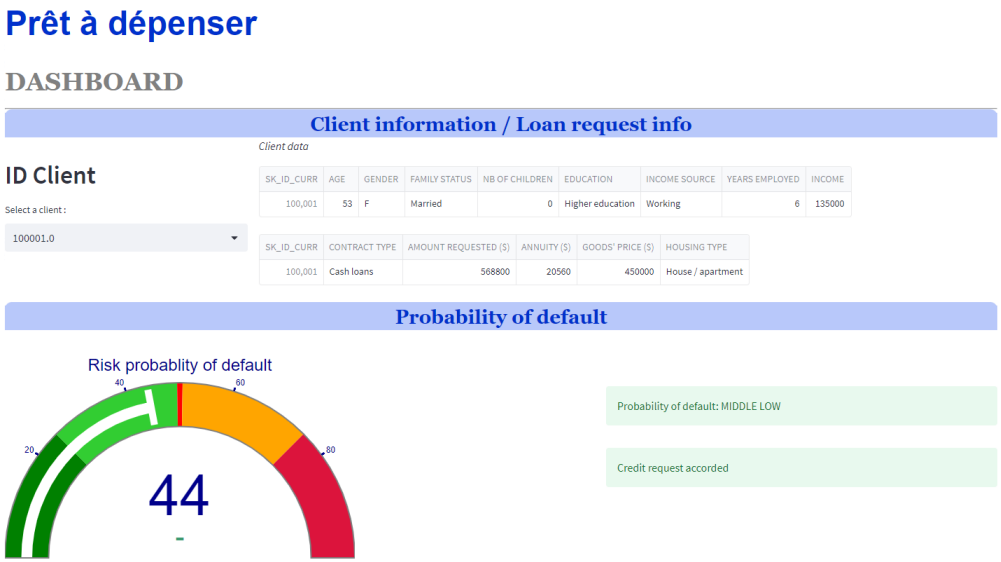
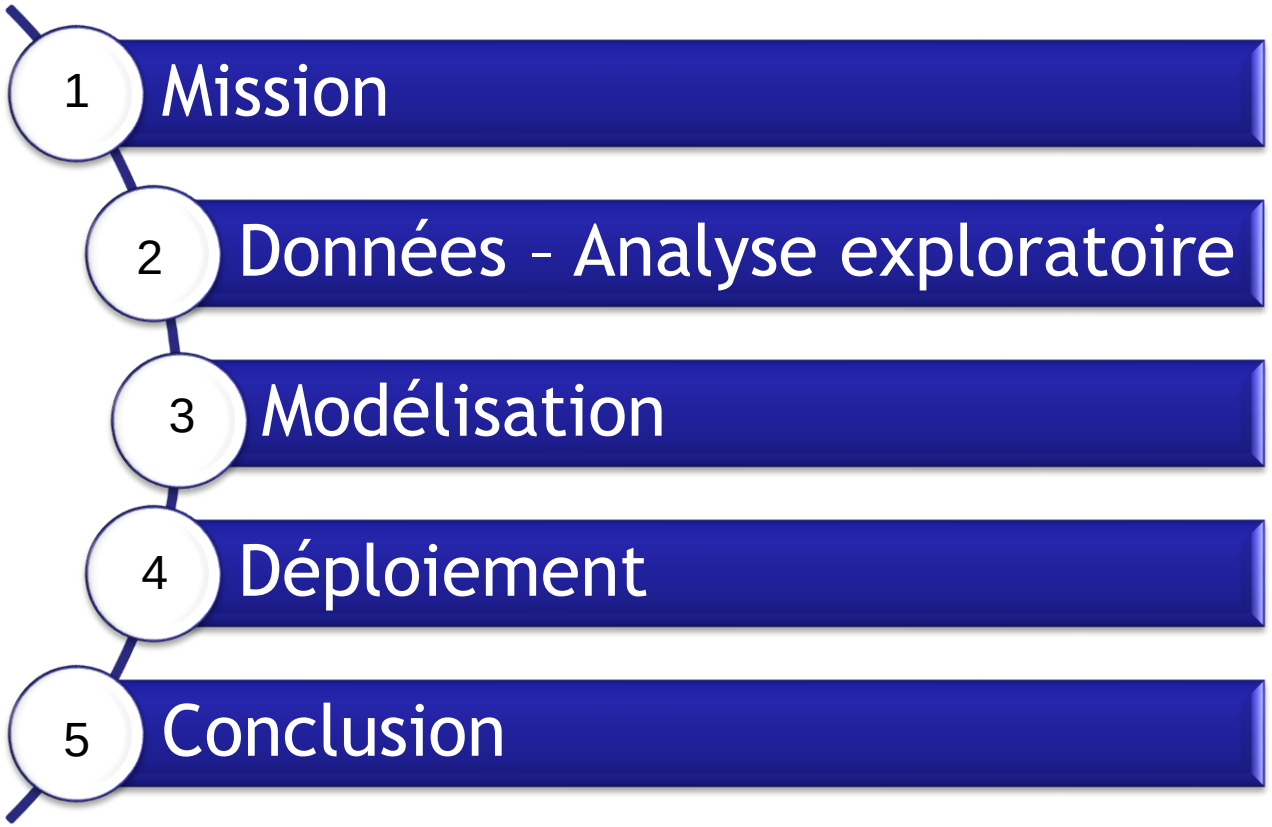
Prêt à dépenser

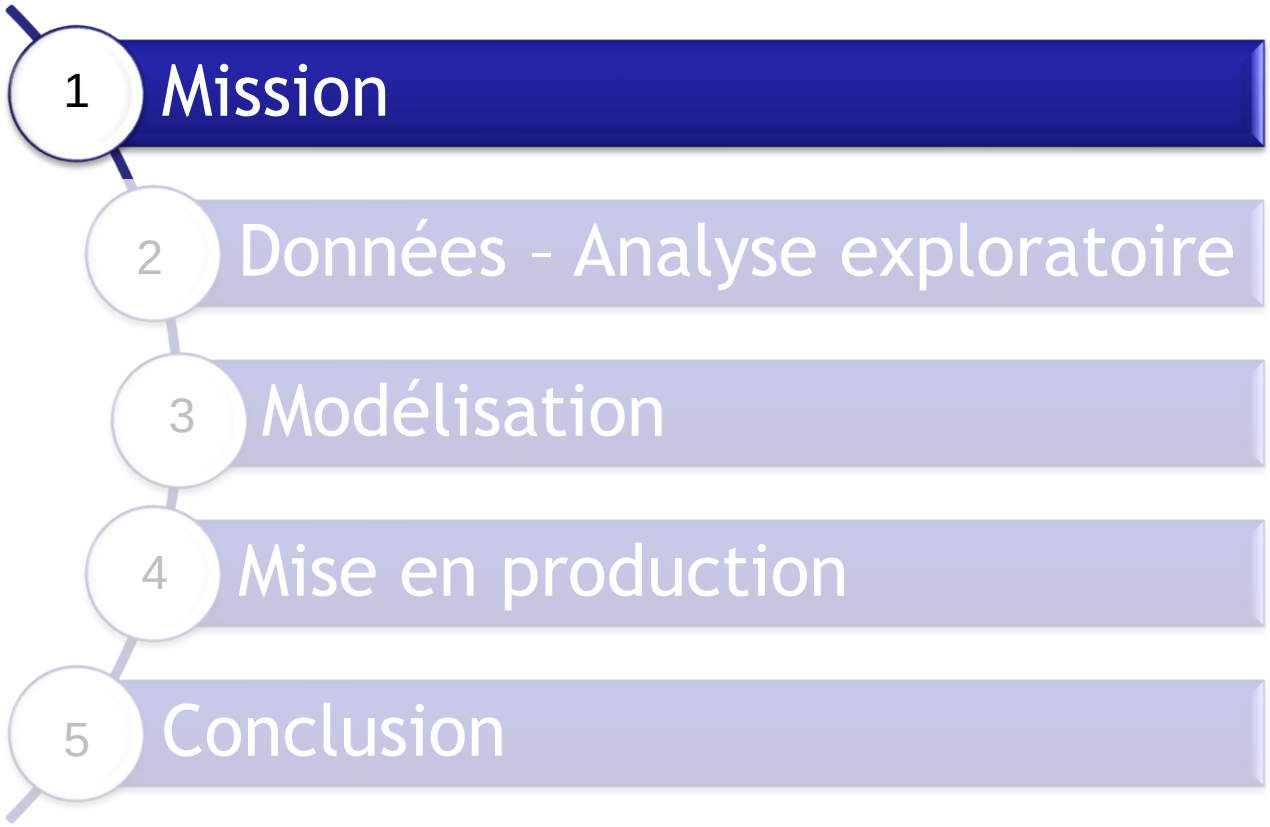
Implémentez un modèle de scoring



Sommaire


1
Mission
2
Données – Analyse exploratoire
3
Modélisation
4
Déploiement
5
Conclusion

Sommaire


1
Mission
2
Données – Analyse exploratoire
3
Modélisation
4
Mise en production
5
Conclusion

1. Mission

Contexte:

Prêt à dépenser est une société qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt

Mission:

Produire un modèle de prédiction de la probabilité de défaut de paiement d'un demandeur de crédit et déployer une api et un dashboard interactif

Objectifs:

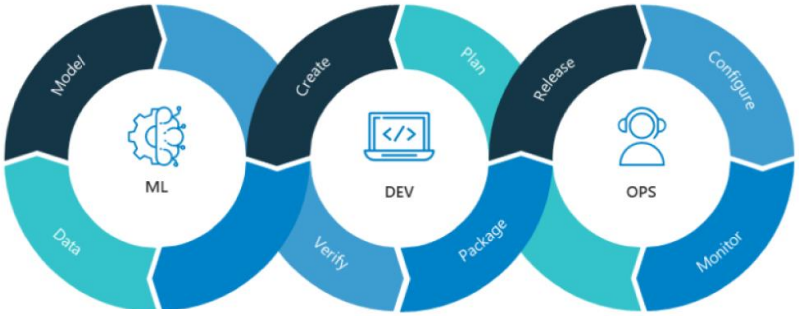
1. Étayer la décision
2. Améliorer la relation client en faisant preuve de transparence
3. Faciliter la compréhension du résultat avec le dashboard

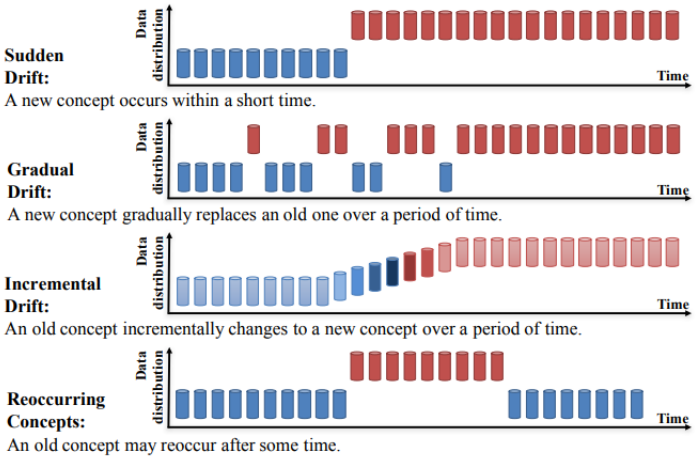
1. Mission

Prêt à dépenser

Les enjeux:

1 Mettre en place une démarche de type MLops pour suivre l'évolution du modèle





2 Mettre en œuvre une analyse du data drift pour anticiper des mises à jour du modèle

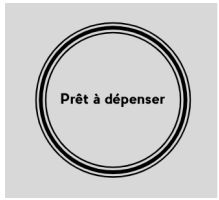
3 Mettre en place un score métier représentatif des enjeux de la mission

Reality

Confusion matrix	Negative : 0	Positive : 1
Negative : 0	True Negative : TN	False Negative : FN
Positive : 1	False Positive : FP	True Positive : TP

Prediction

1. Processus



EDA

kaggle

EDA complète

Pré-traitement

kaggle

Feature engineering
Variables auto
Assemblage

Feature Selection

LGBM
Boruta
BorutaSHAP
RFCE

Nettoyage

Imputation
Valeurs aberrantes
Encodage
Colinéarité
Équilibrage

Modélisation

PYCARET

Pipeline
Choix du modèle

PYCARET

Score métier
Hyperparameter tuning
OPTUNA

GitHub DagsHub

Mlflow
Versioning

PYCARET

Data drift
EVIDENTLY AI

Déploiement

Streamlit

Dashboard

FastAPI

Dashboard

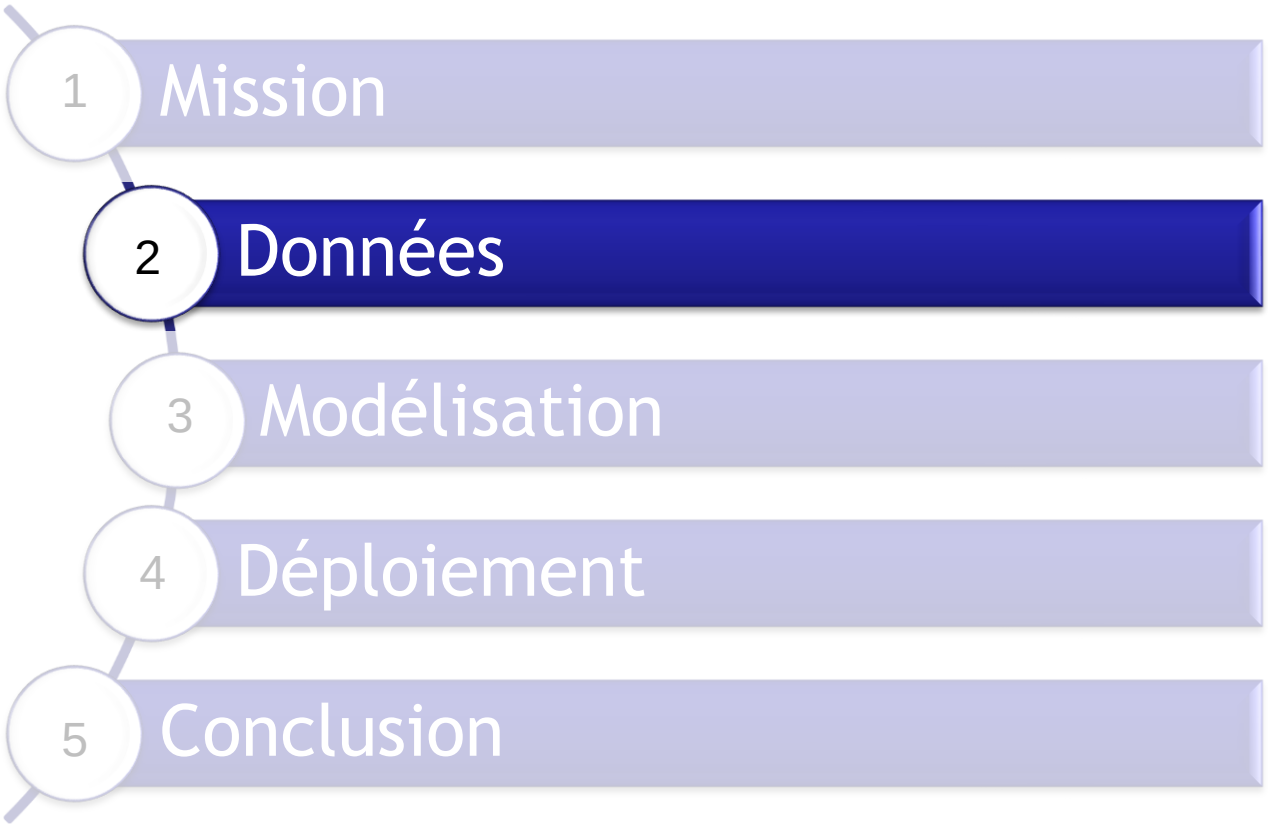
GitHub

Local vers serveur

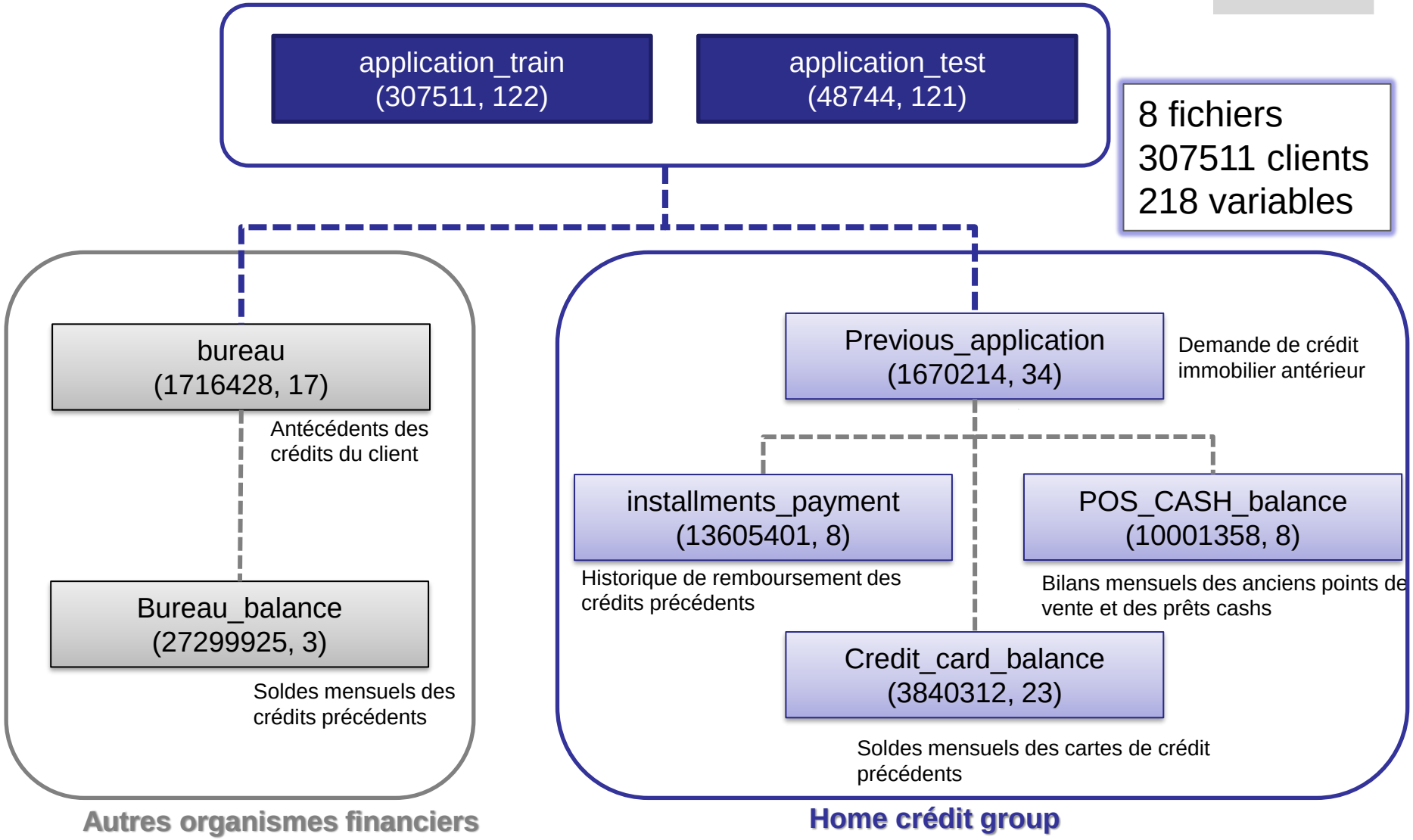
HEROKU

Serveur

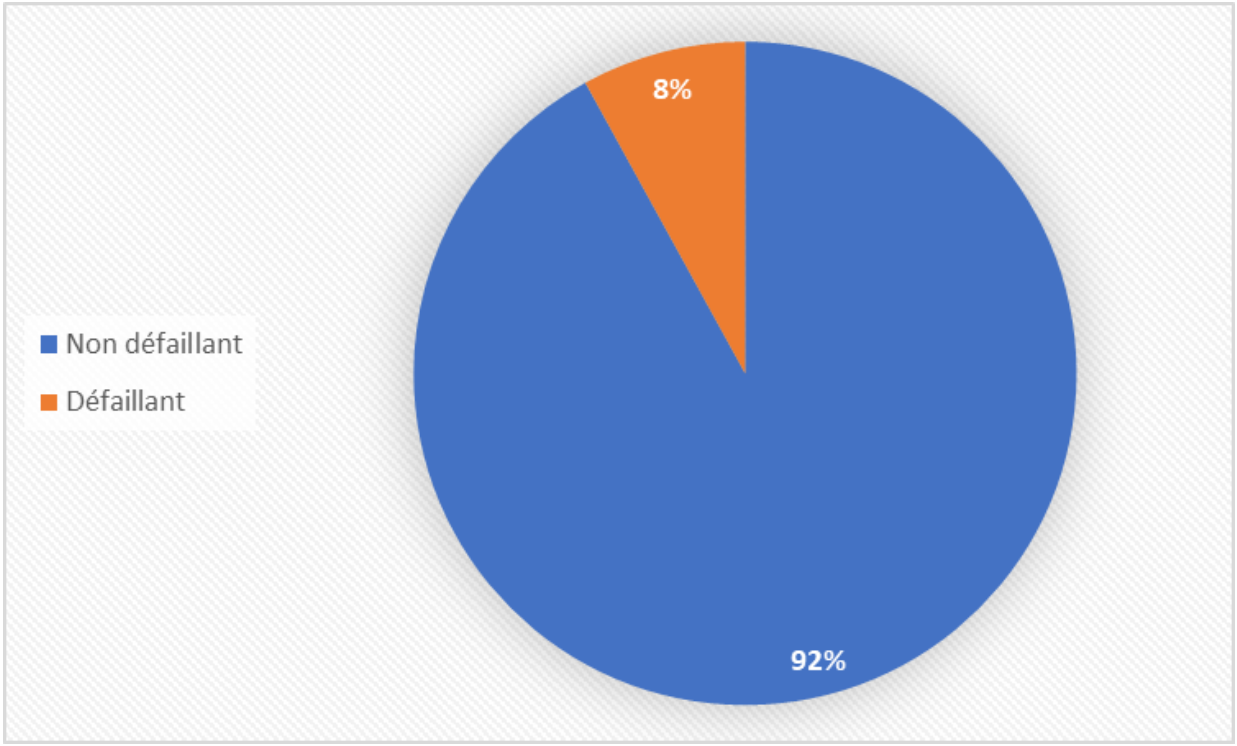
Sommaire


1
Mission
2
Données
3
Modélisation
4
Déploiement
5
Conclusion

2. Les données



2. Les données



2. Pré-traitement

Prêt à dépenser

Valeurs aberrantes (Nettoyage)	Imputation (Feature engineering)	Variables statistiques (Feature engineering)	Encodage (Feature engineering)	Fusion Optimisation (Assemblage)	Feature selection Colinéarités (Réduction de dimension)
Correction des valeurs aberrantes de l'EDA	NaNimputer (verstack + XGboost)	Min, Max, Moyenne, Variance Somme Taille Unique	Variables quantitatives: RobustScaler	Dtypes optimisation	Feature importance Permutation importance Boruta BorutaShap RCFE
		Différence	Variables qualitatives: OneHotEncoding		Colinéarité: coeff de Pearson >0.9

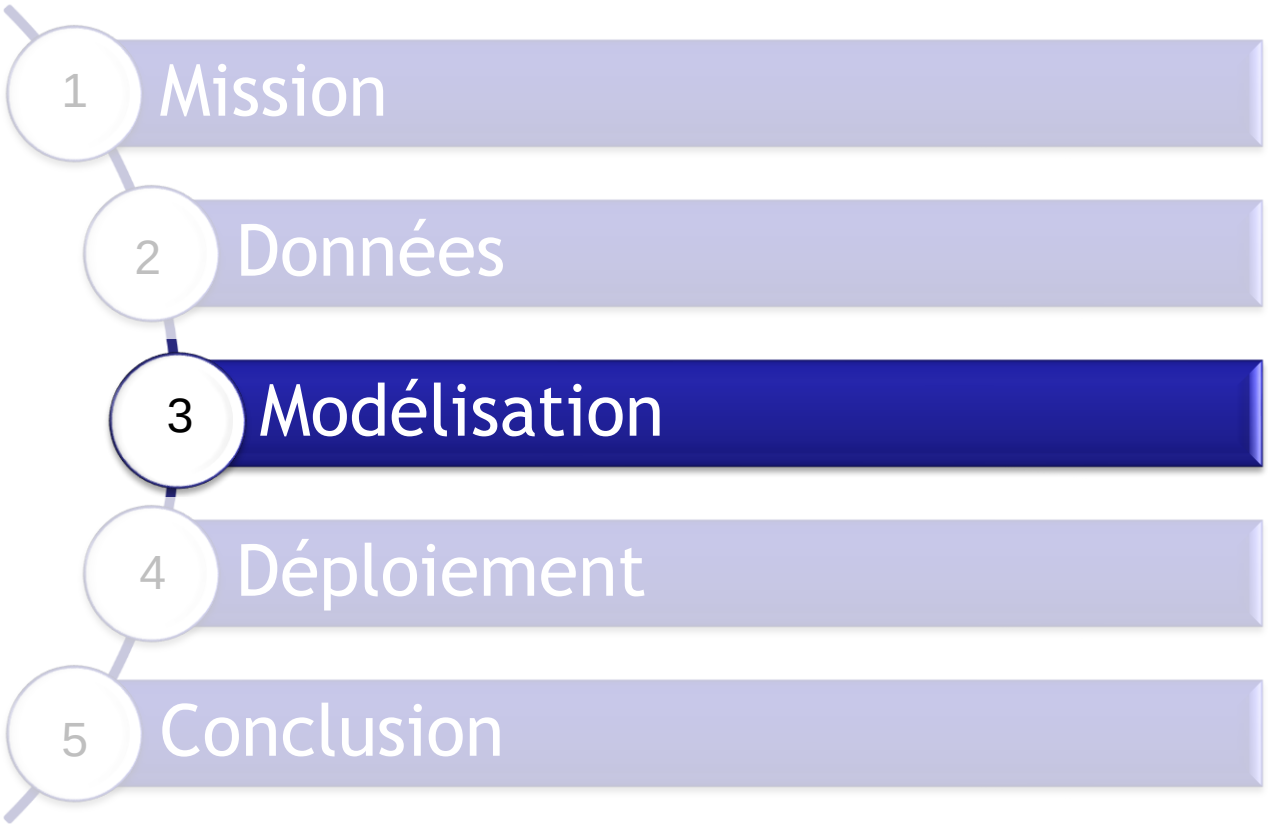
2. Pré-traitement

Prêt à dépenser

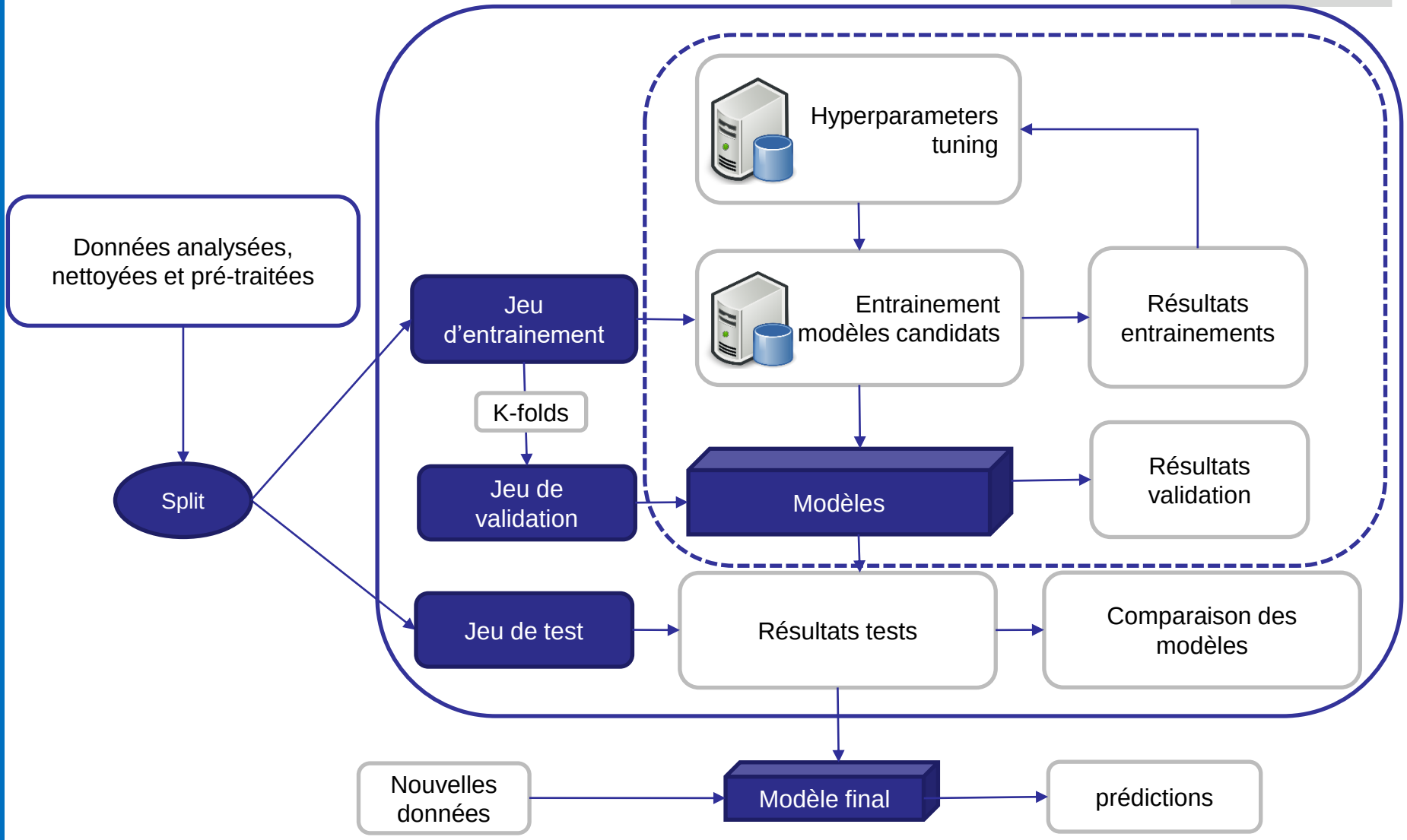
Valeurs aberrantes — Nettoyage	Imputation — Feature engineering	Variables statistiques — Feature engineering	Encodage — Feature engineering	Fusion Optimisation — Assemblage	Feature selection Colinéarités — Réduction de dimension
Correction des valeurs aberrantes de l'EDA	NaNimputer (verstack + XGboost)	Min, Max, Moyenne, Variance Somme Taille Unique	Variables quantitatives: RobustScaler	Dtypes optimisation	Feature importance Permutation importance Boruta BorutaShap RCFE
		Différence	Variables qualitatives: OneHotEncoding		

356 251 lignes – 546 variables
Feature selection nécessaires

Sommaire


1
Mission
2
Données
3
Modélisation
4
Déploiement
5
Conclusion

2. Modélisation



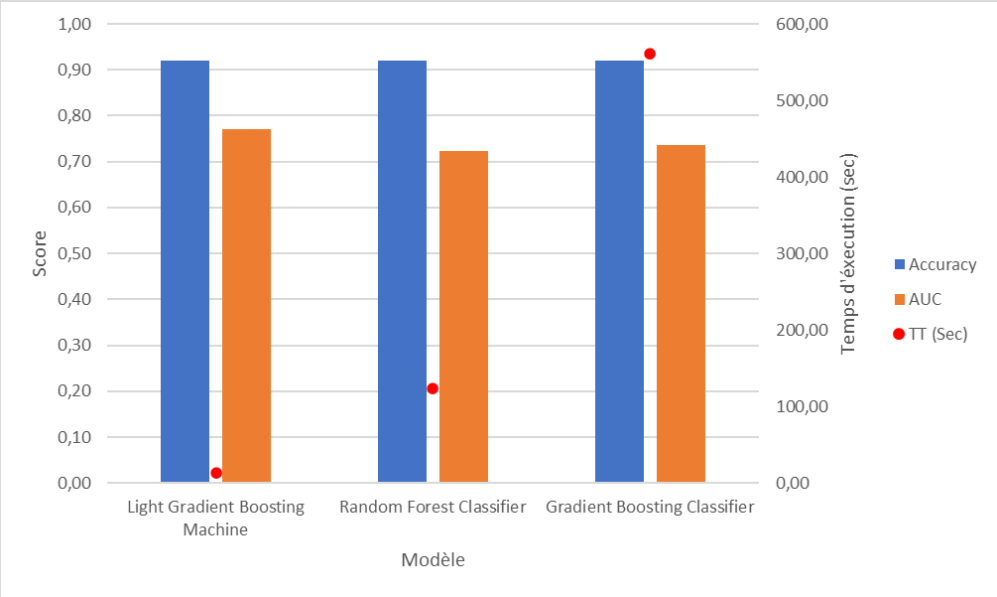
2. Modélisation: choix du modèle

Prêt à dépenser

PYCARET

Classification

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F10	Score Métier	TT (Sec)
Light Gradient Boosting Machine	0,92	0,77	0,02	0,53	0,04	0,04	0,10	0,02	0,54	13,59
Random Forest Classifier	0,92	0,72	0,00	0,53	0,00	0,00	0,03	0,00	0,53	123,68
Gradient Boosting Classifier	0,92	0,74	0,01	0,50	0,01	0,01	0,05	0,01	0,54	561,49
Extra Trees Classifier	0,92	0,72	0,00	0,54	0,01	0,00	0,03	0,00	0,53	67,24
Dummy Classifier	0,92	0,50	0,00	0,00	0,00	0,00	0,00	0,00	0,53	1,97
Extreme Gradient Boosting	0,92	0,77	0,05	0,46	0,10	0,08	0,13	0,06	0,55	358,59
Ada Boost Classifier	0,91	0,70	0,05	0,24	0,08	0,05	0,07	0,05	0,55	106,05
Decision Tree Classifier	0,84	0,54	0,17	0,14	0,15	0,07	0,07	0,17	0,56	29,86
K Neighbors Classifier	0,74	0,58	0,34	0,12	0,17	0,06	0,07	0,29	0,57	491,98
Linear Discriminant Analysis	0,70	0,76	0,69	0,17	0,27	0,16	0,22	0,54	0,70	14,01
Ridge Classifier	0,70	0,00	0,69	0,17	0,27	0,16	0,22	0,54	0,70	3,44
Logistic Regression	0,64	0,62	0,54	0,12	0,20	0,08	0,11	0,41	0,60	73,89
SVM - Linear Kernel	0,59	0,00	0,59	0,11	0,19	0,06	0,10	0,42	0,59	4,52
Quadratic Discriminant Analysis	0,36	0,62	0,79	0,09	0,17	0,03	0,07	0,47	0,54	11,73
Naive Bayes	0,15	0,57	0,94	0,08	0,15	0,00	0,02	0,48	0,48	9,69

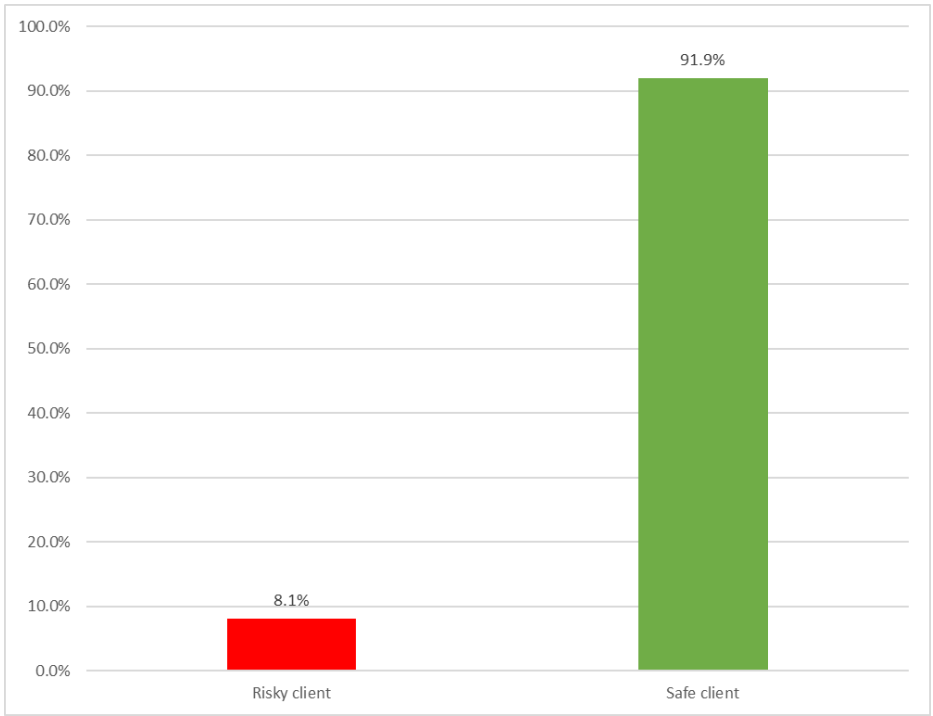


2. Modélisation : traitement du déséquilibre des classes

Données déséquilibrée:

- 92% des clients sont fiables (classe 0)
- 8% des clients ont fait défaut (classe 1)

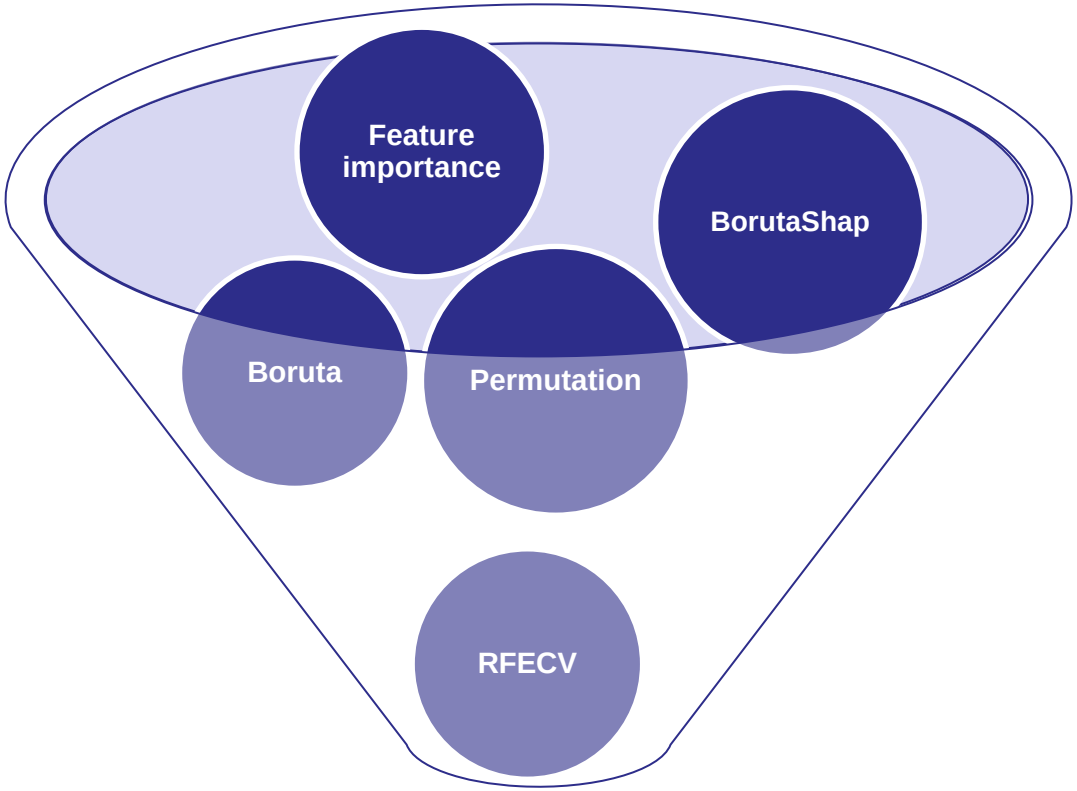
Un modèle mal entrainé va reflèter la distribution de la cible



2. Modélisation – feature selection

LightGBM

546 variables



159 variables

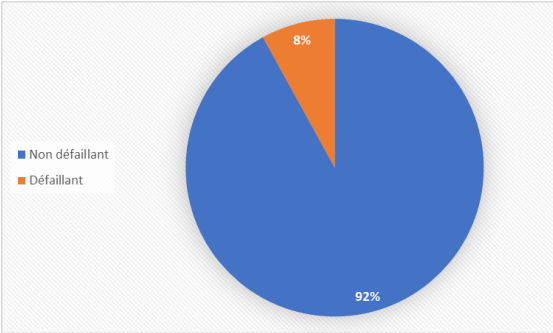
2. Modélisation : traitement du déséquilibre des classes

Prêt à dépenser

Données déséquilibrée:

92% des clients sont fiables (classe 0)

8% des clients ont fait défaut (classe 1)



SMOTE

Créer des données synthétiques de la classe minoritaire

Class weight

Créer un modèle qui attribute des poids différents pour chaque classe pour pénaliser la classe majoritaire

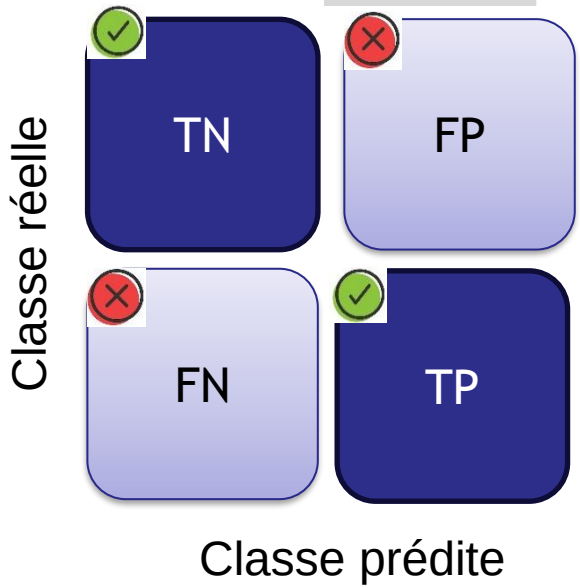
Scale_pos_weight

Modifie le seuil de classification de probabilité de la classe en favorisant une des classes

2. Modélisation - choix des métriques

Métriques générales

- **Accuracy**: pourcentage de client correctement classé
- **AUC-ROC**: capacité du modèle à classer
- **Recall** : pourcentage des clients prédits à risque par rapport aux clients réellement à risque
- **Précision**: pourcentage des clients correctement identifiés à risque par rapport aux clients prédits à risques



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, perte en capital et défaut de remboursement.
FP: client fiable prédit à risque : crédit refuse, manque à gagner.

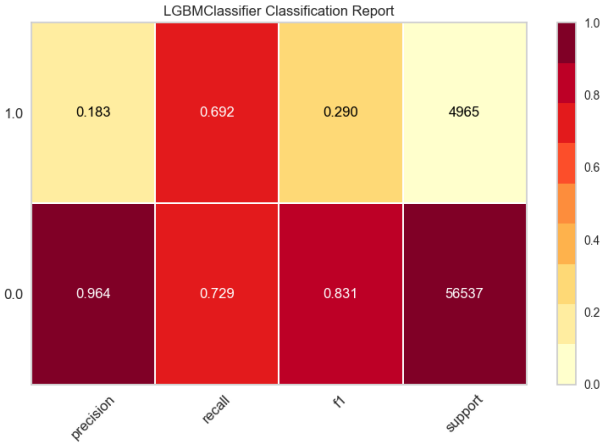
Métriques générales

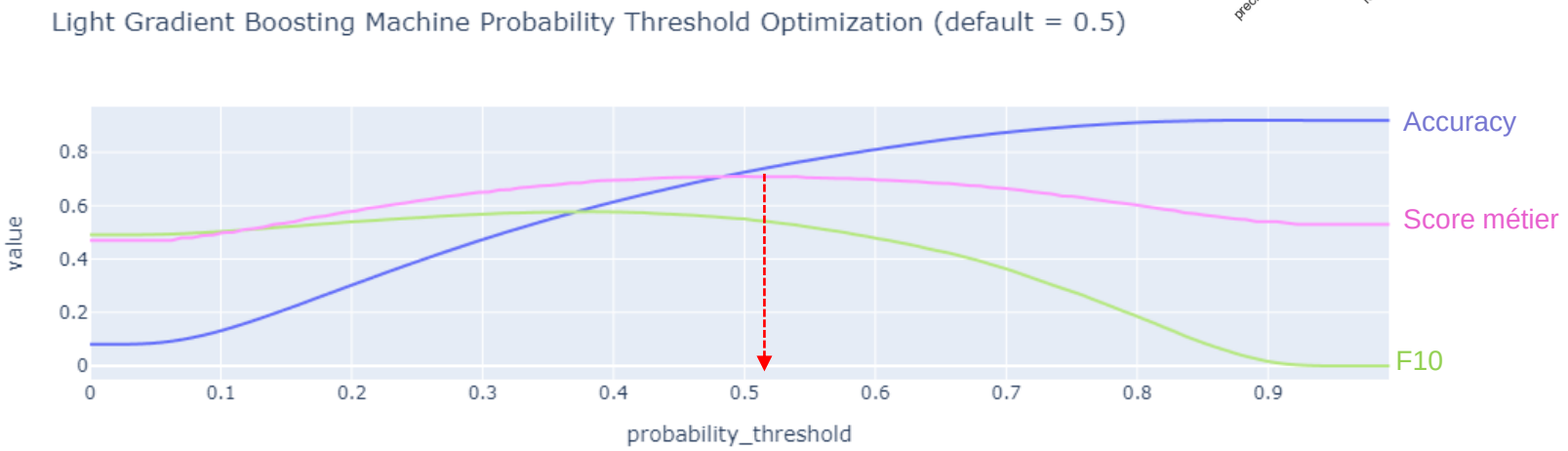
F-bêta : on pénalise les faux négatifs d'un facteur 10 (bêta = 3.16)

Score métier : $1 - \frac{FP+10FN}{N+10P}$

2. Modélisation - optimisation

Optimisation des hyperparamètres:
Optuna ou hyperopt (bayesienne)
Meilleur modèle:
optimisation avec optuna sur la base du score métier
Optimisation du seuil de prédiction
Seuil optimal : 0.539





	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F10	Score Métier
0	Light Gradient Boosting Machine	0.7615	0.7808	0.6405	0.1980	0.3024	0.2043	0.2559	0.5322	0.7100

2. Modélisation - MLflow

Prêt à dépenser

Search public repositories... Issues Pull Requests Resources Explore Pricing Product

sefirotha / OC-DS-P7_mlflow

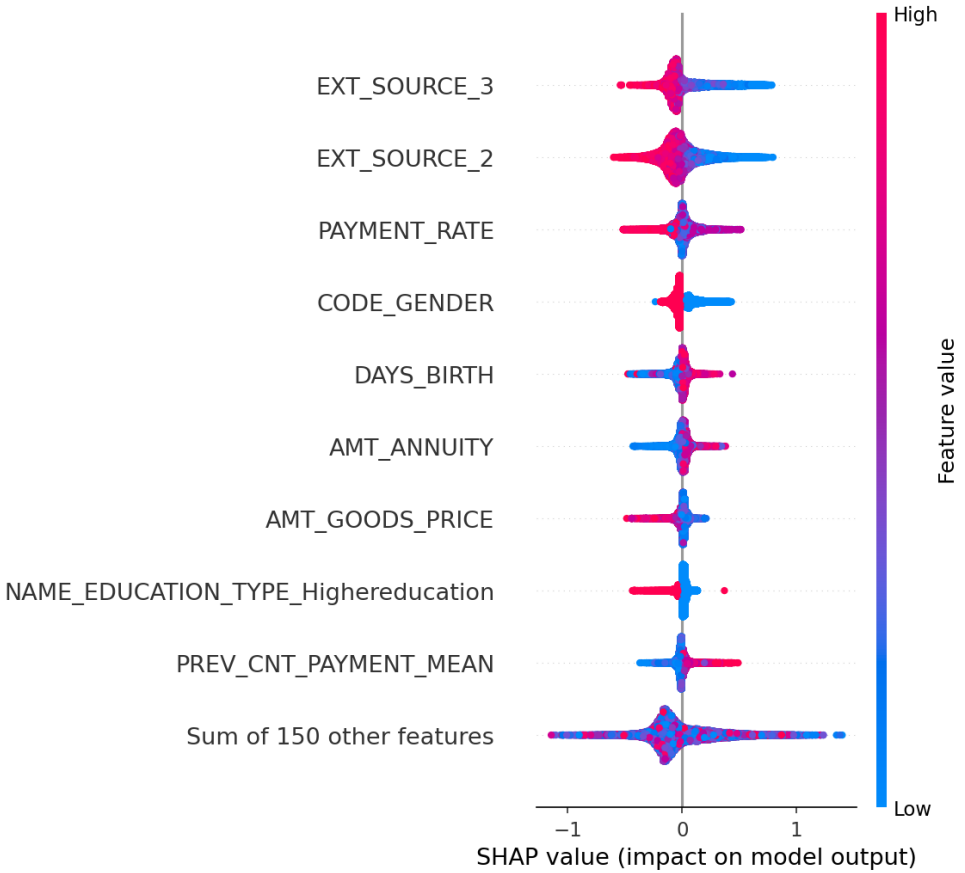
Files | Experiments 58 | Collaboration | Annotations

Compare | Reset filters | Delete | Archive | Labels | Columns

Code	Name	Cre...	Labels	Sou...	Group	boosting_type	class_weight	colsample_b...	AUC	Accuracy	F1
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.776	0.7511	0.2955
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7802	0.7989	0.3145
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7802	0.7989	0.3145
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7727	0.7339	0.2875
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7611	0.7069	0.2722
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7782	0.7255	0.2885
	Session Initialized	7 days ago		mlflow	LGBM opt...						
	bemused-crab-803	7 days ago		mlflow	Default						
	Light Gradient Boo	7 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7549	0.7162	0.272
	Light Gradient Boo	8 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7782	0.7255	0.2885
	Session Initialized	8 days ago		mlflow	LGBM opt...						
	traveling-seal-962	8 days ago		mlflow	Default						
	Light Gradient Boo	8 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7742	0.7119	0.2812
	Light Gradient Boo	8 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7782	0.7255	0.2885
	Session Initialized	8 days ago		mlflow	LGBM opt...						
	fearless-colt-213	8 days ago		mlflow	Default						
	Light Gradient Boo	8 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7769	0.7765	0.3048
	Light Gradient Boo	8 days ago		mlflow	LGBM opt...	gbdt	balanced	1.0	0.7752	0.7452	0.2929

DagsHub https://dagshub.com/sefirotha/OC-DS-P7_mlflow/experiments/#/

2. Modélisation - SHAP



2. Modélisation - Data Drift

Prêt à dépenser

Dataset Drift

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

120	9	0.075
Columns	Drifted Columns	Share of Drifted Columns

Data Drift Summary

Drift is detected for 7.5% of columns (9 out of 120).

Search

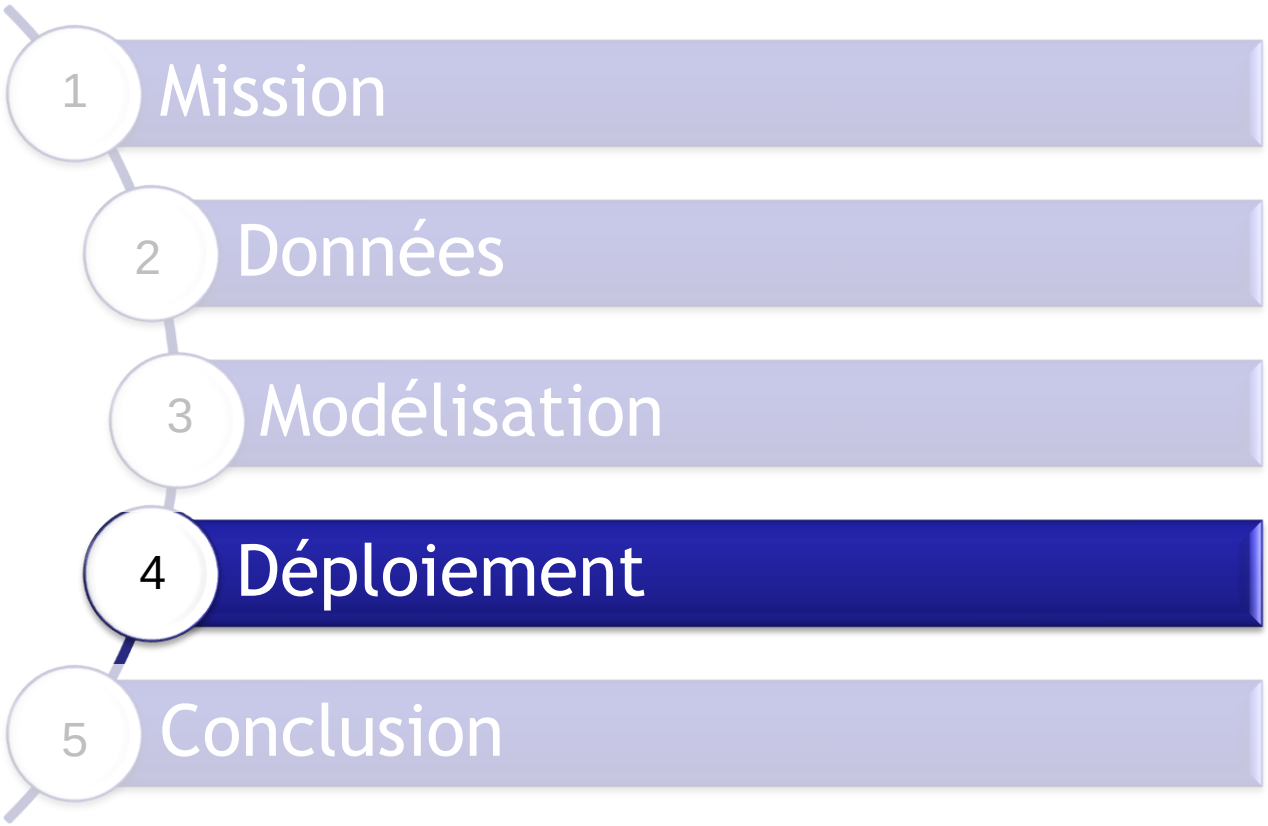
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.359052
AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	0.281765
AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.210785
AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.207334
AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.161102
AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.15426

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
NAME_CONTRACT_TYPE	cat			Detected	Jensen-Shannon distance	0.14755
DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.138977
FLAG_EMAIL	num			Detected	Jensen-Shannon distance	0.122121
FLAG_DOCUMENT_3	num			Not Detected	Jensen-Shannon distance	0.062496

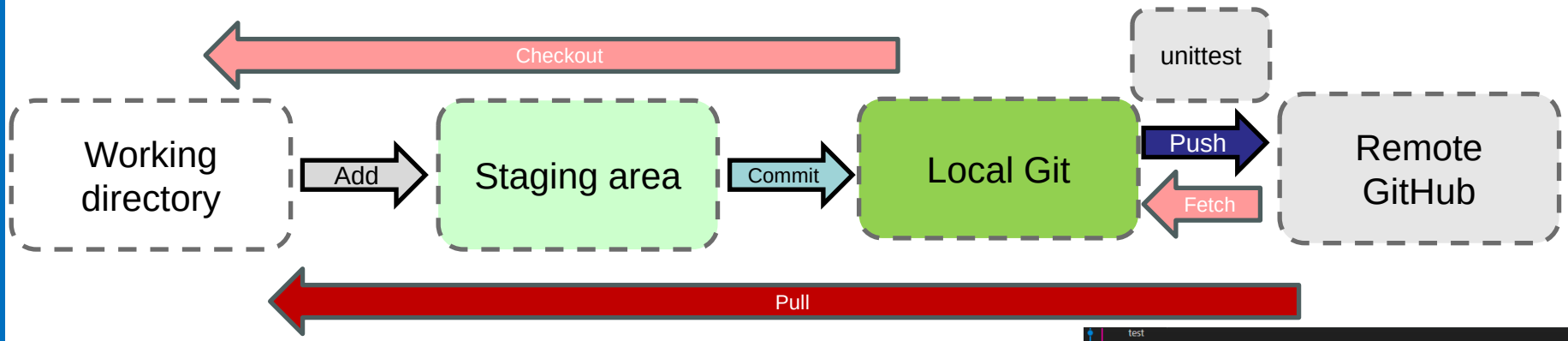
10 rows 1–10 of 120

data_drift_analysis.html

Sommaire


1
Mission
2
Données
3
Modélisation
4
Déploiement
5
Conclusion

3. Développement workflow



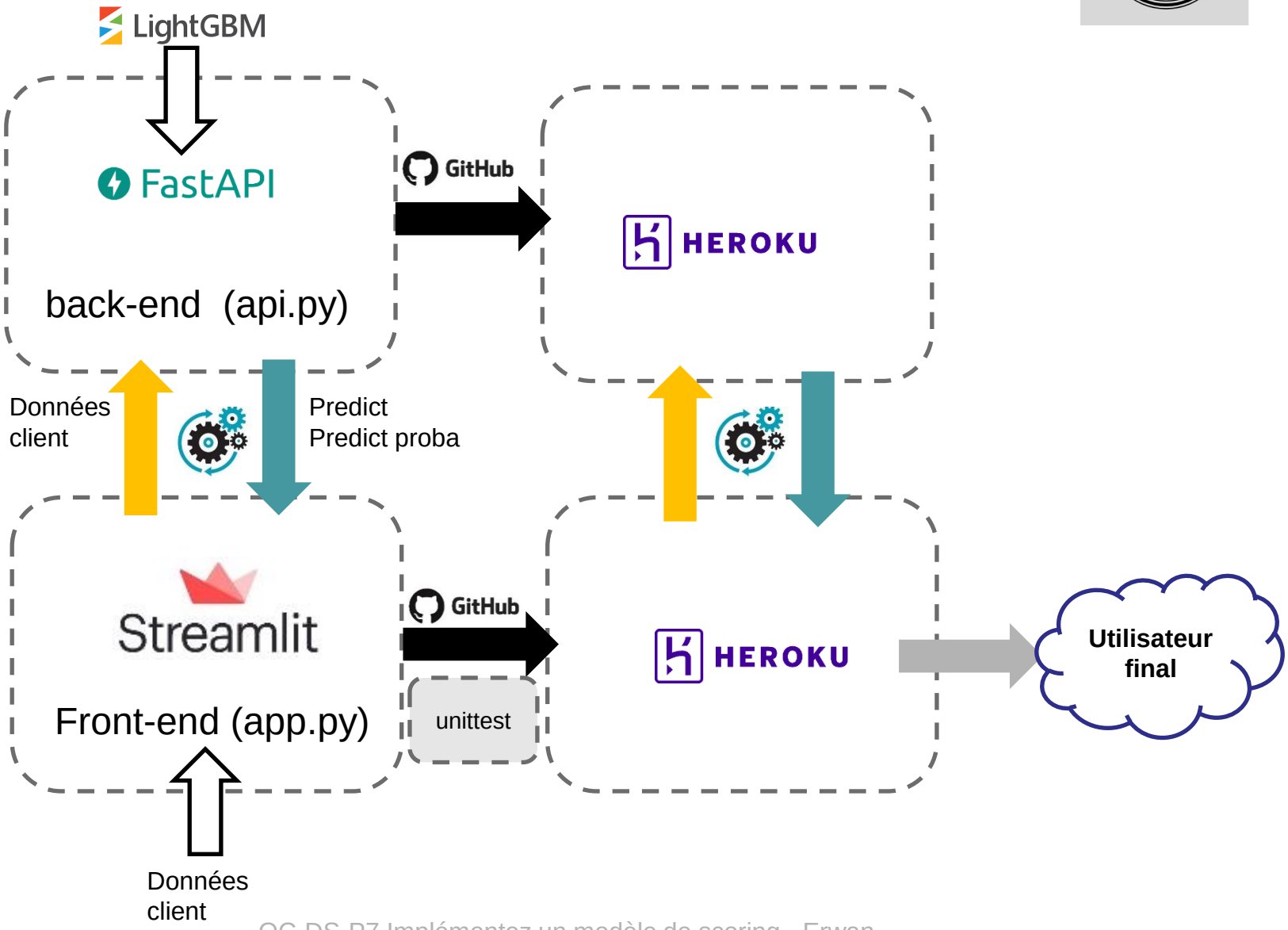

```
(base) PS C:\Users\eberthaud\Documents\GitHub\OC-DS-P7> pytest -s
============================= test session starts =============================
platform win32 -- Python 3.10.9, pytest-7.1.2, pluggy-1.0.0
rootdir: C:\Users\eberthaud\Documents\GitHub\OC-DS-P7
plugins: anyio-3.5.0
collected 9 items

test_app.py Test: Asserting existance of columns_dict.pkl
Test: Asserting existance of 230616_shap_values.pkl
Test: Asserting existance of test_df.pkl
Test: Asserting existance of application_test.pkl
Test: Asserting columns name in  of application_test.pkl
Test: Asserting columns name in  of test_df.pkl
Test: Asserting nb of columns in  of application_test.pkl
Test: Asserting nb of columns in  of test_df.pkl
Test: Asserting nb of columns in  of 230616_shap_values.pkl
============================== 9 passed in 10.80s ==============================
```

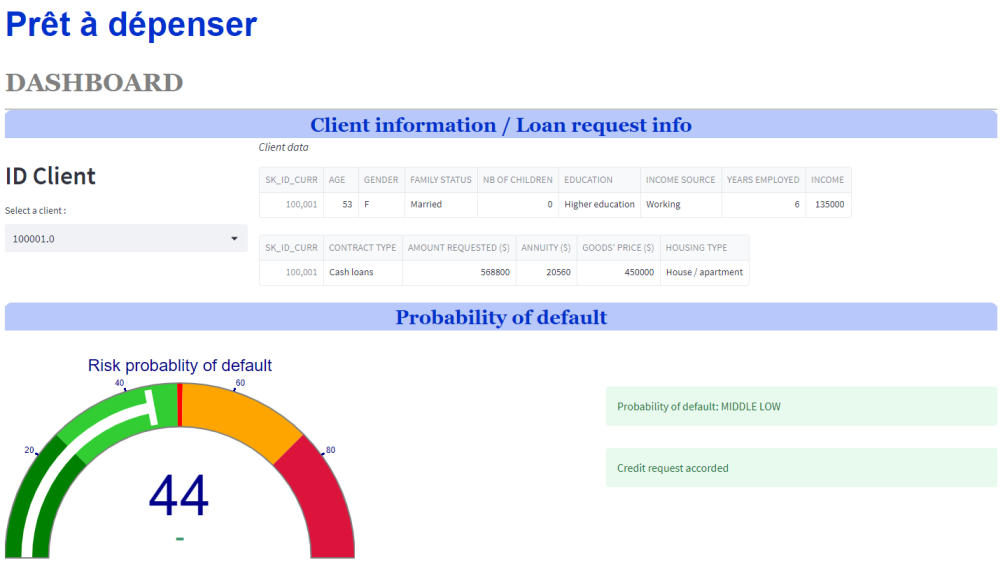
```
test
adding pkl compression
code error
correction
shap install
Adding pandas install
Adding api and shap analysis
adding workflow
adding heroku deployment file
Merge branch 'main' of https://github.com/sefirotha/OC-DS-P7
Adding data
App data
Folder deleted
final version - correction of pydantic models name
correction pydantic models
adding correct version
add index
update code
Update api
change func
updated procfile
updated procfile
change model name
updated procfile
updated requirements
updated procfile
delete port and host
Updated requirements
Updated procfile
Add Procfile
remove procfile
Add runtime
add procfile
change dir
test
Ajout d'un commentaire de controle
first commit
Ajout graphiques comparaison
```

GitHub https://github.com/sefirotha/OC-DS-P7

3. Déploiement - workflow

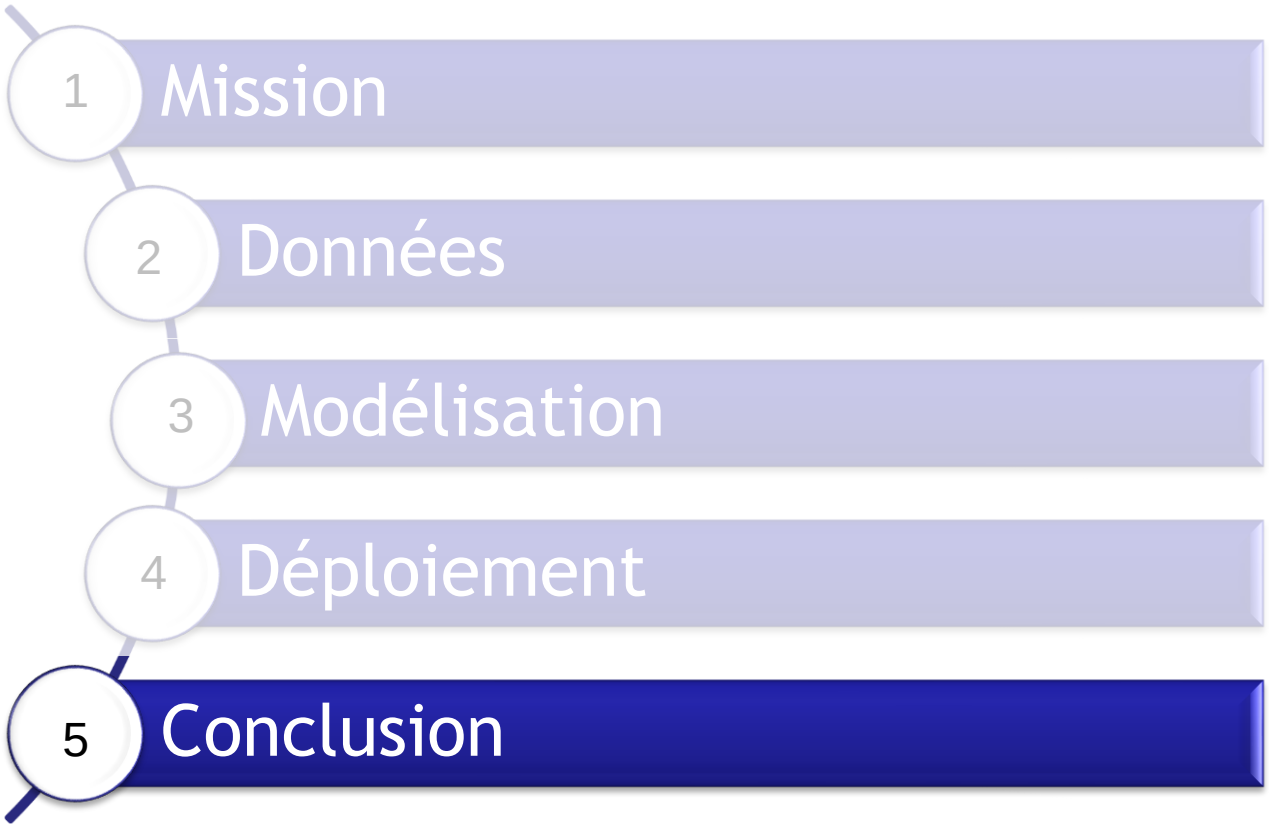


3. Déploiement - workflow



https://credict-score-eb-e593e2243d0a.herokuapp.com/

Sommaire


1
Mission
2
Données
3
Modélisation
4
Déploiement
5
Conclusion

4. Conclusion

Prêt à dépenser

1. **Problématique de classification binaire avec classes déséquilibrées**
2. **Modèle LGBM optimisé sur la base du score métier**

- **Optimisation du dashboard : rapidité, charte graphique**
- **Mise en place d'un suivi pour plan de maintenance**
- **Affiner la métrique métier avec des experts**
- **Questionner le choix de certaines variables (sexe, éthique?)**

Des questions ?