à risques)
- Le tracking complet des modélisations expérimentées permet de générer automatiquement l'historique et facilite la comparaison des résultats.
- L'approche MLops permet un déploiement complètement automatisé et sécurisé par des tests.

QUESTIONS ?