



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

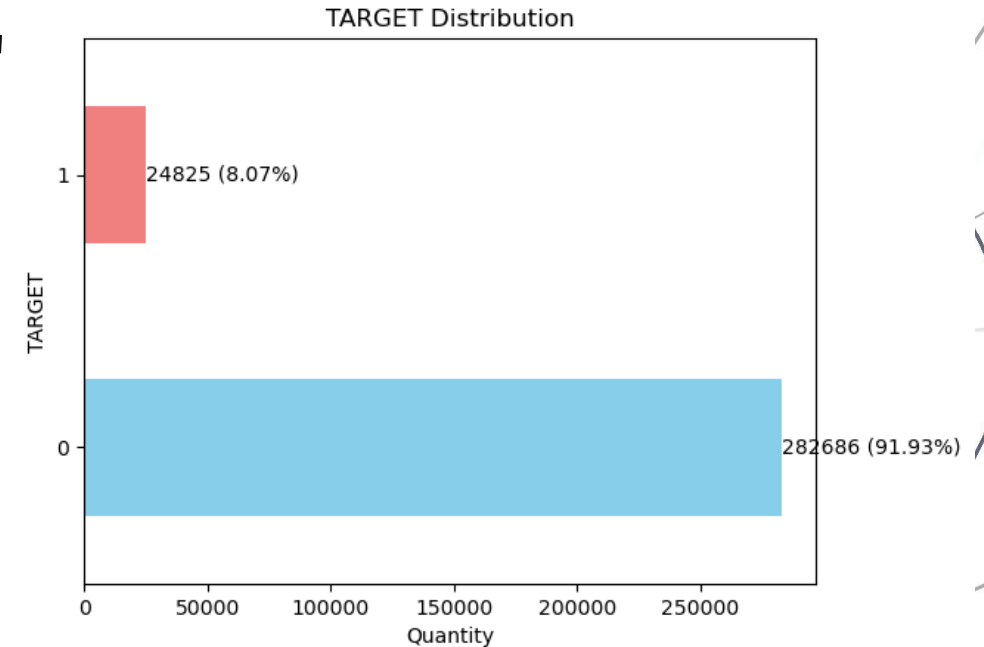
DESCRIPTION DU JEU DE DONNÉES

- 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



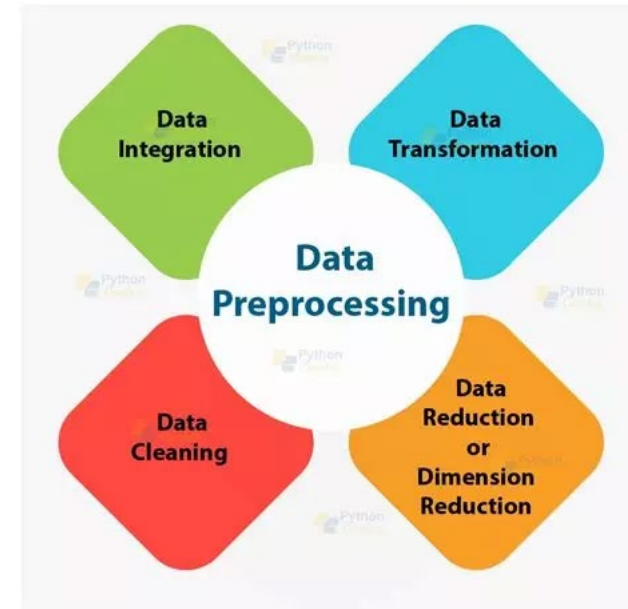
PREPROCESSING

■ Preprocessing :

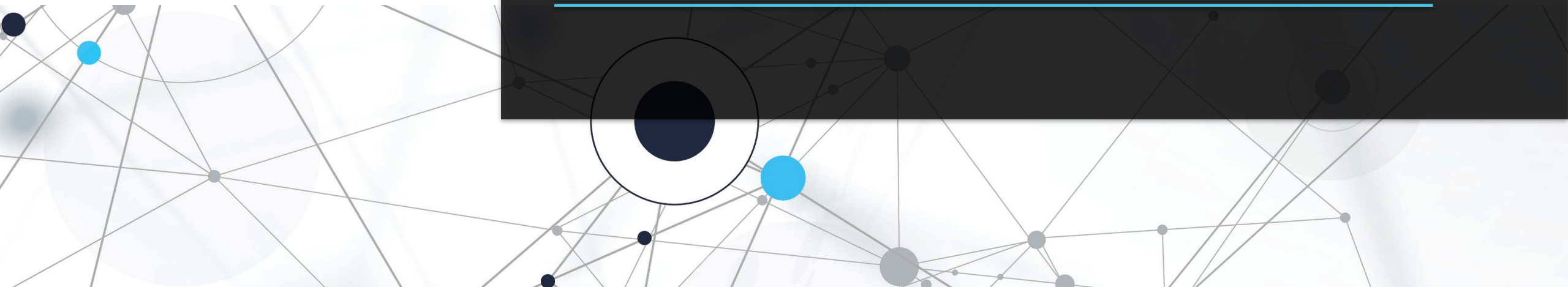
- Agrégation (min, max, sum, mean)
- Encodage : OneHotEncoder, LabelEncoder
- Transformations : StandardScaler
- PCA

■ 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. Entraînements (GPU) et choix modèle final
6. Analyse « *Feature Importance* »

CONSTRAINTES ET SCORING

Contraintes :

- Déséquilibre de la Target
- Perte de coûts : $FN = 10 \times FP$

Evaluation :

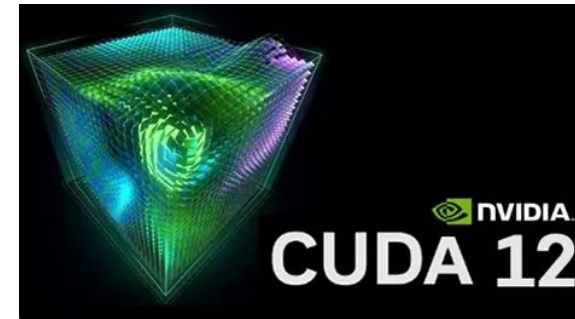
- ✓ Score métier (make_scorer)
- ✓ Matrice de confusion
- ✓ AUC-ROC
- ✓ Temps de calcul



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*)

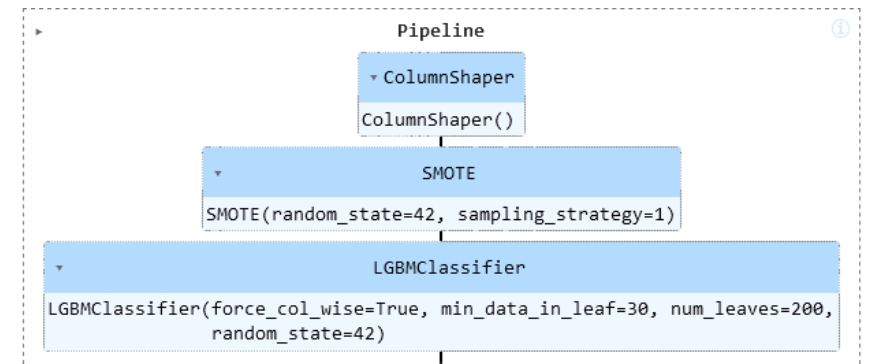
❖ Optimisation : GridSearchCV sur GPU



SYNTHÈSE DES RÉSULTATS

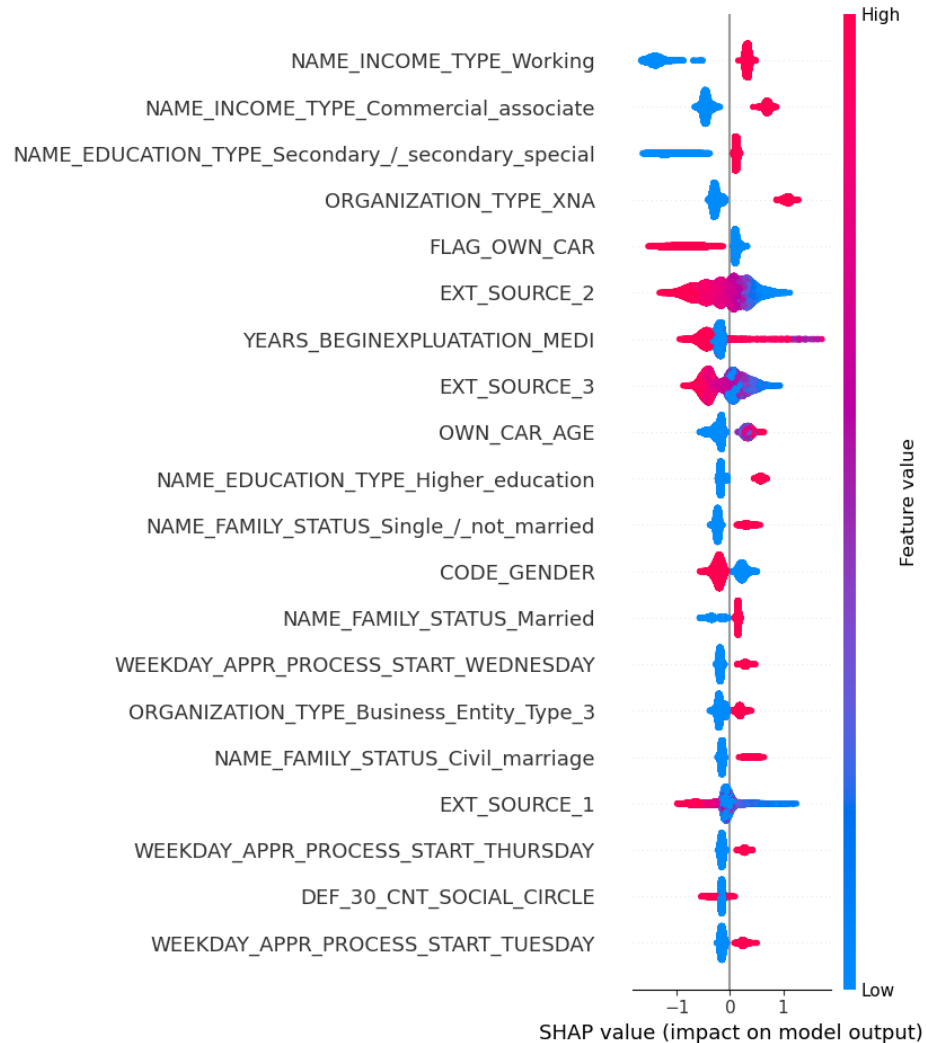
| | CV Accuracy Mean | CV Accuracy StdDev | Confusion Matrix | Business Score | AUC | Training Time (s) |
|--------------------|------------------|--------------------|---------------------------------------|----------------|----------|-------------------|
| Modèle | | | | | | |
| Baseline | 0.919070 | 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**



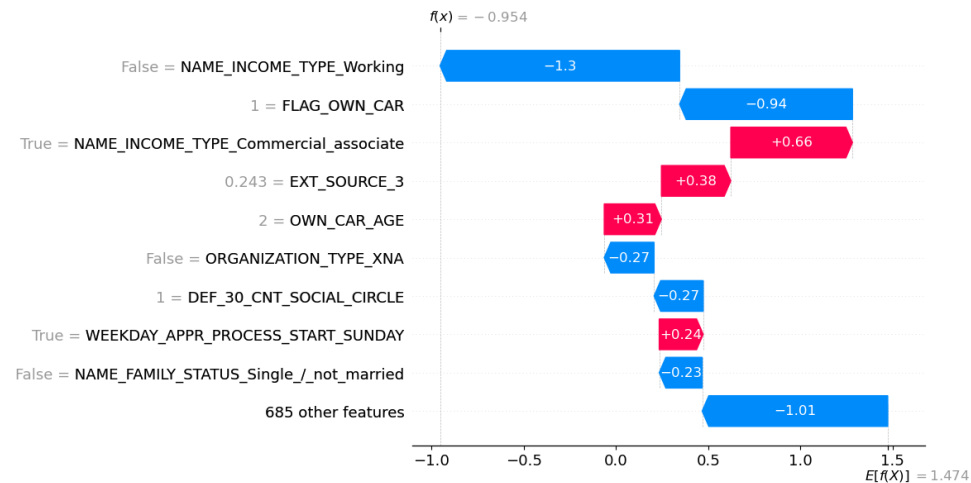
FEATURE IMPORTANCE GLOBALE

- Métier
- Education
- Vie maritale
- ...

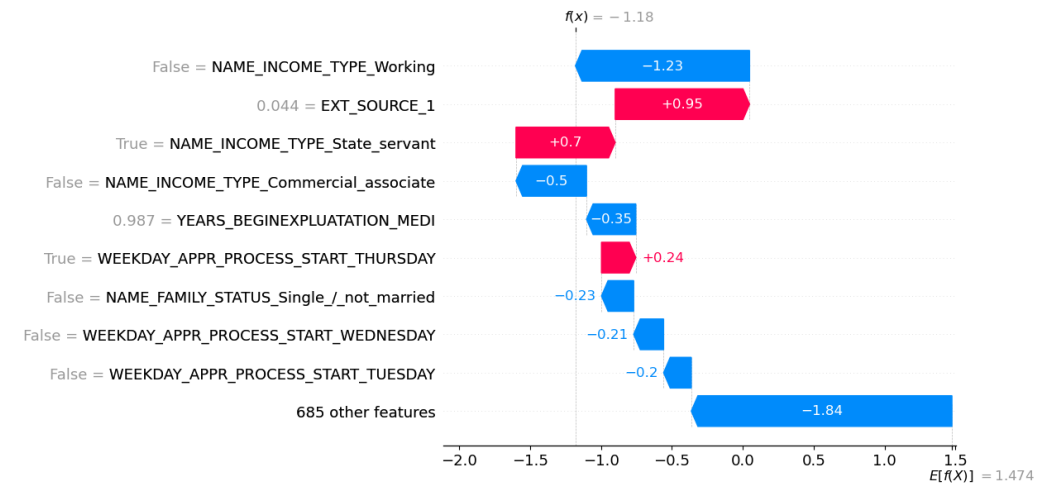


FEATURE IMPORTANCE LOCALE

Exemple - Classe : 0

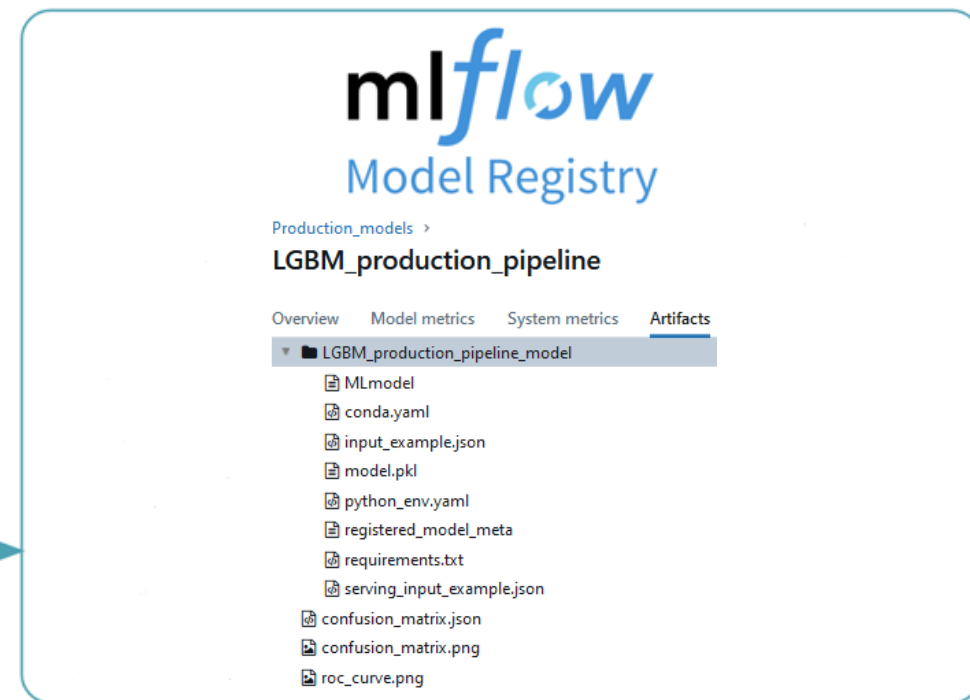
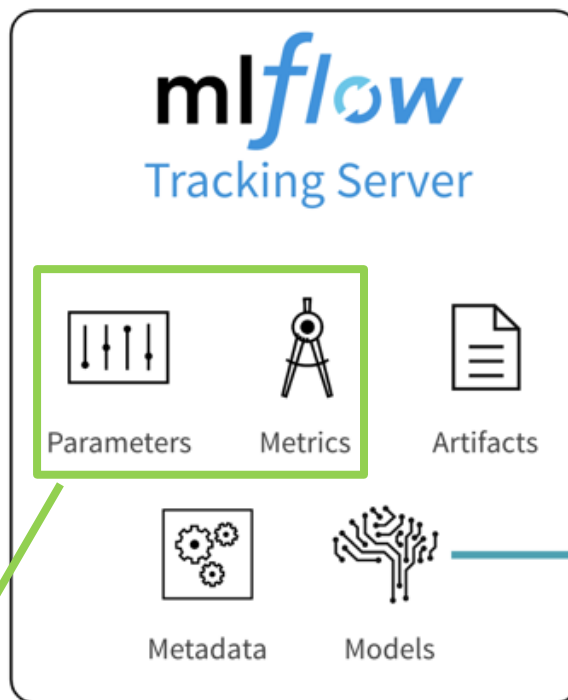


Exemple - Classe : 1



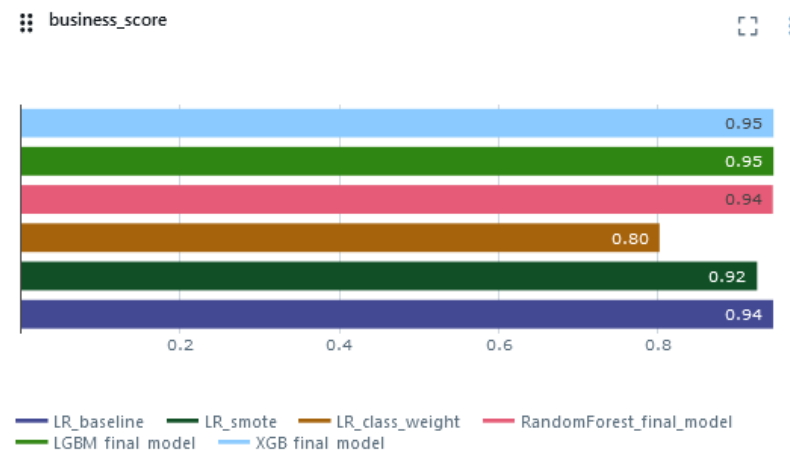
TRACKING

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](#)

mlflow2.17.1

Experiments

Models

+

⌵

Search Experiments

☐

Principal_models

✎

🗑

☐

Baseline_models

✎

🗑

☐

TEST

✎

🗑

☐

Models_evaluation

✎

🗑

☒

Production_models

✎

🗑

Production_models

📄

Provide Feedback

🔗

Share

This is the scoring credit tool project for 'Prêt à dépenser'.This experiment is especially for the production models.

✎

⌵

Runs

Evaluation

Experimental

Traces

Experimental

📄

📈

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

Time created ⌵

State: Active ⌵

Datasets ⌵

⌵ Sort: Created ⌵

⋮

🔄

+ New run

📄 Columns ⌵

📄 Group by ⌵

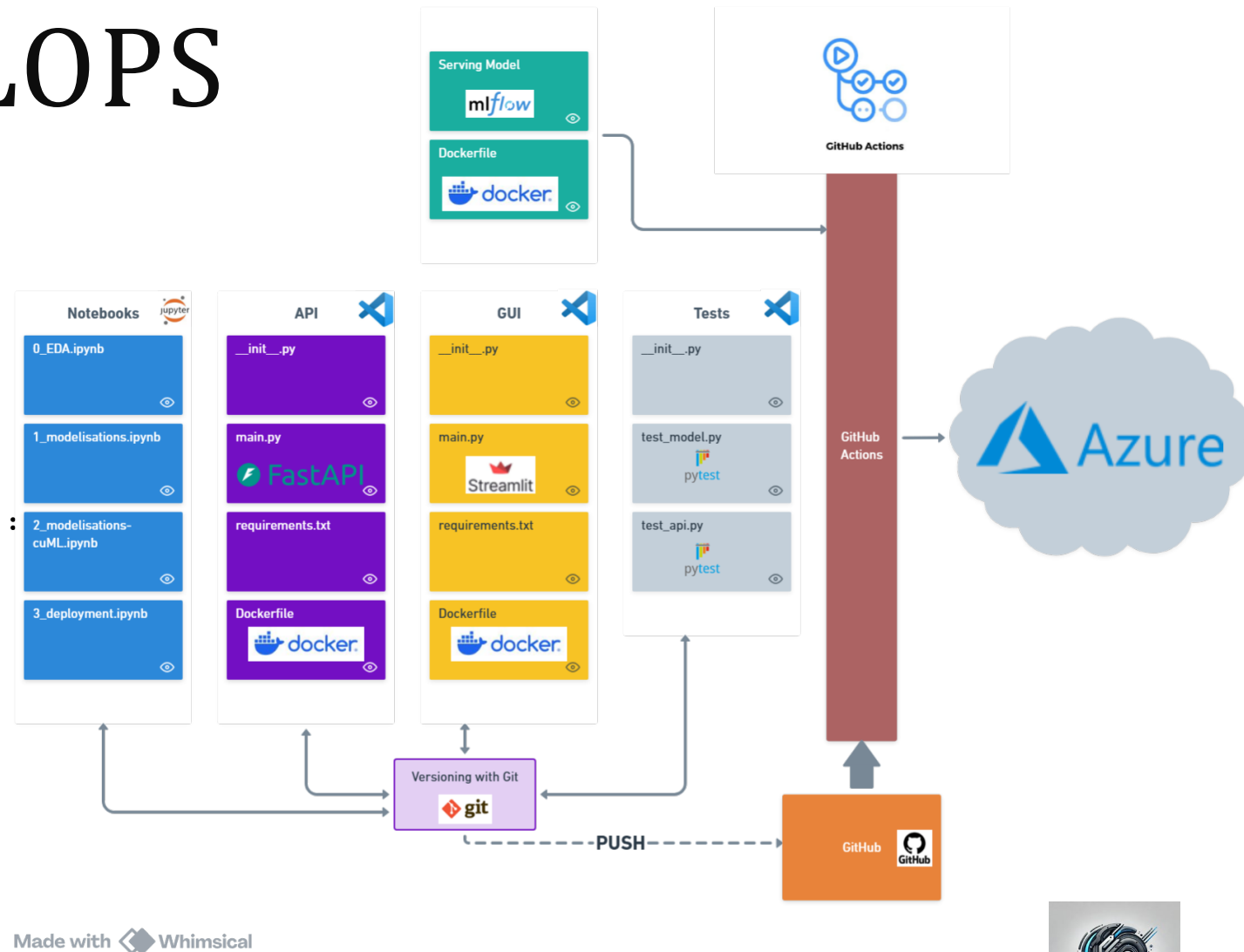
| <input type="checkbox"/> | Run Name | Created ⌵ | Dataset | Duration | Source | Models | |
|--------------------------|----------------------------|---------------|---------|----------|---------------|--------------------------|--------------------------------|
| <input type="checkbox"/> | ● LGBM_production_pipeline | ✅ 6 days ago | - | 3.5min | 🏠 ipykerne... | 🔗 sklearn | ⊕ Show more columns (12 total) |
| <input type="checkbox"/> | ● LGBM_production_pipeline | ✅ 11 days ago | - | 4.9min | 🏠 ipykerne... | 🔗 Pipeline_production v2 | |
| <input type="checkbox"/> | ● LGBM_production_pipeline | ✅ 12 days ago | - | 4.6min | 🏠 ipykerne... | 🔗 sklearn | |
| <input type="checkbox"/> | ● LGBM_production_pipeline | ✅ 12 days ago | - | 3.2min | 🏠 ipykerne... | 🔗 Pipeline_production v1 | |

DÉPLOIEMENT SUR LE CLOUD :

The background of the slide features a complex network of thin grey lines connecting various nodes. Some nodes are represented by small grey dots, while others are larger circles in dark blue or light blue. The overall aesthetic is modern and technological, typical of cloud computing presentations.

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



GITHUB

- Mise à disposition du code sur GitHub
- Commits

Lien :

https://github.com/jeromelegal/Projet_07

Commits

History for [Projet_07](#) / [tests](#) / [test_api.py](#) on [main](#)

🔍 All users

📅 All time

Commits on Dec 1, 2024

Transfer API to Docker container because of difficulties between Azure Web App and Steamlit

● garthcrow committed 2 days ago

fee2c0a

📄 📁 <>

Transfer API to Docker container because of difficulties between Azure Web App and Steamlit

● garthcrow committed 2 days ago

f183f87

📄 📁 <>

Commits on Nov 30, 2024

Modifiat Azure resource-group

● garthcrow committed 3 days ago

ae940ab

📄 📁 <>

Commits on Nov 29, 2024

Delete an unnecessary library

● garthcrow committed 4 days ago

25e07c9

📄 📁 <>

Test file to test API with Pytest

● garthcrow committed 4 days ago

906e901

📄 📁 <>

End of commit history for this file

jeromelegal

Projet_07

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings

🔍 Type to search

+ ▾

🕒

🔄

📧

🌐

Projet_07

Public

📌 Pin

👁 Unwatch 1 ▾

🍴 Fork 0 ▾

★ Star 0 ▾

🌿 main ▾

🌿 1 Branch

🏷 0 Tags

🔍 Go to file

📄 Add file ▾

<> Code ▾

● garthcrow Add error response if not a csv file ✓ 6f67b12 · 18 hours ago 147 Commits

📁 .github/workflows

Issue by inversion api/gui ports in workflow 20 hours ago

📁 api

Update ports and path in Dockerfiles 19 hours ago

📁 data/source

Added TEST dataset for unit testing last week

📁 gui

Add error response if not a csv file 18 hours ago

📁 mlflow

Docker requirements files path correction in Dockerfile yesterday

📁 modules

Dissociate API and gui apps yesterday

📁 notebooks

Dissociate API and gui apps yesterday

📁 tests

Modifiat path in tes_api.py yesterday

📄 .gitignore

Dissociate folders api_docker/ to api/ + gui/ yesterday

📄 README.md

New versions of models, add first API file version 2 weeks ago

📄 jupyterhub_config.py

Update path issue in Dockerfile - gui yesterday

📄 pytest.ini

ini file to eliminate warnings about future update of pytest last week

📄 setup.py

New versions of models, add first API file version 2 weeks ago

📖 README

📄 📁

About

📖 Readme

🔔 Activity

★ 0 stars

👁 1 watching

🍴 0 forks

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Contributors 2

● garthcrow

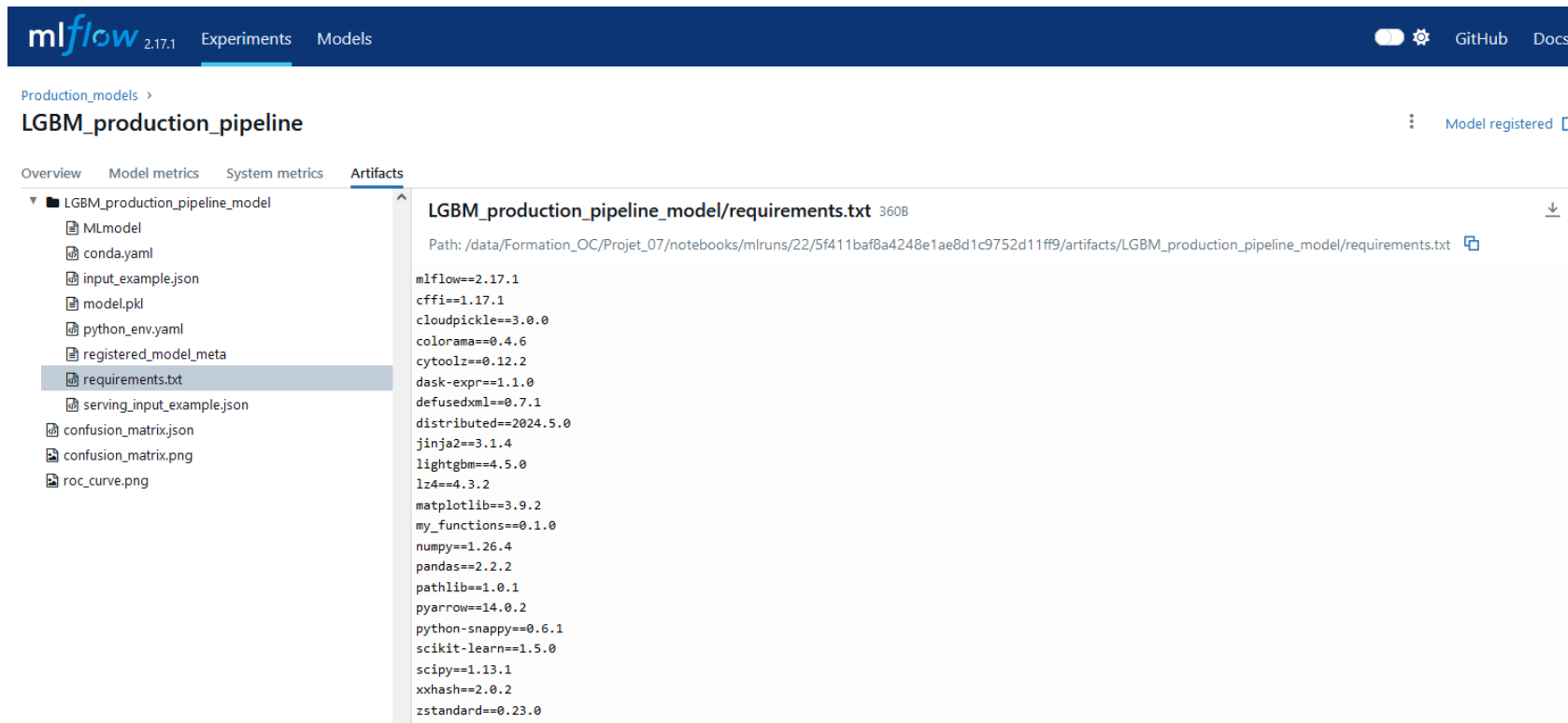
🌐 jeromelegal Jérôme LE GAL

Deployments 44

🟢 Production 4 days ago

SUIVI DES REQUIREMENTS

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



The screenshot displays the Mlflow web interface. At the top, the navigation bar includes the 'mlflow' logo (version 2.17.1), 'Experiments', and 'Models' tabs. On the right, there are links for 'GitHub' and 'Docs'. Below the navigation bar, the breadcrumb 'Production_models >' is shown, followed by the model name 'LGBM_production_pipeline'. To the right of the model name, a status indicator shows 'Model registered' with an external link icon. Below this, a tabbed interface shows 'Overview', 'Model metrics', 'System metrics', and 'Artifacts'. The 'Artifacts' tab is active, displaying a file tree on the left. The file 'requirements.txt' is selected and highlighted. The main content area shows the full text of the 'requirements.txt' file, which lists various Python packages and their versions. The path to the file is also displayed above the text.

mlflow 2.17.1 Experiments Models

Production_models > LGBM_production_pipeline Model registered

Overview Model metrics System metrics Artifacts

LGBM_production_pipeline_model

- MLmodel
- conda.yaml
- input_example.json
- model.pkl
- python_env.yaml
- registered_model_meta
- requirements.txt
- serving_input_example.json
- confusion_matrix.json
- confusion_matrix.png
- roc_curve.png

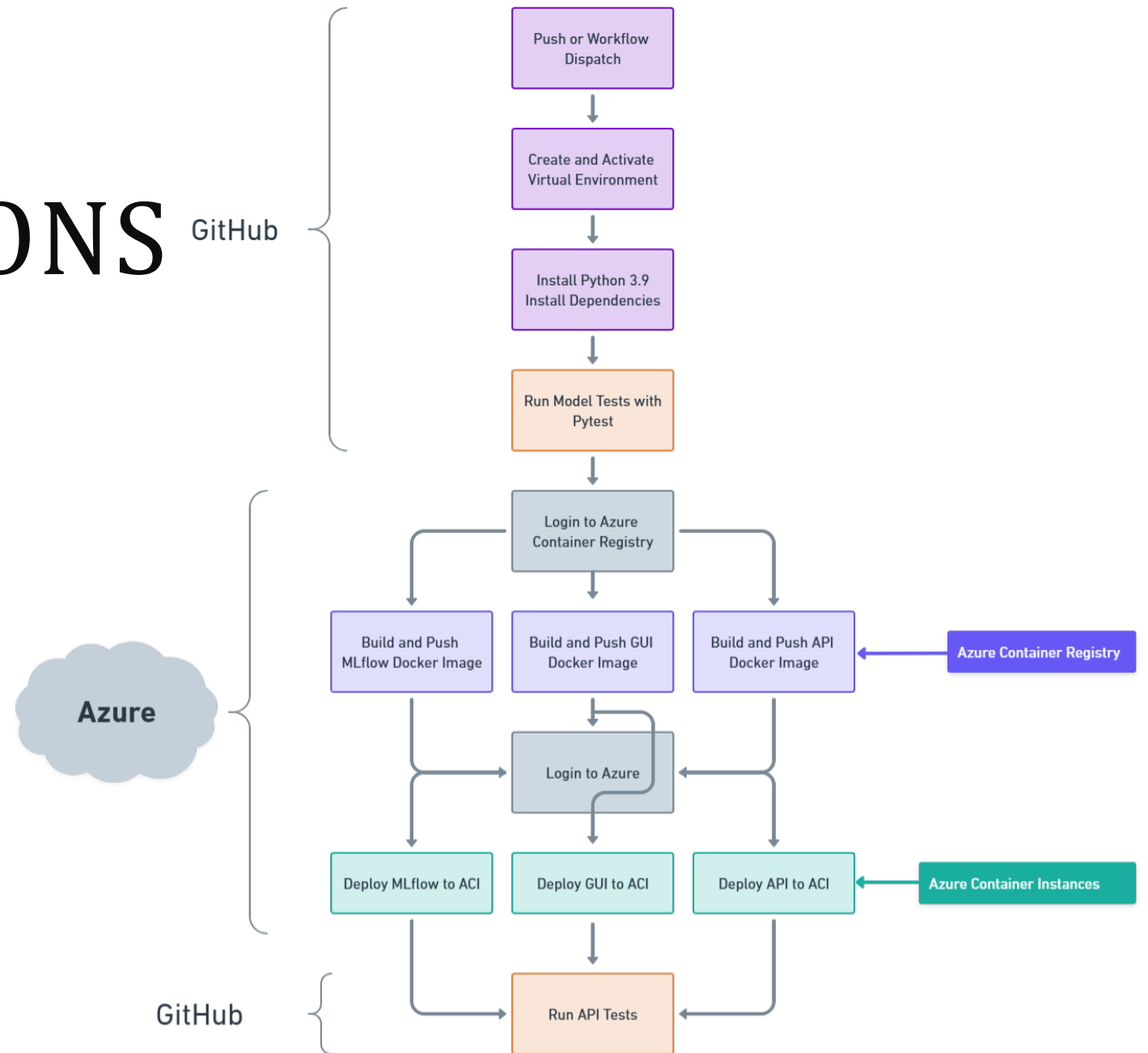
LGBM_production_pipeline_model/requirements.txt 360B

Path: /data/Formation_OC/Projet_07/notebooks/mlruns/22/5f411baf8a4248e1ae8d1c9752d11ff9/artifacts/LGBM_production_pipeline_model/requirements.txt

```
mlflow==2.17.1
cffi==1.17.1
cloudpickle==3.0.0
colorama==0.4.6
cytoolz==0.12.2
dask-expr==1.1.0
defusedxml==0.7.1
distributed==2024.5.0
jinja2==3.1.4
lightgbm==4.5.0
lz4==4.3.2
matplotlib==3.9.2
my_functions==0.1.0
numpy==1.26.4
pandas==2.2.2
pathlib==1.0.1
pyarrow==14.0.2
python-snappy==0.6.1
scikit-learn==1.5.0
scipy==1.13.1
xxhash==2.0.2
zstandard==0.23.0
```

WORKFLOW GITHUB ACTIONS

- Déclenchement sur « push »



TESTS

- Test du modèle avant déploiement

Run API tests

```
1 ▼ Run export PYTHONPATH=$PYTHONPATH:$(pwd)/..
2 export PYTHONPATH=$PYTHONPATH:$(pwd)/..
3 pytest -v test_api.py
4 shell: /usr/bin/bash -e {0}
5 env:
6   pythonLocation: /opt/hostedtoolcache/Python/3.9.20/x64
7   PKG_CONFIG_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib/pkgconfig
8   Python_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
9   Python2_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
10  Python3_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
11  LD_LIBRARY_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib
12  DOCKER_CONFIG: /home/runner/work/_temp/docker_login_1733505987747
13 ===== test session starts =====
14 platform linux -- Python 3.9.20, pytest-8.3.4, pluggy-1.5.0 -- /opt/hostedtoolcache/Python/3.9.20/x64/bin/python
15 cachedir: .pytest_cache
16 rootdir: /home/runner/work/Projet_07/Projet_07
17 configfile: pytest.ini
18 plugins: anyio-4.7.0
19 collecting ... collected 5 items
20
21 test_api.py::test_gui_container_started PASSED [ 20%]
22 test_api.py::test_api_container_started PASSED [ 40%]
23 test_api.py::test_model_container_start PASSED [ 60%]
24 test_api.py::test_format_data_for_api PASSED [ 80%]
25 test_api.py::test_predict PASSED [100%]
26
27 ===== 5 passed in 24.62s =====
```

Run model tests with Pytest

```
1 ▼ Run pytest -v test_model.py
2 pytest -v test_model.py
3 shell: /usr/bin/bash -e {0}
4 env:
5   pythonLocation: /opt/hostedtoolcache/Python/3.9.20/x64
6   PKG_CONFIG_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib/pkgconfig
7   Python_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
8   Python2_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
9   Python3_ROOT_DIR: /opt/hostedtoolcache/Python/3.9.20/x64
10  LD_LIBRARY_PATH: /opt/hostedtoolcache/Python/3.9.20/x64/lib
11 ===== test session starts =====
12 platform linux -- Python 3.9.20, pytest-8.3.4, pluggy-1.5.0 -- /opt/hostedtoolcache/Python/3.9.20/x64/bin/python
13 cachedir: .pytest_cache
14 rootdir: /home/runner/work/Projet_07/Projet_07
15 configfile: pytest.ini
16 plugins: anyio-4.6.2.post1
17 collecting ... collected 1 item
18
19 test_model.py::test_model_prediction PASSED [100%]
20
```

- 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>

The screenshot displays the GitHub Actions interface for a workflow named 'Build and deploy Python app to Azure Web App - OCR-projet07-JLG'. The workflow is currently in a 'Completed' state, indicated by a green checkmark. The main section shows a list of jobs, with the 'build' job selected and expanded. The 'build' job is marked as 'succeeded' and shows a detailed log of steps including setting up the job, logging in to Azure, installing dependencies, and pushing Docker images to Azure Container Registry. The log also shows the deployment of the MLflow API and GUI to Azure Container Instances. The interface includes a search bar for logs, a 'Re-run all jobs' button, and a 'Latest #2' dropdown menu.

jeromelegal / Projet_07

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Build and deploy Python app to Azure Web App - OCR-projet07-JLG

✓ Add error response if not a csv file #119

Re-run all jobs Latest #2

Summary

Jobs

✓ build

Run details Usage Workflow file

Annotations
1 warning

build
succeeded 16 hours ago in 6m 20s

Search logs

| Step | Duration |
|--------------------------------------------|----------|
| Set up job | 8s |
| Pre Login to Azure | 4s |
| Run actions/checkout@v4 | 3s |
| Set up Python version | 0s |
| Create and start virtual environment | 3s |
| Install dependencies | 31s |
| Run model tests with Pytest | 3s |
| Login to Azure Container Registry | 0s |
| Build and Push MLflow Docker Image | 1m 59s |
| Build and Push API Docker Image | 1m 18s |
| Build and Push GUI Docker Image | 43s |
| Login to Azure | 2s |
| Deploy MLflow to Azure Container Instances | 3s |
| Deploy API to Azure Container Instances | 4s |
| Deploy GUI to Azure Container Instances | 42s |
| Run API tests | 30s |
| Post Set up Python version | 0s |
| Post Run actions/checkout@v4 | 0s |
| Post Login to Azure | 0s |
| Complete job | 0s |

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.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 ?