



IMPLÉMENTEZ UN MODÈLE DE SCORING

FORMATION DATA SCIENTIST – PROJET 7

OCTAVE POUILLOT

AOÛT 2023

SOMMAIRE

- Mission & Dataset
- Architecture globale
- Modélisation
 - Démarche
 - MLFlow
 - Résultats
- Déploiement
 - API
 - Dashboard
- Data Drift
- Démonstration

Prêt à dépenser



MISSION

- Société financière "Prêt à dépenser", propose des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt.

Objectifs :

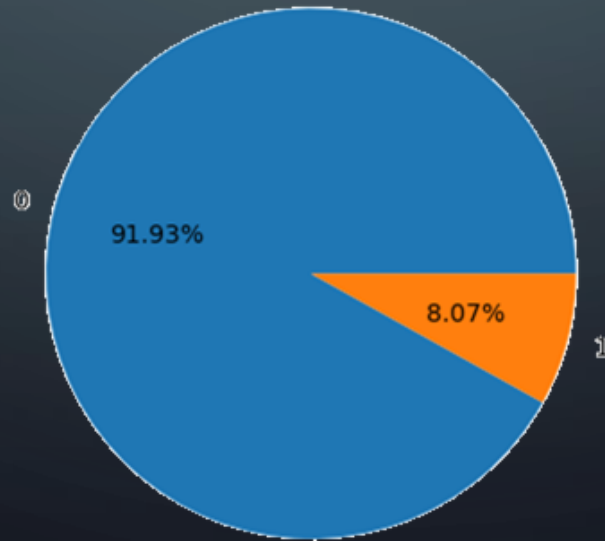
- Construire un modèle de scoring pour prédire la probabilité de faillite d'un client
- Construire un dashboard interactif pour les gestionnaires de la relation client
- Mettre en production le modèle de scoring à l'aide d'une API, ainsi que le dashboard interactif qui appelle l'API pour les prédictions



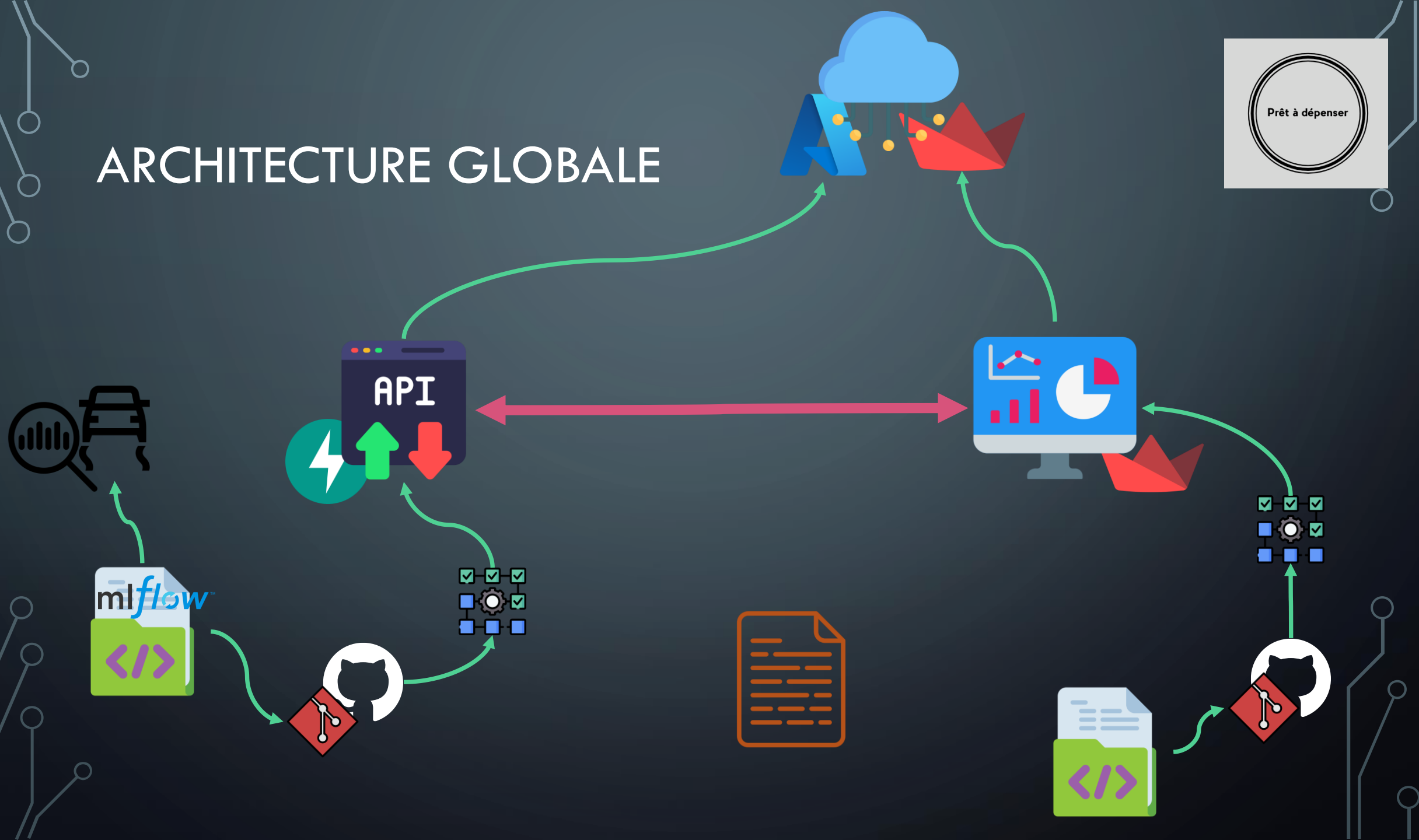
DATASET

Deux dataset utilisés, disponible sur Kaggle :

Application_train	Application_test
Avec target	Sans target
307 511 lignes (clients)	48 744 lignes (clients)
122 features	121 features



ARCHITECTURE GLOBALE

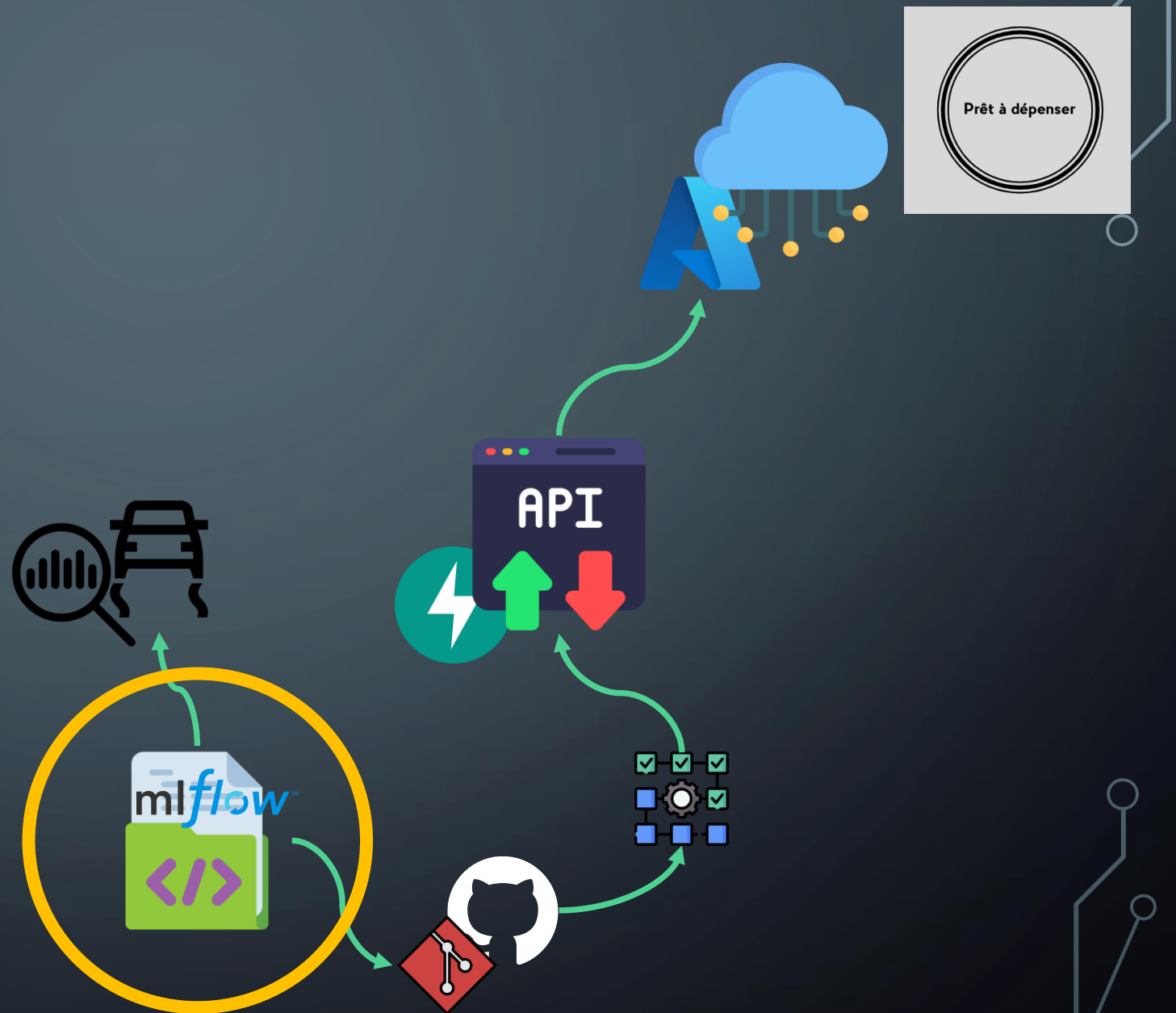


MODÉLISATION

DÉMARCHE

MLFLOW

RÉSULTATS



MODÉLISATION - DÉMARCHE



- Utiliser et adapter des kernel Kaggle
 - analyse exploratoire
 - préparation des données
 - feature engineering
- Kernel utilisés :
 - credit-default-risk
 - predict-score-credit
- Equilibrage :
 - Smote
 - class_weight
- Métriques :
 - $\text{Job_score} = (10 * \text{FNR}) + \text{FPR}$
 - ROC AUC
 - Training time
- Score Global :
 $\text{time} + 2 * \text{roc auc} + 2 * \text{job_score}$

MODÉLISATION - DÉMARCHE

Prêt à dépenser

1ère sélection

- Logistic Regression
- Decision Tree
- Linear Discriminant Analysis
- Gradient Boosting
- XGBoost
- LightBoost
- Adaboost
- Random Forest

Optimisation

- Hyperparamètres
- Pipeline
- GridSearch
- CrossValidation
- Calcul scores
- Logistic Regression
- XGBoost
- LightBoost

Analyse

- Interprétabilité
- Comparaison score

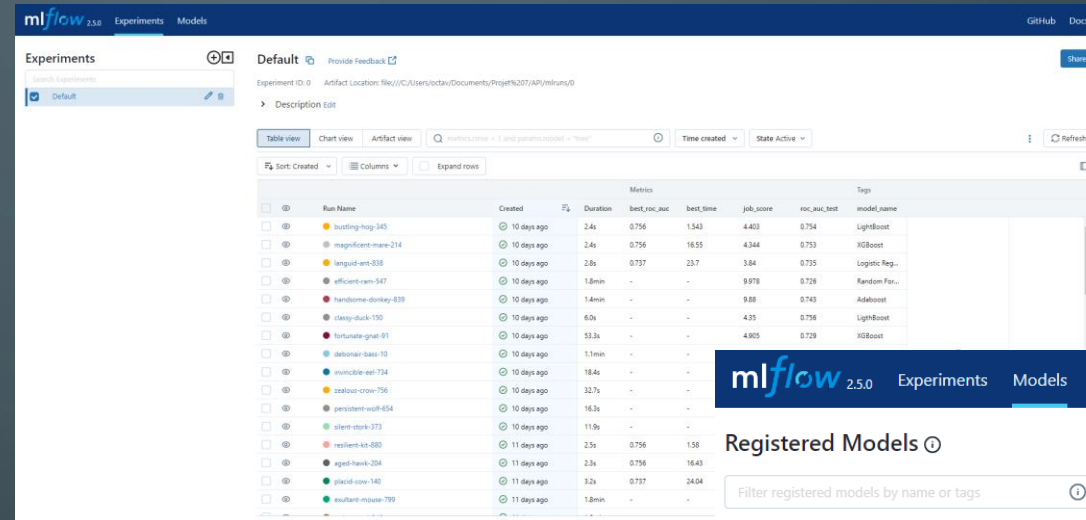
Sélection du modèle

- Calcul du score global
- Enregistrement du modèle automatique

MODÉLISATION - mlflow

Prêt à dépenser

- Fonction tracking :
 - Model
 - Metrics
- Fonction Register:
 - Save model



The screenshot shows the mlflow 2.5.0 Experiments interface. It displays a table of experiment runs with columns for Run Name, Created, Duration, and various metrics. The table is sorted by Created date, showing runs from 10 days ago to 11 days ago. The metrics include best_val_auc, best_time, job_score, and auc_val_test. The interface also includes a search bar, a filter for 'Time created', and a 'State Active' dropdown.

Run Name	Created	Duration	best_val_auc	best_time	job_score	auc_val_test	model_name
buckling-hog-345	10 days ago	2.4s	0.756	1.543	4.403	0.754	LightBoost
magnificent-mare-214	10 days ago	2.4s	0.756	16.55	4.344	0.753	XGBoost
longshot-art-858	10 days ago	2.8s	0.737	23.7	3.84	0.735	Logistic Reg...
efficient-ran-547	10 days ago	1.8min	-	-	9.978	0.728	Random For...
handsome-donkey-838	10 days ago	1.8min	-	-	9.88	0.743	AdaBoost
classy-duck-150	10 days ago	6.0s	-	-	4.35	0.756	LightBoost
fortunate-grail-91	10 days ago	53.3s	-	-	4.905	0.729	XGBoost
debonair-baso-10	10 days ago	1.7min	-	-	-	-	-
invincible-eel-734	10 days ago	18.4s	-	-	-	-	-
zealous-crow-756	10 days ago	32.7s	-	-	-	-	-
persistent-wolf-854	10 days ago	16.3s	-	-	-	-	-
stern-stork-373	10 days ago	11.9s	-	-	-	-	-
restless-kn-880	11 days ago	2.3s	0.756	1.58	-	-	-
aged-hawk-204	11 days ago	2.3s	0.756	16.43	-	-	-
placid-cow-340	11 days ago	3.2s	0.737	24.64	-	-	-
esultant-mouse-789	11 days ago	1.8min	-	-	-	-	-

Registered Models

Filter registered models by name or tags

Name

Latest version

LightBoost

Version 2

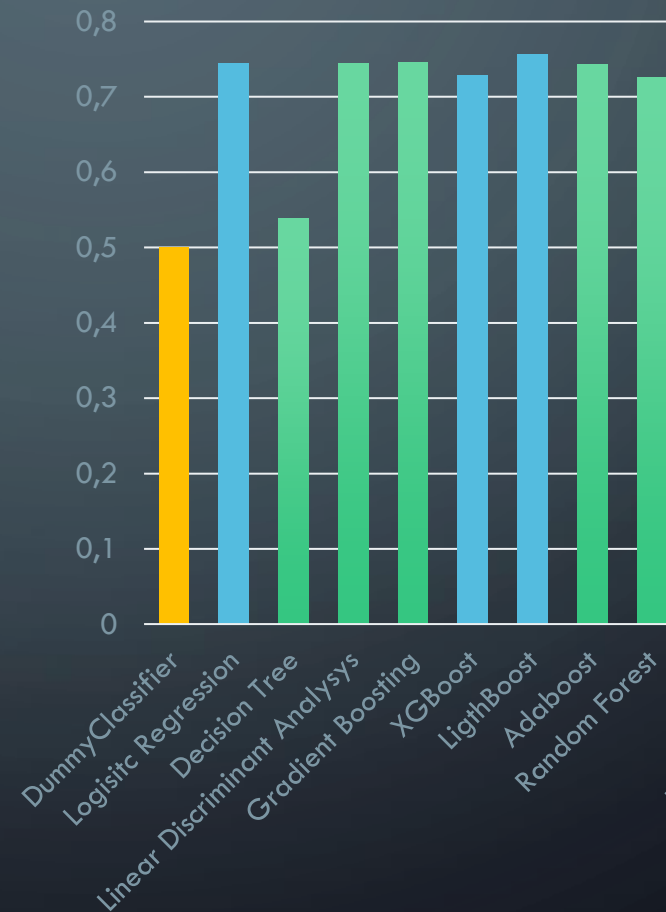
Serveur et stockage en local



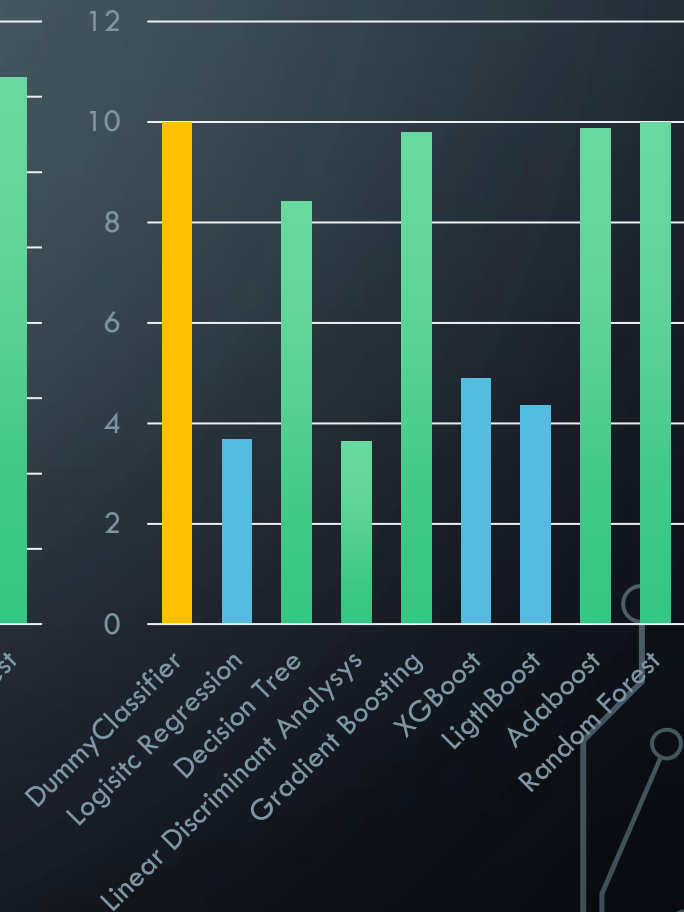
MODÉLISATION - RÉSULTATS

Model Name	Roc auc score	Job score	Time
DummyClassifier	0.5	10	11.9s
Logisitic Regression	0.744	3.687	16.3s
Decision Tree	0.539	8.41	32.7s
Linear Discriminant Analysys	0.745	3.636	18.4s
Gradient Boosting	0.746	9.794	1.1 min
XGBoost	0.729	4.905	53.3s
LigthBoost	0.756	4.35	6.0s
Adaboost	0.743	9.88	1.4min
Random Forest	0.726	9.978	1.8min

Roc auc score



Job score

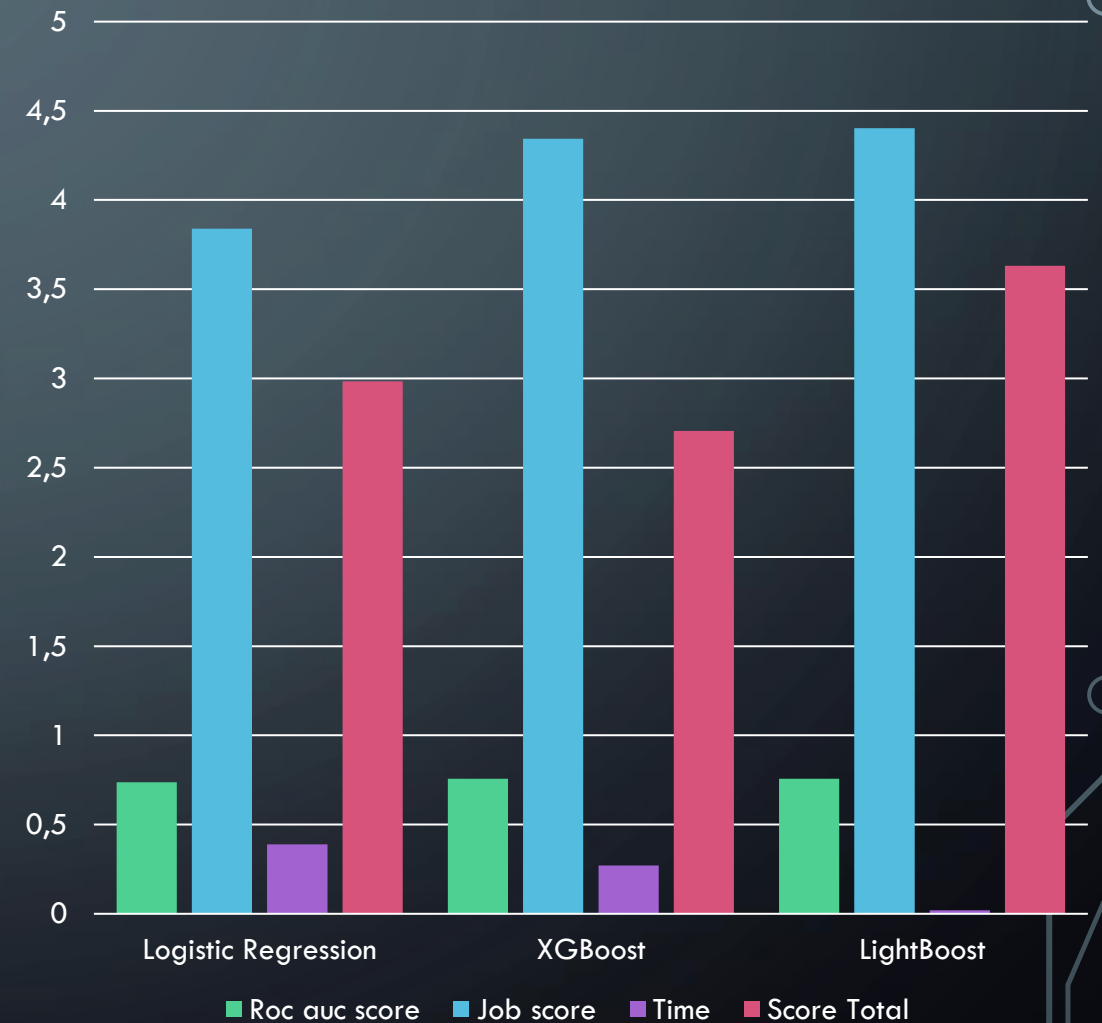


MODÉLISATION - RÉSULTATS

Prêt à dépenser

Score final

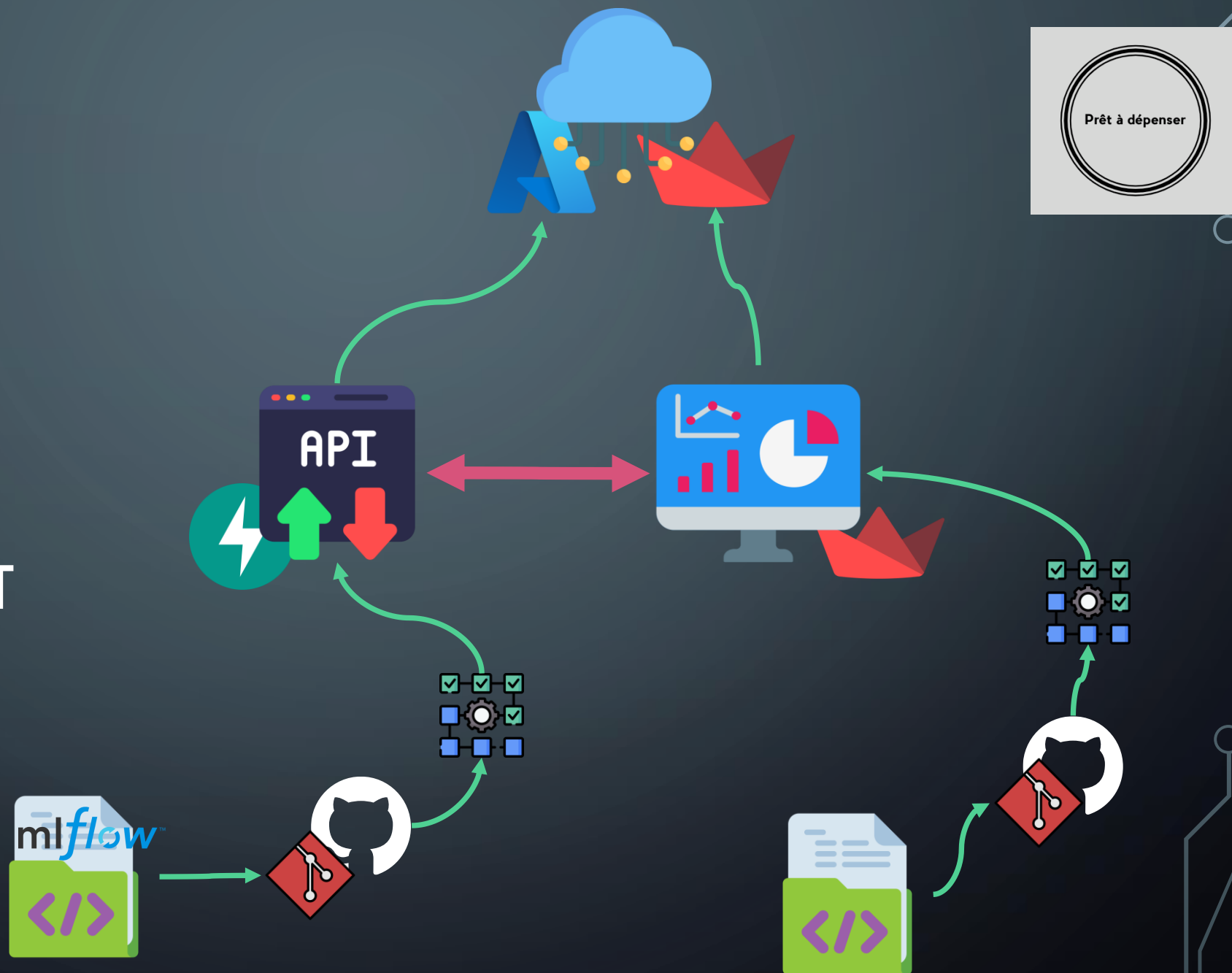
Model Name	Roc auc score	Job score	Time	Score Total
Logistic Regression	0.737	3.84	23.7	2.983
XGBoost	0.756	4.344	16.55	2.706
LightBoost	0.756	4.403	1.543	3.632



DÉPLOIEMENT

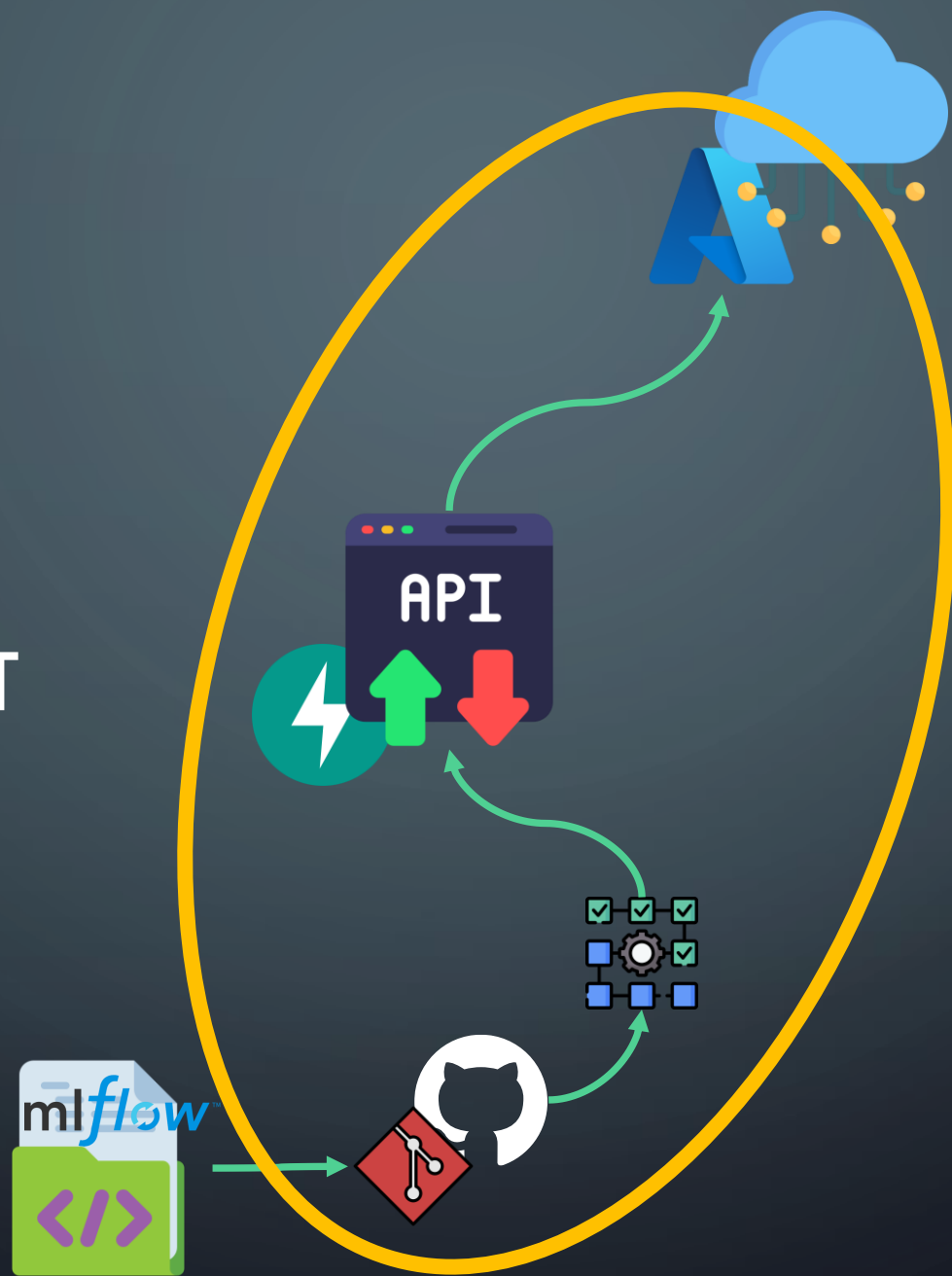
API

DASHBOARD



DÉPLOIEMENT API

FASTAPI
GITHUB
PYTEST
ACTIONS
AZURE



Prêt à dépenser

DÉPLOIEMENT - API

- FastAPI :
 - Chargement des données
 - Chargement du model
 - 3 Fonctions
- Pytest
- GitHub

The screenshot shows the GitHub repository page for 'API_P7' by user 'OPouillot'. The repository is public and has 1 branch and 0 tags. The file list includes: .github/workflows, .ipynb_checkpoints, Model, tests, .gitignore, EDA_notebook.ipynb, First_modeling.ipynb, README.md, df_test_cleaned_3000.csv, main.py, model.pkl, model_selection.py, requirements.txt, and startup.sh. The README.md file is open, showing the title 'Projet7' and a description: 'Ce dépôt GitHub fait parti du Projet 7 de la formation Data Scientist OpenClassRooms. Le projet est composé de plusieurs partie :'. The right sidebar shows repository statistics: 0 stars, 1 watching, 0 forks. It also has sections for Releases, Packages, Deployments (54 total, with 19 in Production), and Languages (Jupyter Notebook 99.8%, Other 0.2%).

Prêt à dépenser

FastAPI 0.1.0 OAS 3.1
/openapi.json

group Return all client prediction and a specific feature

GET /group/ Get Group

feat_imp Return features importance

GET /feat_imp/ Get Shap

customer Return prediction, probabilities and info about a specific client

GET /customer/ Get Predict

DÉPLOIEMENT - API

GitHub Actions

- Build
 - Install env
 - Tests
- Deploy
 - if build passed
 - Azure deploy

Prêt à dépenser

Commits

main

Commits on Sep 18, 2023

fix: test for feat_imp funct

OPouillot committed 17 minutes ago ✓

fix: add comment to functions

OPouillot committed 20 minutes ago ✗

Commits on Sep 13, 2023

update readme

OPouillot committed 5 days ago ✓

Last changes & readme edit

OPouillot committed 5 days ago ✓

Commits on Sep 5, 2023

little test update

OPouillot committed 2 weeks ago ✓

add test script for api

OPouillot committed 2 weeks ago ✓

Summary

Jobs

- ✓ build
- ✓ deploy

Run details

- Usage
- Workflow file

build

succeeded 17 minutes ago in 1m 6s

Test with pytest

```
18 Downloading pluggy-1.3.0-py3-none-any.whl (18 kB)
19 Installing collected packages: pluggy, iniconfig, pytest
20 Successfully installed iniconfig-2.0.0 pluggy-1.3.0 pytest-7.4.2
21 ===== test session starts =====
22 platform linux -- Python 3.11.5, pytest-7.4.2, pluggy-1.3.0
23 rootdir: /home/runner/work/API_P7/API_P7
24 plugins: anyio-3.7.1
25 collected 8 items
26
27 tests/test_main.py .... [ 50%]
28 tests/test_model_selection.py .... [100%]
29
30 ===== warnings summary =====
31 ../../../../opt/hostedtoolcache/Python/3.11.5/x64/lib/python3.11/site-packages/httpx/_models.py:1
32 /opt/hostedtoolcache/Python/3.11.5/x64/lib/python3.11/site-packages/httpx/_models.py:1: DeprecationWarning: 'cgi' is deprecated and slated for removal in Python 3.13
33 import cgi
34
35 tests/test_main.py::test_get_group
36 tests/test_main.py::test_get_predict
37 tests/test_main.py::test_get_predict
38 /opt/hostedtoolcache/Python/3.11.5/x64/lib/python3.11/site-packages/pandas/core/dtypes/cast.py:1641: DeprecationWarning: np.find_common_type is deprecated. Please use 'np.result_type' or 'np.promote_types'.
39 See https://numpy.org/devdocs/release/1.25.0-notes.html and the docs for more information. (Deprecated NumPy 1.25)
40 return np.find_common_type(types, [])
41
42 tests/test_main.py::test_get_predict
43 /home/runner/work/API_P7/API_P7/main.py:60: DeprecationWarning: Conversion of an array with ndim > 0 to a scalar is deprecated, and will error in future. Ensure you extract a single element from your array before
44 performing this operation. (Deprecated NumPy 1.25.)
45 prediction = int(model_pipe.predict(id_features))
46
47 -- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
48 ===== 8 passed, 5 warnings in 3.13s =====
```

- ✓ Upload artifact for deployment jobs
- ✓ Post Run actions/checkout@v2
- ✓ Complete job

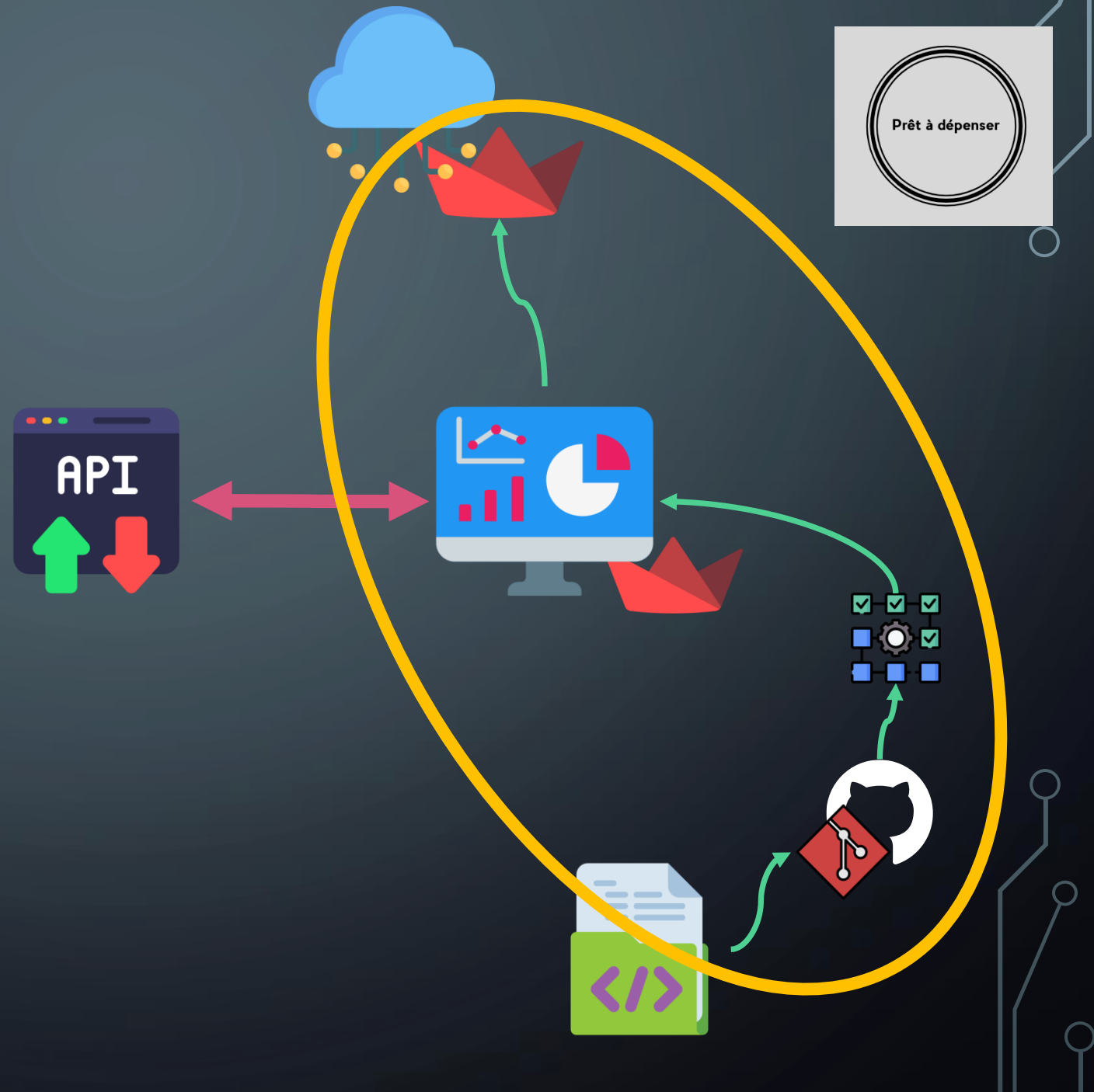
DÉPLOIEMENT DASHBOARD

STREAMLIT

GITHUB

PYTEST

ACTIONS



Prêt à dépenser

DÉPLOIEMENT - DASHBOARD

- Streamlit :
 - Info prêt
 - Données client
 - Ensemble clients
- Pytest
- GitHub

The screenshot shows the GitHub interface for a repository named 'Dashboard_P7' (Public). The repository is on the 'main' branch, has 2 branches, and 0 tags. It includes buttons for 'Go to file', 'Add file', and 'Code'. The commit history shows a merge of updates into main by OPouillot, with 59 commits. The file list includes .github/workflows, .streamlit, __pycache__, tests, README.md, dashboard.py, pad.PNG, and requirements.txt. The right sidebar shows repository statistics: 0 stars, 1 watching, and 0 forks. The 'About' section describes the project as 'Dashboard - Projet 7 Formation Data Scientist OpenClassRooms'. The 'Releases' section indicates no releases are published. The 'Packages' section shows no packages published. The 'Languages' section shows Python at 100.0%. The 'README.md' file is open, showing the project title 'Projet7' and a description of the project as part of the OpenClassRooms Data Scientist Formation Project 7. It lists the components as API and Dashboard. The description states that the project is a credit scoring tool developed by the financial company 'Prêt à dépenser'.

Dashboard_P7 Public

main 2 branches 0 tags

Go to file Add file <> Code

OPouillot Merge updates into main c871061 26 minutes ago 59 commits

.github/workflows	verification merge deploy streamlit	2 weeks ago
.streamlit	first commit of Dashboard repo	3 weeks ago
__pycache__	add test and github action routine	2 weeks ago
tests	add test and github action routine	2 weeks ago
README.md	update readme	5 days ago
dashboard.py	fix: rename api url	27 minutes ago
pad.PNG	first commit of Dashboard repo	3 weeks ago
requirements.txt	update req.txt for adding scipy	2 weeks ago

README.md

Projet7

Ce dépôt GitHub fait parti du Projet 7 de la formation Data Scientist OpenClassRooms. Le projet est composé de plusieurs partie :

- API
- Dashboard

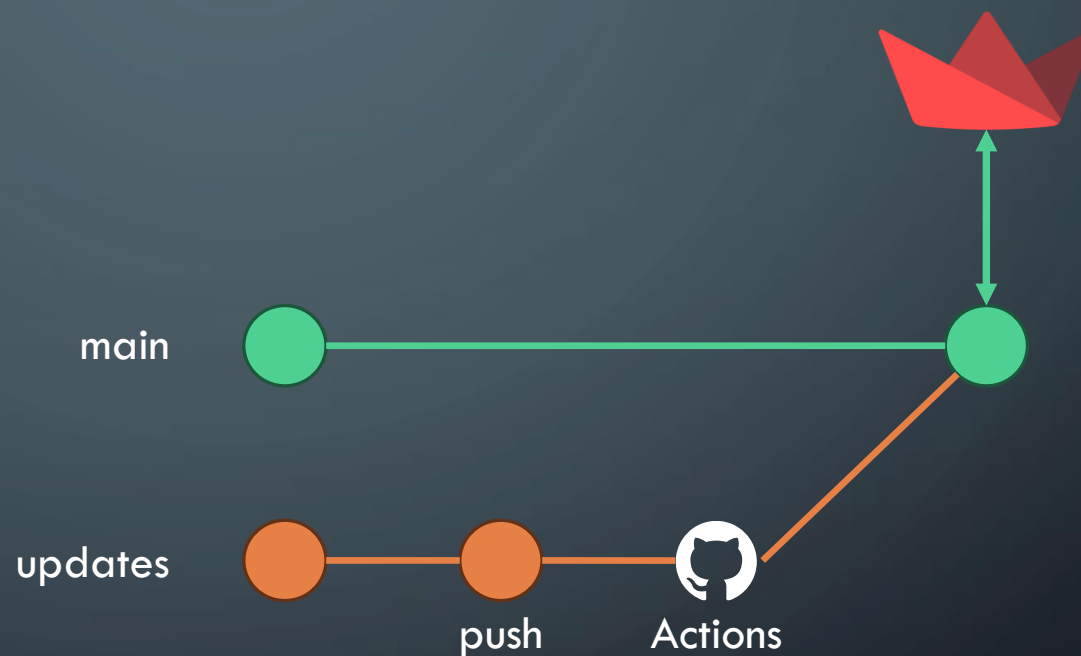
Description du projet

La société financière "Prêt à dépenser", propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt. L'entreprise souhaite mettre en œuvre un outil de scoring crédit qui calcule la probabilité qu'un client rembourse son crédit, puis classe la demande en crédit accordé ou refusé. Elle souhaite donc développer un algorithme de classification en s'appuyant sur des sources de données variées (données

DÉPLOIEMENT - DASHBOARD

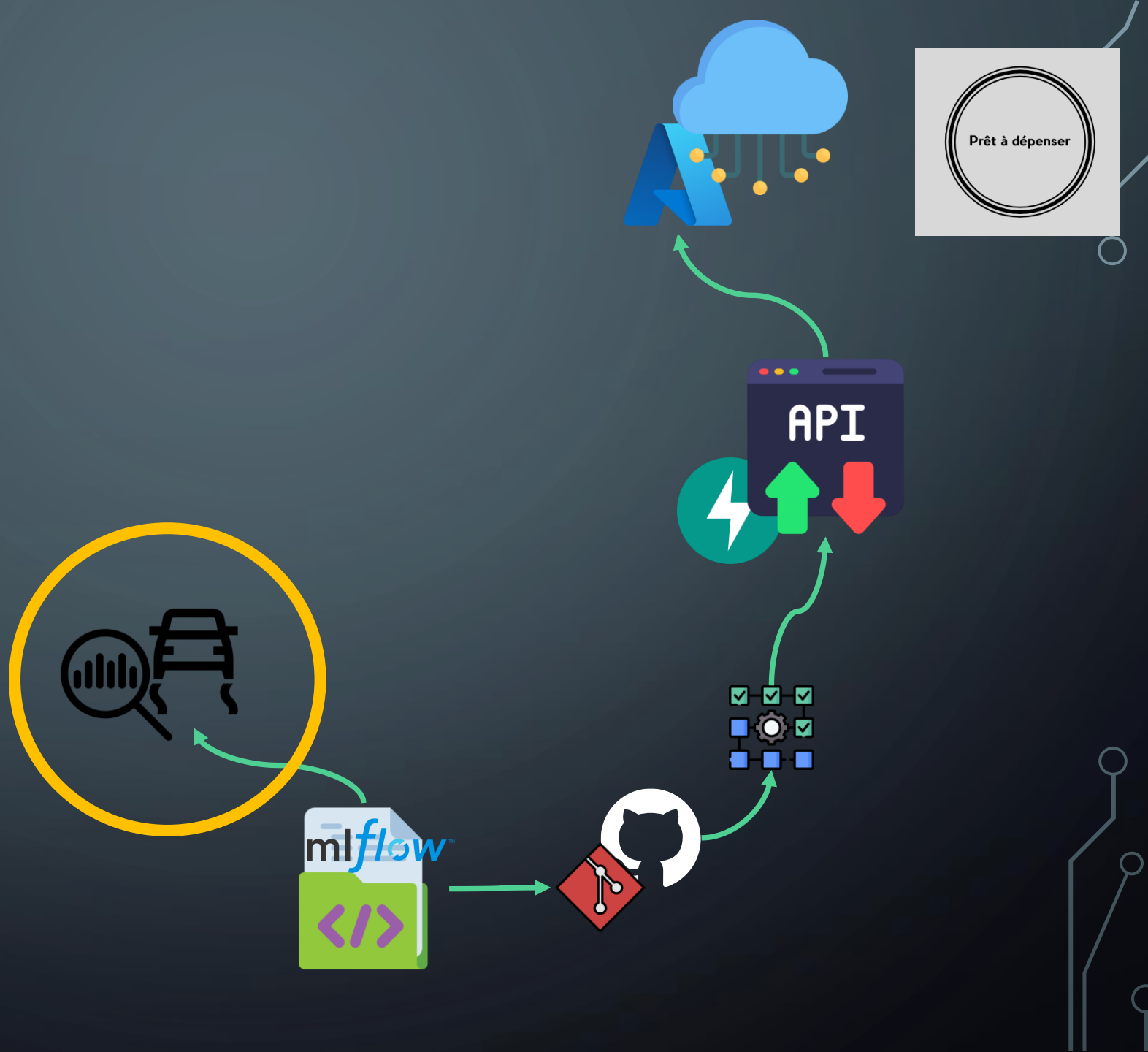
Push updates :

- Build
 - Install env
 - Tests
- Merge on main
 - if build passed



DATA DRIFT

EVIDENTLY AI



DATA DRIFT

- Dataset train et test cleaned (207 features)
- Sample : 10 000
- Drift : 5.8% du dataset (12 features)
- Niveau de vie et pouvoir d'achat
 - revenus annuels
 - Crédit demandé
 - Patrimoine
 - ...
- Seuil maximum → actualiser la modélisation par rapport à des variables comme l'inflation et le PIB.

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> DEBT_TO_INCOME_RATIO	num			Detected	Wasserstein distance (normed)	0.292271
> INCOME_TO_CREDIT	num			Detected	Wasserstein distance (normed)	0.278425
> AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.221456
> AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.218926
> FLAG_DOCUMENT	num			Detected	Jensen-Shannon distance	0.164015
> EXT_SOURCE_1	num			Detected	Wasserstein distance (normed)	0.157719
> NAME_CONTRACT_TYPE_Cash loans	num			Detected	Jensen-Shannon distance	0.151032
> NAME_CONTRACT_TYPE_Revolving loans	num			Detected	Jensen-Shannon distance	0.151032
> AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.148694
> FLAG_EMAIL	num			Detected	Jensen-Shannon distance	0.128973
> DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.120622
> AMT_INCOME_TOTAL	num			Detected	Wasserstein distance (normed)	0.115021

DÉMONSTRATION !

➔ VERS LE [DASHBOARD](#)

➔ VERS L'[API](#)

