

Modèle de scoring crédit



Prêt à dépenser

Modèle de scoring crédit

Sommaire

- Rappel du contexte et des objectifs
- Présentation de la modélisation
- Présentation du pipeline de déploiement
- Présentation de l'analyse de data drift
- Présentation du dashboard
- Conclusion



Rappel du contexte et des objectifs

Sommaire

- Rappel du contexte et des objectifs
 - Objectif & méthode
 - Overview des données
- Rappel du contexte et des objectifs
- Présentation de la modélisation
- Présentation du pipeline de déploiement
- Présentation de l'analyse de data drift
- Présentation du dashboard
- Conclusion



Rappel du contexte et des objectifs

Objectif & Méthode

- **Objectifs :**

- Décision de prêt à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt
- Besoin de transparence et de gestion des données personnelles (RGPD)

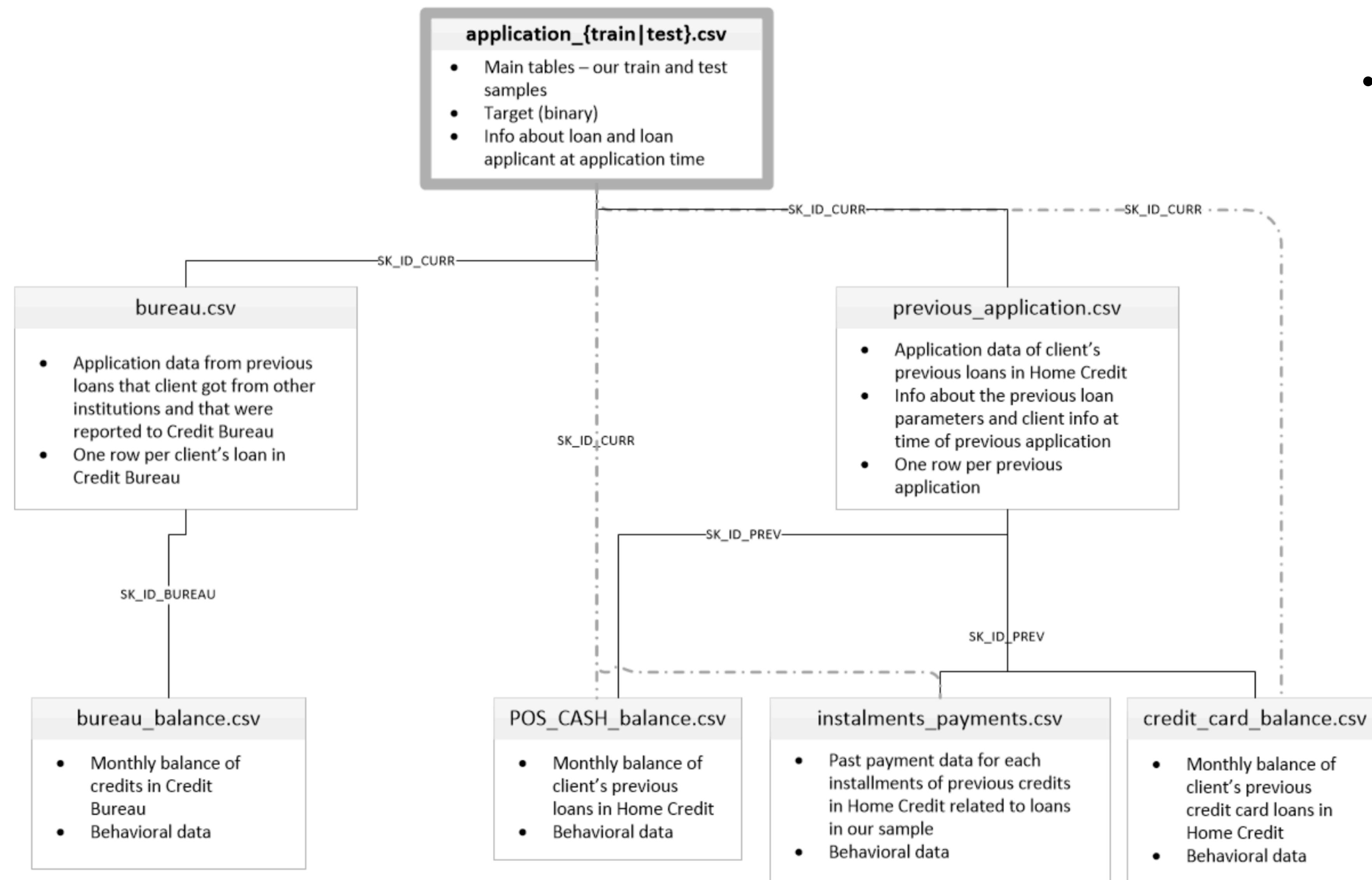
- **Méthode :**

- Outil de « scoring crédit » pour calculer la probabilité de remboursement de crédit puis classification de la demande
- Dashboard interactif



Rappel du contexte et des objectifs

Overview de données



- **307 511 + 48 744 demandes:**
- *Target*
- *Données personnelles*
- *Information de la demande*
- *Anciennes demandes prêt à dépenser*
- *Données d'autres institutions financières*



Présentation de la modélisation

Sommaire

- Rappel du contexte et des objectifs
- **Présentation de la modélisation**
 - Overview de la qualité des données
 - Overview de l'analyse exploratoire
 - Etape des modélisation
 - Architecture MLOps
 - Suivi MLFlow
 - Résultat des modèles
 - Analyse du modèle sélectionné
- Présentation du pipeline de déploiement
- Présentation de l'analyse de data drift
- Présentation du dashboard
- Conclusion



Rappel du contexte et des objectifs

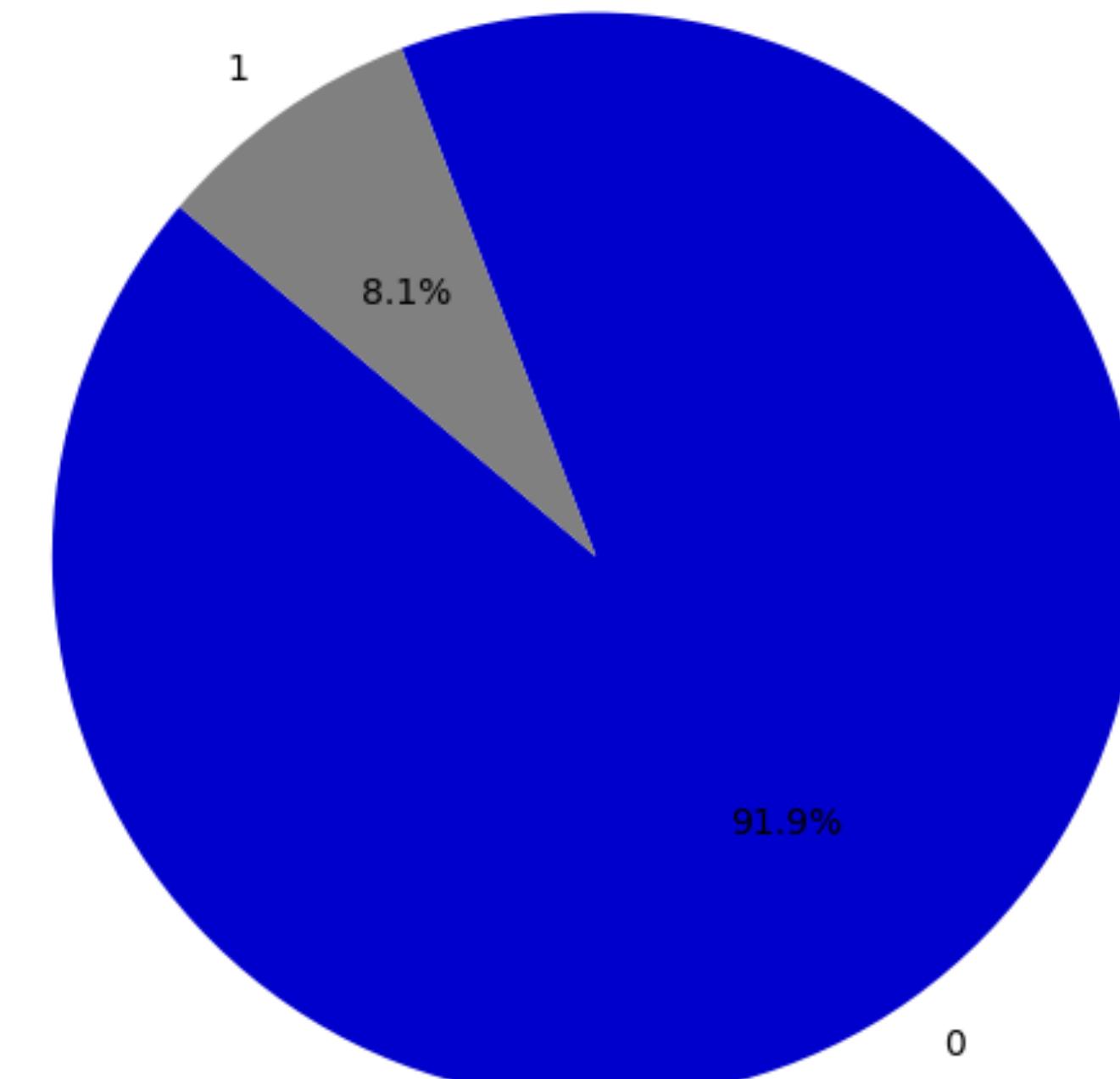
Overview de la qualité des données

QUALITE DES DONNEES :

- *Doublons* : 0 doublons
- *Erreurs de type* : 0 erreurs de type
- *Outliers* : gestion de certaines classes (exemple : ‘XNA’ de ‘CODE_GENDER’ mais axe d'amélioration)
- *Valeurs manquantes* : variable (jusqu'à 70%), traitement des valeurs manquantes en fonction des modèles et des variables sélectionnées

DESEQUILIBRE DES CLASSES

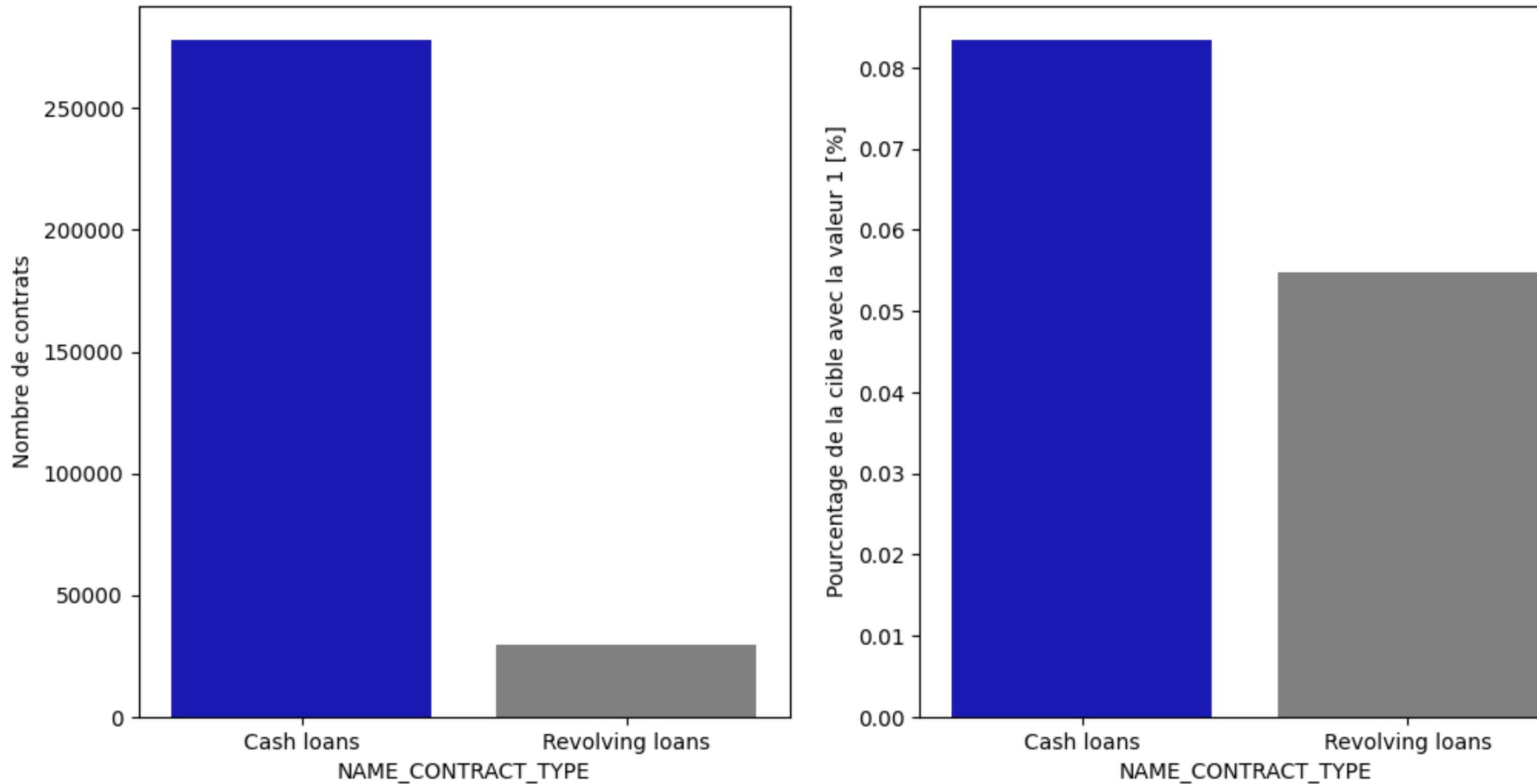
Taux de remboursement des prêts de l'application - ensemble de données d'entraînement



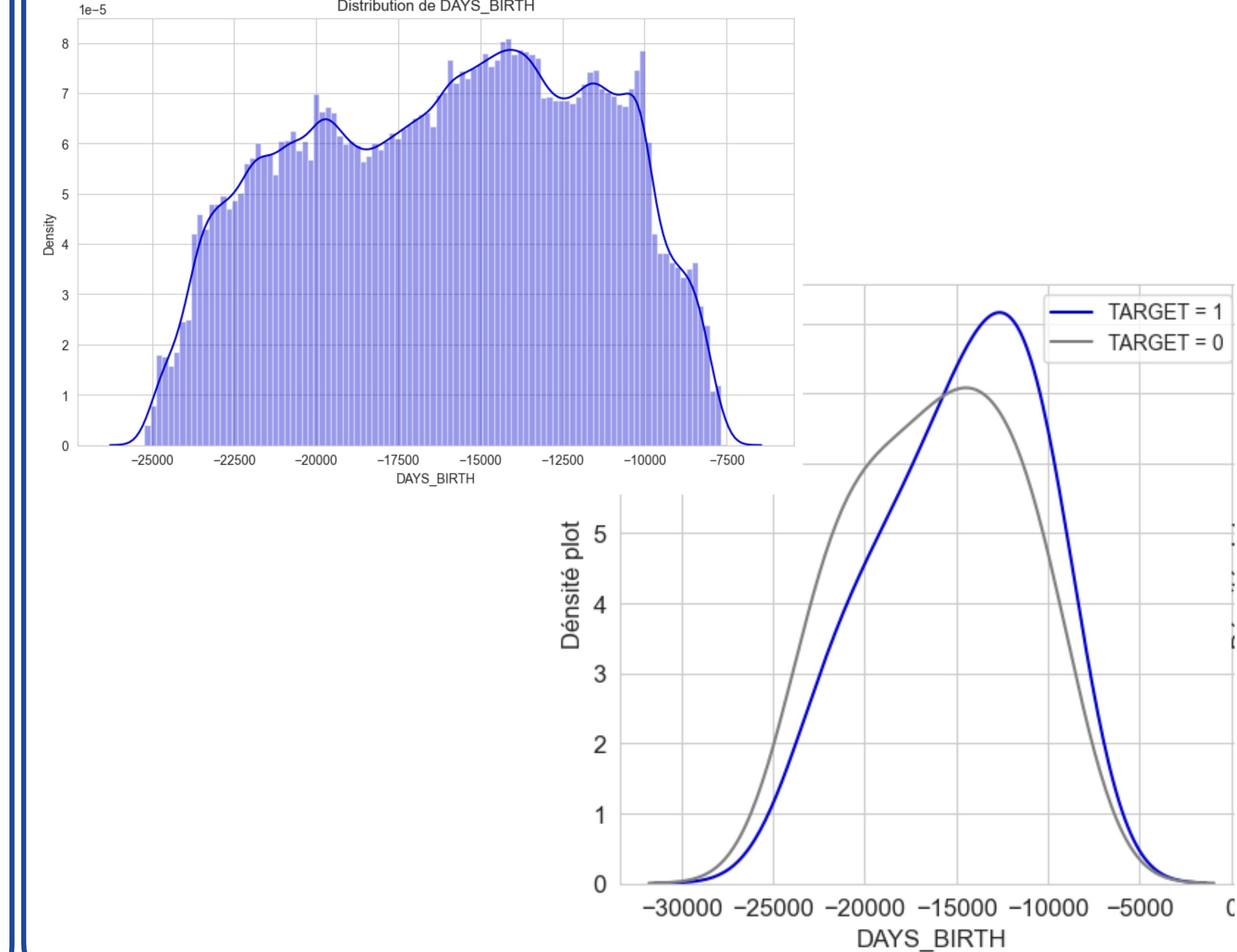
Rappel du contexte et des objectifs

Overview de l'analyse exploratoire

Analyse distribution - Données catégorielles

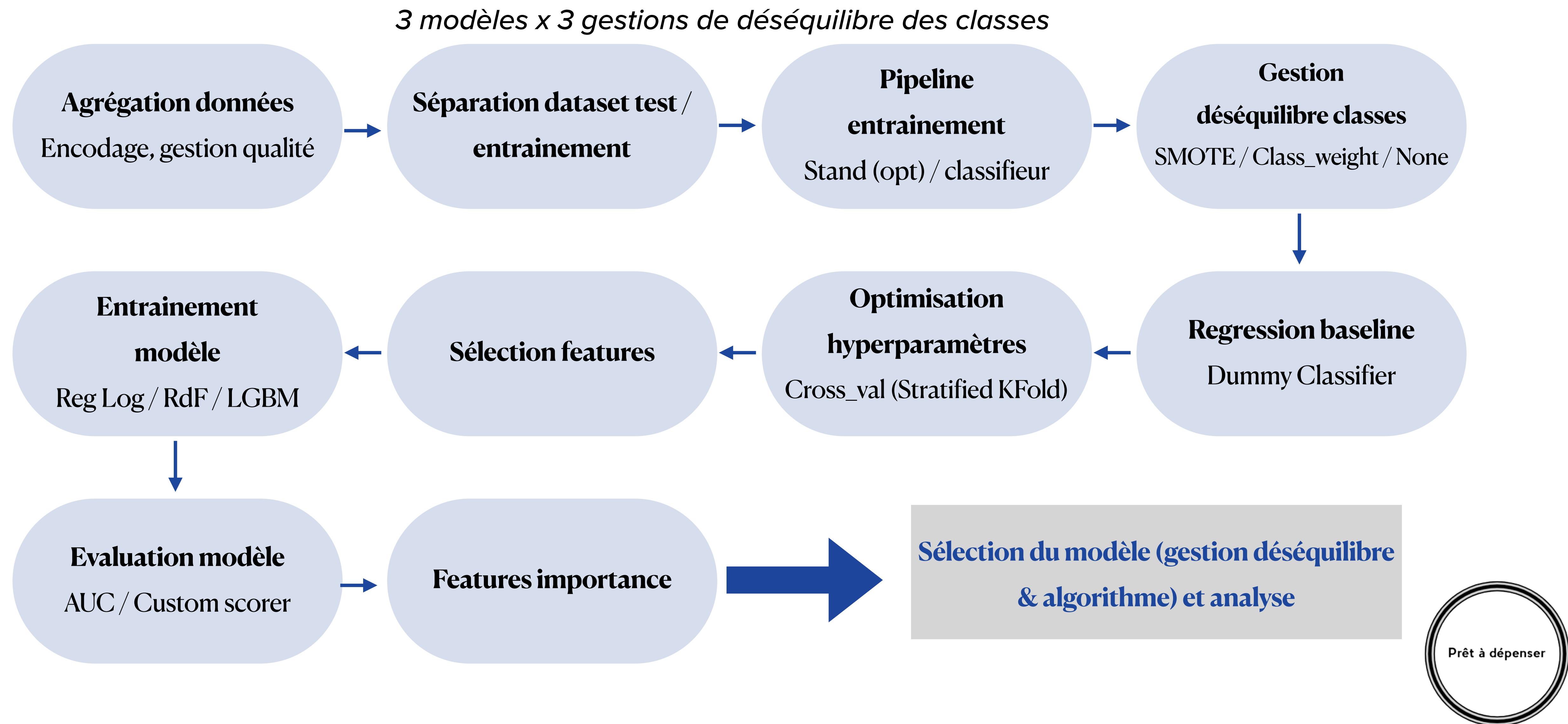


Analyse distribution - Données numériques



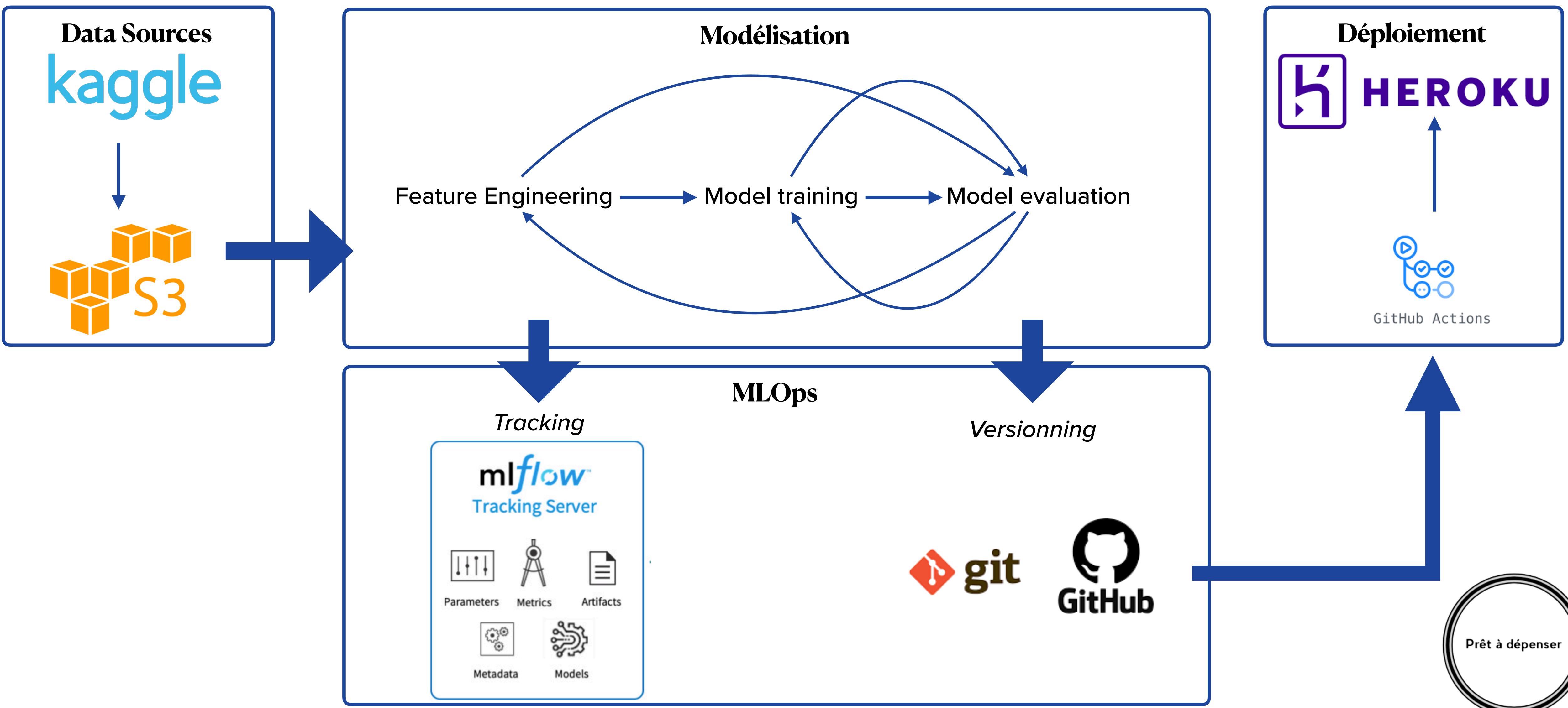
Présentation de la modélisation

Etapes de modélisation



Présentation de la modélisation

Architecture MLOps



Présentation de la modélisation

Suivi MLFlow

Tracking

AUC

Comparing first 10 runs

Run	AUC
1	0.49
2	0.71
3	0.71
4	0.68
5	0.51
6	0.51
7	0.52
8	0.63
9	0.64
10	0.66

Score Personnalisé

Comparing first 10 runs

Run	Score Personnalisé
1	0.83
2	0.50
3	0.50
4	0.55
5	0.84
6	0.85
7	0.87
8	0.63
9	0.61
10	0.59

Model Register

Parameters (5)

Name	Value
Best param	{'classifier__learning_rate': 0.05, 'classifier__n_estimators': 200, 'classifier__num_leaves': 100}
Gestion déséquilibre des classes	None
Nombre de features sélectionnées	547
Threshold	0.09090909090909091
Type de modèle	LGBM

Metrics (5)

Tags

Artifacts

trained_model

- MLmodel
- conda.yaml
- model.pkl
- python_env.yaml
- requirements.txt

Full Path: file:///Users/beatricepin/Documents/2023/Data%20Science/Projet_7_Modele_API/mlruns/0/55cce32ec93a4a... [Register Model](#)

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. You can also [register it to the model registry](#) to version control

Model schema

No schema. See [MLflow docs](#) for how to include input

Make Predictions

Predict on a Spark DataFrame:

```
import mlflow
from pyspark.sql.functions import struct, col
logged_model = 'runs:/55cce32ec93a4adf8505a472d12bd9af/trained_model'
```

$$\text{score_personalized} = \frac{10 \times \text{fn} + \text{fp}}{\text{num_individuals}}$$



Présentation de la modélisation

Résultats des modèles

Modèle	Gestion déséquilibre des classes	AUC - Test	Accuracy - Test	Score personnalisé - Test	Temps d'entraînement (s)	Temps d'inférence (s)
Baseline	NAN	0,49	0,85	0,83	1	0,003
Regression Logistique	None	0,51	0,48	0,85	17	0,2
	Class_weight	0,52	0,24	0,87	25	0,2
	SMOTE	0,51	0,44	0,85	45	0,8
Random Forest	None	0,63	0,67	0,63	1975	1,0
	Class_weight	0,66	0,68	0,59	2391	0,8
	SMOTE	0,64	0,65	0,61	4151	3,8
Light GBM	None	0,71	0,76	0,50	3757	5,0
	Class_weight	0,71	0,74	0,50	2909	8,9
	SMOTE	0,68	0,72	0,54	1272	1,7

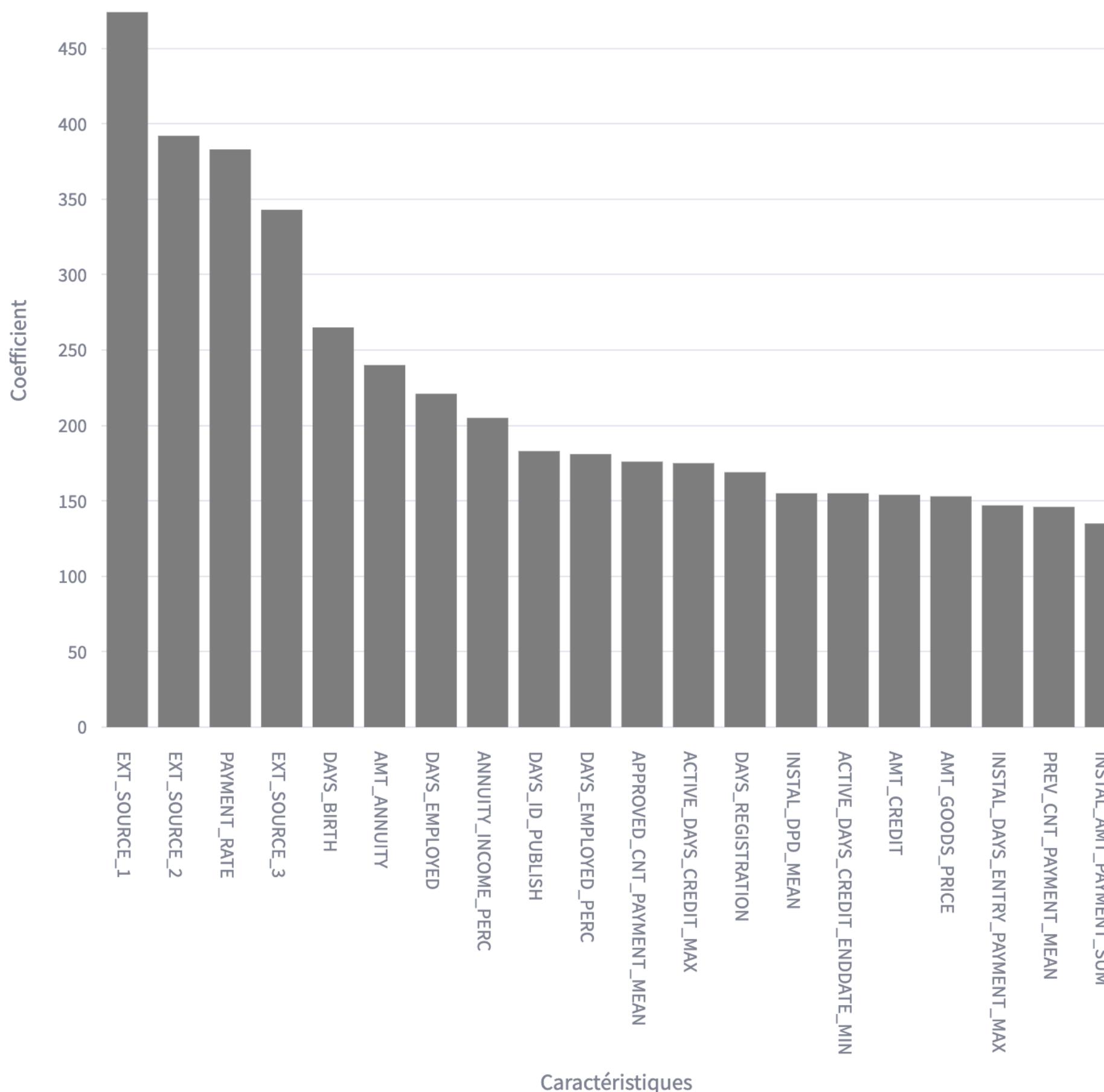


Présentation de la modélisation

Analyse du modèle sélectionné

LGBM & Pas de gestion déséquilibre de classes

Top 20 des caractéristiques les plus importantes:



Résultats :

Baseline - DummyClassifier :
AUC : 0,49
Custom score : 0,83

LGBM :
Meilleurs hyperparamètres :
{'classifier__learning_rate': 0.05,
'classifier__n_estimators': 200,
'classifier__num_leaves': 100}

Threshold : 0.09

Nb features : 547

Score de test :
AUC : 0,71
Custom score : 0,50



Présentation du pipeline de déploiement

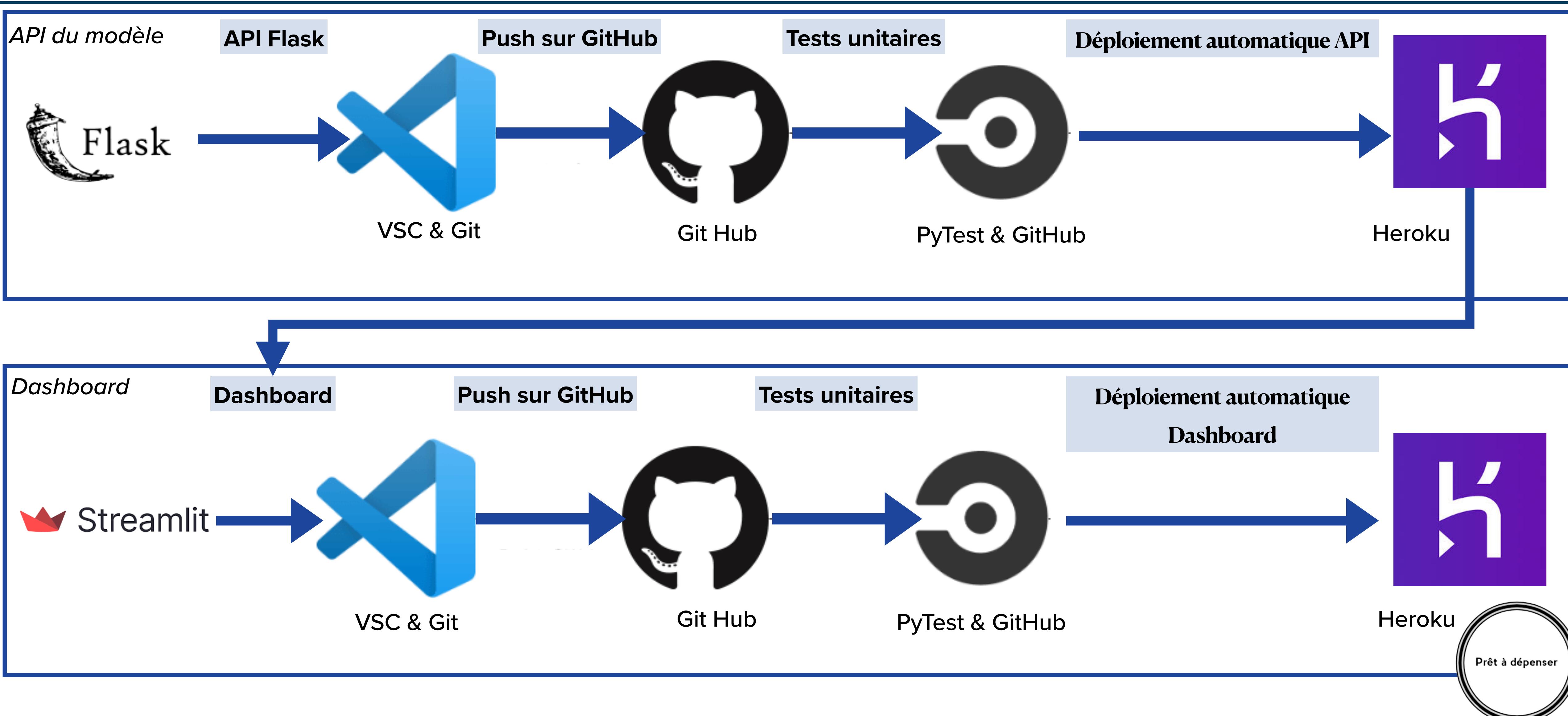
Sommaire

- Rappel du contexte et des objectifs
- Présentation de la modélisation
- Présentation du pipeline de déploiement
 - Architecture de déploiement
 - GitHub & Tests unitaires
 - Heroku
- Présentation de l'analyse de data drift
- Présentation du dashboard
- Conclusion



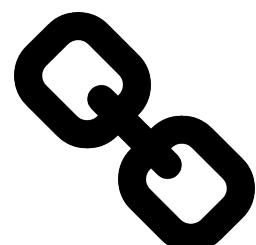
Présentation du pipeline de déploiement

Architecture de déploiement



Présentation du pipeline de déploiement

GitHub & Tests unitaires



Dépots GitHub :

GitHub Modèle & API : <https://github.com/Emeline2104/Projet-7-Models-API>

GitHub Dashboard : <https://github.com/Emeline2104/Projet-7-Dashboard>

1

Commit

Commits 35 Files changed 2

Commits on Dec 13, 2023

- Merge pull request #1 from Emeline2104/models ... Verified 2712985
- Create node.js.yml Verified 01372c2
- Create tests.yml Verified 9307817
- Merge pull request #2 from Emeline2104/models ... Verified c3907ed

2

Push sur le Main - Branch GitHub

Add more commits by pushing to the [dashboard](#) branch on [Emeline2104/Projet-7-Dashboard](#).

This branch has not been deployed
No deployments

Some checks haven't completed yet
1 in progress check

Run Tests and Deploy / test (pull_request) In progress — This check has started...
Details

This branch has no conflicts with the base branch
Merging can be performed automatically.

Merge pull request

You can also [open this in GitHub Desktop](#) or view [command line instructions](#).

3

Validation des TU - Merging

Require approval from specific reviewers before merging
Rulesets ensure specific people approve pull requests before they're merged.
Add rule

All checks have passed
1 successful check

This branch has no conflicts with the base branch
Merging can be performed automatically.

Merge pull request

You can also [open this in GitHub Desktop](#) or view [command line instructions](#).

Présentation du pipeline de déploiement

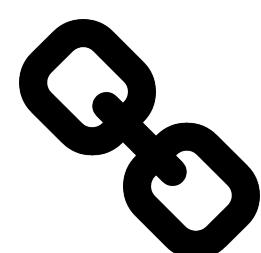
Heroku

4

Déploiement automatique sur Heroku

The screenshot shows the Heroku dashboard for the application 'projet-7-dashboard'. The top navigation bar includes the Heroku logo, a search bar, and links for 'Jump to Favorites, Apps, Pipelines'. Below the navigation, the application path 'Personal > projet-7-dashboard' is shown, along with a GitHub integration link for 'Emeline2104/Projet-7-Dashboard' and a 'main' button. A blue arrow points from the 'Activity' tab in the navigation bar to the deployment log below. The 'Activity Feed' section lists two recent events: a deployment by 'emeline.tapin@gmail.com' at 9:22 PM and a build success by the same user at 9:19 PM.

The screenshot shows a client information page for the application 'projet-7-dashboard-d395108bbc0c.herokuapp.com'. The title 'Recherche client' is visible. Below it are four items: 'Informations client' (with a bar chart icon), 'Informations crédit' (with a credit card icon), 'Informations comparaison' (with a line graph icon), and 'Informations modèle' (with a lightbulb icon). To the right of the page, a large text 'Bienvenue sur le tableau de' is displayed, followed by a partial view of a circular graphic.



Adresses url déployées :

Modèle & API : <https://projet-7-dashboard-d395108bbc0c.herokuapp.com/>

Dashboard : <https://projet-7-38cdf763d118.herokuapp.com/>

Prêt à dépenser

Présentation de l'analyse de data drift

Sommaire

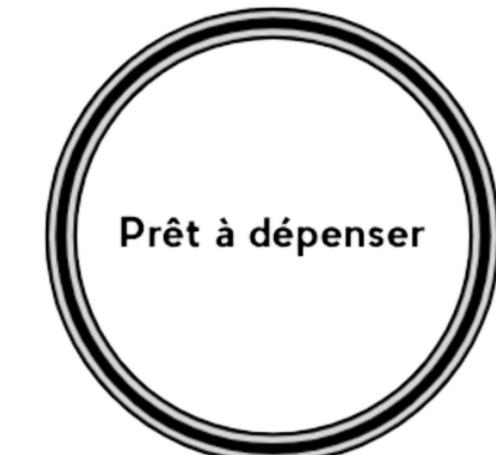
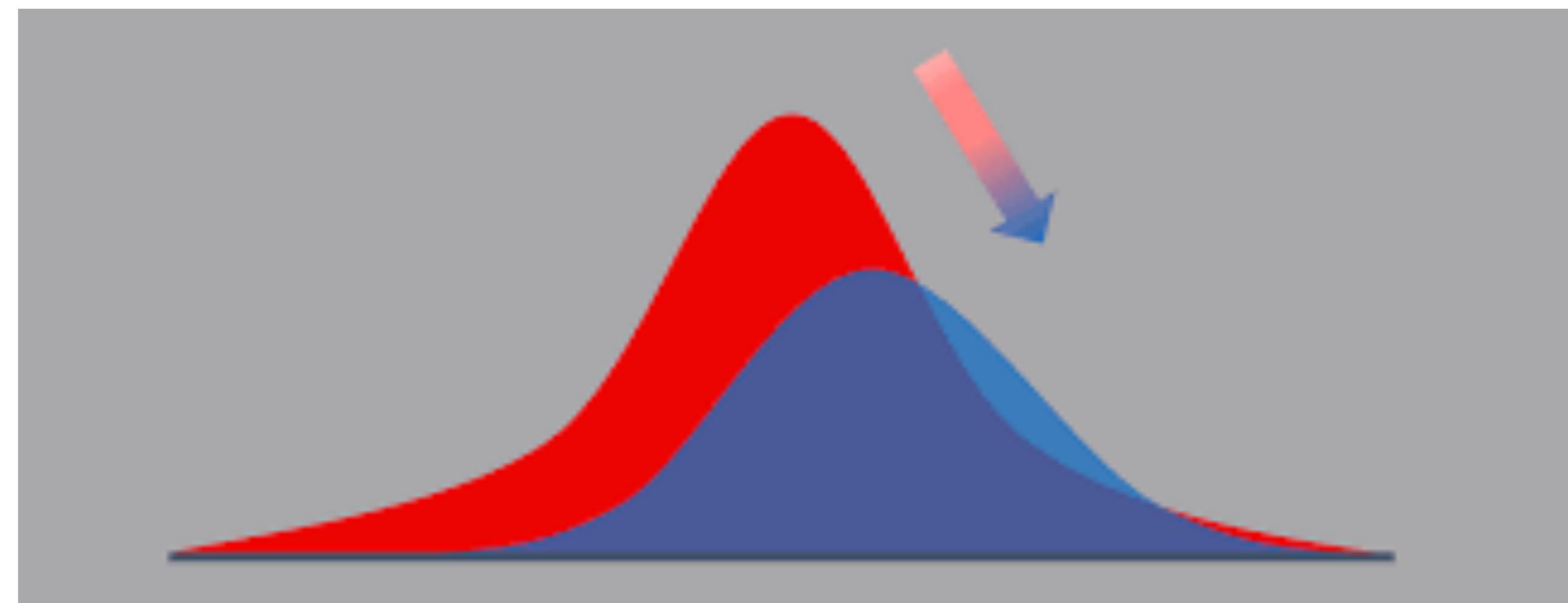
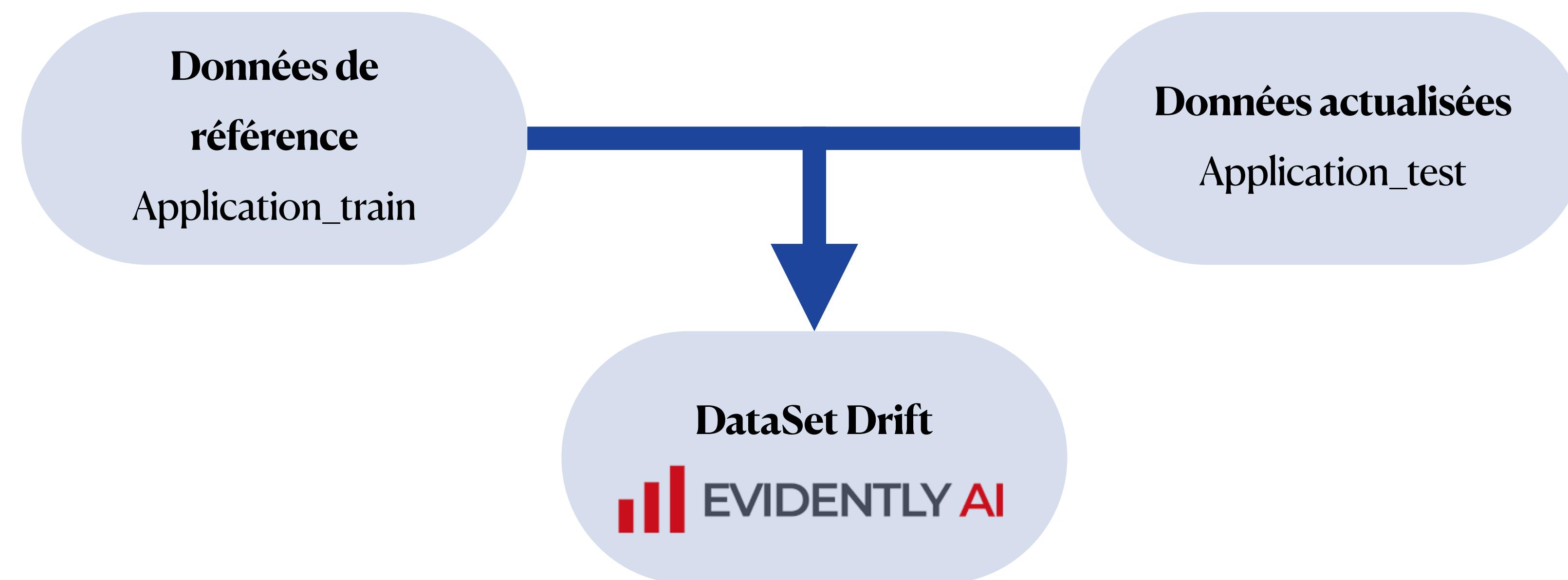
- Rappel du contexte et des objectifs
- Présentation de la modélisation
- Présentation du pipeline de déploiement
- **Présentation de l'analyse de data drift**
 - Méthode
 - Résultats
- Présentation du dashboard
- Conclusion



Prêt à dépenser

Présentation de l'analyse de data drift

Méthode



Présentation de l'analyse de data drift

Résultats (1/2)

Dataset Drift

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

121

Columns

11

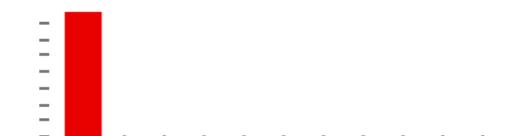
Drifted Columns

0.0909

Share of Drifted Columns

Data Drift Summary

Drift is detected for 9.091% of columns (11 out of 121).

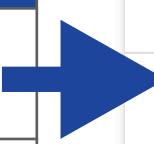
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	2.326826
> AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.584319
> AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.411233



Présentation de l'analyse de data drift

Résultats (2/2)

Caractéristiques avec une Forte Distorsion	
AMT_REQ_CREDIT_BUREAU_MON (Numerique)	Wasserstein distance (normed) : 2.33
AMT_REQ_CREDIT_BUREAU_WEEK (Numerique)	Wasserstein distance (normed) : 0.58
AMT_REQ_CREDIT_BUREAU_QRT (Numerique)	Wasserstein distance (normed) : 0.41
NAME_CONTRACT_TYPE (Catégorique)	Jensen-Shannon distance : 0.14
Caractéristiques avec une Distorsion Modérée	
AMT_GOODS_PRICE, AMT_CREDIT (Numeriques)	Wasserstein distances (normed) : 0.23
AMT_ANNUITY (Numerique)	Wasserstein distance (normed) : 0.15
DAYS_LAST_PHONE_CHANGE (Numerique)	Wasserstein distance (normed) : 0.13
FLAG_EMAIL (Numerique)	Jensen-Shannon distance : 0.12
AMT_REQ_CREDIT_BUREAU_DAY (Numerique)	Wasserstein distance (normed) : 0.11



Column: AMT_REQ_CREDIT_BUREAU_MON

Type: num

Reference Distribution: (red bar)

Current Distribution: (black bar)

Data Drift: Detected

Stat Test: Wasserstein distance (normed)

Drift Score: 2.326826

DATA DRIFT **DATA DISTRIBUTION**



abs perc

current reference

Actions à mettre en place :

- Surveiller de près les caractéristiques avec une distorsion significative.
- Adapter les modèles pour tenir compte des changements détectés.
- Réévaluer la qualité des prédictions et, si nécessaire, re-entrainer le modèle.
- Mettre en œuvre un processus de suivi continu du data drift pour assurer la robustesse du modèle sur le long terme.

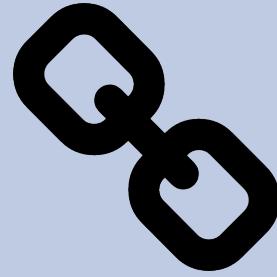
Présentation du dashboard

Sommaire

- Rappel du contexte et des objectifs
- Présentation de la modélisation
- Présentation du pipeline de déploiement
- Présentation de l'analyse de data drift
- **Présentation du dashboard**
 - Page d'accueil
 - Page client
 - Page crédit
 - Page comparaison
 - Page modèle
- Conclusion

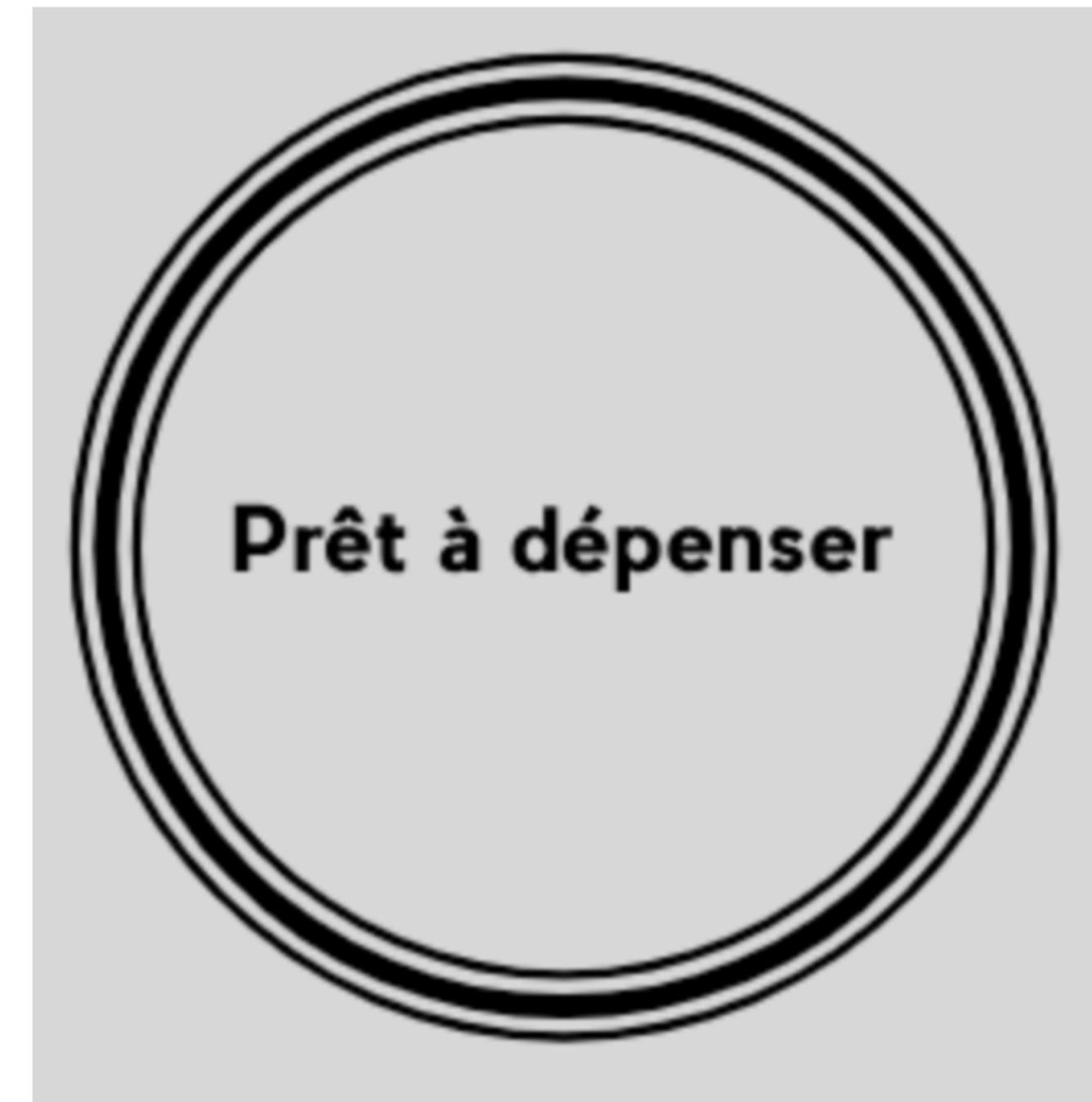
Présentation du dashboard

Page d'accueil



<https://projet-7-38cdf763d118.herokuapp.com/>

Bienvenue sur le tableau de bord



Bienvenue dans le tableau de bord interactif de Prêt à Dépenser. Les onglets sont disposés de la manière suivante :

- Informations client
- Informations sur la demande de crédit
- Informations de comparaison avec des groupes proches
- Informations sur le modèle

Entrez l'identifiant du client pour obtenir des informations:

ID du client:

100002

[Valider et accéder aux informations du client](#)

You pouvez accéder aux autres pages !



Présentation du dashboard

Page client



<https://projet-7-38cdf763d118.herokuapp.com/>

Page d'informations sur le client

ID du client : 100002

Informations personnelles:

Choisir une information:

Nombre d'enfants

Valeur : 0

Unité :

Description : *Aucune description disponible*

Informations concernant les données des institutions financières:

Choisir une information:

Statut Crédit

Valeur : *Closed*

Unité :

Description : *Statut des crédits signalés par le Bureau de crédit*

Informations sur les anciennes demandes:

Choisir une information:

Statut du Contrat

Valeur : *Approved*

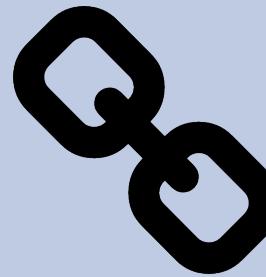
Unité :

Description : *Statut du contrat de la demande précédente (approuvé, annulé, ...)*



Présentation du dashboard

Page crédit (1/2)



<https://projet-7-38cdf763d118.herokuapp.com/>

Page d'informations crédit

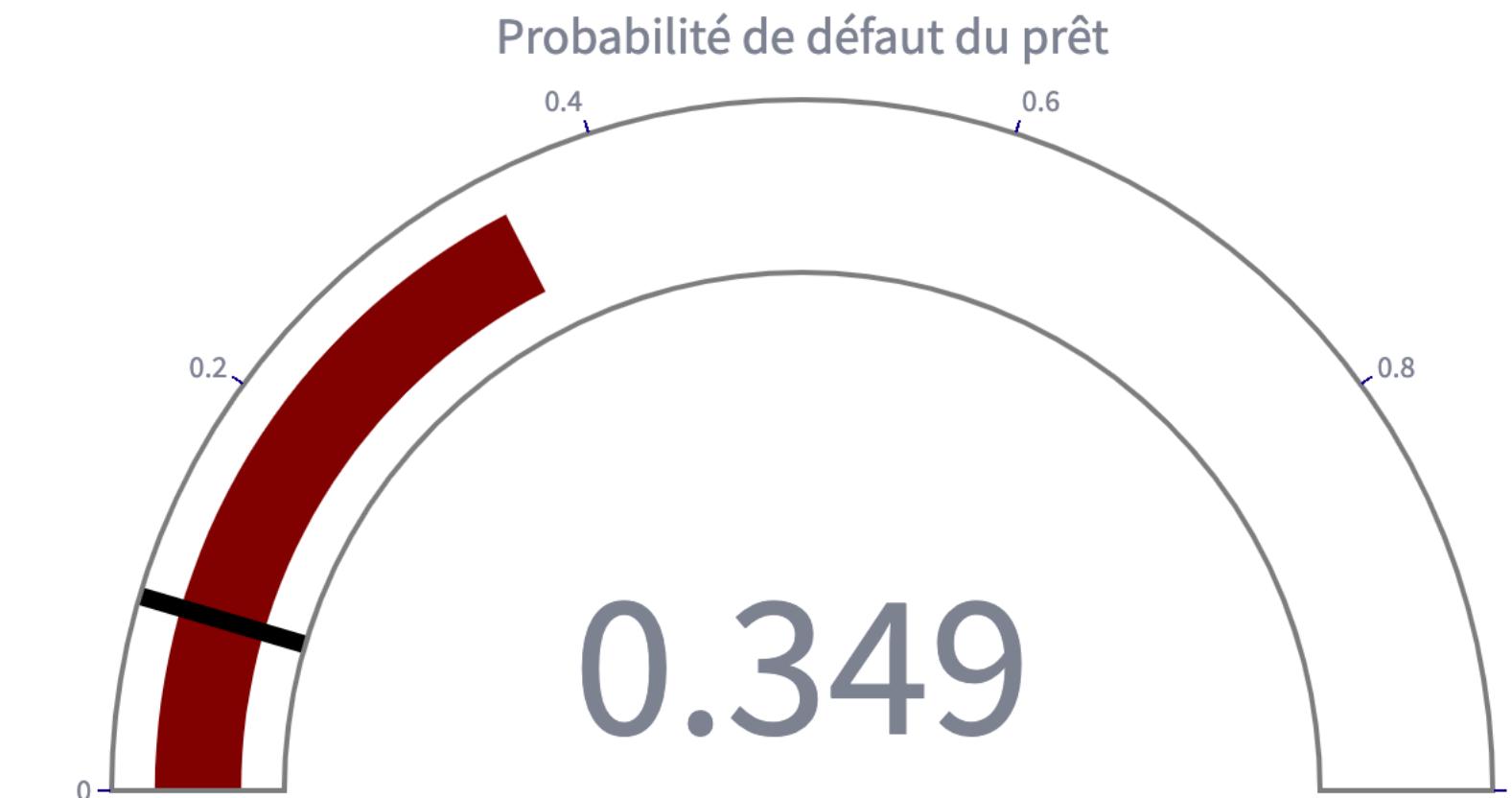
Prédictions pour le client:

Le client n'a pas obtenu son prêt.

Probabilité de défaut du prêt : 0.35

Seuil maximal : 0.09

Probabilité minimale au-delà duquel la demande de prêt est refusée.



Informations crédit:

Type de prêt: Cash loans

Montant du crédit demandé (€): 406597.5

Montant des annuités du crédit (€): 24700.5

Montant des biens pour lequel le crédit est octroyé (€): 351000.0

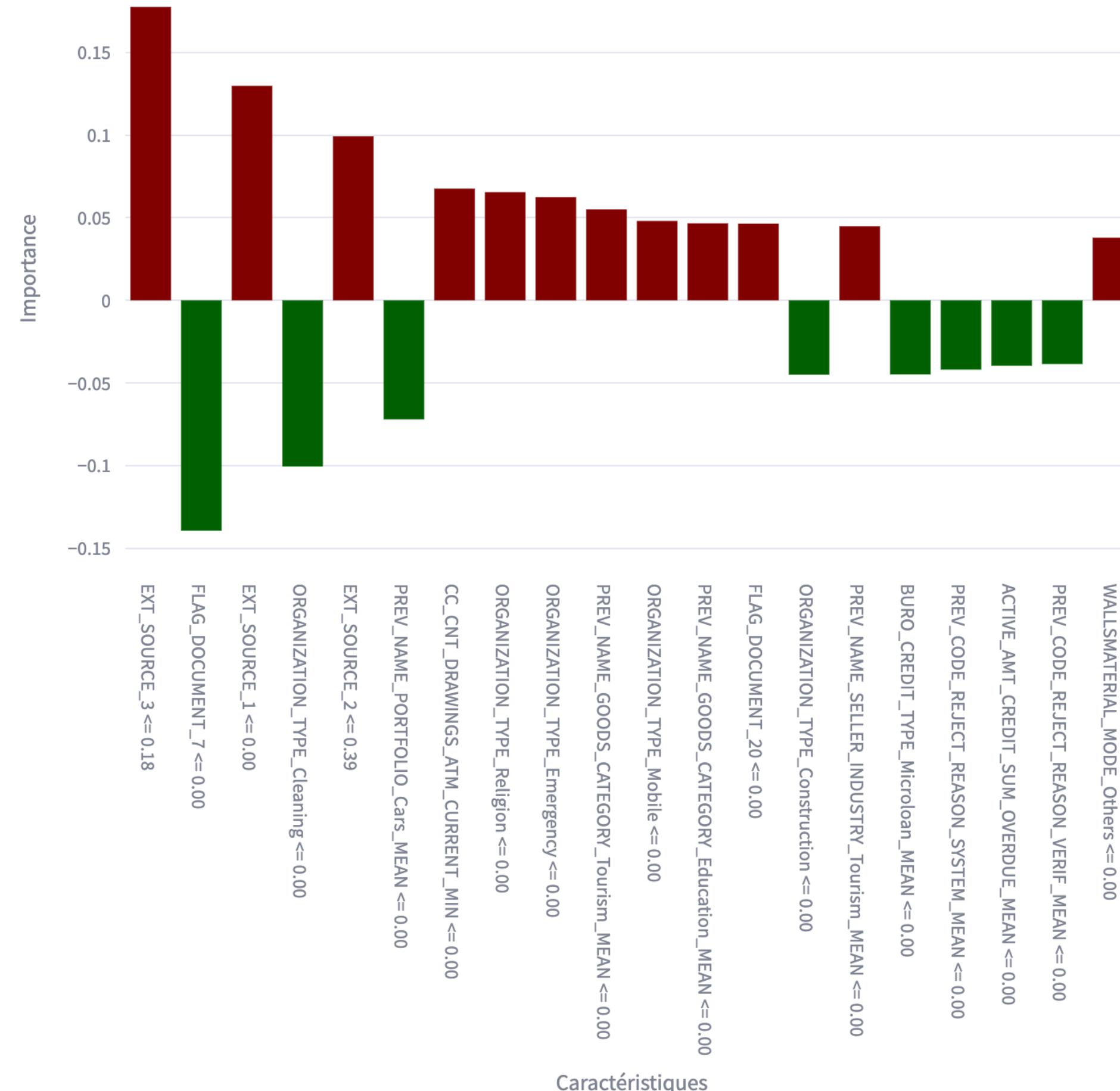


Présentation du dashboard

Page crédit (2/2)

Explication des caractéristiques relatives au choix de l'octroi du prêt:

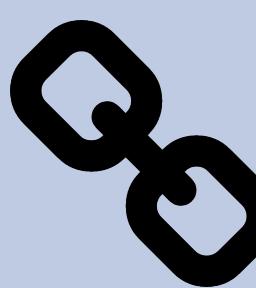
Top 20 des caractéristiques les plus importantes



Ce graphique explique la prédition de défaut de crédit concernant le client spécifié. Les caractéristiques qui ont le plus influencé la décision du modèle concernant la décision d'octroi du crédit.

- **Échelle de l'importance :** L'axe vertical représente l'importance de chaque caractéristique. Une barre plus haute signifie que cette caractéristique a une influence plus forte sur la décision du modèle.
- **Caractéristiques individuelles :** Chaque barre sur le graphique correspond à une caractéristique particulière du client, comme le revenu, l'âge, ou le montant du prêt.
- **Direction de l'impact :** La direction de la barre indique si la caractéristique a une influence positive ou négative sur la probabilité de refus de crédit. Une barre pointant vers le haut signifie une influence positive (rouge), et une barre pointant vers le bas signifie une influence négative (vert).
- **Interprétation des barre :** En analysant ces barres, vous pouvez déterminer quelles caractéristiques ont le plus contribué à la probabilité de refus de crédit. *Par exemple, si la barre la plus haute représente le revenu, cela signifie que le revenu a eu la plus grande influence sur la probabilité de refus.*

En résumé, ce graphique aide à comprendre pourquoi le modèle a pris la décision qu'il a prise pour ce client spécifique. Plus la barre est haute, plus la caractéristique est importante, et la direction de la barre indique si cette caractéristique a eu un impact positif ou négatif sur la probabilité de refus de crédit.

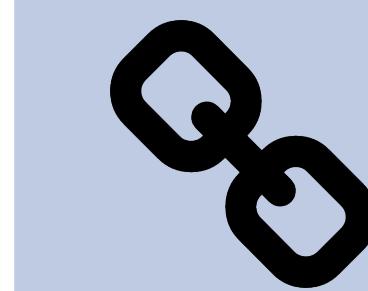


<https://projet-7-38cdf763d118.herokuapp.com/>



Présentation du dashboard

Page comparaison



<https://projet-7-38cdf763d118.herokuapp.com/>

Page de comparaison avec d'autres clients

Mode de comparaison:

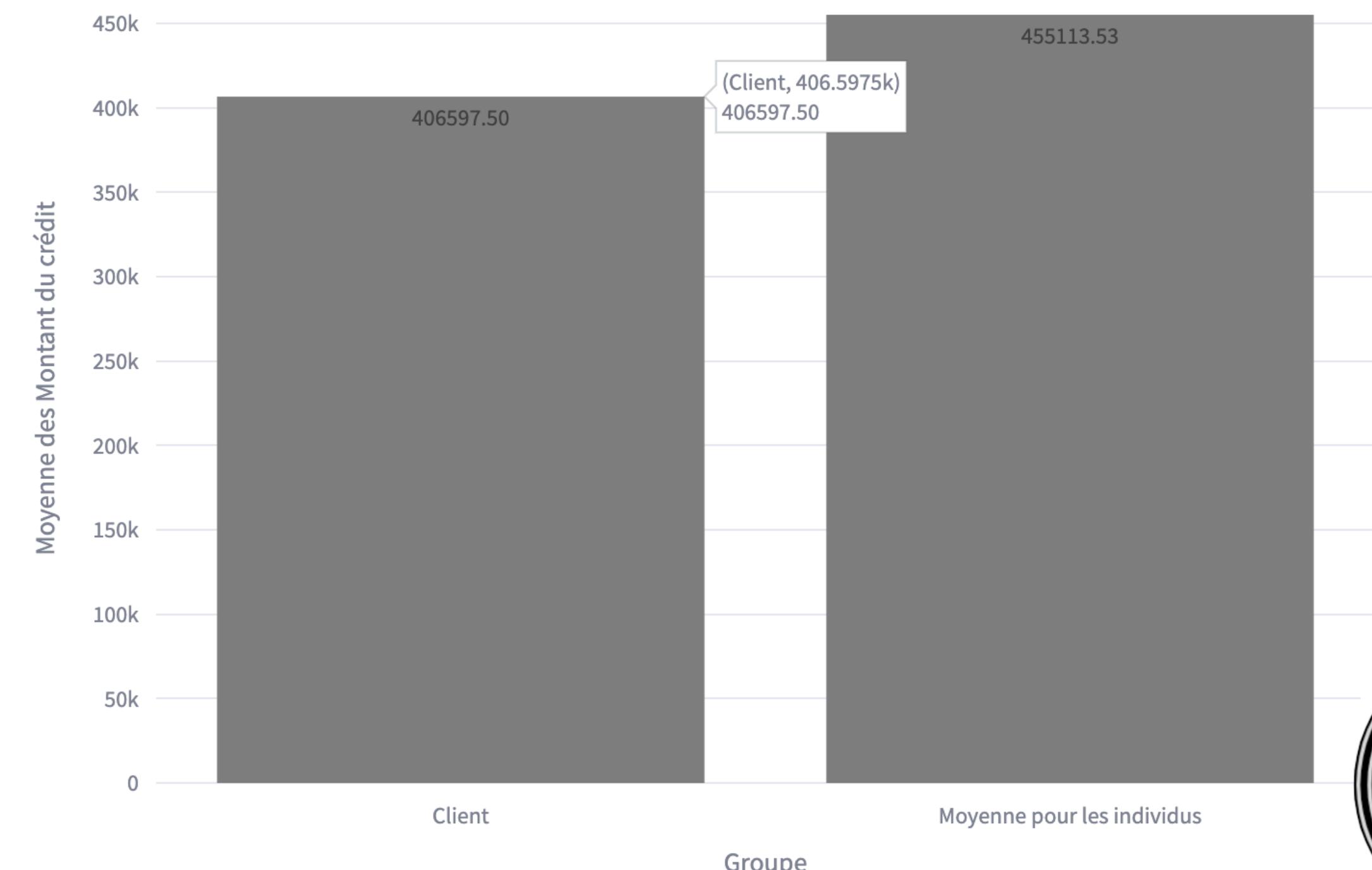
Par groupe d'âge

Donnée associée du client: Moins de 30 ans

Indicateur de comparaison:

Montant du crédit

Comparaison des Montant du crédit du client et des individus de comparaison



Présentation du dashboard

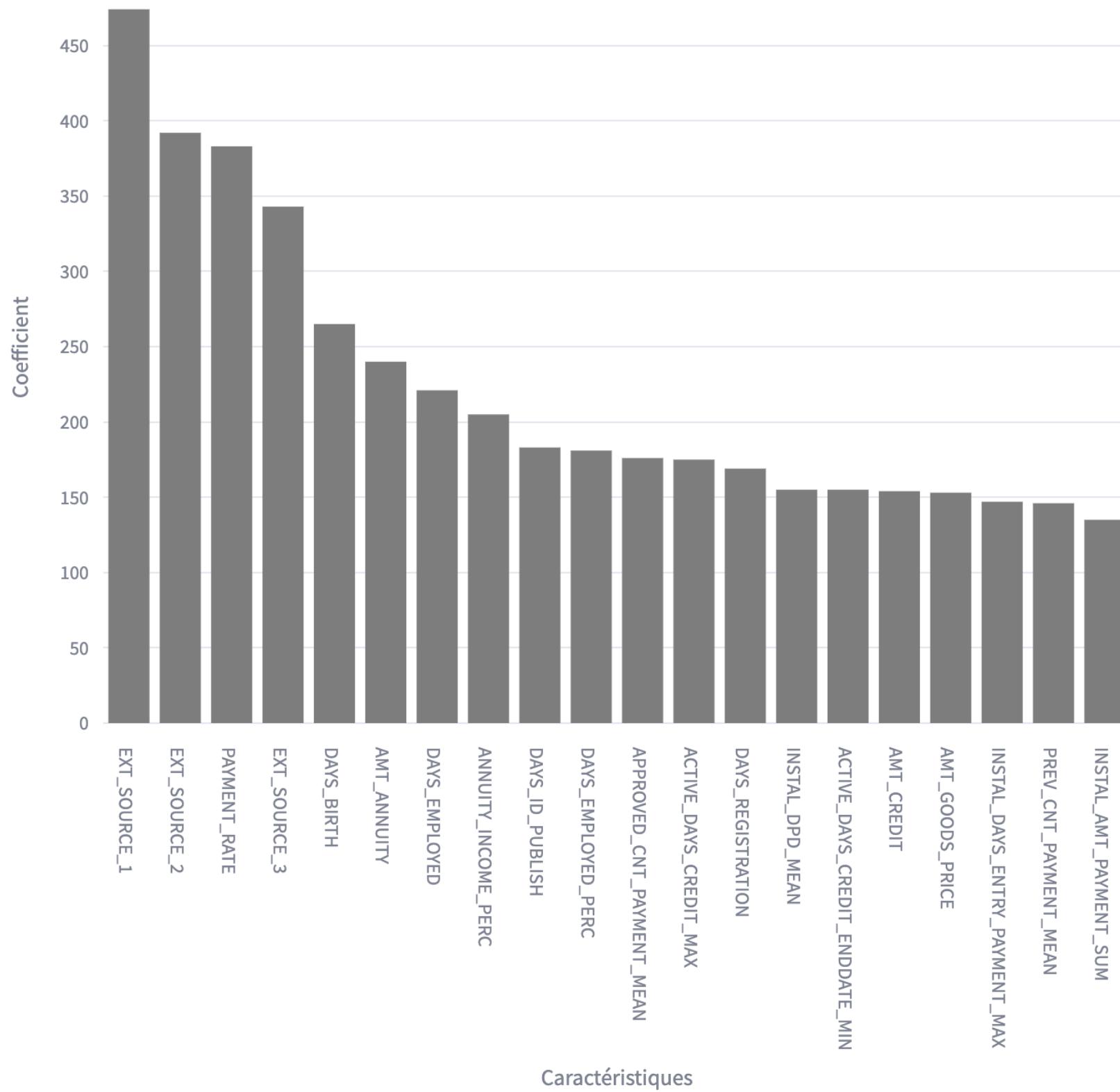
Page modèle

Page d'information sur le modèle

Un modèle d'apprentissage automatique a été utilisé pour évaluer l'importance des différentes informations clients dans la prédition du risque de défaut de crédit.

Caractéristiques principales du modèle:

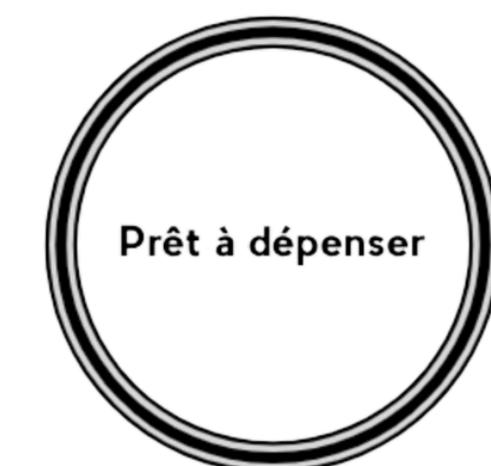
Top 20 des caractéristiques les plus importantes:



Le graphique affiche les 20 facteurs les plus influents pour le modèle. Voici quelques points clés à retenir sur les informations ayant le plus d'importance pour la décision d'octroi du crédit :

- **Sources externes de notation** (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) : Ces scores, provenant de sources externes, sont des indicateurs cruciaux pour évaluer la fiabilité financière des clients.
 - **Taux de paiement** (PAYMENT_RATE) : La façon dont les paiements sont effectués est un facteur déterminant. Un taux de paiement élevé est associé à une meilleure capacité de remboursement.
 - **Âge du client** (DAYS_BIRTH) : L'âge du client joue un rôle crucial. Des clients plus jeunes peuvent être considérés comme plus risqués.
 - **Mensualité du prêt** (AMT_ANNUITY) : La mensualité du prêt est importante. Des mensualités plus élevées par rapport au revenu peuvent indiquer un risque plus élevé.
 - **Ancienneté de l'emploi** (DAYS_EMPLOYED) : Plus le client est longtemps employé, plus il est stable financièrement.
 - **Proportion annuité/revenu** (ANNUITY_INCOME_PERC) : Cette proportion peut indiquer le niveau de confort financier du client.
 - **Délai depuis le dernier changement d'identité** (DAYS_ID_PUBLISH) : Des changements fréquents peuvent être associés à un risque plus élevé.
 - **Proportion de jours employés par rapport à l'âge** (DAYS_EMPLOYED_PERC) : Mesure la stabilité de l'emploi tout au long de la vie.
 - **Nombre moyen de paiements approuvés** (APPROVED_CNT_PAYMENT_MEAN) : Un indicateur de la gestion des paiements approuvés.
 - **Nombre maximal de jours de crédit actifs** (ACTIVE_DAYS_CREDIT_MAX) : Un historique de crédit actif plus long peut indiquer une stabilité financière.
 - **Délai depuis la dernière inscription** (DAYS_REGISTRATION) / **Délai minimal de crédit actif** (ACTIVE_DAYS_CREDIT_ENDDATE_MIN) / **Moyenne des retards de paiement** (INSTAL_DPD_MEAN) : Ces facteurs contribuent à évaluer la stabilité financière et la gestion du crédit.
 - **Montant du crédit** (AMT_CREDIT) / **Prix des biens** (AMT_GOODS_PRICE) / **Montant total des paiements d'acompte** (INSTAL_AMT_PAYMENT_SUM) : Des indicateurs importants pour évaluer la capacité du client à gérer les montants financiers associés au crédit.
 - **Nombre maximal de jours d'entrée de paiement** (INSTAL_DAYS_ENTRY_PAYMENT_MAX) / **Nombre moyen de paiements précédents** (PREV_CNT_PAYMENT_MEAN) : Des indicateurs liés aux paiements précédents.

Ces facteurs permettent de mieux comprendre comment le modèle prend ses décisions.



Conclusion

Sommaire

- Rappel du contexte et des objectifs
- Présentation de la modélisation
- Présentation du pipeline de déploiement
- Présentation de l'analyse de data drift
- Présentation du dashboard
- Conclusion



Prêt à dépenser

Conclusion

Résultats & Limites

Résultats :

- Modèle de scoring et classification octroi des prêts
- Dashboard interactif pour améliorer l'accessibilité des résultats et la transparence

Limites des modèles :

- Modèle : Score du modèle mitigé
- Déploiement : Limite de ressources disponibles avec les solutions Cloud gratuites

Axes d'amélioration - Modèle :

- Amélioration gestion outliers et des données manquantes
- Traitement de déséquilibre de classes (ex: ADASYN)
- Sélection de variables optimisées (ex: ACP, RFE, tests statistiques avancés, etc.)
- Optimisation des hyperparamètres (ex: optimisation bayésienne)
- Test autre modèle (ex: ensemble de modèles)

Axes d'amélioration - Déploiement / Dashboard :

- Déploiement dans une solution Cloud plus performante (ex: AWS, Azure)
- Amélioration des fonctionnalités du dashboard interactif et de l'esthétique

