

P7 - Implémentez un modèle de scoring

Deployment of a scoring API on internet

Summary

- Preliminary
- Exploratory Analysis
- Modelization
- Recording experiments and models with MLFlow
- API with FASTAPI and dashboard with streamlit
- Testing with Pytest
- Sharing code and versioning with Github
- Deployment of the API on internet with Heroku
- Drift analysis
- Conclusion

Preliminary

- **Mission** : deploy on internet a **classification algorithm** to predict a **solvability probability** for potential customers of « **Prêt à dépenser** » applying for a loan.
- **Requirements** : ensure the **transparency** of the classification for the customers on an **interactive dashboard**.

Exploratory Data Analysis

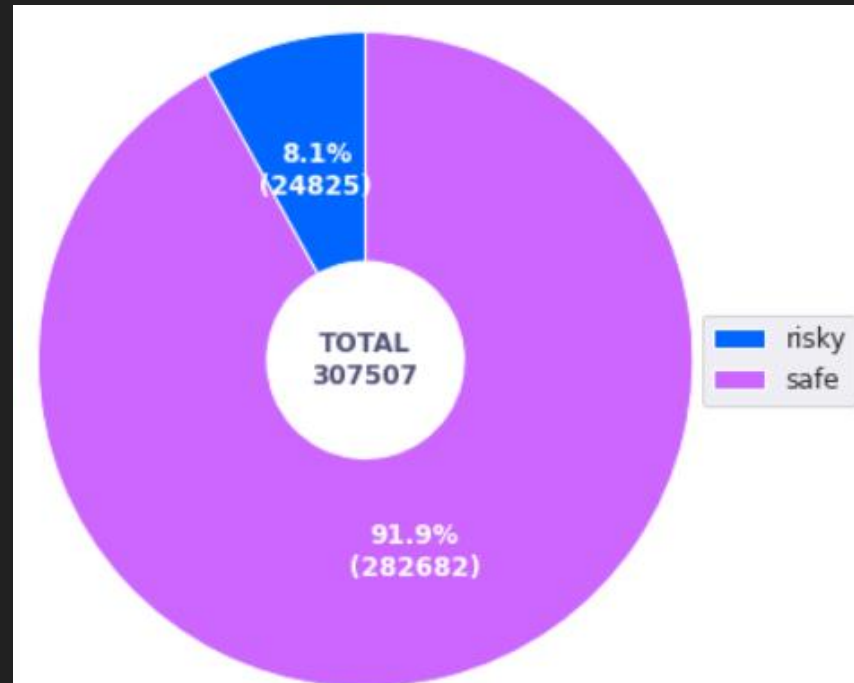
- Data are dispatched among ten files :
 - HomeCredit_columns_description.csv
 - POS_CASH_balance.csv
 - application_test.csv
 - application_train.csv
 - bureau.csv
 - bureau_balance.csv
 - credit_card_balance.csv
 - installments_payments.csv
 - previous_application.csv
 - sample_submission.csv

Exploratory Data Analysis

- As suggested, we use the cleaning and merging job done on :
 - <https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script>
- The final resulting dataset :
 - 506 features
 - 356251 rows :
 - 307507 for modelization
 - 48744 for new customers

Exploratory Data Analysis

- We check how balanced the data for modelization are:



Exploratory Data Analysis

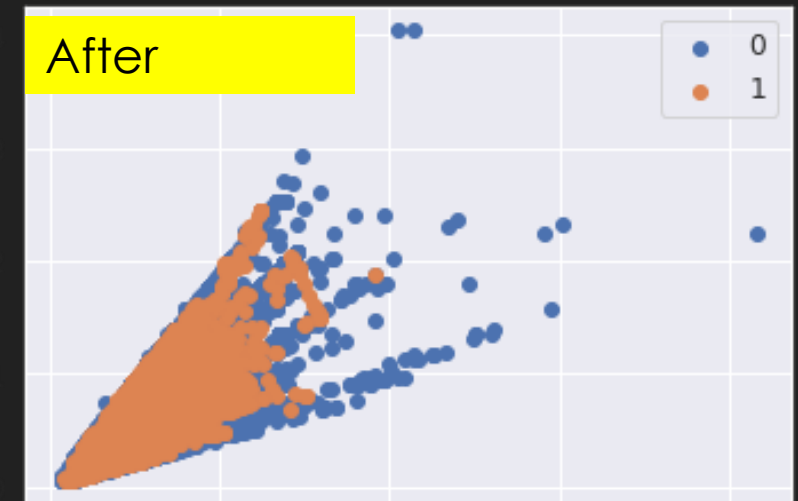
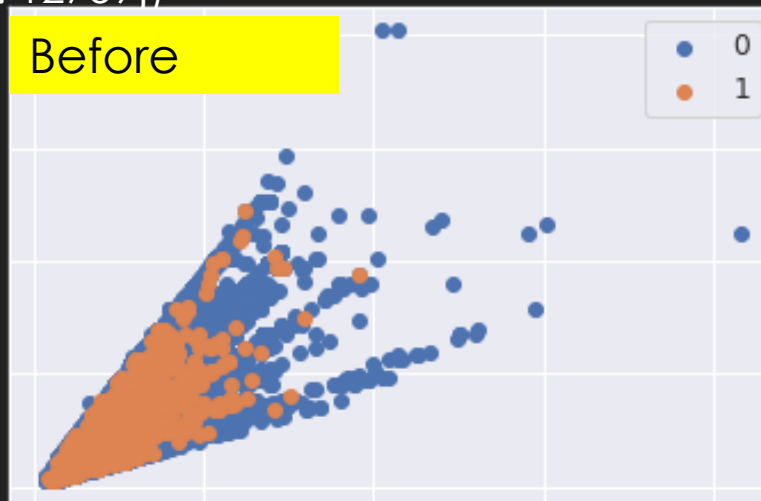
- Further, we remove data_leaks
 - features EXT_SOURCE : scores highly related with TARGET
- And we remove features contributing to bivariate correlations above 0.90.
 - 506 features => 434 features



Extract of the correlation matrix

Modelization

- Balancing data with BorderlineSMOTE
 - *BorderlineSMOTE is a data augmentation technique used in machine learning to balance imbalanced datasets by generating synthetic minority samples .*
 - Before : Counter({0: 12759, 1: 1116})*
 - After : Counter({0: 12759, 1: 12759})*



* "0" stands for customer with no solvability risk and "1" stands for customer with a solvability risk

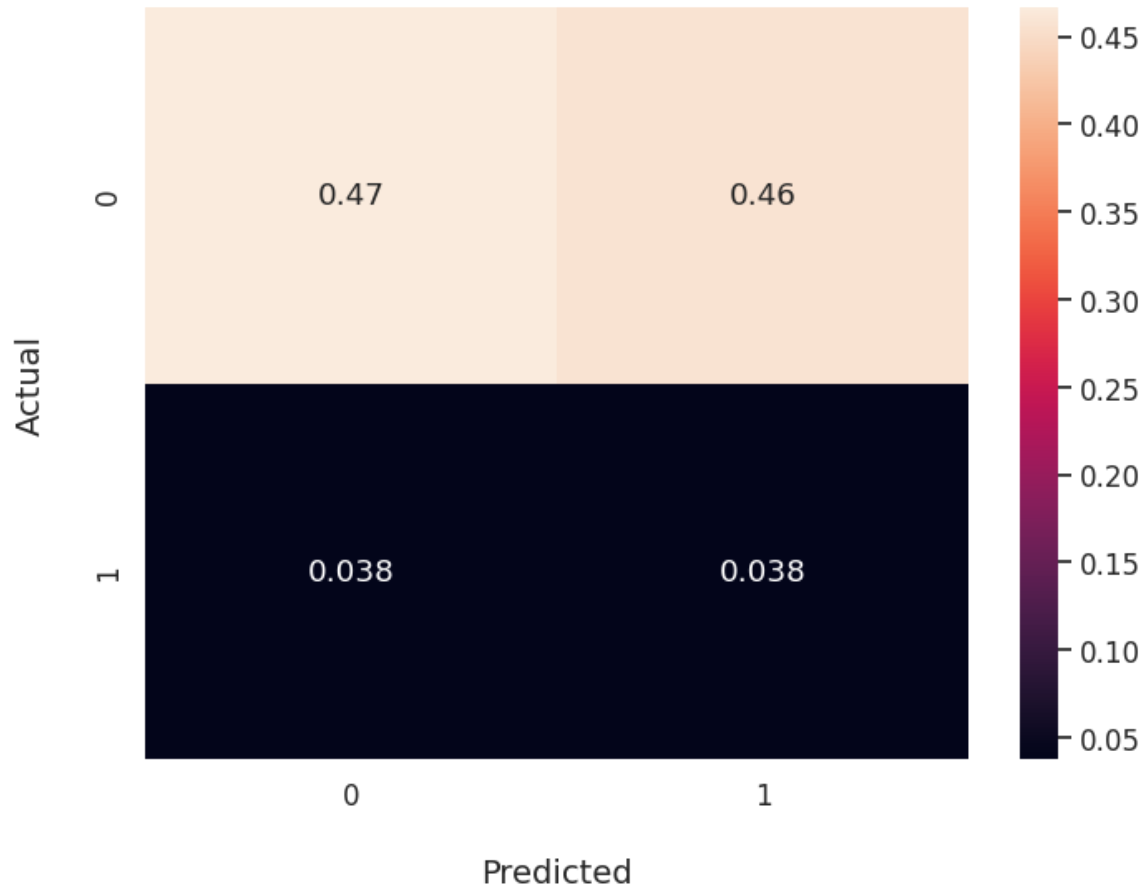
Modelization

- **RandomizedSearch** is a hyperparameter tuning technique in machine learning that randomly samples from a defined search space to find the optimal combination of hyperparameters for a given model
- We cross validate four models with RandomizedSearch :
 - **DummyClassifier**
 - **LogisticRegression**
 - **RandomForestClassifier**
 - **LGBMClassifier**
- We select the best model according to its minimal custom metric on the test sample:
 - **Custom Metric = 10FN + FP**

Modelization

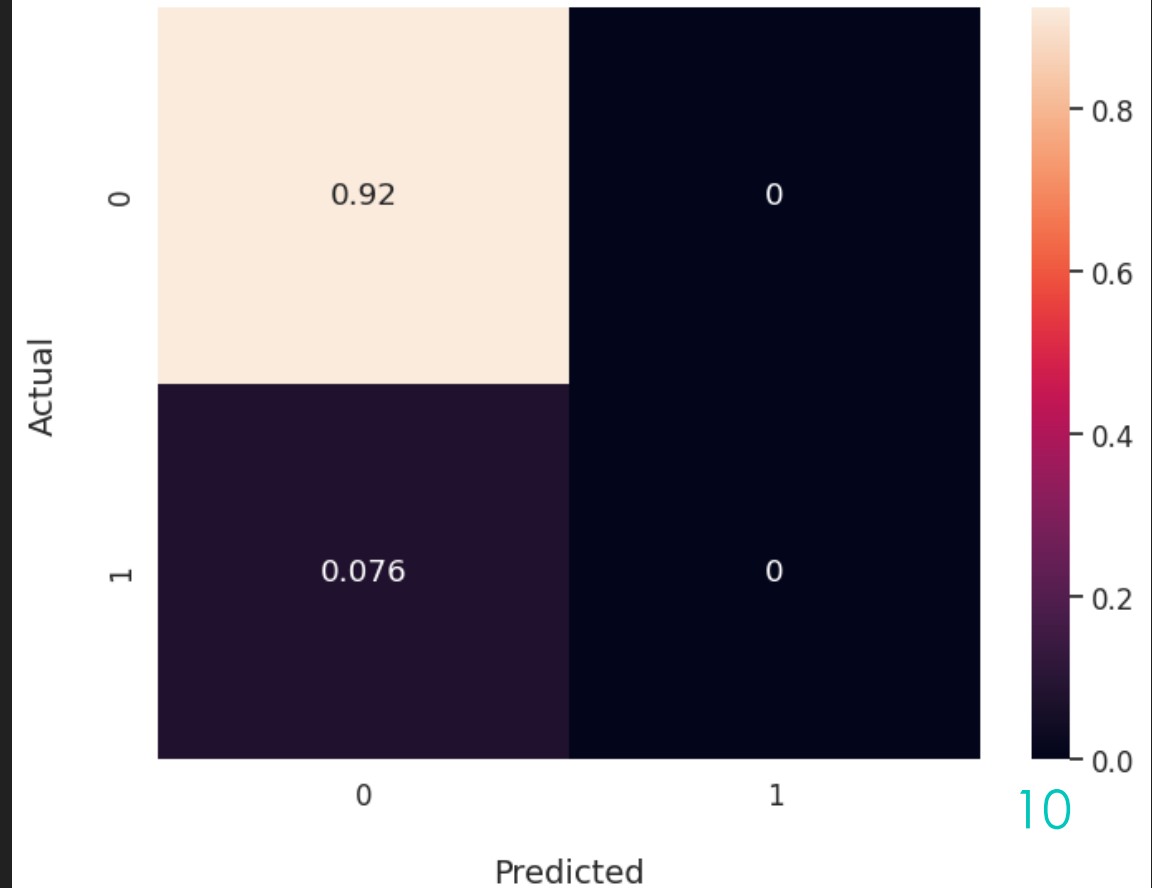
DummyClassifier

Confusion Matrix for credit default



LogisticRegression

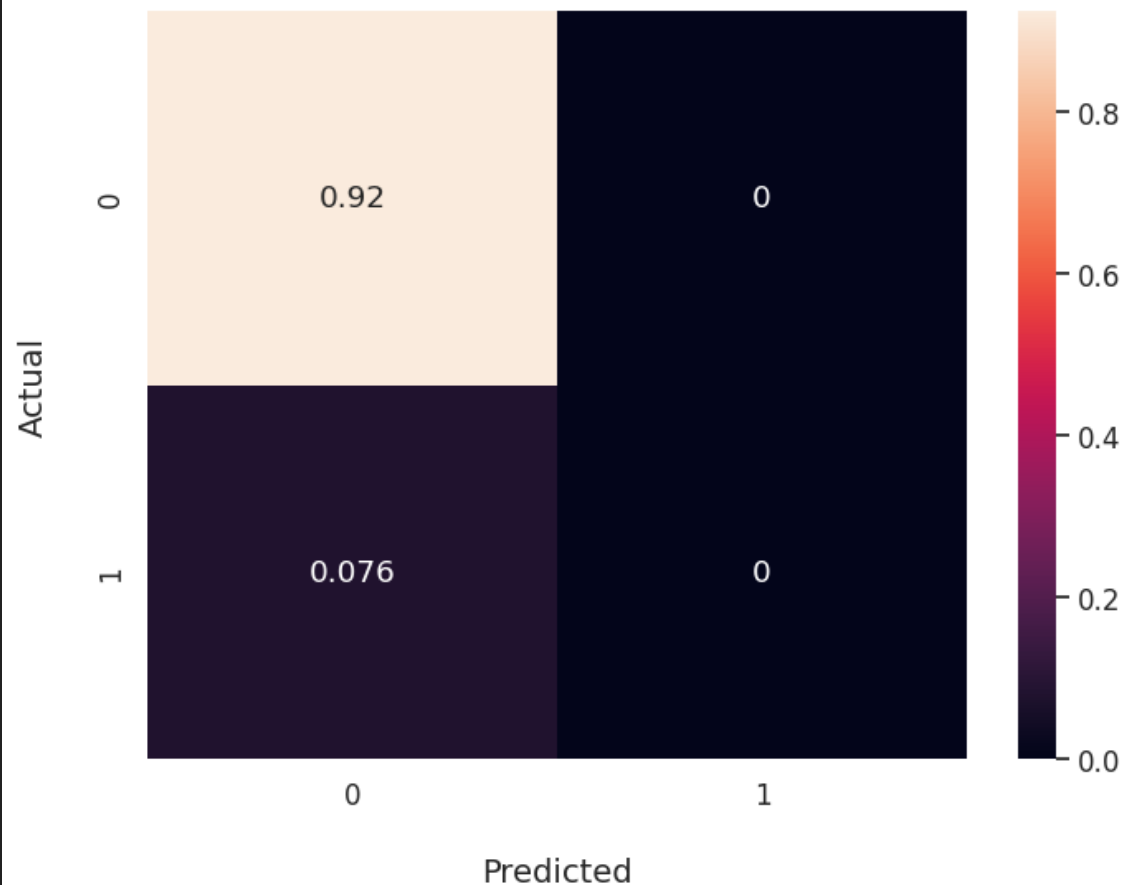
Confusion Matrix for credit default



Modelization

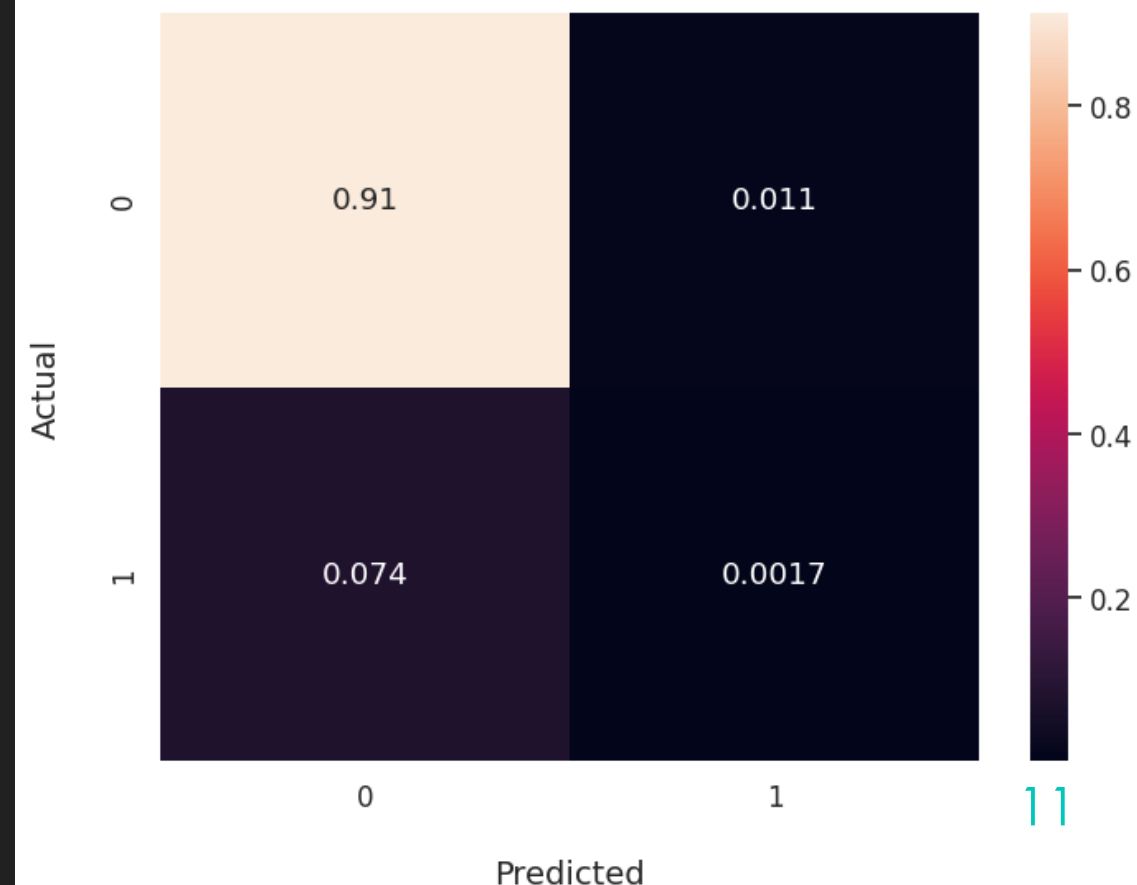
RandomForestClassifier

Confusion Matrix for credit default

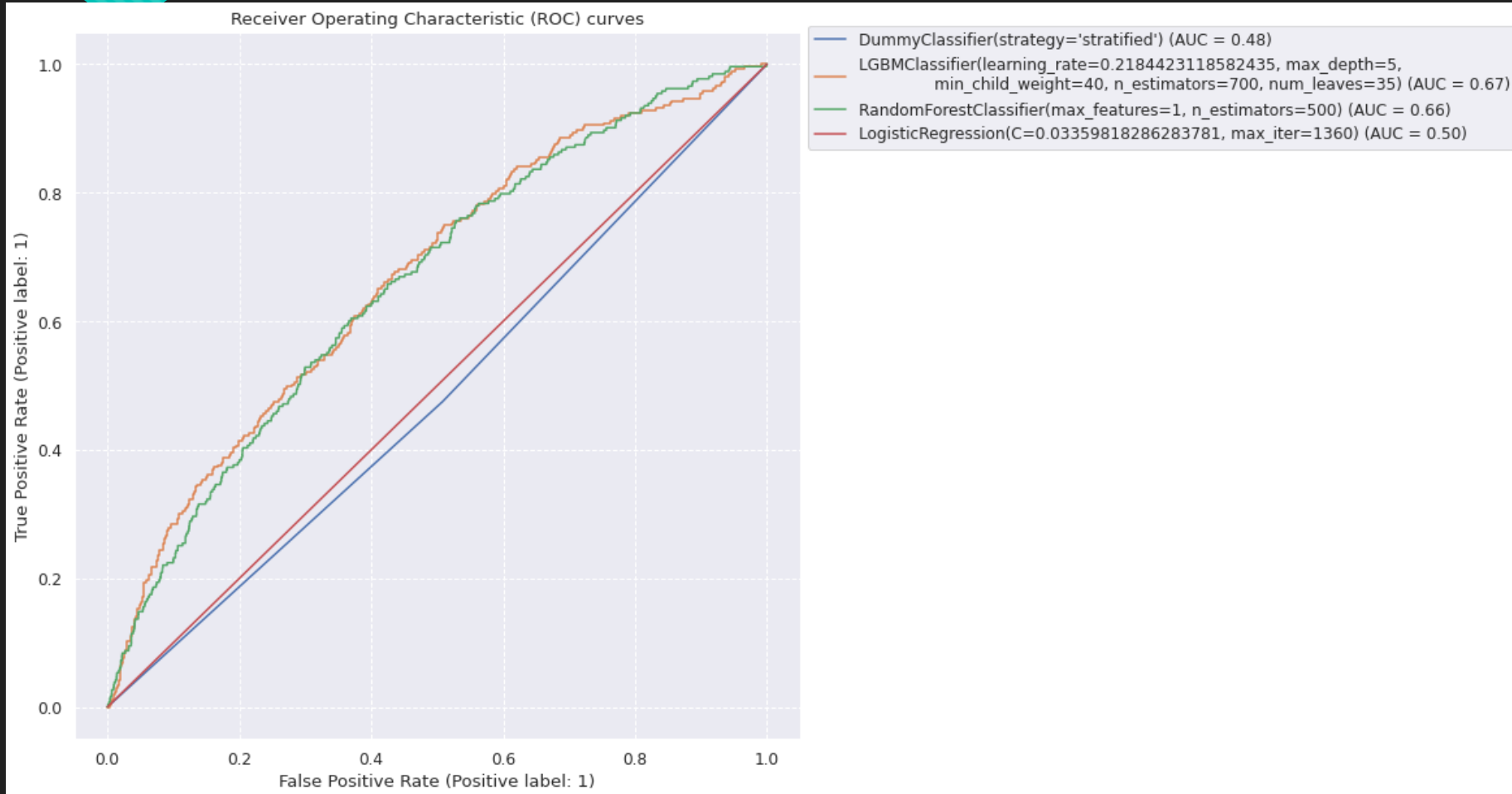


LGBMClassifier

Confusion Matrix for credit default



Modelization



Modelization

models	Custom Cost train	Custom Cost validation	Accuracy
DummyClassifier(strategy='stratified')	2.75	0.837994	0.504468
LGBMClassifier(learning_rate=0.218442311858243...	0.41	0.751802	0.914961
(DecisionTreeClassifier(max_features=1, random...	0.37	0.758144	0.924186
LogisticRegression(C=0.03359818286283781, max_...	5.00	0.758144	0.924186

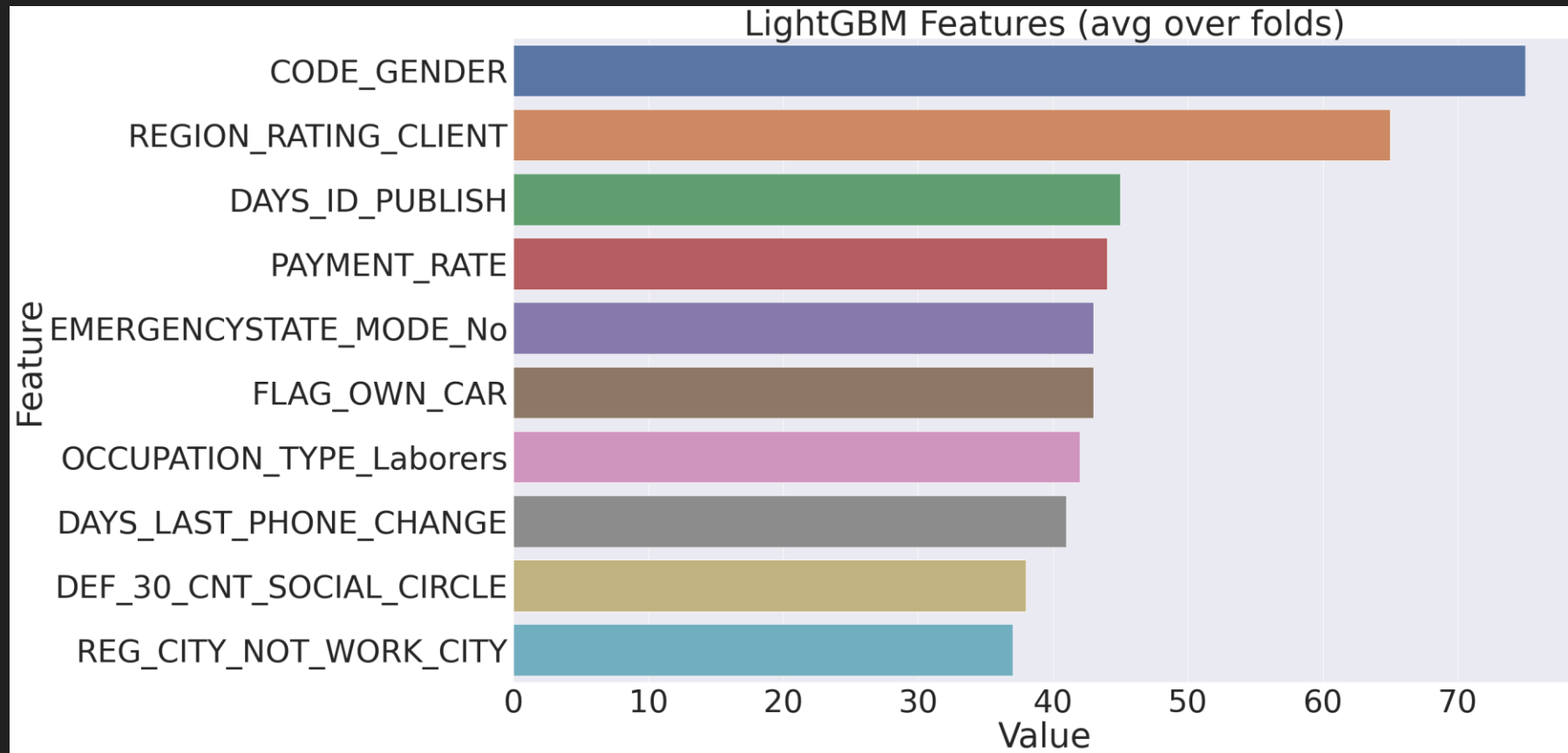
○ Best model : LGBMClassifier

Modelization

- The score obtained in the previous table are for a probability threshold of 0.5
- Now we measure the score for 100 threshold values between 0 and 1
- Probability threshold optimization :
 - Best Cost on test sample : 0.626
 - **Best Threshold : 0.09**

Modelization

○ Feature importance



Recording experiments and models with MLFlow

- We record models in Microsoft Azure studio :

Répertoire par défaut > formation > Models

Model List

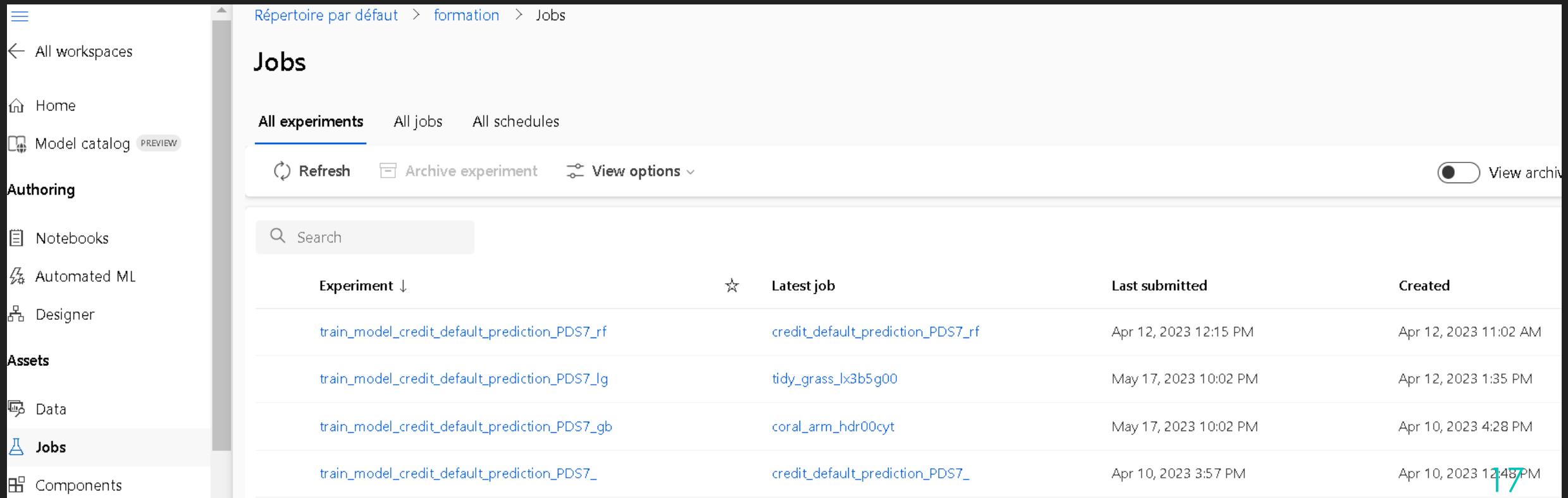
+ Register ▾ Refresh Delete Archive Deploy ▾ Compare (preview) ▾ View options ▾ ☒ Show latest versions only ☐ Include archived

Search Filter Columns

Name	☆	Version	Type	Source	Experiment	Job (Run ID)	Created
Best_model		6	MLFLOW	This workspace	Default	fa9d8e35-e4ce-4af3-87ac-ca8c8...	May 26, 2024
Best_model		5	MLFLOW	This workspace	Default	1cdd406f-b173-48f0-add2-026...	May 19, 2024

Recording experiments and models with MLFlow

- We record experiments in Microsoft Azure studio :



The screenshot displays the Microsoft Azure ML Studio interface, specifically the 'Jobs' section. The left sidebar contains navigation links for 'All workspaces', 'Home', 'Model catalog', 'Authoring' (Notebooks, Automated ML, Designer), 'Assets' (Data, Jobs, Components), and 'Components'. The main content area shows the 'Jobs' page with a breadcrumb trail 'Répertoire par défaut > formation > Jobs'. Below the title, there are tabs for 'All experiments' (selected), 'All jobs', and 'All schedules'. Action buttons include 'Refresh', 'Archive experiment', and 'View options'. A search bar is present above a table of experiments. The table has columns for 'Experiment', 'Latest job', 'Last submitted', and 'Created'. A 'View archived' toggle is on the right.

Experiment ↓	☆ Latest job	Last submitted	Created
train_model_credit_default_prediction_PDS7_rf	credit_default_prediction_PDS7_rf	Apr 12, 2023 12:15 PM	Apr 12, 2023 11:02 AM
train_model_credit_default_prediction_PDS7_lg	tidy_grass_lx3b5g00	May 17, 2023 10:02 PM	Apr 12, 2023 1:35 PM
train_model_credit_default_prediction_PDS7_gb	coral_arm_hdr00cyt	May 17, 2023 10:02 PM	Apr 10, 2023 4:28 PM
train_model_credit_default_prediction_PDS7_	credit_default_prediction_PDS7_	Apr 10, 2023 3:57 PM	Apr 10, 2023 12:48 PM

API with FASTAPI and Dashboard with Streamlit

The image shows a REST client interface with the following components:

- URL:** `https://monapp.herokuapp.com/docs#/default/predict_credit_predict_post`
- Method:** `POST` (labeled `/predict` Predict Credit)
- Parameters:** No parameters
- Request body:**

```
{  
  "SK_ID_CURR": 100002  
}
```
- Execute:** A button to execute the request.
- Responses:**
 - Code:** 200
 - Response body:**

```
{  
  "prediction": "Crédit refusé (prédiction > 0.09)",  
  "risque_default": 0.42  
}
```
 - Response headers:**

```
connection: keep-alive  
content-length: 74  
content-type: application/json  
date: Thu, 08 Jun 2023 09:01:03 GMT  
server: uvicorn  
via: 1.1 vegur
```

Red arrows indicate the flow of data: from the request body to the response code, and from the response code to the response body.

API with FASTAPI and Dashboard with Streamlit

1-Enter loan number

http://localhost:8501

Gmail - Boîte de réc... Planning CAF - Connexion Pôle emploi Compte ameli - mo... Équations de l'hype... OpenClassRooms Formulaire demand... Engagement de For...

Veillez sélectionner un numéro de demande de prêt

100002

situation familiale : Single / not married

Nombre d'enfant(s) : 0

Age : 26

```
if st.button("Prediction"):  
    resultat = requests.post(url="http://monapp.herokuapp.com/predict", data=json.dumps(input_data))
```

Projet 7 - Implémentez un modèle de scoring

Données client :

	CODE_GENDER	REGION_RATING_CLIENT	DAYS_ID_PUBLISH	PAYMENT_RATE	EMERGENCYSTATE_MODE_No	FLAG
	0	0	2	-2,120	0.0607	1

Prediction

2-Display basic customer info

3-Display all customer info

4-When pushed, gives the status of the customer (accepted/refused) and the failure probability

Prediction

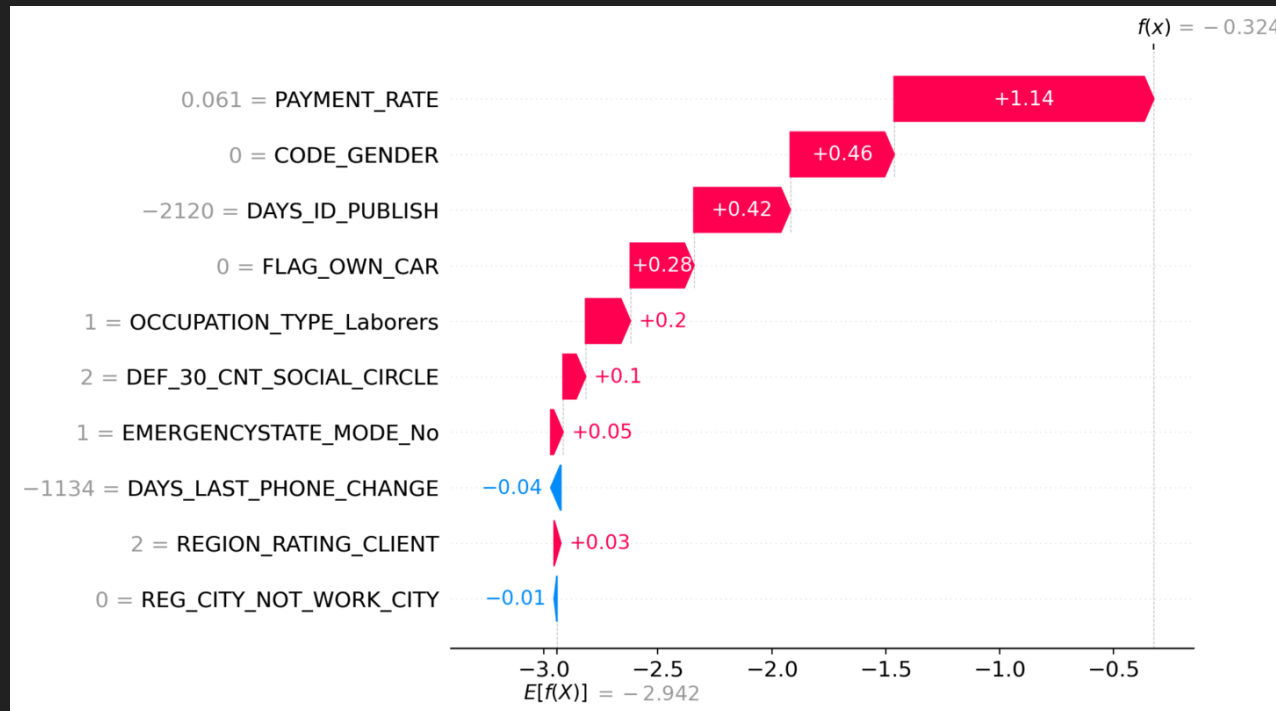
Crédit refusé

19

Probabilité de défaillance (limite 0.09): 0.42

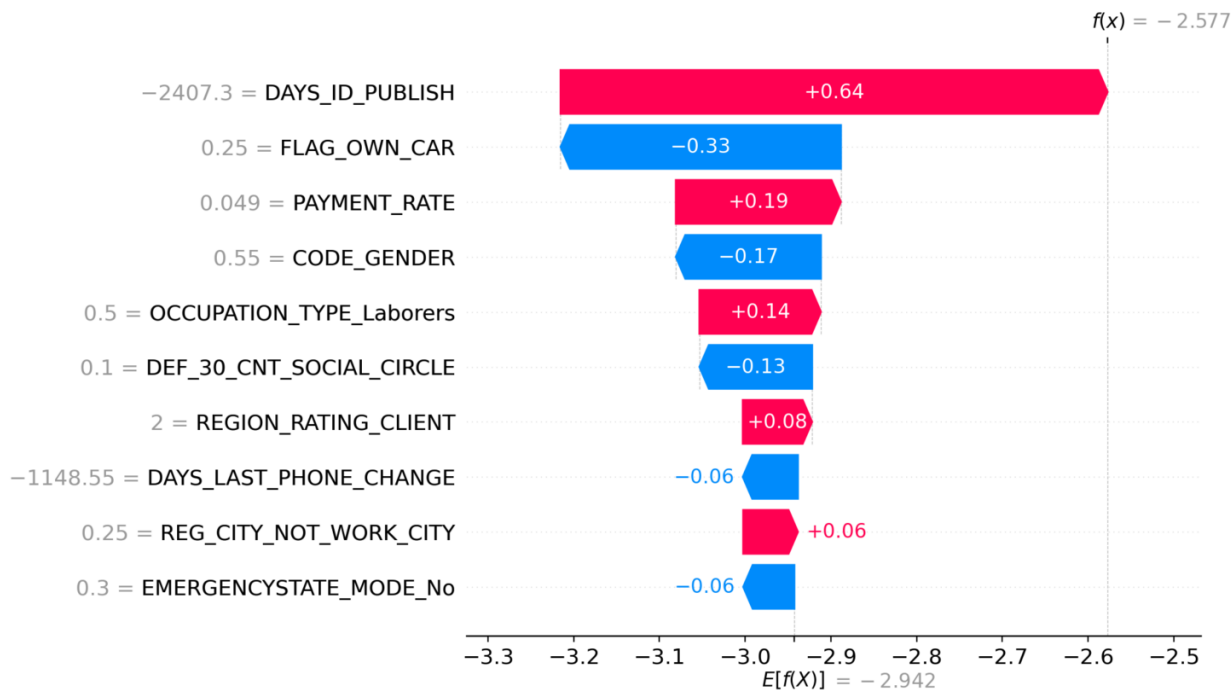
API with FASTAPI and Dashboard with Streamlit

Customer

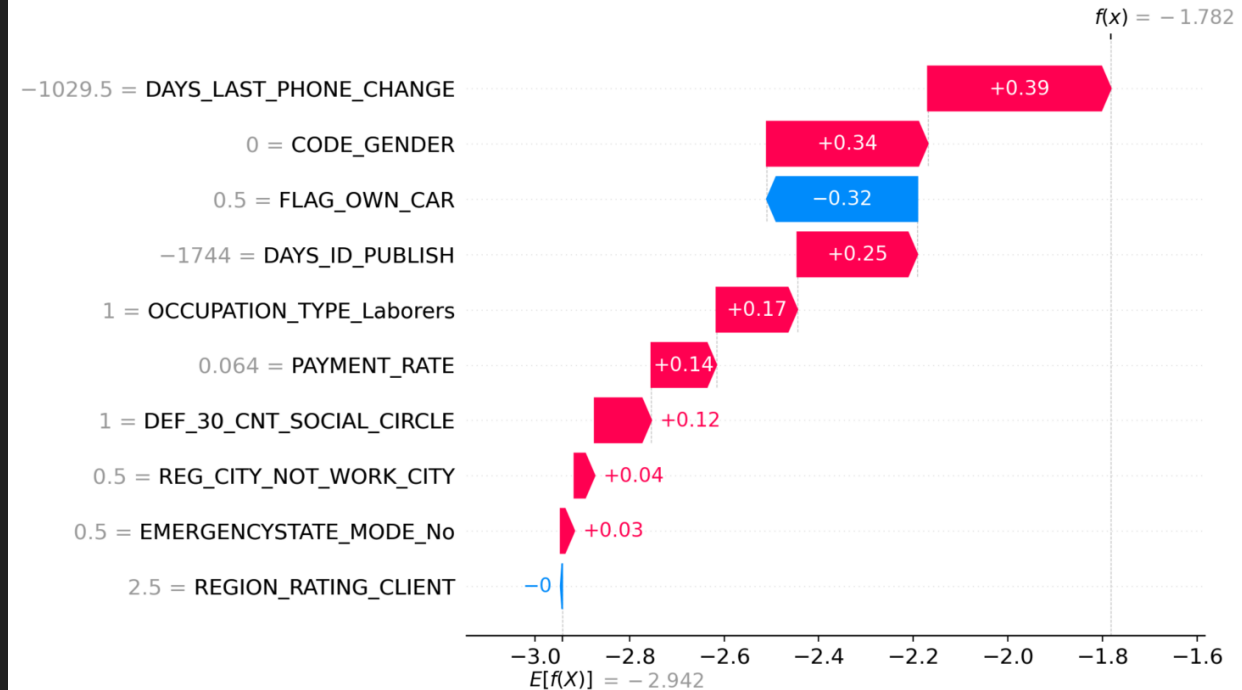


API with FASTAPI and Dashboard with Streamlit

Global population



Similar population (age, sex)



Testing with Pytest

- Two tests are implemented in the file « tests/test_P7.py » checking the ability of the algorithm to correctly predict an « accepted » and a « refused » customer.
- See the implementation of the tests below :

```
(venv) C:\Users\John\Desktop\Formation\7-Implémentez un modèle de scoring\tests>pytest
```

```
===== 2 passed, 6 warnings in 4.93s =====
```

```
(venv) C:\Users\John\Desktop\Formation\7-Implémentez un modèle de scoring\tests>pytest
===== test session starts =====
platform: win32 -- Python: 3.9.13, pytest: 7.1.3, pluggy: 1.0.0
rootdir: C:\Users\John\Desktop\Formation\7-Implémentez un modèle de scoring\tests

..... warnings summary .....
..\venv\lib\site-packages\pkg_resources\__init__.py:121
  C:\Users\John\Desktop\Formation\venv\lib\site-packages\pkg_resources\__init__.py:121: DeprecationWarning
    warnings.warn("pkg_resources is deprecated as an API", DeprecationWarning)

..\venv\lib\site-packages\pkg_resources\__init__.py:2870
..\venv\lib\site-packages\pkg_resources\__init__.py:2870
  C:\Users\John\Desktop\Formation\venv\lib\site-packages\pkg_resources\__init__.py:2870: DeprecationWarning
    Implementing implicit namespace packages (as specified in PEP 420) is preferred to 'pkg_resources.declar
    keywords.html#keyword-namespace-packages
    declare_namespace(pkg)

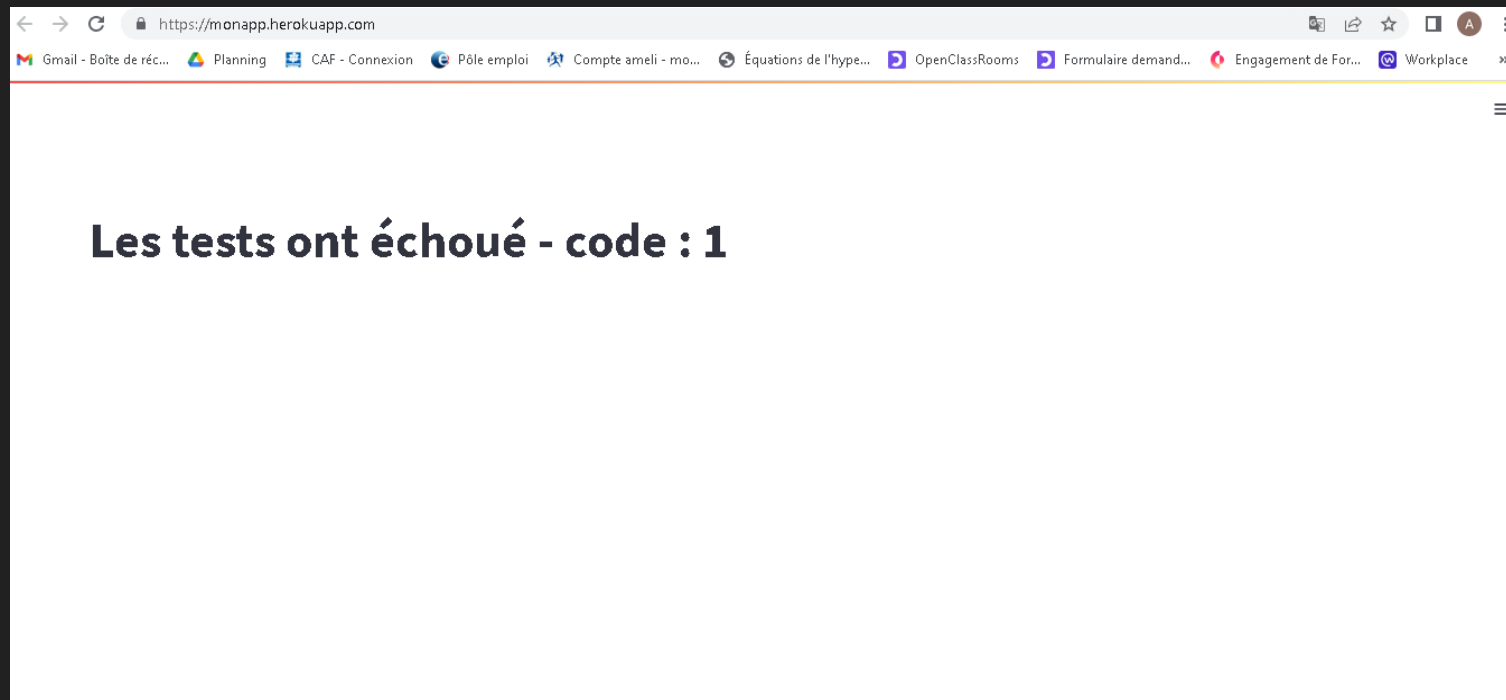
..\venv\lib\site-packages\pkg_resources\__init__.py:2870
  C:\Users\John\Desktop\Formation\venv\lib\site-packages\pkg_resources\__init__.py:2870: DeprecationWarning
    Implementing implicit namespace packages (as specified in PEP 420) is preferred to 'pkg_resources.declar
    keywords.html#keyword-namespace-packages
    declare_namespace(pkg)

test_P7.py::test_predict_accepted
test_P7.py::test_predict_refused
  C:\Users\John\Desktop\Formation\venv\lib\site-packages\sklearn\base.py:318: UserWarning: Trying to unpick
    2.2. This might lead to breaking code or invalid results. Use at your own risk. For more info please refe
    https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
    warnings.warn(

-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
===== 2 passed, 6 warnings in 4.93s =====
(venv) C:\Users\John\Desktop\Formation\7-Implémentez un modèle de scoring\tests>
```

Deployment of the API on internet with Heroku

- The tests are integrated in the API
 - In case of failure the user gets a message



Sharing code and versioning with Github

← → ↺ https://github.com/Antoine1608/OC-DS-P7

Gmail - Boîte de réc... Planning CAF - Connexion Pôle emploi Compte ameli - mo... Équations de l'hype... OpenClassRooms Form

🔍 Search or jump to... Pull requests Issues Codespaces Marketplace Explore

Antoine1608 / OC-DS-P7 (Public)

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags Go to file Add file <> Code About

Antoine1608 Update README.md 0b5ec53 now 24 commits

app	revue arborescence	last week
data	revue df limité aux features importantes	last week
model	new model and main improved on basic info	2 minutes ago
tests	new model and main improved on basic info	2 minutes ago
ui	new model and main improved on basic info	2 minutes ago
.gitignore	reinit gitignore	2 weeks ago
Naudy_Antoine_0_EDA_032023.ip...	mise à jour arborescence	last week
Naudy_Antoine_2_dossier_code_...	mise à jour arborescence	last week
Profile	pytest	last week
README.md	Update README.md	now
note introductive.docx	inclusion des tests directement dans main.py	4 days ago
note méthodologique.docx	maj drif detection	last week
requirements.bt	arborescence refaite	2 weeks ago

README.md

Dashboard Streamlit de scoring déployé sur Heroku

Projet 7 du parcours Data Scientist

Un dashboard streamlit est déployé sur Heroku.

Il affiche des données et un graphe relatif à un client choisi :

https://github.com/Antoine1608/OC-DS-P7

Commit history

Commits

master

Commits on May 26, 2023

Update README.md

Antoine1608 committed 3 minutes ago

Verified

0b5ec53

<>

new model and main improved on basic info

Antoine1608 committed 5 minutes ago

2e0efda

<>

Commits on May 22, 2023

maj tests

Antoine1608 committed 4 days ago

2770f4e

<>

testing tests

Antoine1608 committed 4 days ago

8cfe735

<>

test de test

Antoine1608 committed 4 days ago

8db4dcf

<>

supprimer 'import pytest'

Antoine1608 committed 4 days ago

9a29559

<>

inclusion des tests directement dans main.py

Antoine1608 committed 4 days ago

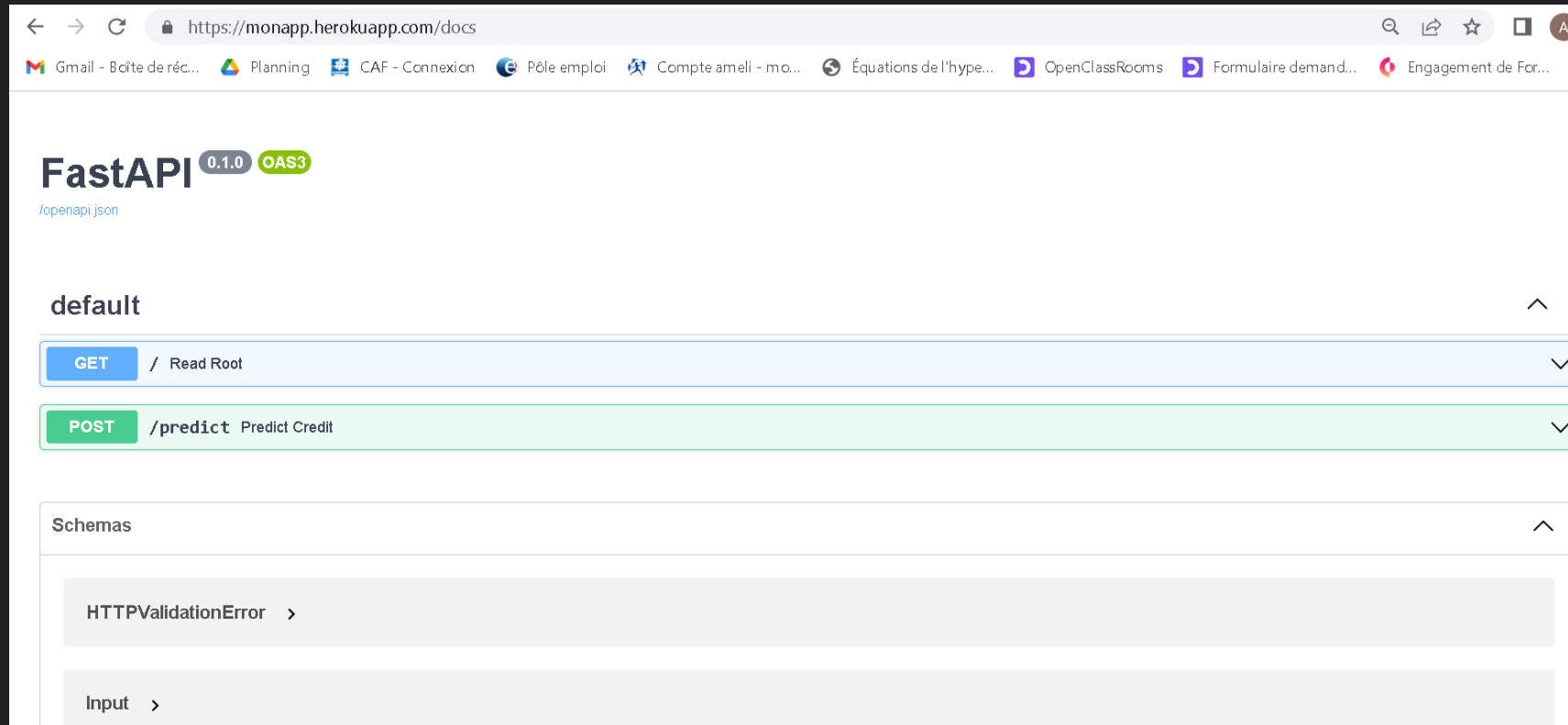
ddcf54

<>

24

Deployment of the API on internet with Heroku

- The backend API can be accessed on the url :
- <https://monapp.herokuapp.com/docs>



Deployment of the API on internet with Heroku

- The dashboard can be accessed on the url :
- <https://mondash.herokuapp.com>

The screenshot shows a web browser at <https://mondash.herokuapp.com>. The page title is "Projet 7 - Implémentez un modèle de scoring". On the left, a sidebar contains input fields: "Veuillez sélectionner un numéro de demande de prêt" (dropdown with "100002"), "Situation familiale : Single / not married", "Nombre d'enfant(s) : 0", and "Age : 26". The main area displays "Données client :" followed by a table of client data. Below the table is a "Prediction" button. The prediction result is "Crédit refusé (prédiction > 0.09)" with a "probabilité de défaillance 0.42". At the bottom, a "Client" section shows a bar chart for "0.061 = PAYMENT_RATE" with a value of "+1.14" and a function $f(x) = -0.324$.

	CODE_GENDER	REGION_RATING_CLIENT	DAYS_ID_PUBLISH	PAYMENT_RATE	EMERGENCYSTATE_MODE_No	FLA
	0	0	2	-2,120	0.0607	1

Drift analysis

- The drift analysis shows two features drifting between the « application_train » data and the « application_test » data :
 - DAYS_ID_PUBLISH with a K-S p_value of 0.03776
 - PAYMENT_RATE with a K-S p_value of 0
- See :
 - "tests\drift_report.html"

Dataset Drift

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

10

Columns

2

Drifted Columns







0.2

Share of Drifted Columns

Data Drift Summary

Drift is detected for 20.0% of columns (2 out of 10).

Search

Column	Type	Reference Distribution	Current Distribution	Data Drift ↑	Stat Test	Drift Score
> DAYS_ID_PUBLISH	num			Detected	K-S p_value	0.03776
> PAYMENT_RATE	num			Detected	K-S p_value	0
> DEF_30_CNT_SOCIAL_CIRCLE	num			Not Detected	K-S p_value	0.969001

Conclusion

- FastAPI, Streamlit, Github and Heroku enable us to deploy swiftly an API with a dashboard on internet.
- By selecting 10 important features in the dataset we can show and explain easily the result of the prediction to the customer.
- However we have seen that data are drifting over time. It's necessary to re-train the model periodically to keep the good performance of the model.
- It's also possible, if time is available, to test other models with additional parameters to go further in the optimisation of the algorithms.