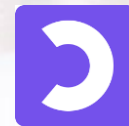


Implémentez un modèle de scoring



Rim BAHROUN

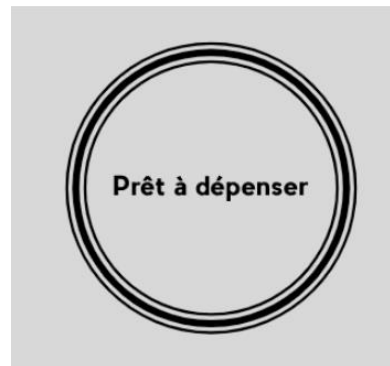
Parcours Data Scientist | projet 7

Juin 2023

Problématique & mission



Prêt à dépenser : société financière de crédit à la consommation



Mission

1. Construire un **modèle de scoring** pour la prédiction de la probabilité de remboursement d'un crédit
2. Construire un **Dashboard interactif** à destination des gestionnaires de relation client
3. **Mettre en production** le modèle à l'aide d'une **API**, ainsi que le **Dashboard**



Objectifs

Faciliter le travail des gestionnaires de relation client
Permettre plus de transparence vis-à-vis du client
Réduire les pertes financières relatives à une fausse prédiction



Plan de la présentation

01 Préparation des données

02 Modélisation

03 Pipeline de déploiement

04 Data drift

05 Dashboard

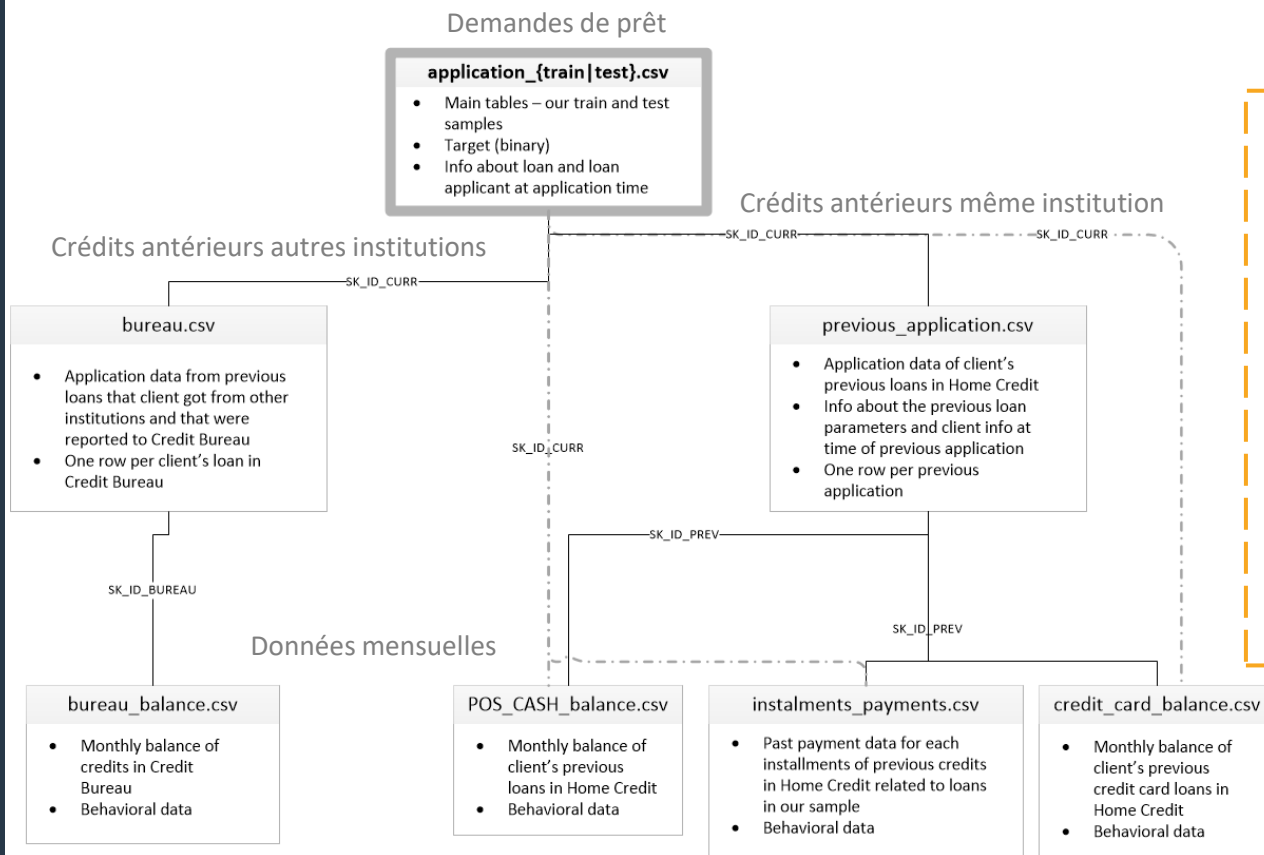


1. Préparation des données



Données à disposition : 8 fichiers .csv

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



- Jointure des tables selon les clés primaires
- Imputation des valeurs manquantes
- Correction des valeurs aberrantes
- Création de nouvelles variables métier
- Encodage des variables catégoriques
- Agrégation des données **par client**

kaggle

Jeu agrégé par client
797 variables
307 507 clients

Kernel Kaggle: <https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script>



1. Préparation des données



797 variables

Réduction de dimensionnalité



1. Suppression des variables à plus de 75% de valeurs manquantes
2. Suppression des variables corrélées à plus de 90%
3. Suppression des variables à faible importance ($< 5\%$ /arbre de décision)
4. Suppression des variables à moins de 2% de variance
5. Sélection des variables par SelectKbest
6. Sélection des variables par Recursive Feature Elimination



- Gain matériel: espace de stockage
- Gain en temps de calcul
- Simplification du modèle
- Meilleure interprétabilité des prédictions

Jeu de données
35 variables
307 507 clients



Plan de la présentation

01 Préparation des données

02 Modélisation

03 Pipeline de déploiement

04 Data drift

05 Dashboard



2. Modélisation

Démarche de modélisation



*Jeu de données
nettoyé*



35 variables
307 507 clients

Prétraitement des données



Encodage/Standardisation
**Traitement du déséquilibre
des classes**

y: vecteur target de 0/1

X: Données des clients

Modélisation

- Implémentation des modèles de **classification**
- Choix des métriques d'évaluation
- Optimisation des hyperparamètres

Evaluations des performances

- Comparaison des modèles
- Choix du modèle final pour la prédiction

**Prédiction de la probabilité
de remboursement**

**Prédiction de la classe du
client: client fiable 0/
client à risque 1**



2. Modélisation

Traitement du déséquilibre des classes

Problème de classification binaire:

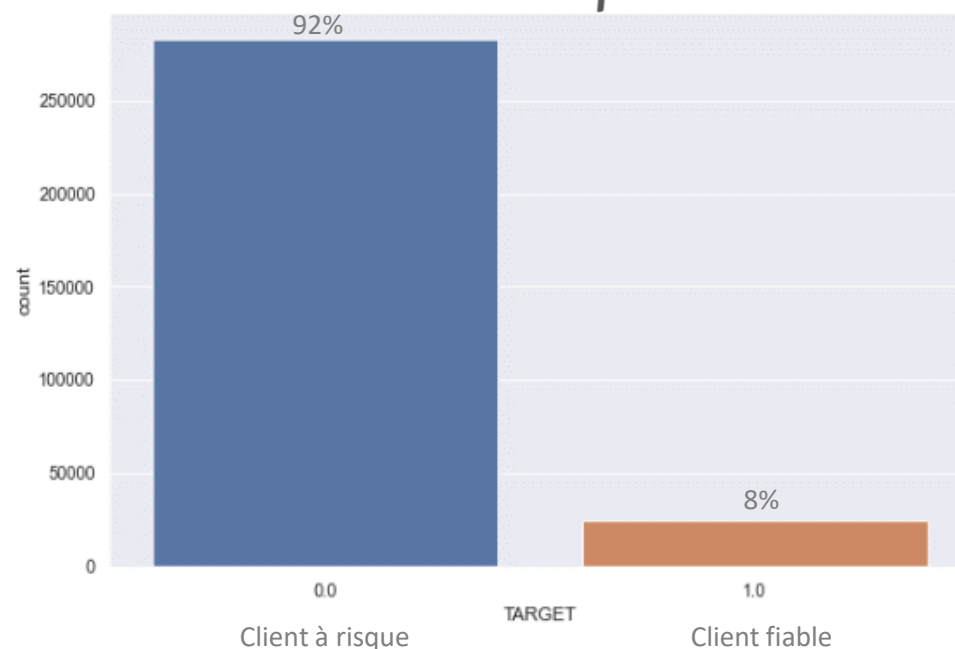
- Classe 0: client fiable : **92%** des clients
- Classe 1: client à risque : **8%** des clients



Déséquilibre entre les deux classes



Un **modèle naïf** prédisant systématiquement le client en fiable aurait une accuracy de **0.92 !!**



2. Modélisation

Traitement du déséquilibre des classes



Sous échantillonnage

Retirer aléatoirement des clients de la classe majoritaire

Perte d'information

Sur échantillonnage SMOTE

Créer des données synthétiques des clients de la classe minoritaire

Augmentation de la taille des données

Class weight

Créer un modèle pénalisé en attribuant des poids différents pour chaque classe

Méthode retenue pour la suite du projet



2. Modélisation

Choix des métriques d'évaluation



Métriques généraux

- **AUC-ROC:** indicateur de capacité du modèle a bien classé les clients
- **Accuracy:** pourcentage de client correctement classé
- **Recall:** pourcentage de client à risque identifié correctement par rapport au client réellement à risque
- **Precision:** pourcentage de client à risque identifié correctement par rapport au client prédit à risque



Hypothèse métier

Le coût d'un faux négatif FN est dix fois supérieur au coût d'un faux positif FP.

FN: client à risque prédit fiable : crédit accordé et perte en capital.

FP: client fiable prédit à risque : crédit refusé et manque à gagner en marge.

		Classe prédite	
		Client fiable	Client à risque
Classe Réelle	Client fiable	Vrais négatifs / TN	Faux positifs / FP
	Client à risque	Faux négatif / FN	Vrais positifs / TP



2. Modélisation

Choix des métriques d'évaluation

Le coût d'un faux négatif FN est dix fois supérieur au coût d'un faux positif FP.

FN: client à risque prédit fiable : crédit accordé et perte en capital.

FP: client fiable prédit à risque : crédit refusé et manque à gagner en marge.



- **F_beta:** combine à la fois la précision et le rappel en utilisant un poids relatif beta $\beta^2=10$ et donc $\beta = 3.16$.

$$f_{\beta} = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + (\beta^2 FN + FP)}$$

- **Score métier :** compris entre 0 et 1. Plus il est grand, meilleur est le modèle.



$$\text{Costum_score} = 1 - \frac{1*FP + 10*FN}{1*N + 10*P}$$

Coût de la prédiction

Coût maximal: tous les négatifs ont été prédits en positifs

		Classe prédite	
		Client fiable	Client à risque
Classe Réelle	Client fiable	Vrais négatifs / TN	Faux positifs / FP
	Client à risque	Faux négatif / FN	Vrais positifs / TP



Métriques
spécifiques

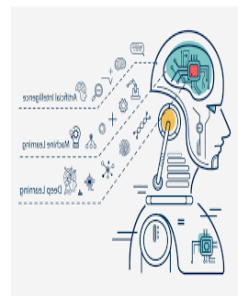


2. Modélisation

Algorithmes de classification

Algorithmes d'apprentissage supervisé à tester

- DummyClassifier
- LogisticRegression
- RandomForestClassifier
- ExtraTreesClassifier
- LGBMClassifier



Optimisation des hyperparamètres

GridSearchCV - **score métier** - validation croisée sur 10 plis.

Optimisation du seuil de prédiction

pour la méthode **predict_proba** pour le modèle sélectionné.

Metrics														
Parameters														
Run Name	Created	Duration	AUROC	Accuracy	Precision	Recall	Score_métier	class_weight	criterion	learning_rate	max_depth	max		
Extra Tree	1 month ago	1.1h	0.73	0.68	0.15	0.65	0.67	balanced	gini	-	7	s		
Random Forest	1 month ago	1.5h	0.73	0.72	0.16	0.61	0.67	balanced	entropy	-	9	lc		
LGBM	1 month ago	7.6min	0.74	0.7	0.16	0.65	0.68	balanced	-	0.02	8	-		
LogisticReg	1 month ago	2.1min	0.73	0.68	0.15	0.65	0.67	balanced	-	-	-	-		
Dummy	1 month ago	30.5s	0.5	0.92	0	0	0.53	-	-	-	-	-		

mlflow ui --backend-store-uri sqlite:///mlflow.db



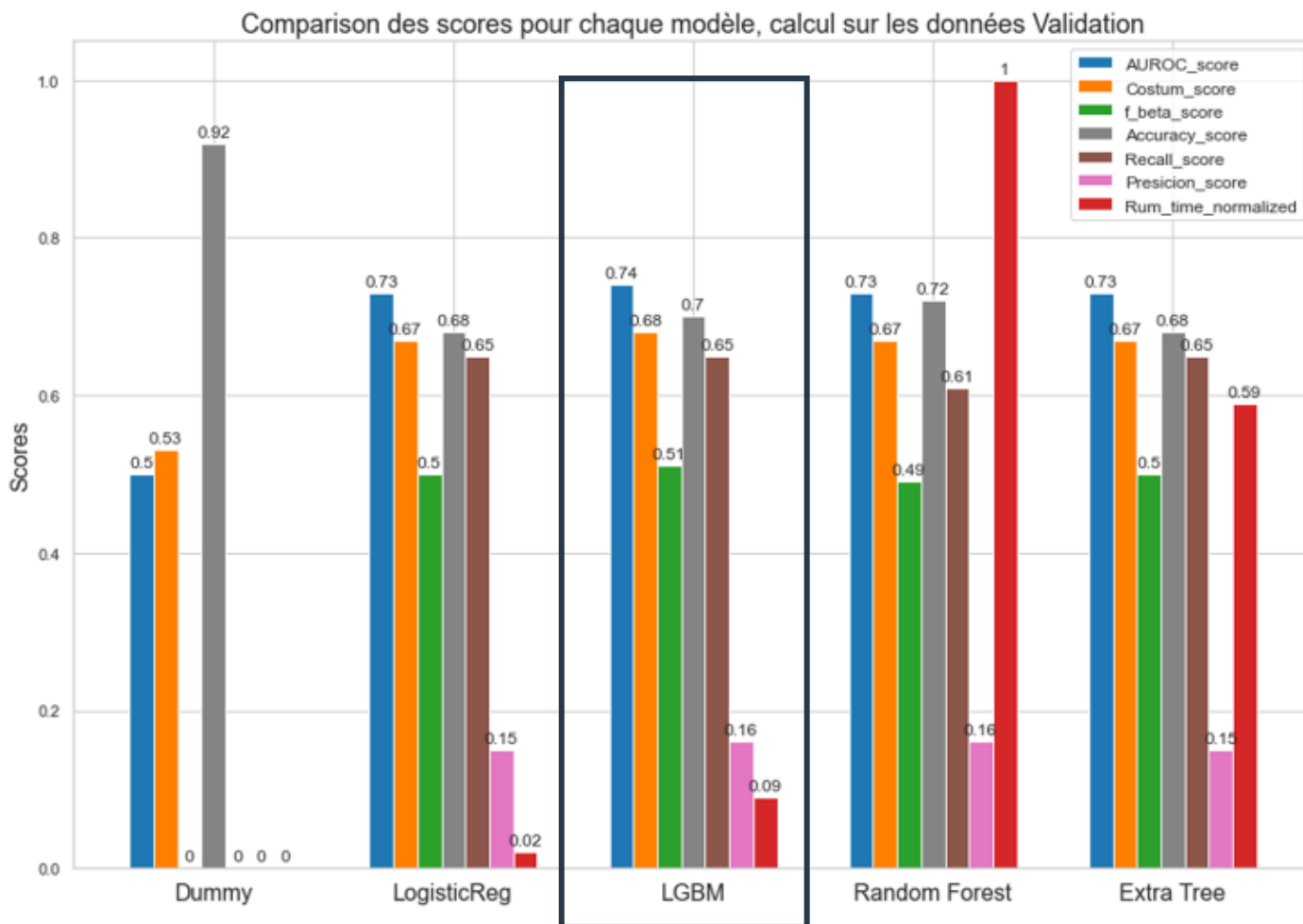
2. Modélisation

Algorithmes de classification

Synthèse des résultats



 LightGBM

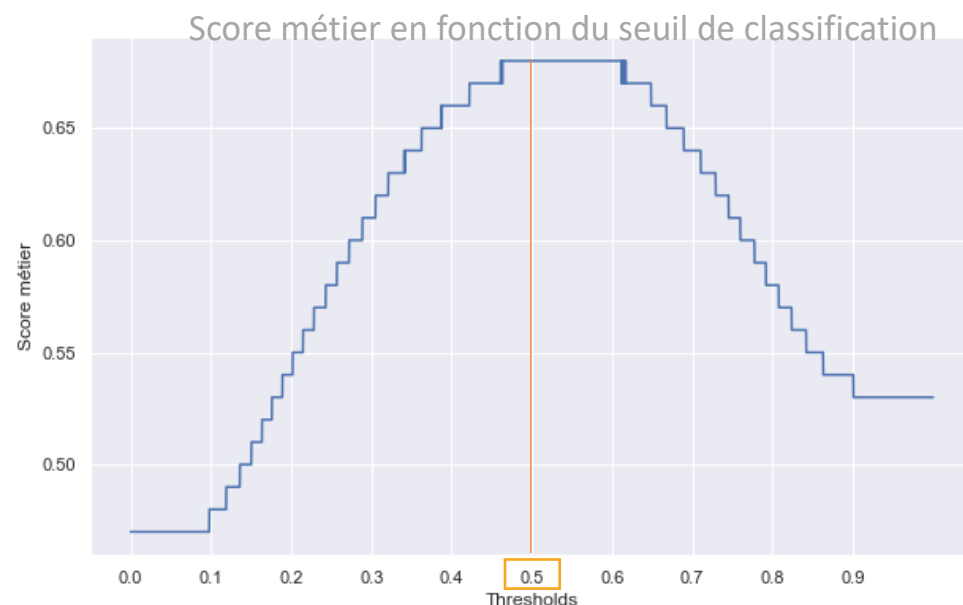


2. Modélisation

Algorithmes de classification

Synthèse des résultats

Seuil de classification



 **LightGBM**

	Train set	Test set
AUROC	0.77	0.74
Score_métier	0.70	0.68
f_beta	0.54	0.51
Accuracy	0.69	0.69
Recall	0.70	0.66
Presicion	0.17	0.16

mlflow
Model Registry

mlflow 2.3.1 Experiments Models [GitHub](#) [Docs](#)

Registered Models

[Share and manage machine learning models. Learn more](#)

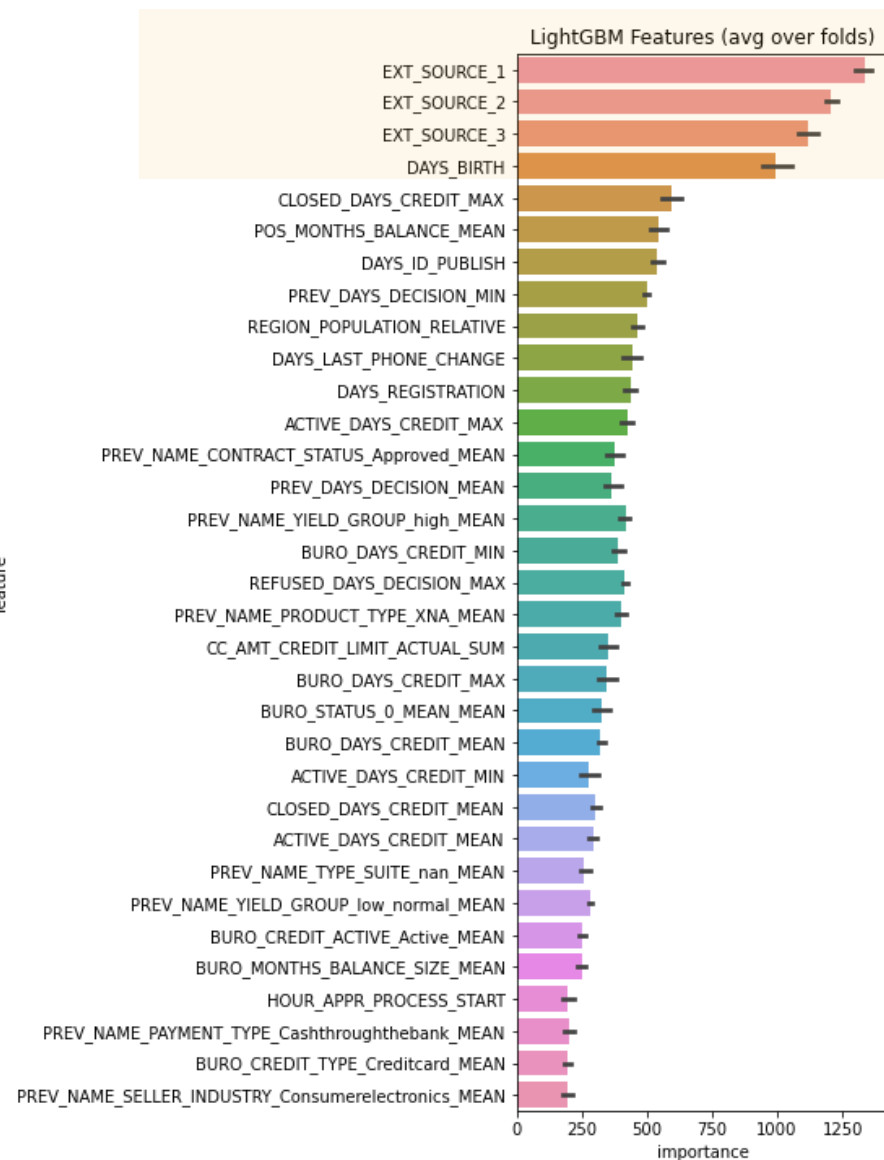
[Create Model](#)

[Search](#) [Clear](#)

Name	Latest Version	Staging	Production	Last Modified	Tags
LGBM_final	Version 1	-	Version 1	2023-06-05 13:03:37	-

2. Modélisation

Interprétabilité globale du modèle



2. Modélisation

Interprétabilité locale du modèle

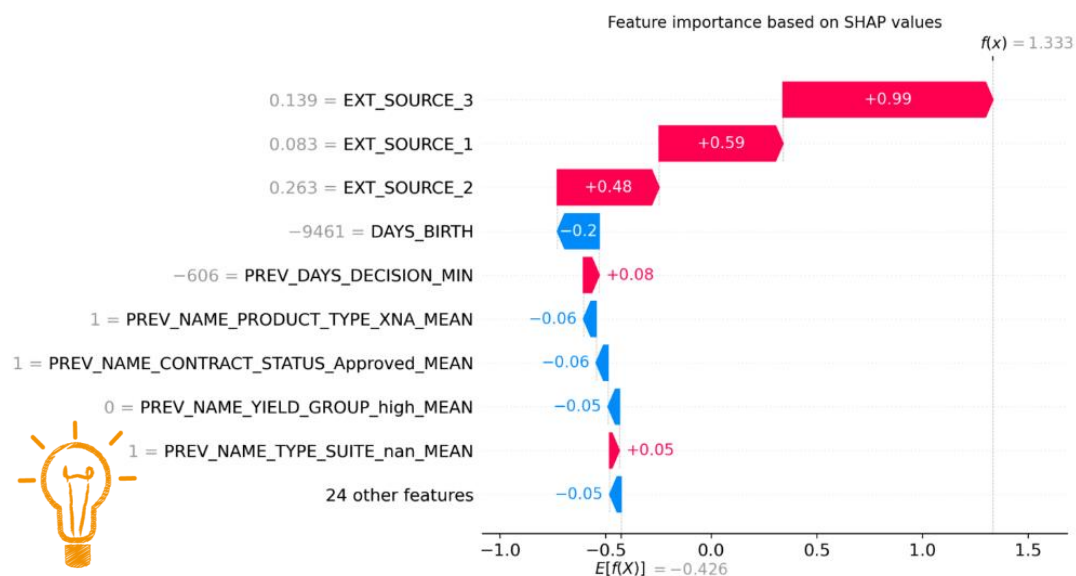
Exemples de clients

Client à risque

Probabilité de remboursement **21%**

Crédit refusé

Client: 1000024



Client fiable

Probabilité de remboursement **79%**

Crédit accordé

Client: 100004



Plan de la présentation

01 Préparation des données

02 Modélisation

03 Pipeline de déploiement

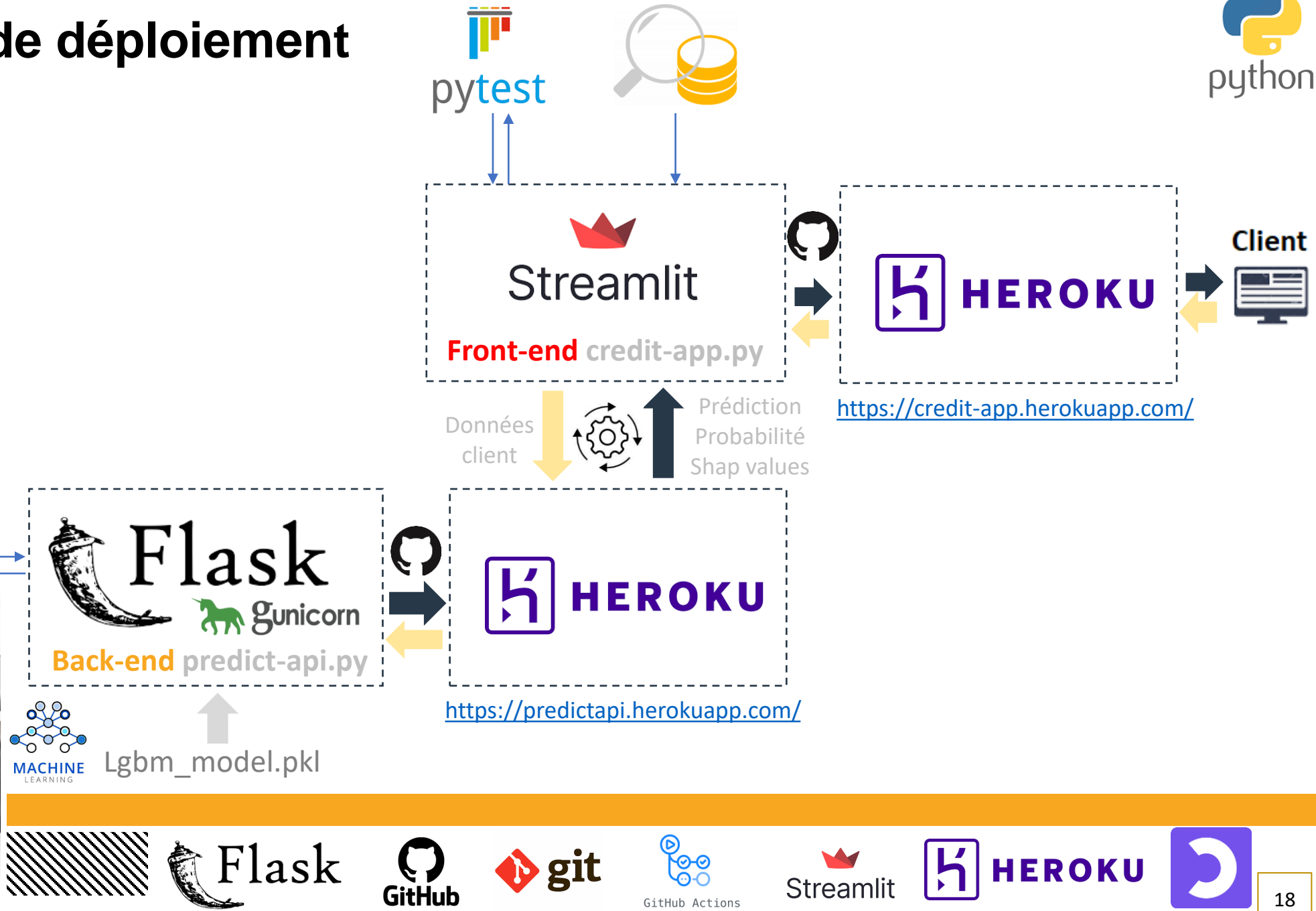
04 Data drift

05 Dashboard



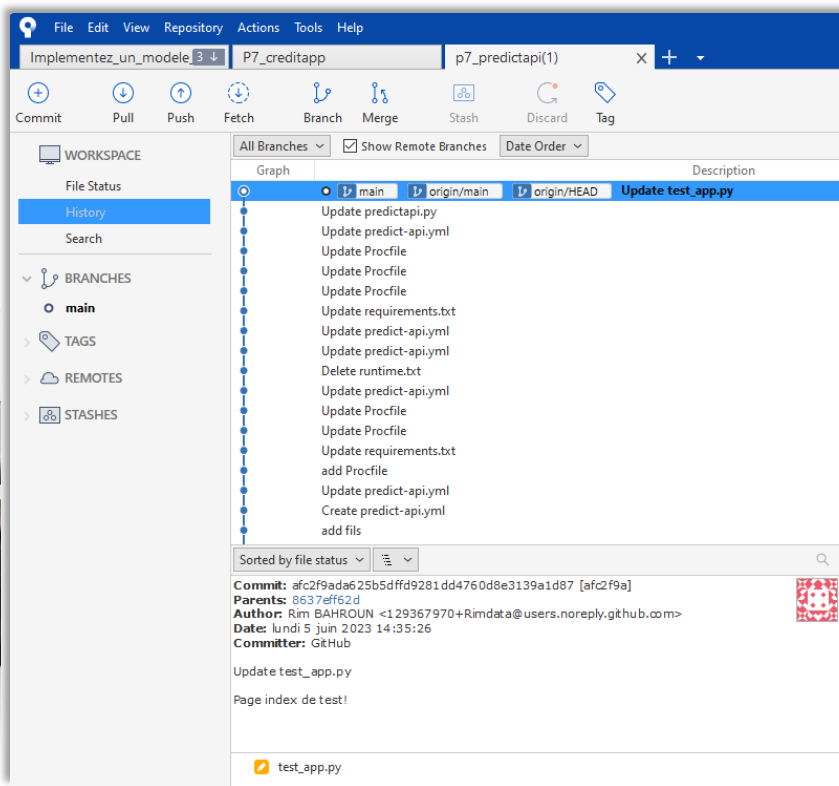
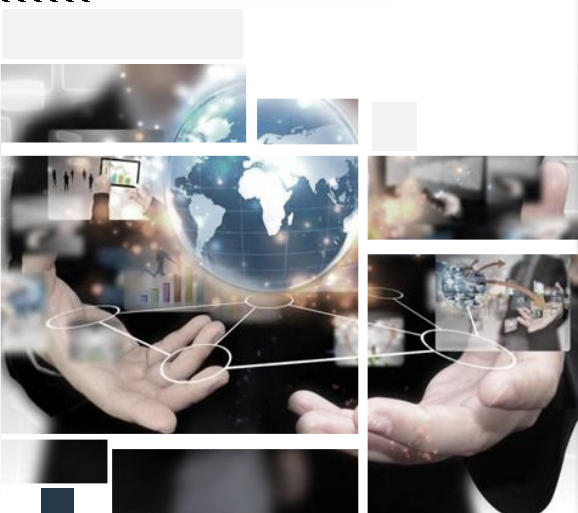
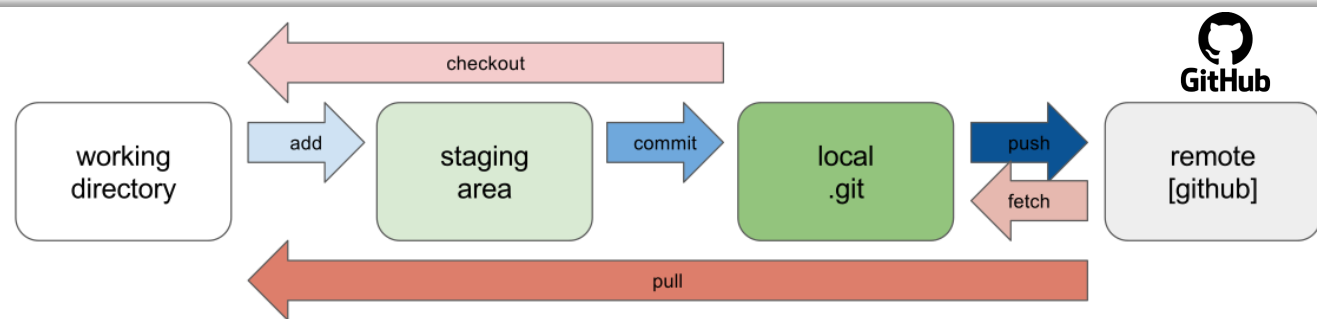


3. Pipeline de déploiement



3. Pipeline de déploiement

Logiciel de version de code



```
MINGW64:/d/Documents/00penclasse/room/P7_DS_03_04_2023/p7_predictapi

rimla@DESKTOP-PFF0ATG MINGW64 /d/Documents/00penclasse/room/P7_DS_03_04_2023/p7_p
redictapi (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
        modified:   test_app.py

Untracked files:
  (use "git add <file>..." to include in what will be committed)
        __pycache__/
        df_credit_dash_score.csv

no changes added to commit (use "git add" and/or "git commit -a")

rimla@DESKTOP-PFF0ATG MINGW64 /d/Documents/00penclasse/room/P7_DS_03_04_2023/p7_p
redictapi (main)
$ git add test_app.py
warning: in the working copy of 'test_app.py', LF will be replaced by CRLF the next ti
me Git touches it

rimla@DESKTOP-PFF0ATG MINGW64 /d/Documents/00penclasse/room/P7_DS_03_04_2023/p7_predict
api (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.

Changes to be committed:
  (use "git restore --staged <file>..." to unstage)
        modified:   test_app.py

Untracked files:
  (use "git add <file>..." to include in what will be committed)
        __pycache__/
        df_credit_dash_score.csv

rimla@DESKTOP-PFF0ATG MINGW64 /d/Documents/00penclasse/room/P7_DS_03_04_2023/p7_predictapi (main)
$ git commit -m "fonction test prediction ajoutée"
[main 9ef4a67] fonction test prediction ajoutée
1 file changed, 14 insertions(+), 14 deletions(-)
```

26 mai 2023 15:42	Rimdata <rimbahr	Za3d7c3
26 mai 2023 15:28	Rim BAHROUN <1	93f4e0e
26 mai 2023 15:18	Rim BAHROUN <1	b8c99e0
26 mai 2023 15:15	Rimdata <rimbahr	a0a62bf

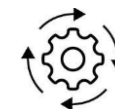


3. Pipeline de déploiement

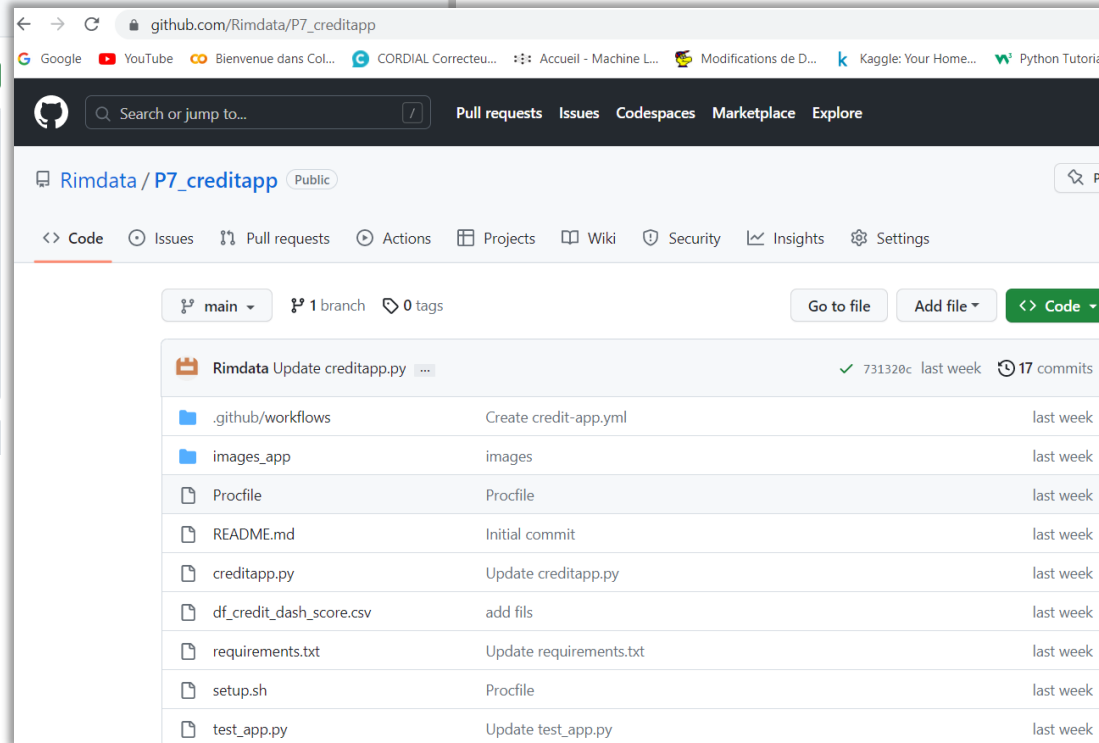
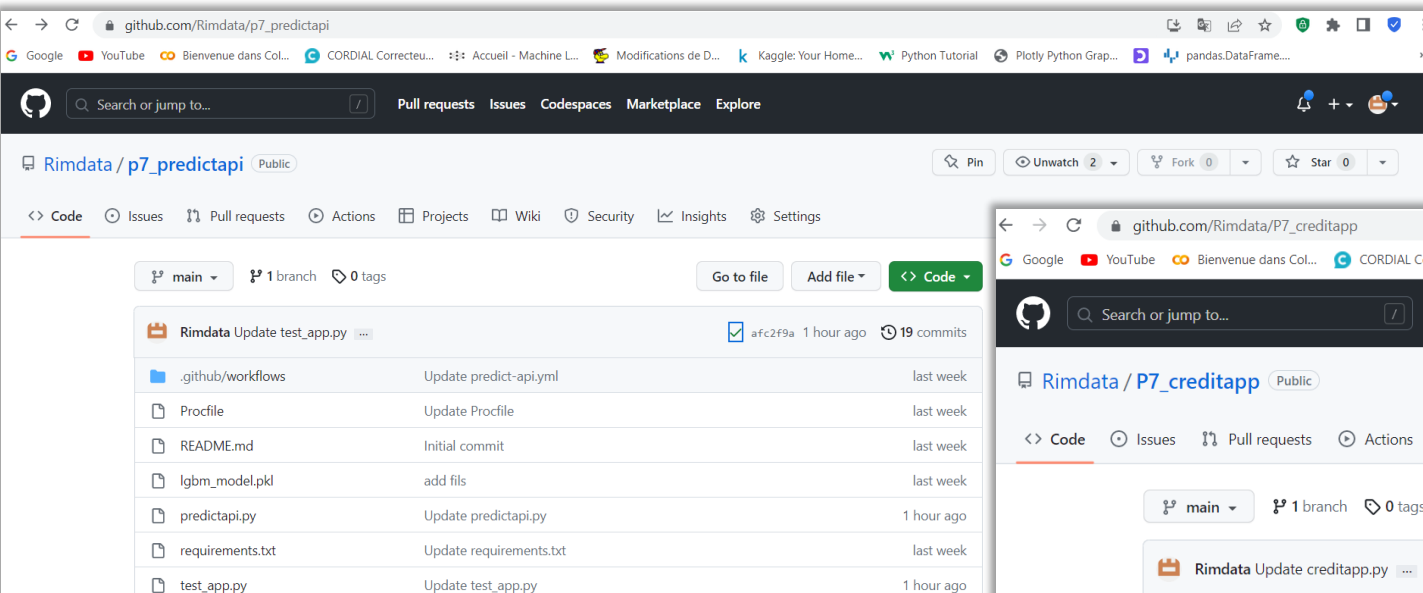
Stockage et partage sur le cloud



Déploiement continu et automatisé



Front-end `credit-app.py`
Back-end `predict-api.py`



build

succeeded last week in 3m 59s

- > Set up job
- > Run actions/checkout@v3
- > Run git fetch --prune --unshallow
- > Set up Python 3.10
- > Install dependencies
- > Lint with flake8
- > Test with pytest
- > Deploy to Heroku
- > Post Set up Python 3.10
- > Post Run actions/checkout@v3
- > Complete job



https://github.com/Rimdata/p7_predictapi/tree/main

https://github.com/Rimdata/P7_creditapp



3. Pipeline de déploiement

Tests unitaires



```
Invite de commandes
-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
===== short test summary info =====
FAILED test_app.py::test_predict - requests.exceptions.JSONDecodeError: Expecting value: line 1 column 1 (char 0)
===== 1 failed, 1 passed, 24 warnings in 5.88s =====
```

```
(env_python) D:\Documents\00penclasse room\P7_DS_03_04_2023\p7_predictapi>pytest
===== test session starts =====
platform win32 -- Python 3.9.12, pytest-7.3.1, pluggy-1.0.0
rootdir: D:\Documents\00penclasse room\P7_DS_03_04_2023\p7_predictapi
collected 2 items

test_app.py ..

===== warnings summary =====
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\utils\_clustering.py:35: 1 warni
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\utils\_clustering.py:54: 1 warni
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\utils\_clustering.py:63: 1 warni
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\utils\_clustering.py:69: 1 warni
```

TEST FAILED



```
Invite de commandes
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\maskers\_tabular.py:186: 1 warning
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\maskers\_tabular.py:197: 1 warning
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\maskers\_image.py:175: 1 warning
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\explainers\_partition.py:676: 1 warning
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\explainers\_exact.py:179: 1 warning
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\explainers\_exact.py:205: 1 warning
The 'nopython' keyword argument was not supplied to the 'numba.jit' decorator. The implicit default value for this arg
ument is currently False, but it will be changed to True in Numba 0.59.0. See https://numba.readthedocs.io/en/stable/ref
erence/deprecation.html#deprecation-of-object-mode-fall-back-behaviour-when-using-jit for details.
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\plots\_colors\_colorconv.py:617
'np.bool8' is a deprecated alias for 'np.bool_'. (Deprecated NumPy 1.24)
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\shap\plots\_image.py:20
IPython could not be loaded!
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\sklearn\base.py:318
Trying to unpickle estimator StandardScaler from version 1.0.2 when using version 1.2.2. This might lead to breaking c
ode or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\sklearn\base.py:318
Trying to unpickle estimator LabelEncoder from version 1.0.2 when using version 1.2.2. This might lead to breaking cod
e or invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
..\Implementez_un_modele_de_scoring\env_python\lib\site-packages\sklearn\base.py:318
Trying to unpickle estimator Pipeline from version 1.0.2 when using version 1.2.2. This might lead to breaking code or
invalid results. Use at your own risk. For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
===== 2 passed, 24 warnings in 5.98s =====
(env_python) D:\Documents\00penclasse room\P7_DS_03_04_2023\p7_predictapi>
```

PASSED



Plan de la présentation

01 Préparation des données

02 Modélisation

03 Pipeline de déploiement

04 Data drift

05 Dashboard



Data drift

En comparant les distributions des 33 caractéristiques dans l'ensemble d'entraînement et l'ensemble de test, un drift a été détecté sur 6 caractéristiques, ce qui représente 18% du total. Avec un seuil de 50%, le Data Drift n'est donc pas détecté sur notre ensemble de test.

Dataset Drift

Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5

33

Columns








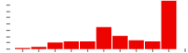
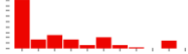
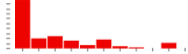


6

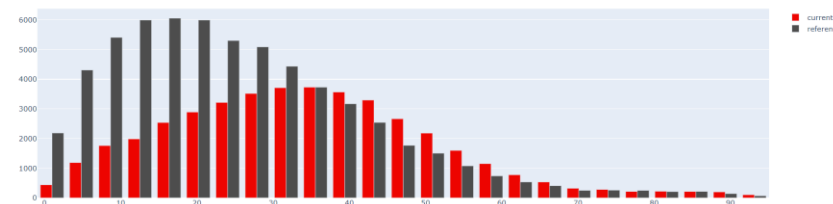
Drifted Columns

0.182

Share of Drifted Columns

Data Drift Summary

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> BURO_MONTHS_BALANCE_SIZE_MEAN	num			Detected	Wasserstein distance (normed)	0.524881
> BURO_STATUS_0_MEAN_MEAN	num			Detected	Wasserstein distance (normed)	0.252943
> DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.139765
> PREV_NAME_CONTRACT_STATUS_Approved_MEAN	num			Detected	Wasserstein distance (normed)	0.118842
> PREV_NAME_YIELD_GROUP_high_MEAN	num			Detected	Wasserstein distance (normed)	0.111511
> PREV_NAME_PAYMENT_TYPE_Cashtroughthebank_MEAN	num			Detected	Wasserstein distance (normed)	0.105636



Les 6 variables détectées ne sont pas les plus importantes pour le modèle sélectionné.

Plan de la présentation

01 Préparation des données

02 Modélisation

03 Pipeline de déploiement

04 Data drift

05 Dashboard



Insérer une page web

Cette application vous permet d'insérer des pages web sécurisées commençant par `https://` dans l'ensemble de diapositives. Pour des raisons de sécurité, les pages web non sécurisées ne sont pas prises en charge.

Veuillez entrer l'URL ci-dessous.

`https://`

Remarque : de nombreux sites web populaires autorisent l'accès sécurisé. Veuillez cliquer sur le bouton d'aperçu pour vérifier si la page web est accessible.

Visionneuse web [Conditions](#) | [Confidentialité et cookies](#)

Aperçu

 HEROKU



Plan de la présentation

01 Préparation des données

02 Modélisation

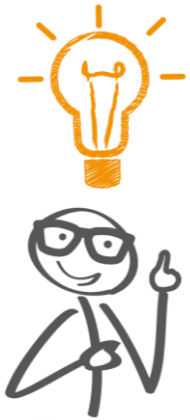
03 Pipeline de déploiement

04 Data drift

05 Dashboard



Conclusion



- Définir un pipeline d'entraînement des modèles adaptés au métier: **MLFlow**
- Utiliser un logiciel de version de code : **Git/Github**
- Evaluer les performances des modèles d'apprentissage supervisé
- Définir une stratégie de suivi : analyse de **Data drift: evidently**
- Déployer un modèle via une **API** dans le web: **Flask/ Heroku**
- Réaliser des tests unitaires automatisés : **Pytest**
- Réaliser et déployer un **Dashboard** : **Streamlit/ Heroku**



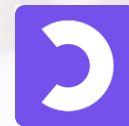
Merci pour votre attention



Rim BAHROUN



Implémentez un modèle de scoring



Rim BAHROUN

Parcours Data Scientist | projet 7
Juin 2023