

Implémenter un modèle de scoring

Felipe PEREIRA DE LIMA

<https://github.com/feliplim/credit-scoring>

Date : 18/12/2023



Prêt à dépenser

Table de matière

1

Contexte, objectif, jeux de données

2

Modélisation

3

Pipeline de déploiement

4

Data drift

5

Dashboard

6

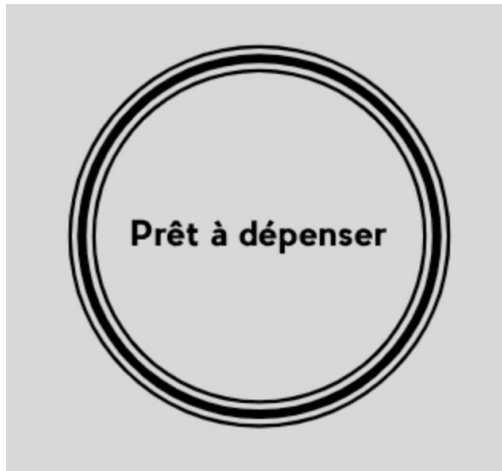
Limites et améliorations



1

Contexte, objectifs
et jeux de données

Contexte et objectifs



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.

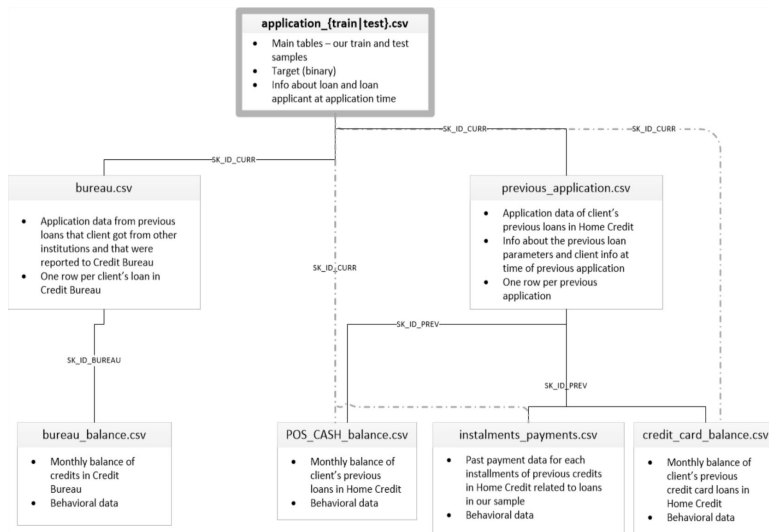
L'entreprise souhaite mettre en oeuvre un outil scoring crédit qui calcule la probabilité qu'un client rembourse ou pas son crédit, puis classifie la demande de crédit en accordée ou refusée

De plus, les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de transparence vis-à-vis des décisions d'octroi de crédit.

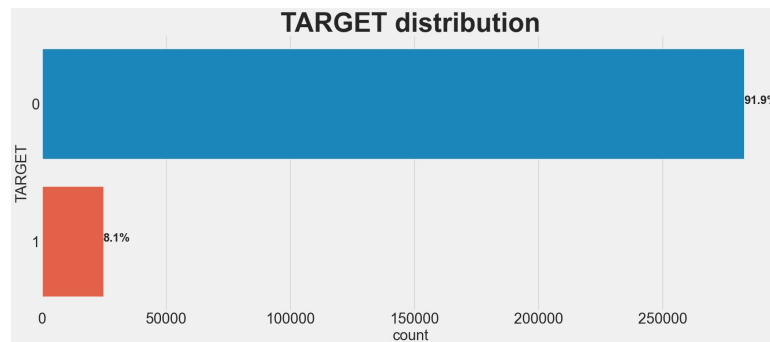
***Prêt à dépenser* décide donc de développer un dashboard interactif pour que les chargés de relation client puissent à la fois expliquer les décisions d'octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.**

Jeux de données

7 fichiers avec > 200 variables



Application “train” regroupe 307.511 clients dont on connaît la décision d’octroi de crédit.



Application “test” regroupe 48.774 clients dont on ne connaît pas la décision d’octroi de crédit.

Historique de prêt dans
d’autres institutions
financières

Historique de prêt chez *Prêt à
dépenser*

Exploration et nettoyage

- Suppression des données incorrectes
- Remplacement des anomalies par NaN
- Encodage des variables catégorielles
- Remplacement des données manquantes par la médiane
- Création des features (ratios)
- Agrégations (somme, moyenne)
- Suppression des variables corrélées



2

Modélisation

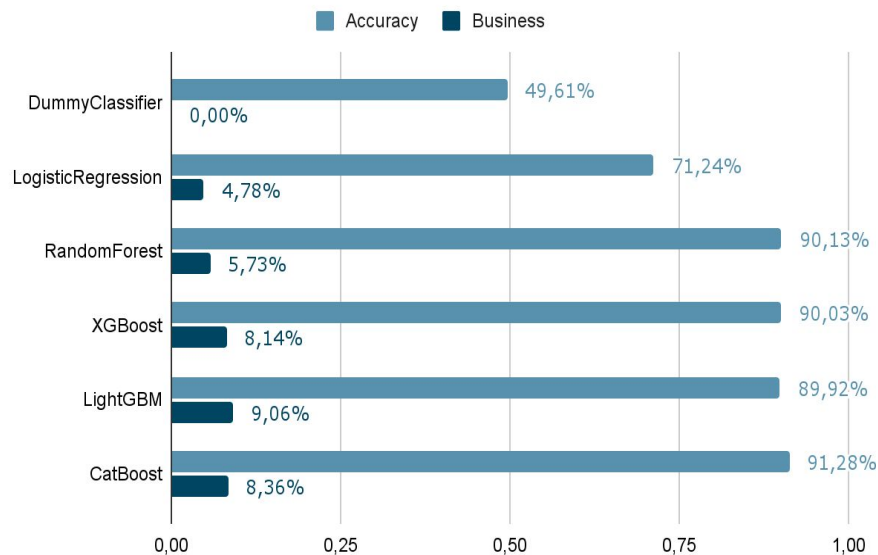
Choix métrique et modèle

La difficulté dans cette étape se base sur le fait que les faux négatifs sont plus impactants que les faux positifs. **Les faux négatifs coûtent plus cher que les faux positifs à l'entreprise.**

Train / Test split : 80/20
Chaque sous-échantillon contient la même distribution de classes

Evaluation des candidats à l'aide d'une validation croisée k-fold = 5

Business score qui pénalise 10x plus les faux négatifs



Suivi MLFlow

Table Chart Evaluation **Experimental**

<input type="checkbox"/>		Run Name	Created	Dataset	Duration	Source	Models
<input type="checkbox"/>		LightGBM_final	2 hours ago	-	3.2s	ipykerne...	sklearr
<input type="checkbox"/>		LightGBM_fine_tuned_2	2 hours ago	-	7.0s	ipykerne...	sklearr
<input type="checkbox"/>		CatBoost_fine_tuned	3 hours ago	-	2.6s	ipykerne...	sklearr
<input type="checkbox"/>		LightGBM_fine_tuned	3 hours ago	-	3.1s	ipykerne...	sklearr
<input type="checkbox"/>		XGBoost_fine_tuned	3 hours ago	-	3.2s	ipykerne...	sklearr
<input type="checkbox"/>		RandomForest_fine_tuned	4 hours ago	-	2.6s	ipykerne...	sklearr
<input type="checkbox"/>		LogisticRegression_fine_tuned	4 hours ago	-	3.1s	ipykerne...	sklearr
<input type="checkbox"/>		CatBoost	4 hours ago	-	1.8s	ipykerne...	sklearr
<input type="checkbox"/>		LightGBM	4 hours ago	-	2.2s	ipykerne...	sklearr
<input type="checkbox"/>		XGBoost	4 hours ago	-	2.0s	ipykerne...	sklearr
<input type="checkbox"/>		RandomForest	4 hours ago	-	2.6s	ipykerne...	sklearr
<input type="checkbox"/>		LogisticRegression	4 hours ago	-	1.9s	ipykerne...	sklearr
<input type="checkbox"/>		DummyClassifier	4 hours ago	-	3.0s	ipykerne...	sklearr

183 matching runs

Suivi MLFlow

credit_scoring [Provide Feedback](#)

Share

Experiment ID: 373801358020512024 Artifact Location: file:///Users/felipelima/Documents/projets/credit-scoring/mlruns/373801358020512024

> Description [Edit](#)

Q metrics.rmse < 1 and params.model = "tree"



Time created ▾

State: Active ▾

Sort: Created ▾



+ New run

Table **Chart** Evaluation **Experimental**

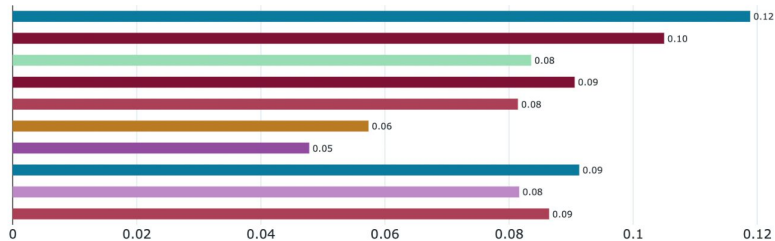
Run Name
LightGBM_final
LightGBM_fine_tuned_2
CatBoost_fine_tuned
LightGBM_fine_tuned
XGBoost_fine_tuned
RandomForest_fine_tuned
LogisticRegression_fine_tun...
CatBoost
LightGBM
XGBoost
RandomForest
LogisticRegression
DummyClassifier

183 matching runs

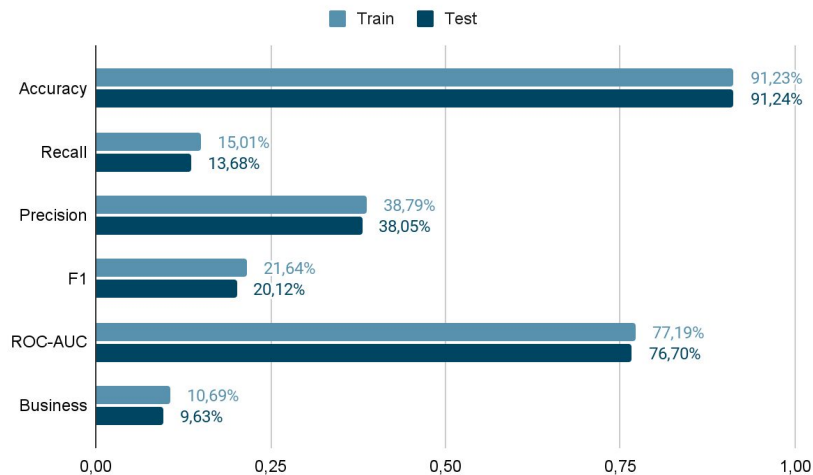
+ Add chart

validation_business_score

Comparing first 10 runs

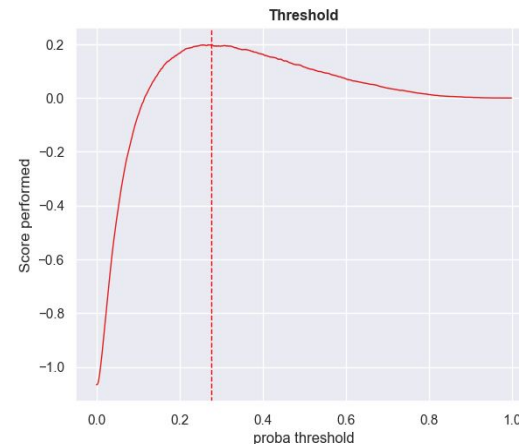


Performance du modèle après optimisation

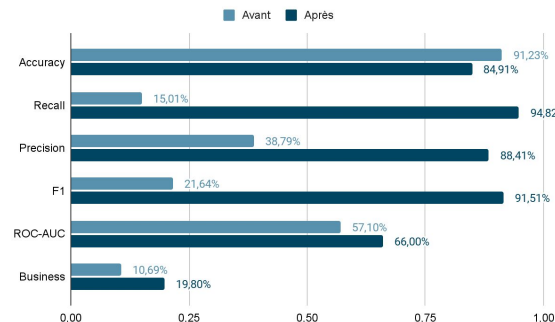


Le business score a augmenté de 9,1% à 10,7% sur le jeu d'entraînement, mais il n'est que de 9,6% sur le jeu de test.

Pour compléter l'optimisation, le seuil qui détermine la classe (1 ou 0) a été optimisé.

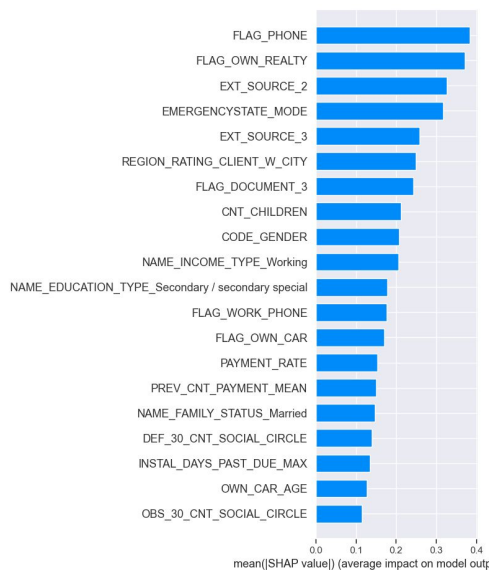
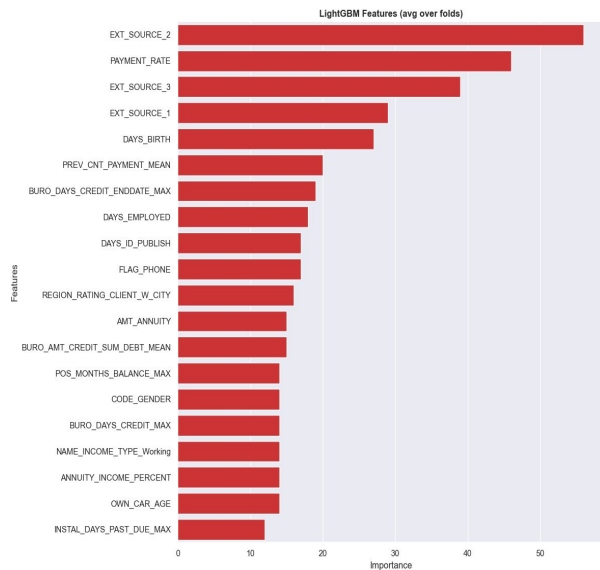


Le business score augmente de 10,7% à 19,8%.



Interprétation globale et locale

Les **valeurs de Shapley** calculent l'importance d'une variable en comparant ce qu'un modèle prédit avec et sans cette variable. Néanmoins, comme l'ordre dans lequel un modèle voit les variables peut impacter ses prédictions, cela est fait de façon aléatoire, afin que les fonctionnalités soient comparées équitablement.





3

Pipeline de
déploiement

Pipeline de déploiement

1. Script de l'API
2. Dockerfile pour l'API
3. Fichier requirements.txt
4. Fichier Github Actions (CI/CD)
5. Push sur dev → merge sur main
6. Déploiement de l'API sur AWS EC2 (grâce à Github Actions)
7. Déploiement du dashboard sur Streamlit Sharing

Dossier Github

The screenshot shows the Github repository page for 'credit-scoring' by user 'feliplim'. The repository is public and has 2 branches and 0 tags. The user has 223 commits. The repository contains several files and folders, including .github/workflows, api, dashboard, data/processed, docs, models, notebooks, tests, .gitignore, README.md, and requirements.txt. The most recent commit is 'Update README.md' by feliplim, committed yesterday.

credit-scoring Public








main 2 Branches 0 Tags

Go to file Add file Code

feliplim Update README.md ✓ 9b98de6 · yesterday 223 Commits

.github/workflows	fix: correct error in file directory	3 days ago
api	feat: add new endpoints to api	2 days ago
dashboard	fix: fix x-axis labels	2 days ago
data/processed	fix: changes in data types	3 days ago
docs	Delete docs/.DS_Store	yesterday
models	feat: change model and data sources	4 days ago
notebooks	fix: changes in data types	3 days ago
tests	feat: changes to improve api	4 days ago
.gitignore	feat: change workflows to connect to ec2 instance	5 days ago
README.md	Update README.md	yesterday
requirements.txt	feat: add package	3 days ago

Push et merge sur Github

Commits on Dec 14, 2023		
Update README.md feliplim committed 2 days ago · ✓ 2 / 2	Verified 25b3425	 <>
Merge pull request #74 from feliplim/dev ... feliplim committed 2 days ago · ✓ 2 / 2	Verified 1c3c644	 <>
fix: fix x-axis labels feliplim committed 2 days ago	ae69e74	 <>
Merge pull request #73 from feliplim/dev ... feliplim committed 2 days ago · ✓ 2 / 2	Verified d00c204	 <>
fix: fix spelling error feliplim committed 2 days ago	7df528a	 <>
feat: change plot types feliplim committed 2 days ago	c70a6be	 <>
feat: add new endpoints to api feliplim committed 2 days ago	5d41c6c	 <>

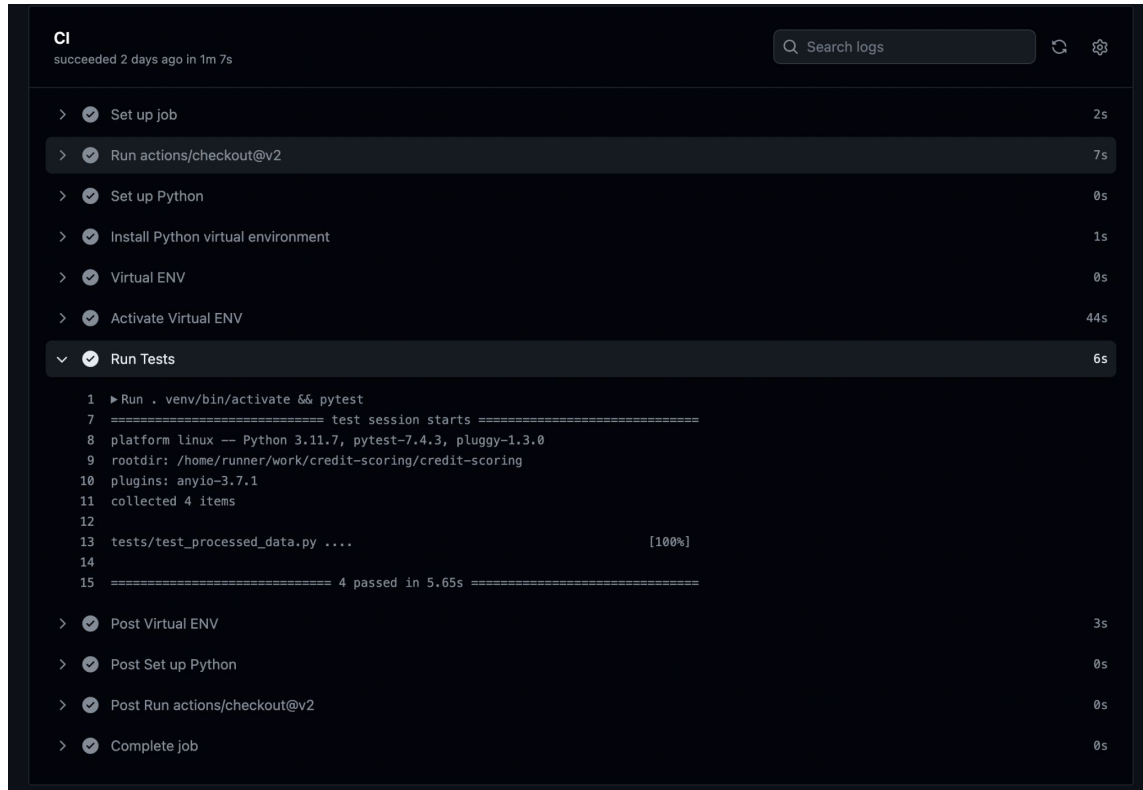
Tests unitaires en local

```
felipelina@MacBook-Air-de-Felipe credit-scoring % pytest
===== test session starts =====
platform darwin -- Python 3.11.6, pytest-7.4.3, pluggy-1.3.0
rootdir: /Users/felipelina/Documents/projets/credit-scoring
plugins: anyio-3.7.1
collected 4 items

tests/test_processed_data.py .... [100%]

===== 4 passed in 0.50s =====
felipelina@MacBook-Air-de-Felipe credit-scoring %
```

Tests unitaires sur Github



The screenshot displays a GitHub Actions CI workflow log. At the top, it indicates the workflow 'ci' succeeded 2 days ago in 1m 7s. A search bar for logs is present. The workflow steps are listed with their durations:

- Set up job (2s)
- Run actions/checkout@v2 (7s)
- Set up Python (0s)
- Install Python virtual environment (1s)
- Virtual ENV (0s)
- Activate Virtual ENV (44s)
- Run Tests (6s)
- Post Virtual ENV (3s)
- Post Set up Python (0s)
- Post Run actions/checkout@v2 (0s)
- Complete job (0s)

The 'Run Tests' step is expanded, showing the following terminal output:

```
1 ▶ Run . venv/bin/activate && pytest
7 ===== test session starts =====
8 platform linux -- Python 3.11.7, pytest-7.4.3, pluggy-1.3.0
9 rootdir: /home/runner/work/credit-scoring/credit-scoring
10 plugins: anyio-3.7.1
11 collected 4 items
12
13 tests/test_processed_data.py .... [100%]
14
15 ===== 4 passed in 5.65s =====
```

Déploiement du dashboard



[Gallery](#) [Components](#) [Community](#) [Docs](#) [Blog](#)

[Analytics](#)



[Settings](#)



feliplim

Your apps

New app



credit-scoring · main · dashboard/1_🏠_Homepage.py



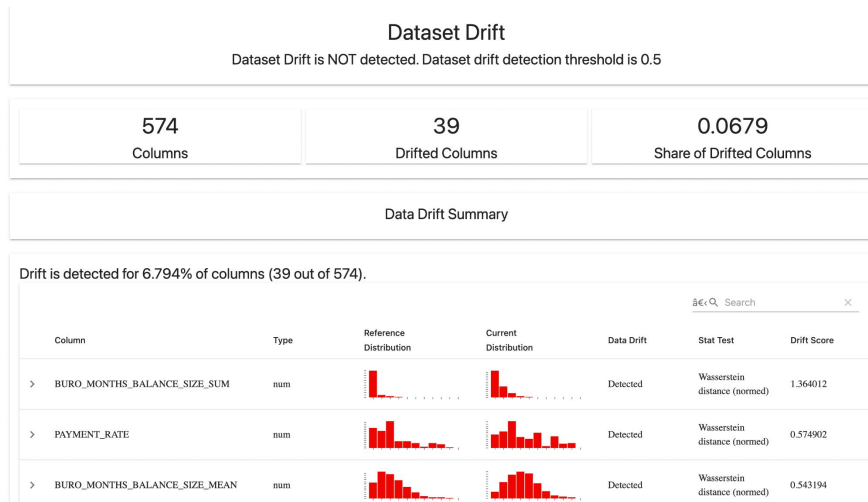


4

Data drift

Data drift

Le **data drift**, ou la dérive des données, est le changement de la distribution de la donnée au fil du temps. Quand les caractéristiques ou propriétés de la donnée changent, cela peut impacter la performance du modèle qui l'utilise et l'exactitude de l'analyse et de la prise de décision.



39 colonnes ont dérivé, ce qui représente **6,8%** des colonnes



5

Dashboard

<https://credit-scoring-felipelim.streamlit.app/>



6

Limites et
améliorations

Limites et améliorations

- Connaissance limitée → vérifier la cohérence des choix de variables
- Définir plus finement la métrique d'évaluation et la fonction coût en accord avec l'équipe métier
- Partie interactive sur le dashboard pour vérifier des scénarios si changement de la valeur de variables du clients.

A woman with blonde hair in a ponytail, wearing a light-colored turtleneck, is standing and gesturing with her right hand while speaking. She is in a modern office setting with wooden desks, white chairs, and dark shelving units in the background. A man is seated at a desk in the background, looking towards the speaker. The image is overlaid with large, dark blue and white geometric shapes, including diamonds and triangles. The text "Merci de votre attention" is displayed in a light orange color within one of the dark blue shapes.

Merci de votre
attention