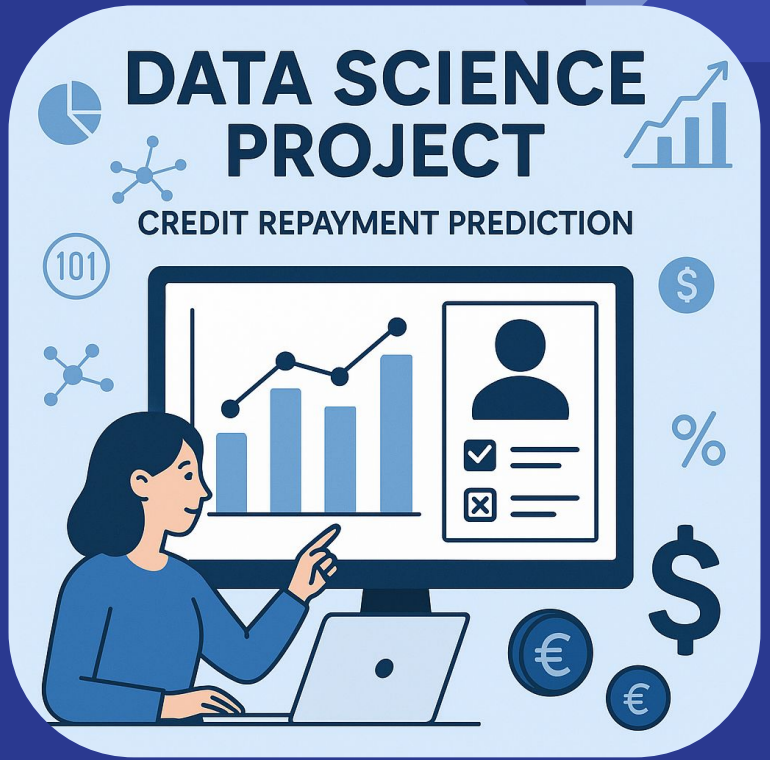
Credit repayment

-

Implémentation d’un modèle de scoring

-

Déploiement dans le Cloud



Owner : Ph. Deflandre

Création : 03/05/2025

Environnement de travail

Environnement Jupyter Lab sur Windows 11 (Edge Version 130.0.2849.56)

Environnement virtuel

Python version 3.11.7

Numpy version 1.26.4

Matplotlib version 3.8.0

Pandas version 2.1.4

Seaborn version 0.12.2

Scikit-learn version 1.5.2

Shap version 0.46.0

Pydantic version 2.11.3

Evidently version 0.7.4

Imblearn version 0.13.0

XGBoost version 3.0.0

Mlflow version 2.21.0

Requests version 2.31.0

Json version 2.0.9

LightGBM version 4.6.0

Uvicorn version 0.34.0

FastAPI version 0.115.12

Streamlit version 1.45.0

Rappel de la mission

- Comparer des modèles de scoring pour déterminer la probabilité qu'un prêt soit remboursé.
- Comprendre les features qui vont être déterminantes dans la prédiction.
- Mettre en production le modèle de scoring le plus performant à l'aide d'une API. Réaliser une interface de test de l'API.
- Créer une interface utilisateur user friendly.
- Analyser un éventuel Data Drift.

Analyse de la base de données

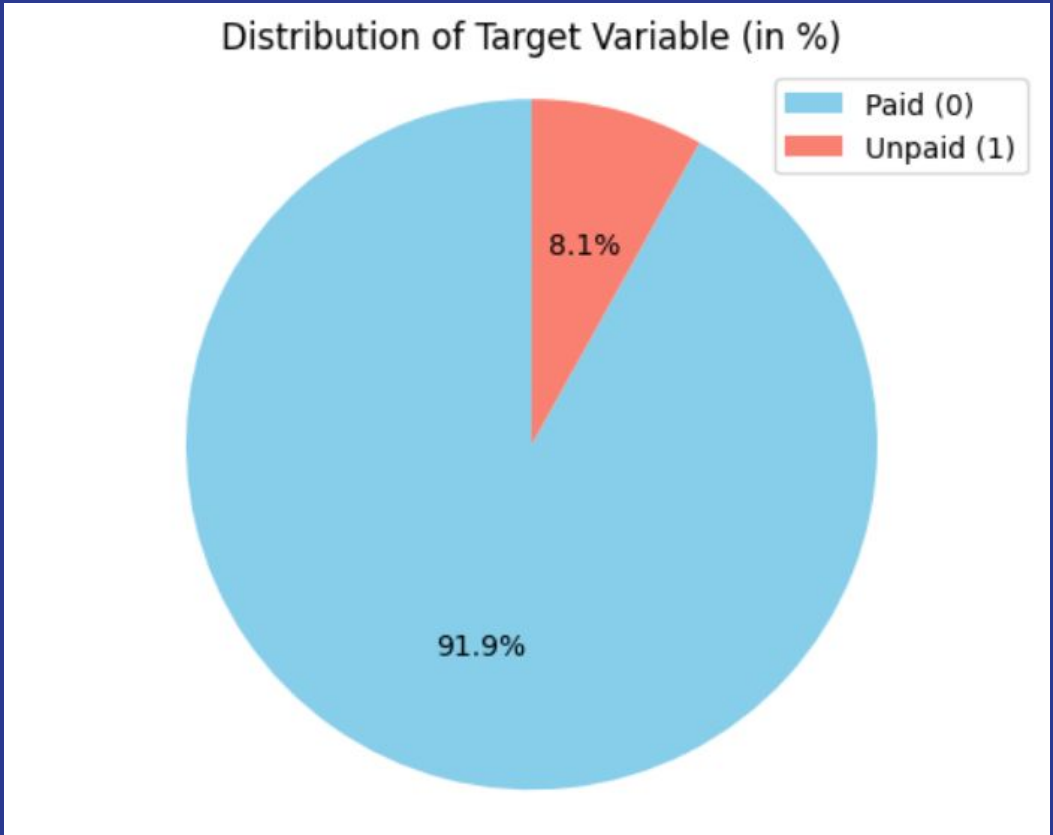
Source : https://www.kaggle.com/c/home-credit-default-risk/data

- application_train.csv
  - 121 features
  - 1 target
  - 307 511 lignes (crédits)

- type de contrat
- informations personnelles
- situation familiale
- informations administratives
- informations relatives au crédit

- application_test.csv
  - 121 features
  - 48 744 lignes (crédits)

Target

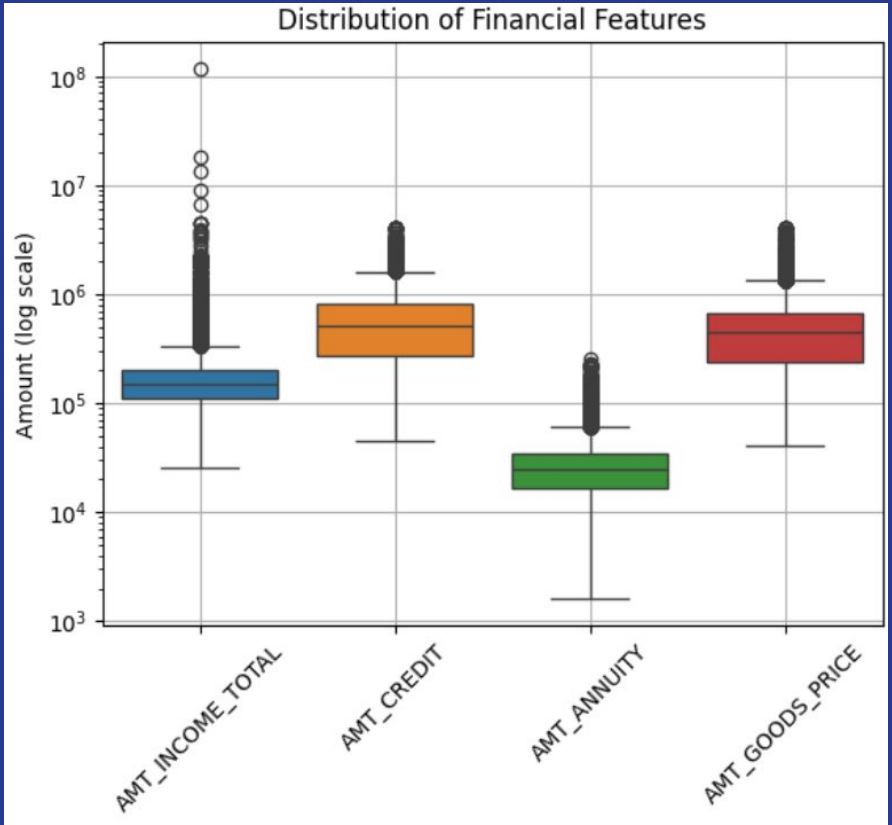


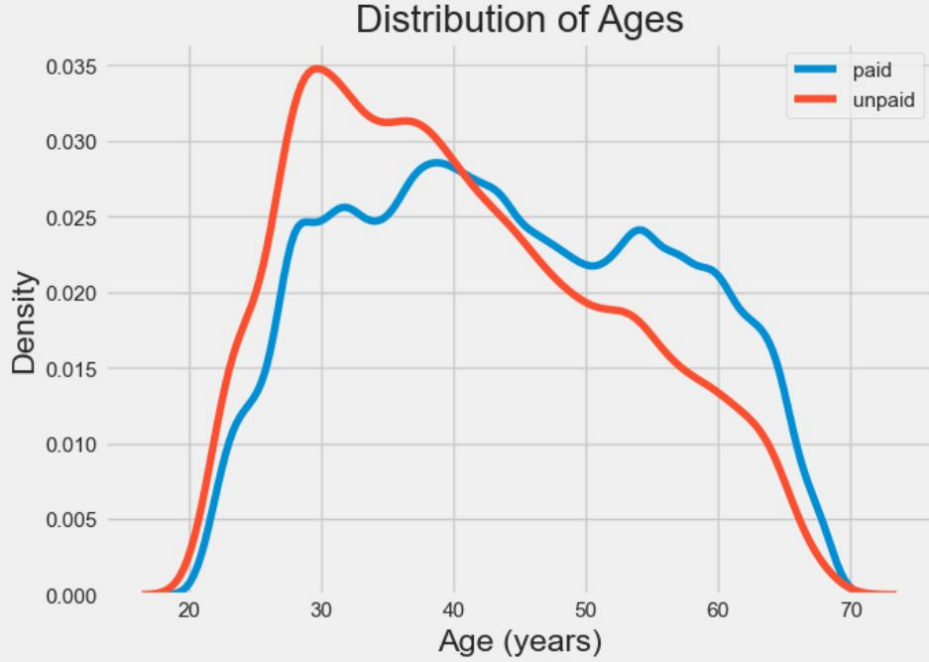
Classification binaire déséquilibrée

0 => le prêt est remboursé

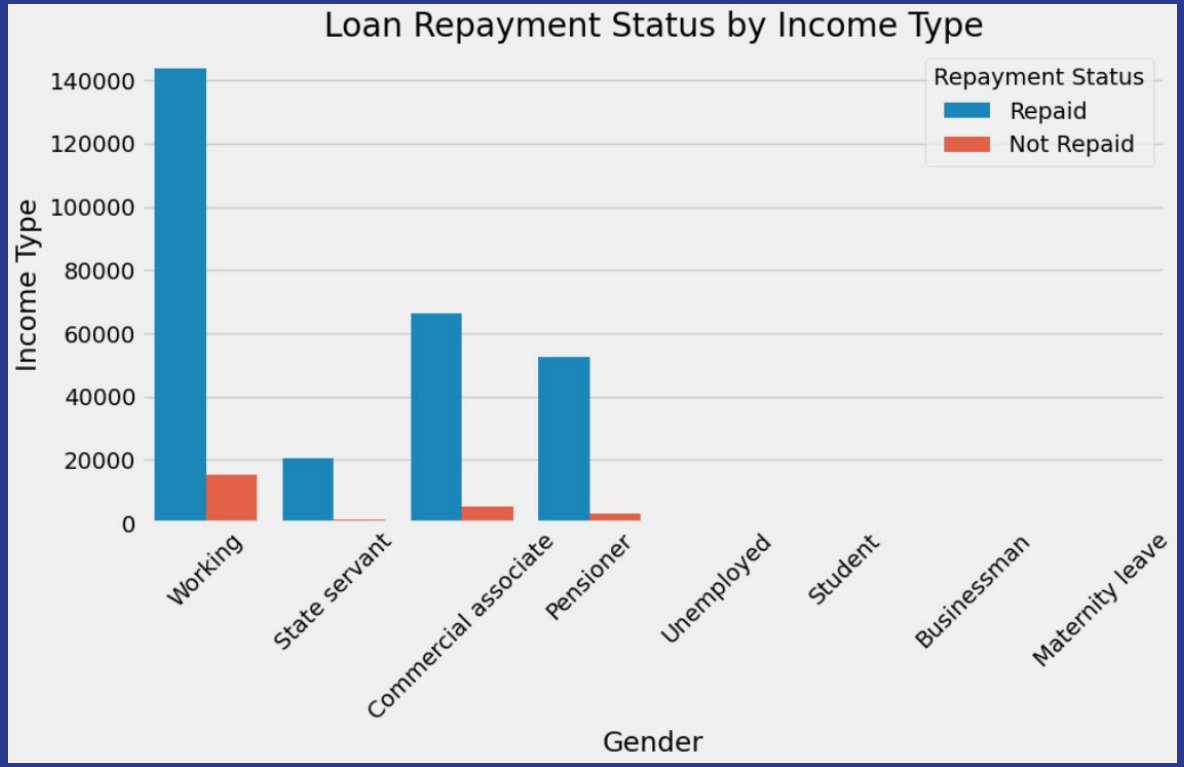
1 => le prêt n’est pas remboursé

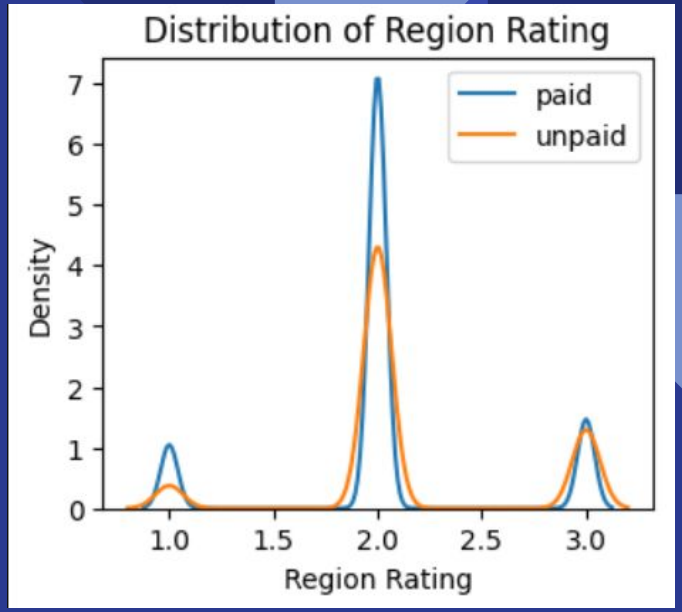
Features

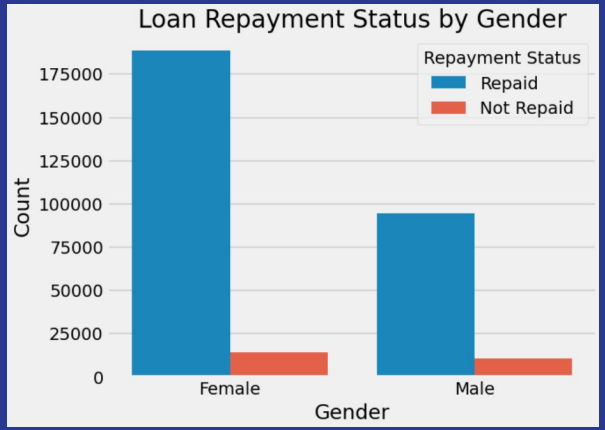

Distribution of Financial Features
Amount (log scale)
10^8
10^7
10^6
10^5
10^4
10^3
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE


Distribution of Ages
paid
unpaid
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
70
Age (years)

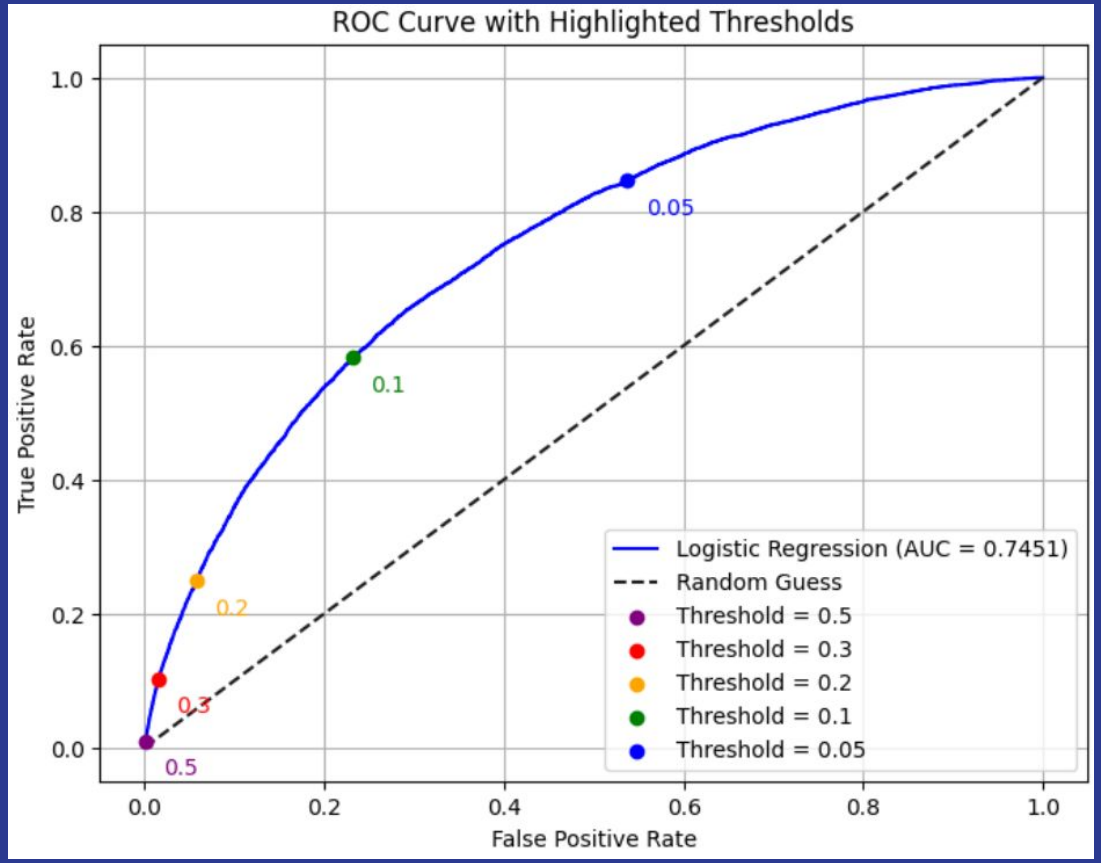
Features


Loan Repayment Status by Income Type
Repayment Status
Repaid
Not Repaid
Income Type
140000
120000
100000
80000
60000
40000
20000
0
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
Gender


Distribution of Region Rating
paid
unpaid
Density
7
6
5
4
3
2
1
0
1.0
1.5
2.0
2.5
3.0
Region Rating


Loan Repayment Status by Gender
Repayment Status
Repaid
Not Repaid
Count
175000
150000
125000
100000
75000
50000
25000
0
Female
Male
Gender

Baseline - Logistic Regression classification



Premier modèle de référence

Comment estimer la qualité de prédiction du modèle ?

```
Train Accuracy: 0.919
Test Accuracy: 0.919
Test Balanced Accuracy: 0.504

Confusion Matrix:
 [[70609    78]
 [ 6132    59]]
```

Métrique business personnalisée

Crédits non accordés qui auraient été remboursés

```
Train Accuracy: 0.919
Test Accuracy: 0.919
Test Balanced Accuracy: 0.504

Confusion Matrix:
 [[70609    78]
 [ 6132    59]]
```

Crédits accordés non remboursés

Feature engineering

Polynomial Features

$$(EXT\_SOURCE\_1)^a \times (EXT\_SOURCE\_2)^b \times (EXT\_SOURCE\_3)^c \times (DAYS\_BIRTH)^d$$

$$(a, b, c, d) \in [\![0; 3]\!]$$

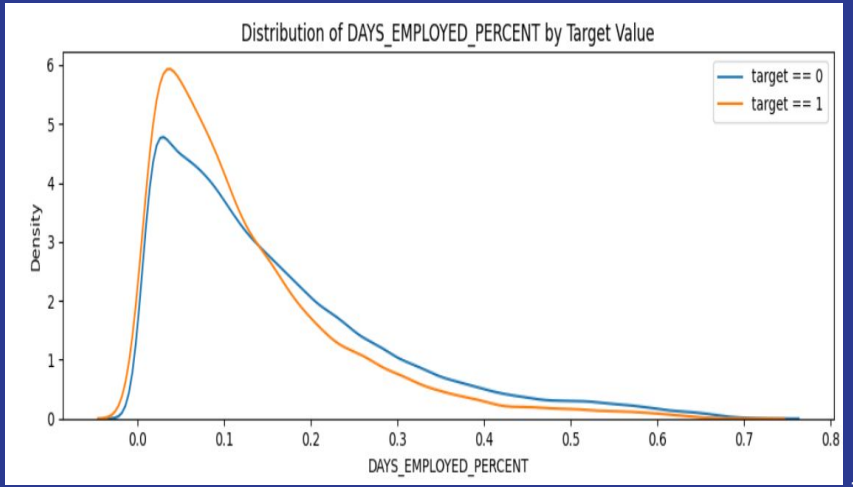
Features “métier”

$$taux\ d'endettement = \frac{montant\ du\ crédit}{revenus}$$

$$taux\ d'effort\ annuel = \frac{annuité}{revenus}$$

$$ratio\ de\ remboursement = \frac{annuité}{montant\ du\ crédit}$$

$$taux\ d'emploi\ relatif = \frac{jours\ travaillés}{âge}$$



Comparaison des modèles

Logistic Regression with features engineering

- class_weight hyperparamètre
- cross-validation
- MLFlow
- utilité du Feature Engineering

	Feature	Coefficient	Absolute_Coefficient
7	AMT_ANNUITY	2.494649	2.494649
8	AMT_GOODS_PRICE	-2.436208	2.436208
80	DEF_60_CNT_SOCIAL_CIRCLE	1.737929	1.737929
278	ANNUITY_INCOME_PERCENT	1.648104	1.648104
78	DEF_30_CNT_SOCIAL_CIRCLE	1.533950	1.533950
10	DAYS_BIRTH	1.145663	1.145663
246	EXT_SOURCE_3_y	-0.954602	0.954602
252	EXT_SOURCE_2 EXT_SOURCE_3	-0.846523	0.846523
33	EXT_SOURCE_3_x	-0.820314	0.820314
93	FLAG_DOCUMENT_13	-0.785311	0.785311

Comparaison des modèles

Random Forest

- class_weight hyperparamètre
- problème de surapprentissage

AdaBoost

- class_weight hyperparamètre
- optimisation des hyperparamètres avec GridSearchCV
- entraînement lent du modèle

GradientBoosting

- optimisation des hyperparamètres avec GridSearchCV
- amélioration de l'AUC
- pas d'hyperparamètre d'équilibrage des classes

Comparaison des modèles

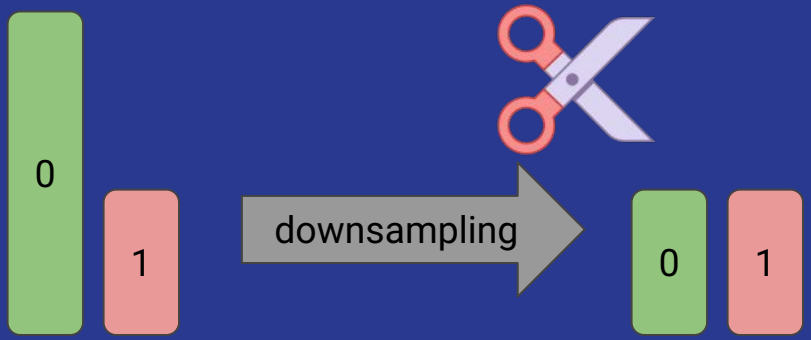
XGBoost

- scale_pos_weight hyperparamètre
- optimisation des hyperparamètres avec GridSearchCV

LightBoosting

- scale_pos_weight hyperparamètre
- optimisation des hyperparamètres avec GridSearchCV
- rapide

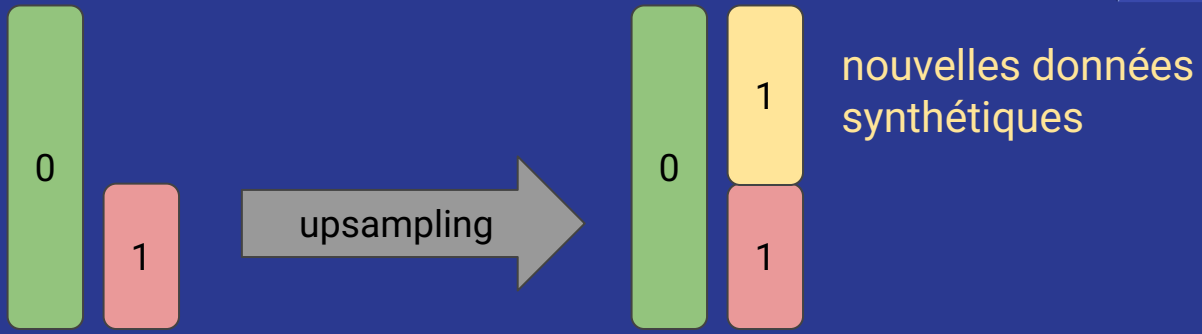
Downsampling - imblearn



XGBoost

- scale_pos_weight = 1
- optimisation des hyperparamètres avec GridSearchCV

Upsampling - SMOTE

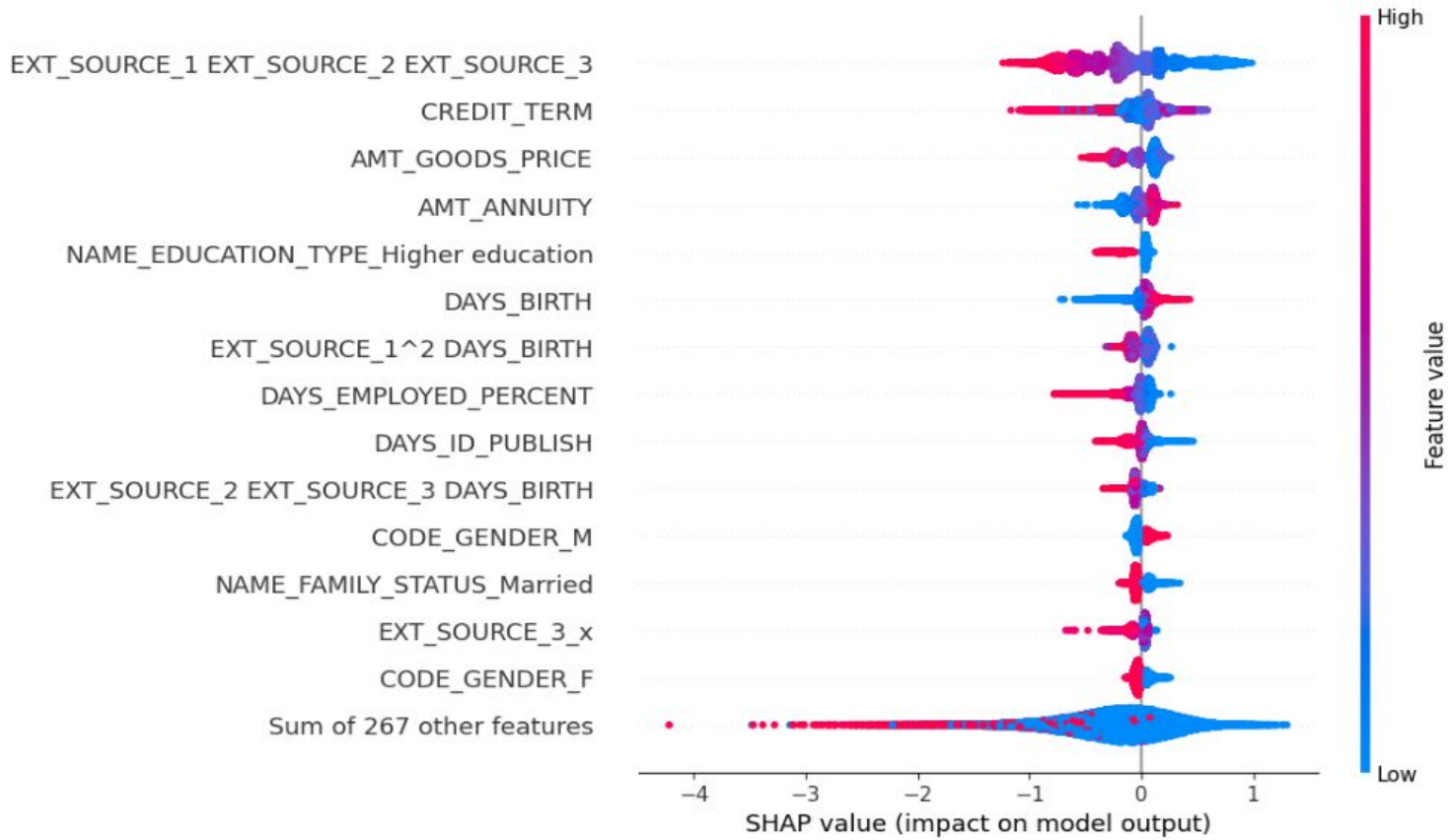


nouvelles données synthétiques

XGBoost

- scale_pos_weight = 1
- hyperparamètres du downsampling

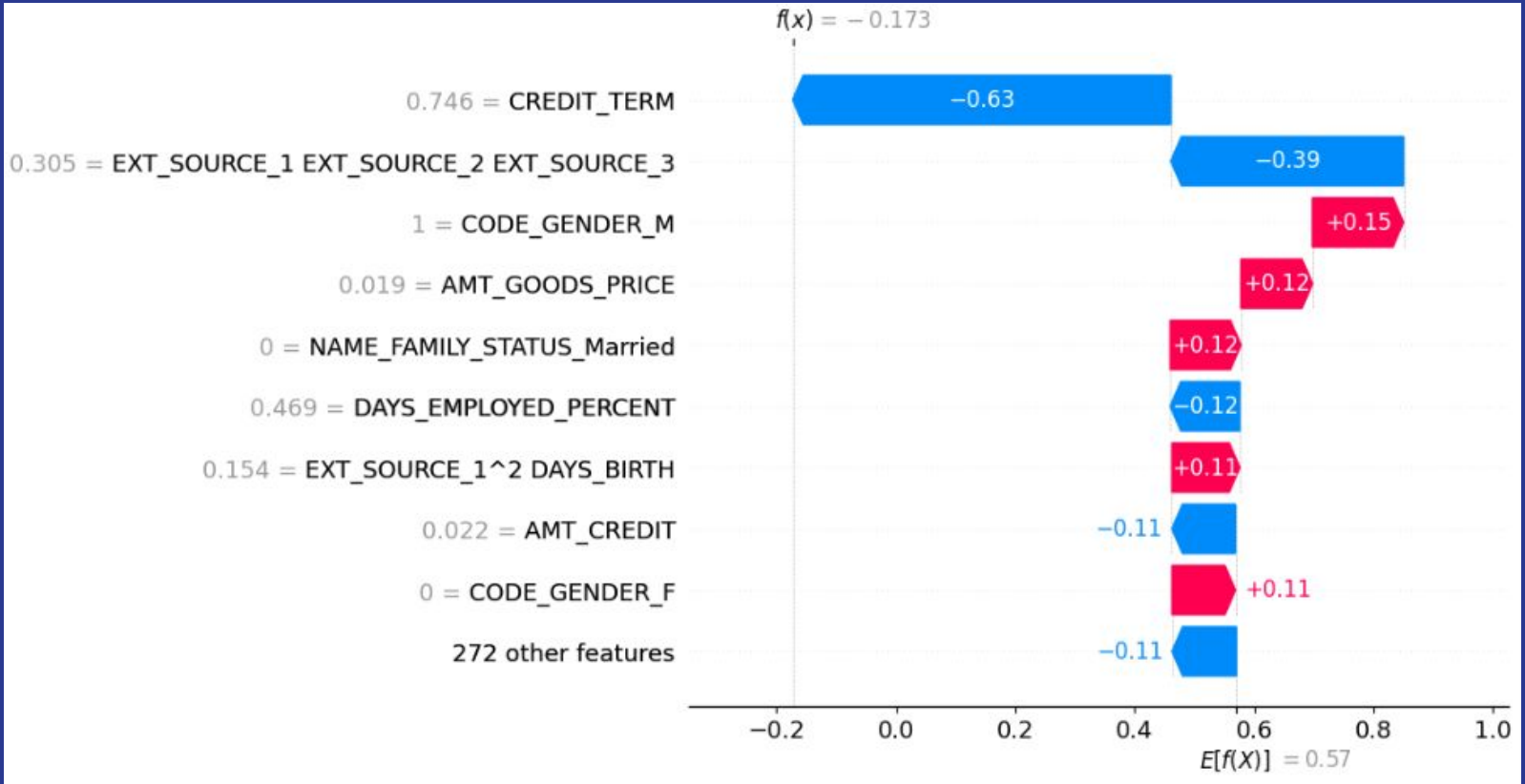
Interprétabilité globale - XGBoost



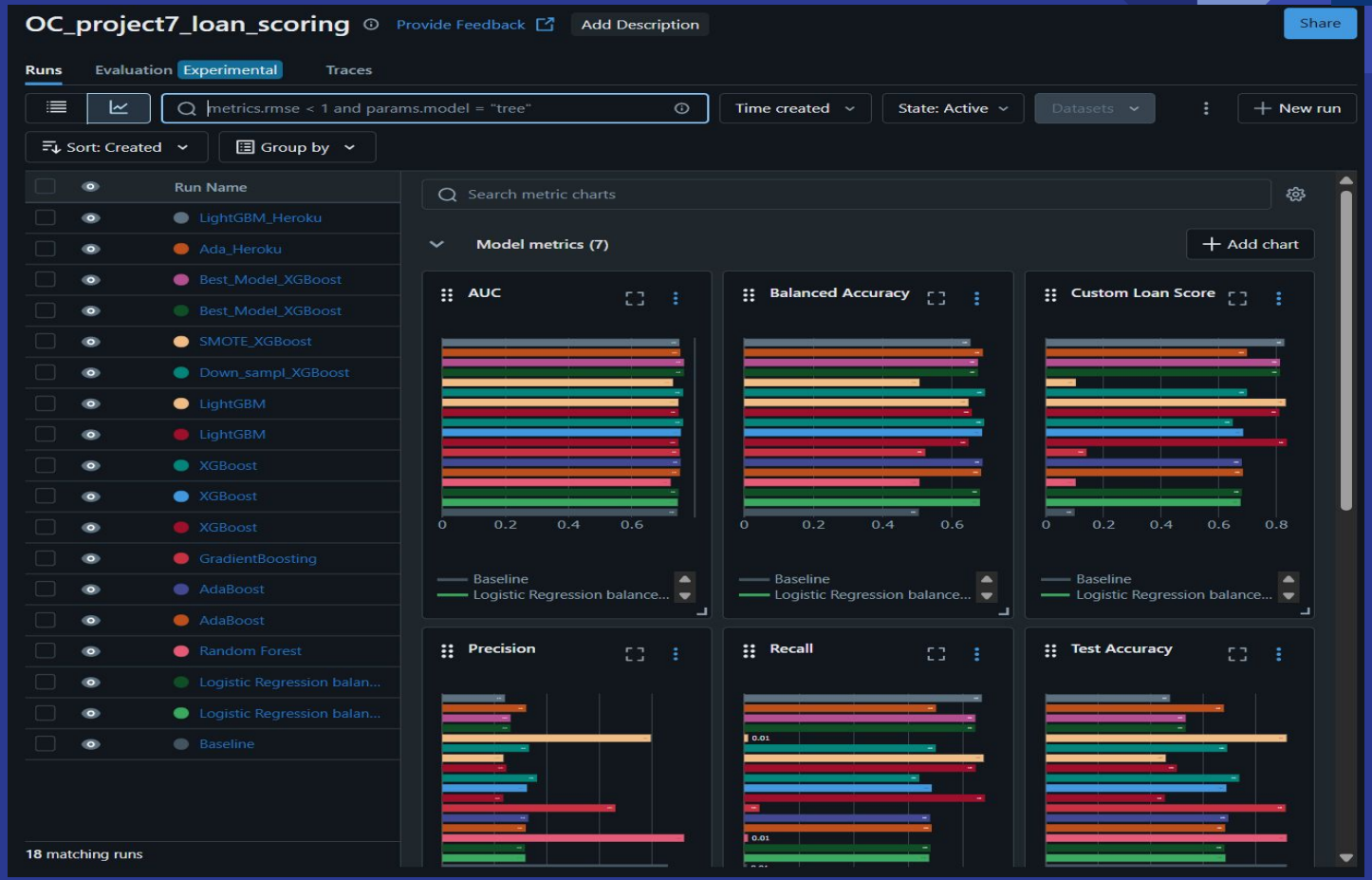
Interprétabilité locale - XGBoost

exemple prêt n°3

SHAP

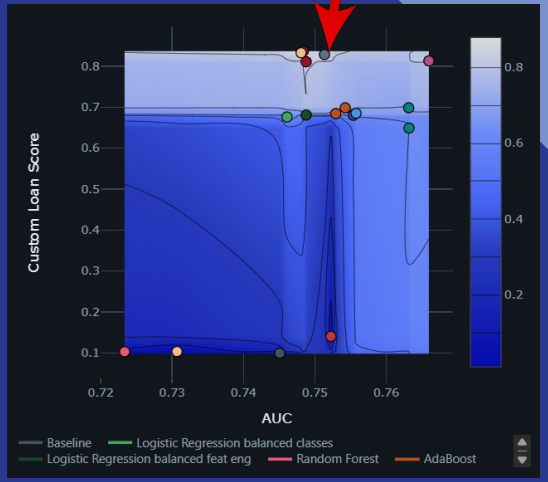


Tracking d'expérimentation - UI MLFlow



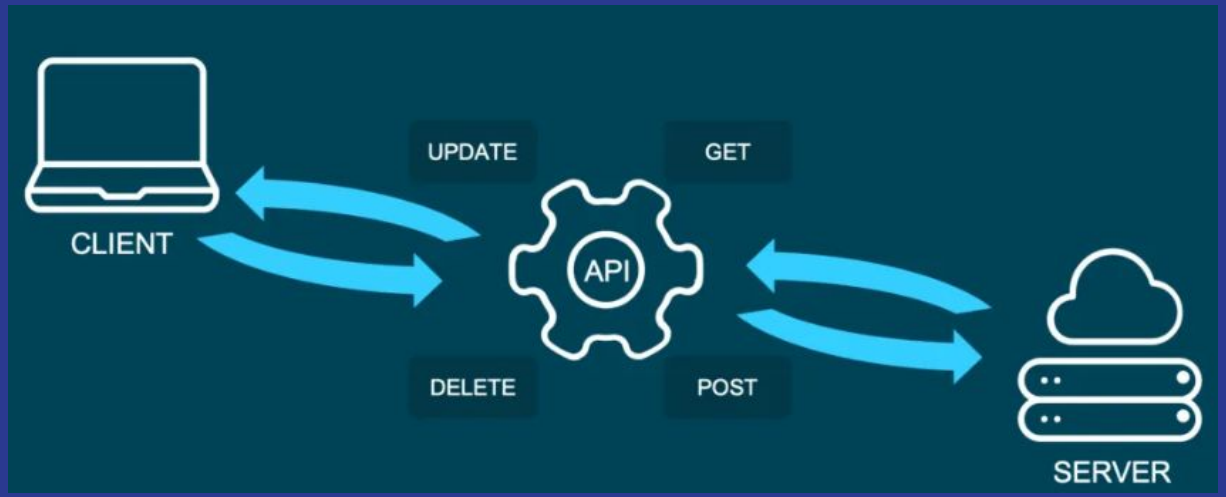
Choix du modèle à déployer en production

Custom Loan Score



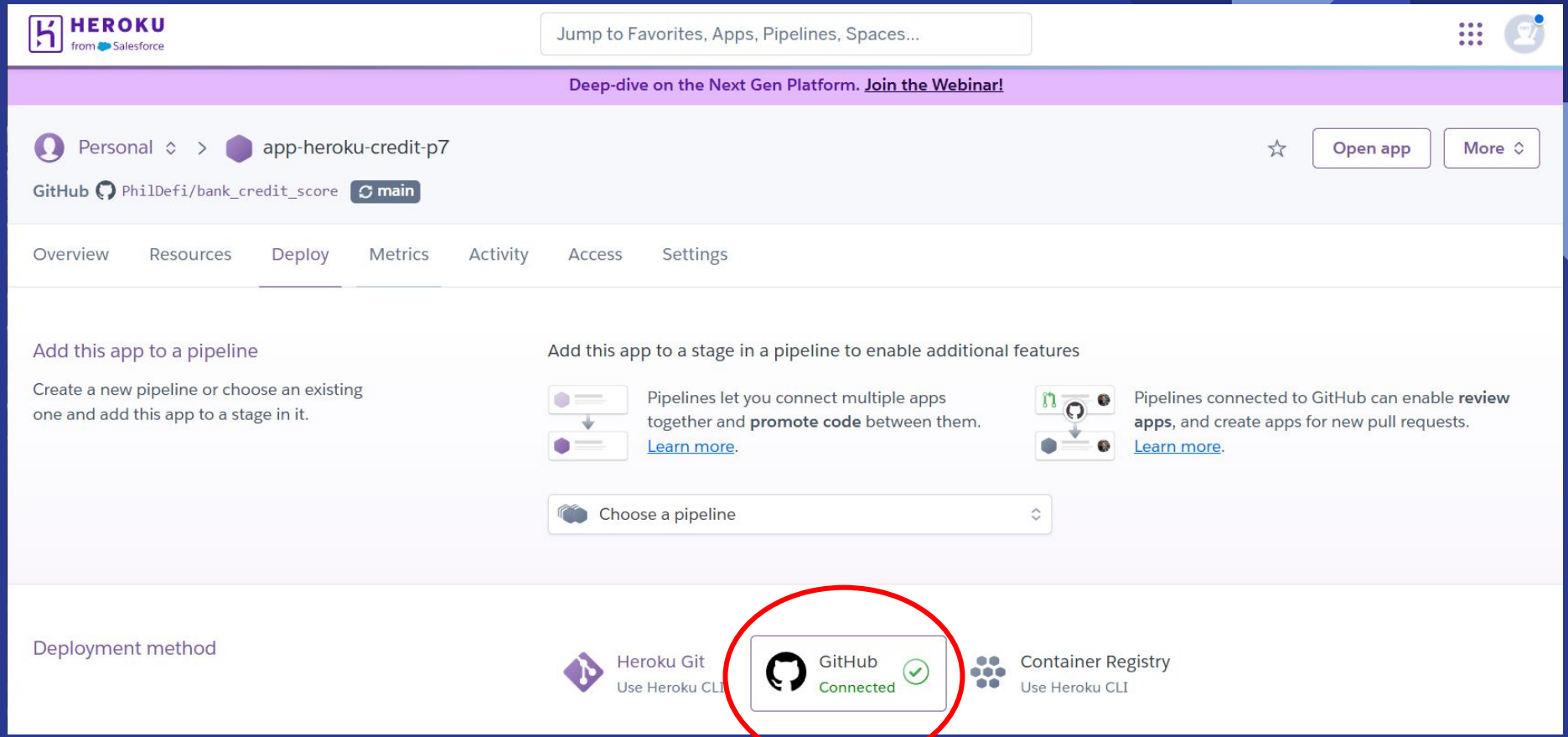
- combinaisons de plusieurs métriques
- taille des dependencies nécessaires sur Heroku !

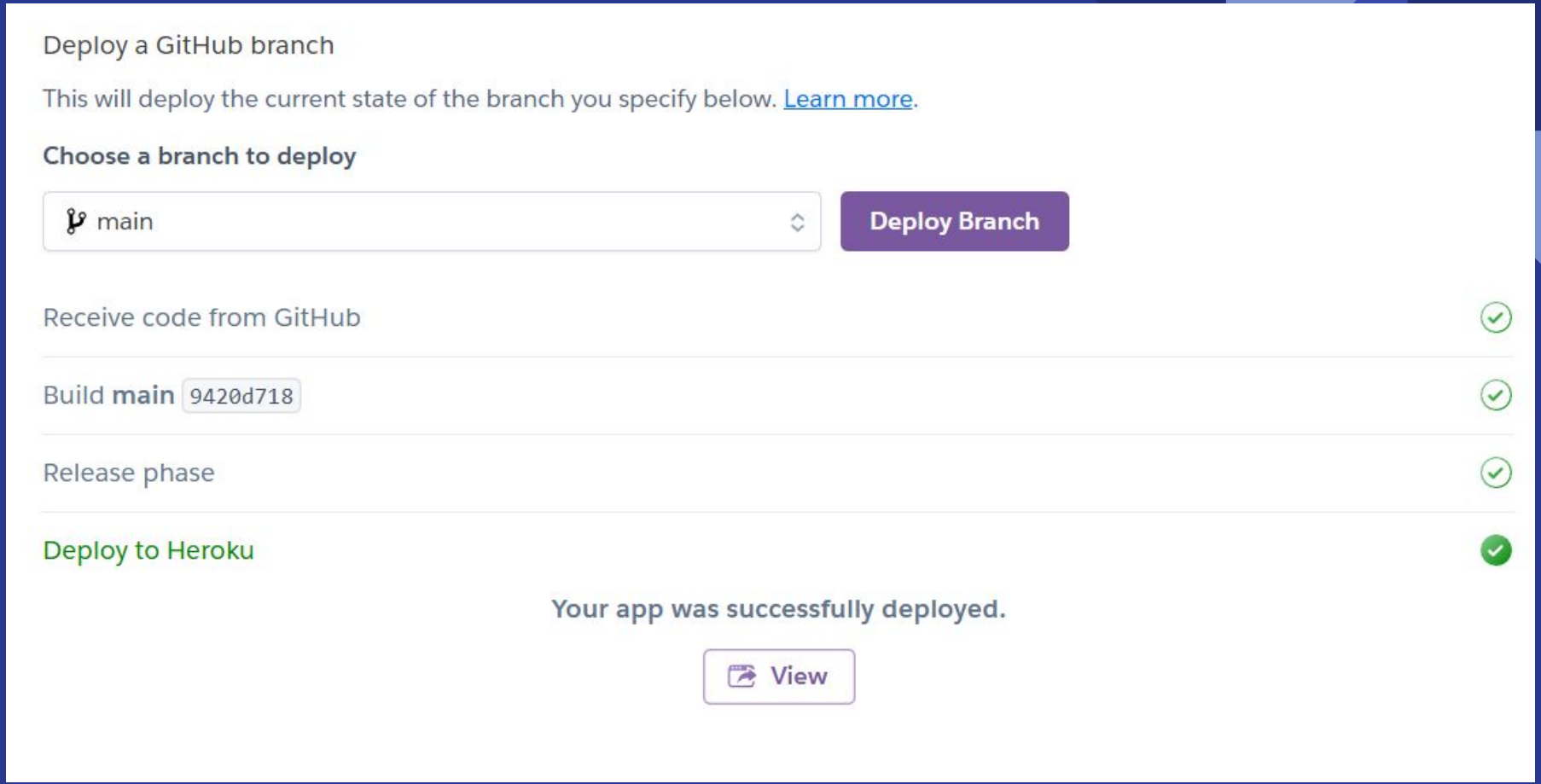
API

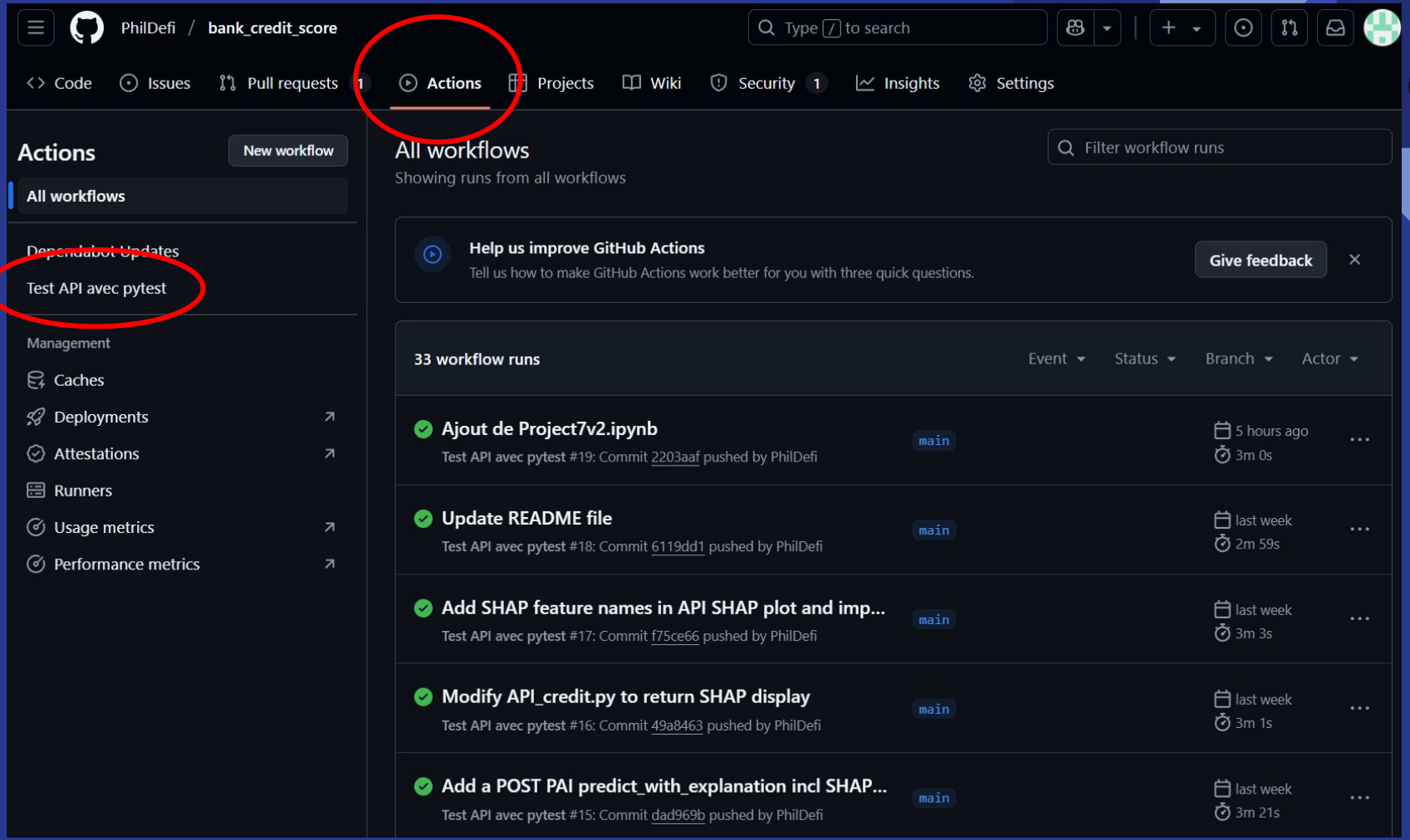


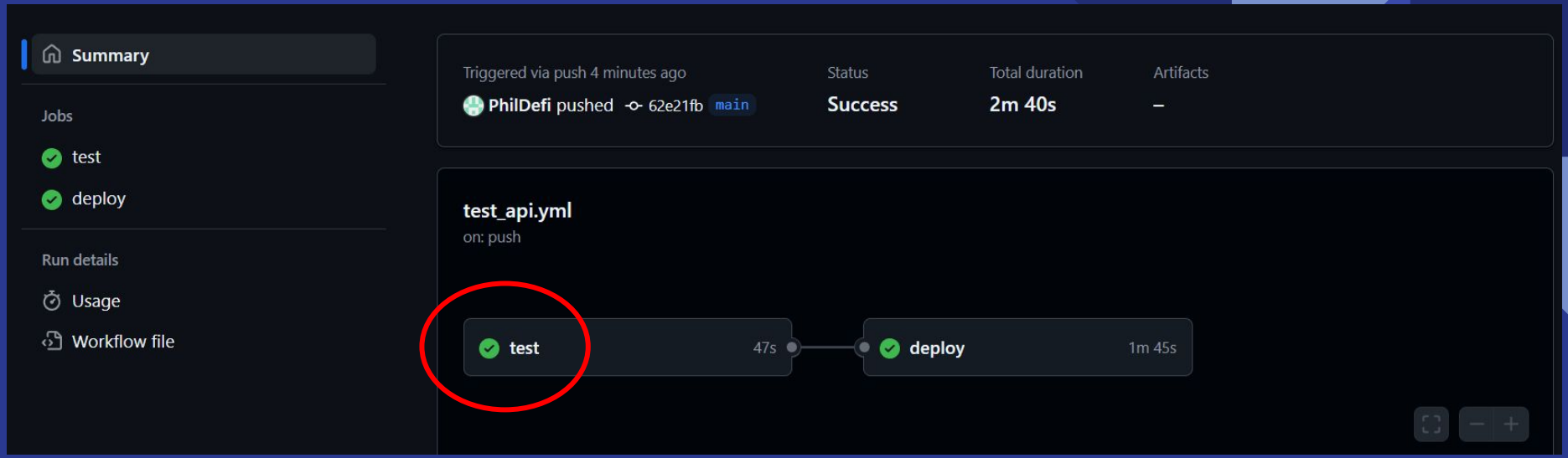
- FastAPI
- Tests de l'API pytest
- Déploiement sur Heroku
- Workflow : déploiement automatique sur Heroku depuis GitHub conditionné au succès des tests unitaires

GitHub Actions


HEROKU
from Salesforce
Jump to Favorites, Apps, Pipelines, Spaces...
Deep-dive on the Next Gen Platform. Join the Webinar!
Personal
app-heroku-credit-p7
Open app
More
GitHub PhilDefi/bank_credit_score
main
Overview
Resources
Deploy
Metrics
Activity
Access
Settings
Add this app to a pipeline
Create a new pipeline or choose an existing one and add this app to a stage in it.
Add this app to a stage in a pipeline to enable additional features
Pipelines let you connect multiple apps together and promote code between them.
Learn more.
Pipelines connected to GitHub can enable review apps, and create apps for new pull requests.
Learn more.
Choose a pipeline
Deployment method
Heroku Git
Use Heroku CLI
GitHub
Connected
Container Registry
Use Heroku CLI

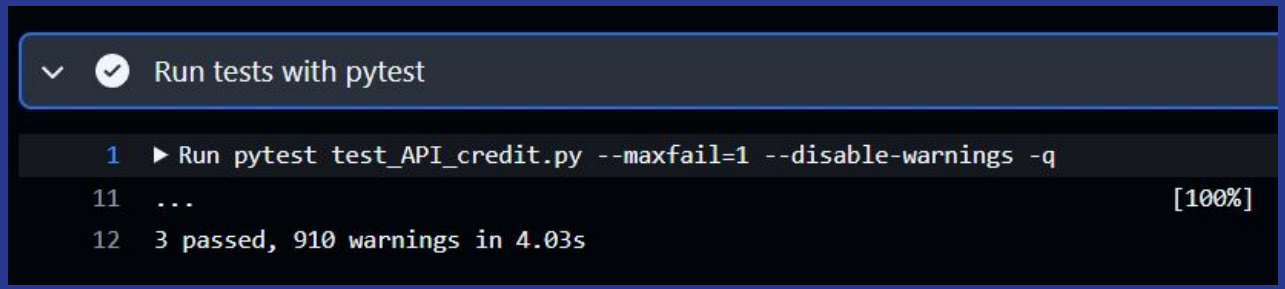

Deploy a GitHub branch
This will deploy the current state of the branch you specify below. Learn more.
Choose a branch to deploy
main
Deploy Branch
Receive code from GitHub
Build main 9420d718
Release phase
Deploy to Heroku
Your app was successfully deployed.
View

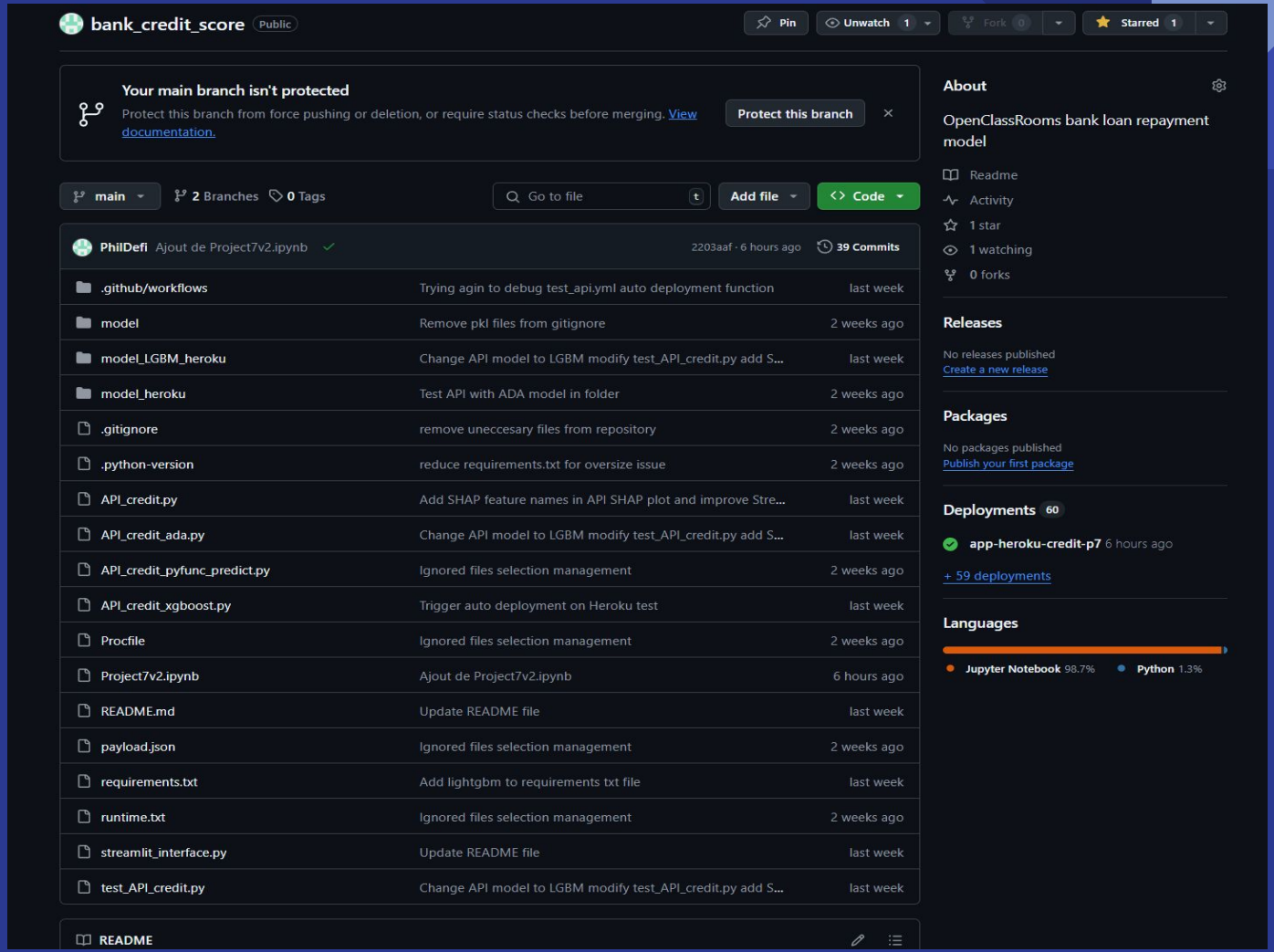

PhilDefi / bank_credit_score
Type / to search
Code
Issues
Pull requests 1
Actions
Projects
Wiki
Security 1
Insights
Settings
Actions
New workflow
All workflows
Test API avec pytest
Management
Caches
Deployments
Attestations
Runners
Usage metrics
Performance metrics
All workflows
Showing runs from all workflows
Filter workflow runs
Help us improve GitHub Actions
Tell us how to make GitHub Actions work better for you with three quick questions.
Give feedback
33 workflow runs
Event
Status
Branch
Actor
Ajout de Project7v2.ipynb
Test API avec pytest #19: Commit 2203aaf pushed by PhilDefi
main
5 hours ago
3m 0s
Update README file
Test API avec pytest #18: Commit 6119dd1 pushed by PhilDefi
main
last week
2m 59s
Add SHAP feature names in API SHAP plot and imp...
Test API avec pytest #17: Commit f75ce66 pushed by PhilDefi
main
last week
3m 3s
Modify API_credit.py to return SHAP display
Test API avec pytest #16: Commit 49a8463 pushed by PhilDefi
main
last week
3m 1s
Add a POST PAI predict_with_explanation incl SHAP...
Test API avec pytest #15: Commit dad969b pushed by PhilDefi
main
last week
3m 21s



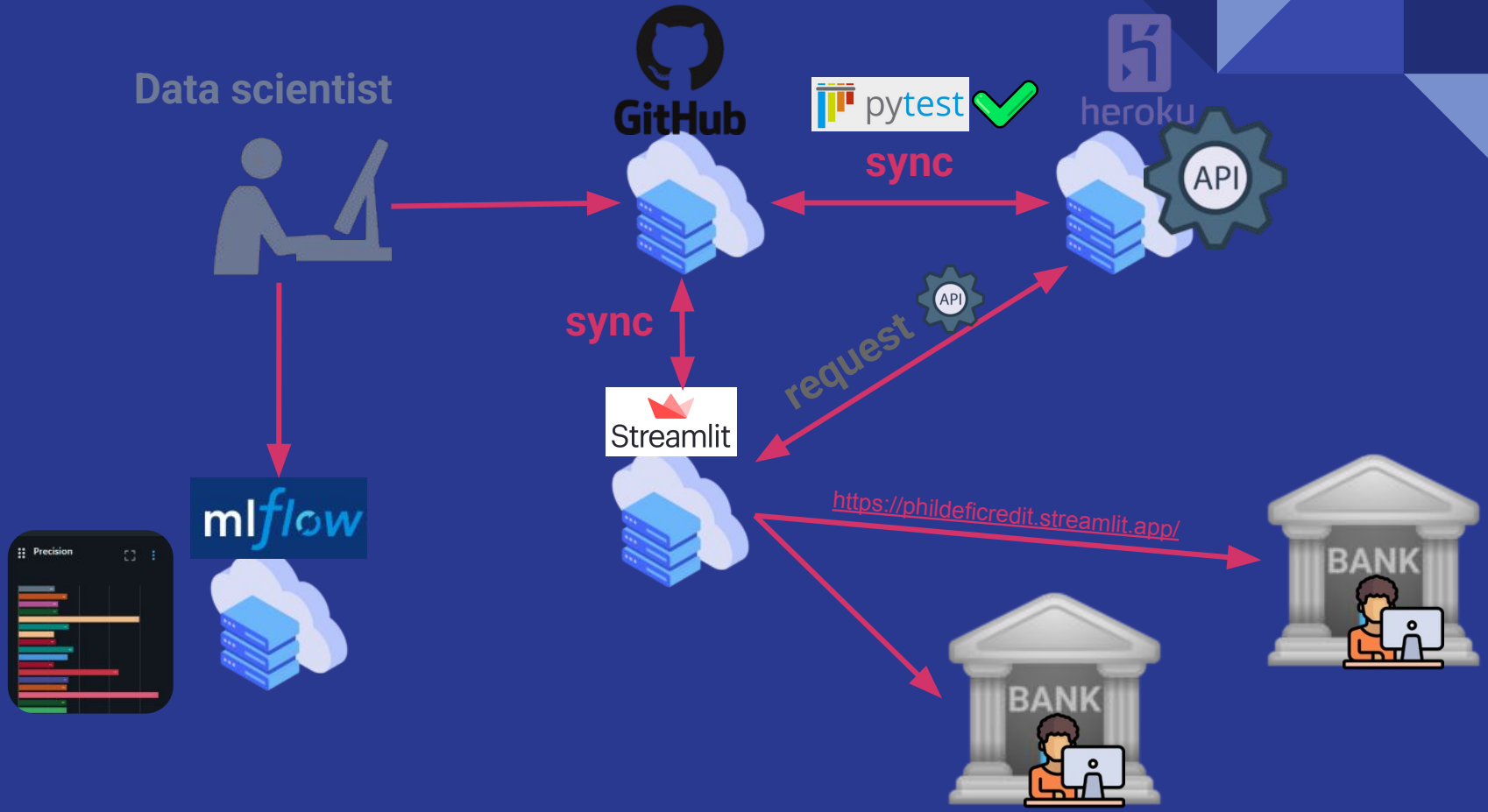
GitHub

fichier .yml




bank_credit_score
Public
Pin
Unwatch 1
Fork 0
Starred 1
Your main branch isn't protected
Protect this branch from force pushing or deletion, or require status checks before merging. View documentation.
Protect this branch
main
2 Branches
0 Tags
Go to file
t
Add file
Code
PhilDefi Ajout de Project7v2.ipynb
2203aaf · 6 hours ago
39 Commits
.github/workflows Trying agin to debug test_api.yml auto deployment function last week
model Remove pkl files from gitignore 2 weeks ago
model_LGBM_heroku Change API model to LGBM modify test_API_credit.py add S... last week
model_heroku Test API with ADA model in folder 2 weeks ago
.gitignore remove uneccesary files from repository 2 weeks ago
.python-version reduce requirements.txt for oversize issue 2 weeks ago
API_credit.py Add SHAP feature names in API SHAP plot and improve Stre... last week
API_credit_ada.py Change API model to LGBM modify test_API_credit.py add S... last week
API_credit_pyfunc_predict.py Ignored files selection management 2 weeks ago
API_credit_xgboost.py Trigger auto deployment on Heroku test last week
Procfile Ignored files selection management 2 weeks ago
Project7v2.ipynb Ajout de Project7v2.ipynb 6 hours ago
README.md Update README file last week
payload.json Ignored files selection management 2 weeks ago
requirements.txt Add lightgbm to requirements txt file last week
runtime.txt Ignored files selection management 2 weeks ago
streamlit_interface.py Update README file last week
test_API_credit.py Change API model to LGBM modify test_API_credit.py add S... last week
README
About
OpenClassRooms bank loan repayment model
Readme
Activity
1 star
1 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Deployments 60
app-heroku-credit-p7 6 hours ago
+ 59 deployments
Languages
Jupyter Notebook 98.7%
Python 1.3%

Architecture MLOps



User interface

Streamlit

https://phildeficredit.streamlit.app

Share

dashboard

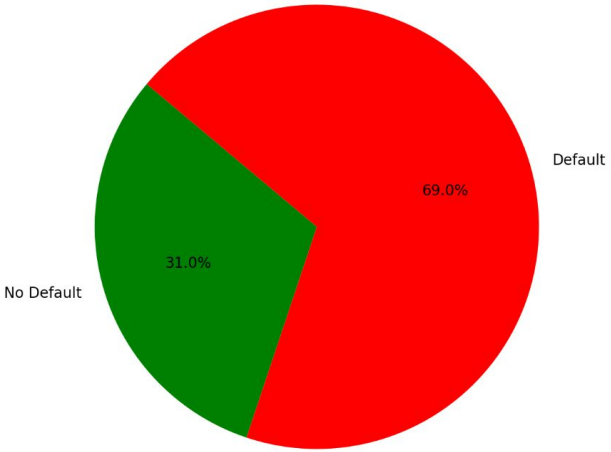
Predict the credit repayment probability using a ML model

Select the credit you want a prediction for :

44

Prediction Chart | Loan Details | Explanation

Loan Default Probability



Manage app

21:28

edit.streamlit.app

Fork

ML model

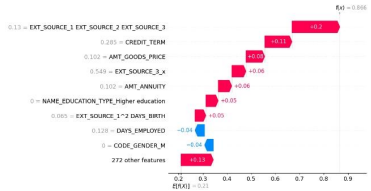
Select the credit you want a prediction for :

42

Prediction Chart | Loan Details | Explanation

Explanation

This section provides SHAP interpretability to explain the model's prediction for the submitted loan application.



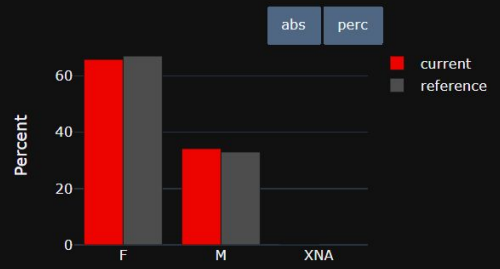
SHAP Waterfall Explanation

Data Drift

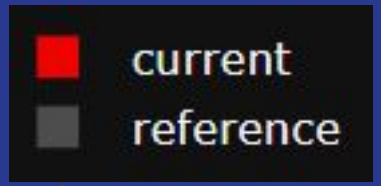
EVIDENTLY AI

CODE_GENDER
cat

	current	reference
count	307511.00	48744.00
missing	0.00 (0.00%)	0.00 (0.00%)
XNA	4.0 (0%)	0.0 (0%)
M	105059.0 (34%)	16066.0 (33%)
F	202448.0 (66%)	32678.0 (67%)



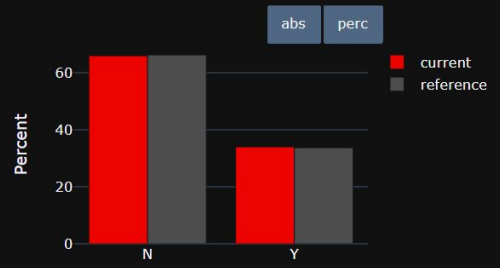
application_test.csv



application_train.csv

FLAG_OWN_CAR
cat

	current	reference
count	307511.00	48744.00
missing	0.00 (0.00%)	0.00 (0.00%)
Y	104587.0 (34%)	16433.0 (34%)
N	202924.0 (66%)	32311.0 (66%)



Conclusion

Elaboration d'un modèle en tenant compte du déséquilibre des classes

Définition d'une métrique métier sur-mesure

Mise en place d'une API déployée dans le Cloud

Mise en place d'une interface utilisateur déployée dans le Cloud

Démonstrateur :
https://phildeficredit.streamlit.app/

Repository GitHub :
https://github.com/PhilDefi/bank_credit_score