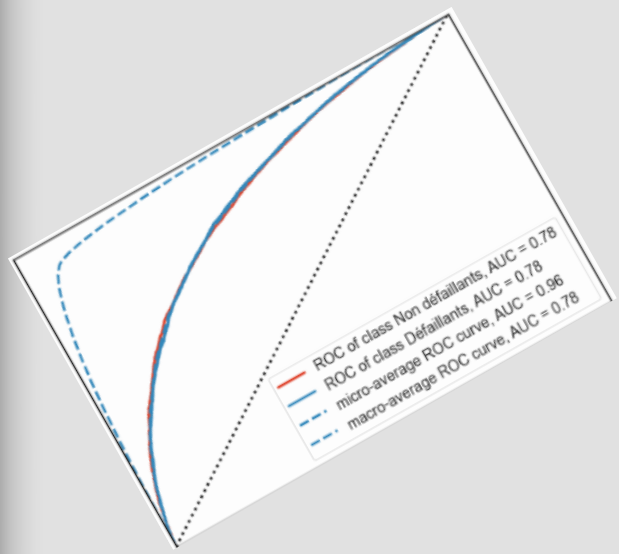
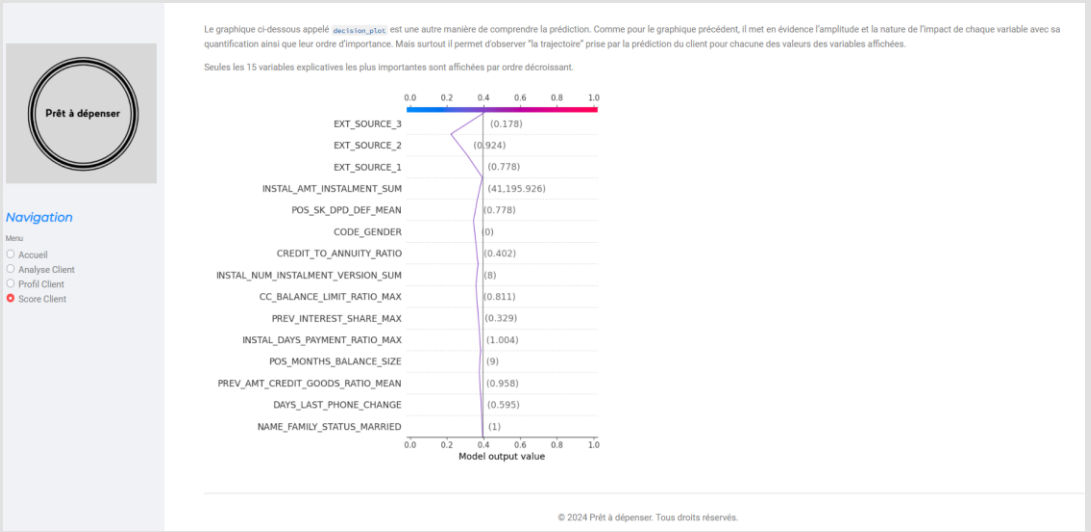
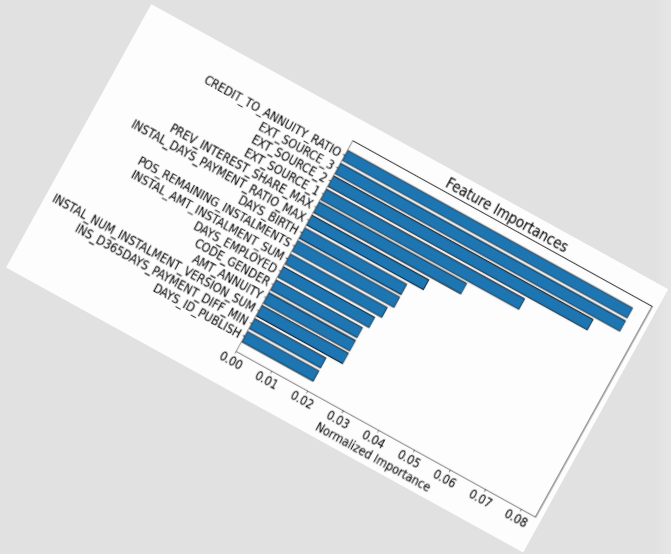
Implémenter un modèle de Scoring







Sommaire

Sommaire

1 Contexte

Prêt à dépenser:

L'entreprise souhaite créer un outil de "scoring crédit" basé sur un algorithme de classification pour prédire la probabilité de remboursement d'un client et prendre des décisions d'accord ou de refus de crédit.

Mission:

- Construire un **modèle de scoring automatisé** pour prédire la **probabilité de faillite** d'un client.
- Développer un **Dashboard interactif** pour les gestionnaires de la relation client, permettant d'interpréter les prédictions du modèle et d'améliorer leur connaissance des clients.

Objectifs:

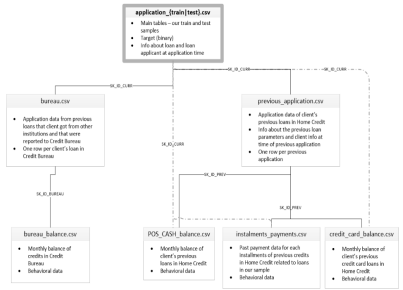
Les chargés de relation client ont souligné l'importance de la **transparence** dans les décisions d'octroi de crédit. Pour répondre à cette demande, l'entreprise prévoit de développer un tableau de bord **interactif** permettant d'expliquer les **décisions de crédit** de manière transparente et de donner aux clients un **accès facile** à leurs informations personnelles.

1 Contexte – déroulement

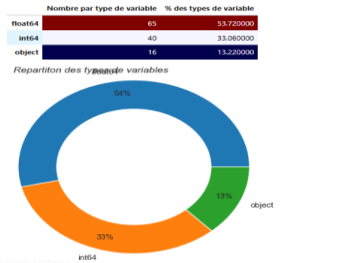
Prêt à dépenser

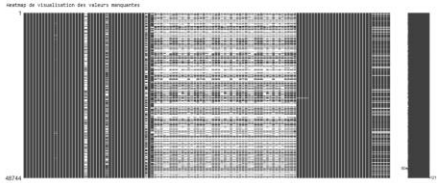
EDA

Traitement des 8 fichiers de prêt.



Stat, typage, données manquantes.





PRE-PROCESSING

Nettoyage des données

Typage, valeurs manquantes, imputation

Feature engineering

- Création nouveaux types de features (min, max, sum, …)
- Encodage (cat encoding, …)
- Elimination des fortes colinéarités
- Fusion des datasets nettoyés

Données test et entrainement

FEATURES SELECTION

LightGBM

- Recherche des features importances avec 2 itérations afin d'éliminer les 0 importances.

Permutation importance

- Permutation-based Importance avec sklearn

- Permutation importance avec eli5

Conservation des variables les plus fréquentes

MODELISATION

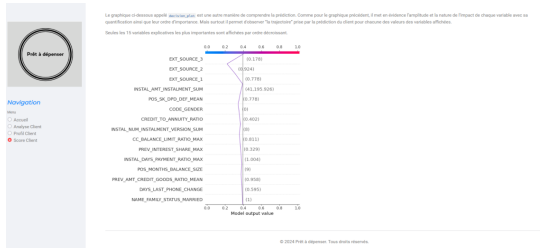
LightGBM

- Séparation du dataset
- Gestion du déséquilibre (SMOTE, …)
- Sélection des métriques (métier, …)
- Optimisation bayésienne des différents modèles (selon les métriques, …):
- Détermination de la probabilité optimale.

Modèle Final

DASHBOARDING

Développement

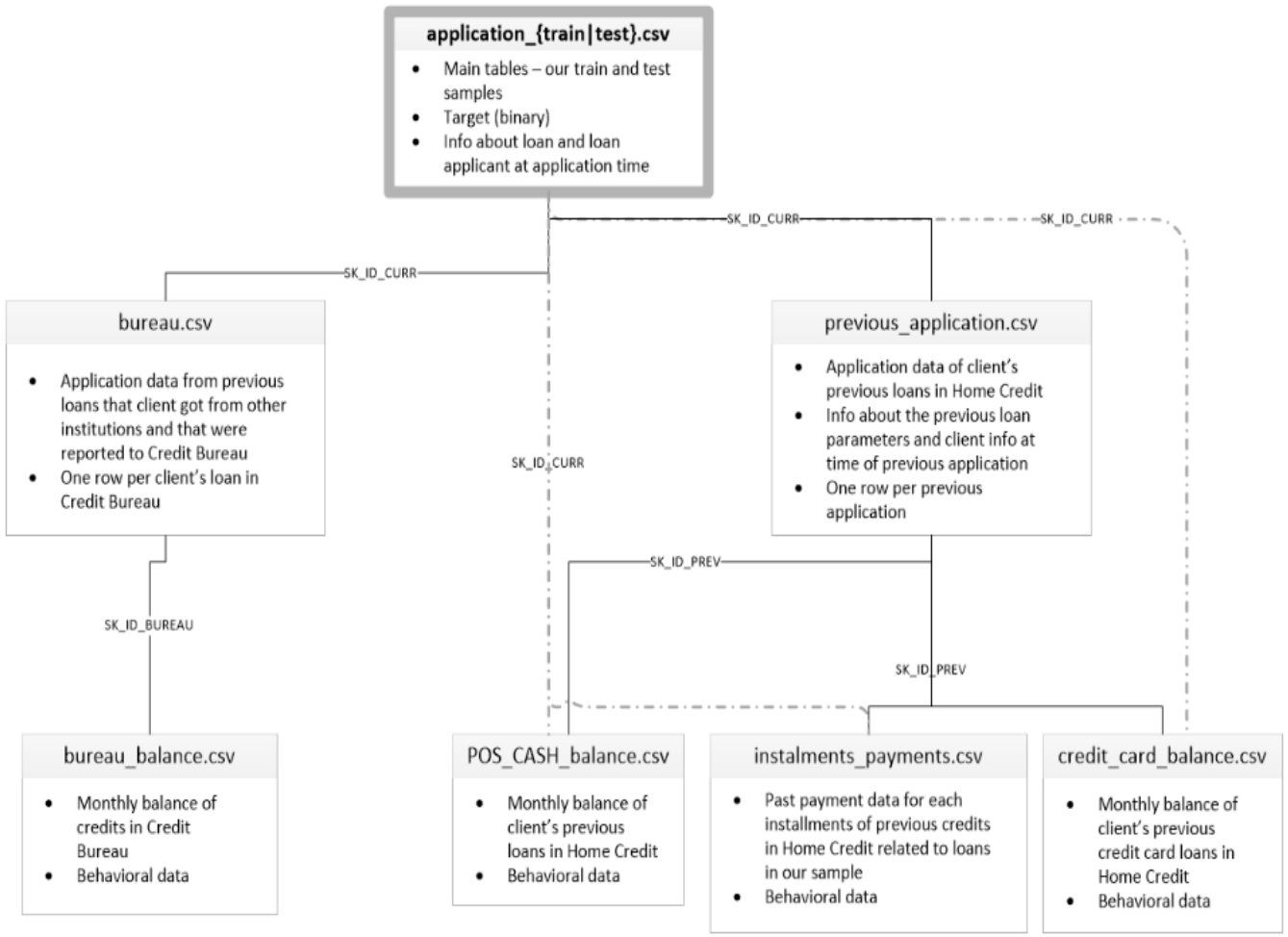


Déploiement

Streamlit

Sommaire

2 Traitement des données – jeu datasets



8 datasets:
307511 clients et 530 variables

application_train.csv (307511,122)

application_test.csv (48744,121)

previous_application.csv (1670214,37)

- Demandes précédentes de prêts immobiliers

POS_CASH_balance.csv (10001358,8)

- Instantanés mensuels du solde des points de vente précédents et des prêts en espèces

instalments_payments.csv (13605401,8)

- Historique de remboursement des crédits précédents.

credit_card_balance.csv (3840312,23)

- Reçu mensuel du solde des cartes de crédits précédents

bureau.csv (3840312,23)

- Antécédents des crédits des clients

bureau_balance.csv (3840312,23)

- Soldes mensuels des crédits précédents

2 Traitement des données – Pré-processing sur les 8 datasets

Nettoyage

- Typage des données
  - - Numérique ordinal en object
    - Optimisation mémoire du dataframe (int64 en int8, …)
- Valeurs aberrantes, valeurs différentes sur les données train/test
  - Transformation des valeurs aberrantes de l'EDA
  - Transformation ou suppression valeurs uniques du train set ≠ test set
- Valeurs manquantes
  - Suppression des nan avec conservation des variables importantes de l'EDA
- Imputation
  - *1ère étape* Imputation par la médiane
  - *2ère étape* Imputation par valeurs nulles 0 Imputation Qualitatives constante XNA
  - *3ère étape* Imputation NaNImputer

Feature engineering

- Variables métiers, statistiques
  - *Manuel:* Ajout variables métiers
  - *Automatique:* Ajouts variables Statistiques (min, max, sum, …)
- Encodage
  - *Quantitatives: StandardScaler(*)*
  - *Qualitatives:* LabelEncoder (binaire) pd.get_dummies (> 2 valeurs .uniques)

Assemblage

- Merge
  - Fusion des 6 datasets avec les données train/test

Colinéarité

- Colinéarité
  - Suppression des variables fortement colinéaires

2 Traitement des données – Feature Engineering

Prêt à dépenser

Automatique

- Encodage features catégoriques
- Création de features statistiques: quantitatives et qualitatives (min, max, sum, mean, count,…)
- Fonction ‘Feature Engineering’ intégrant ces caractéristiques sur chaque dataset

Manuel

- Ratio/différence des variables métiers
- Conversion jours en années: ratio
- Nouveaux features crées par fusion d’autres features (addition, multiplication, …)
- Données externes: ratio, moyenne, min, max

2 Traitement des données – Fusion datasets

Dataset initial	Dimensions initiales	Dimensions Datasets après le process FE	Fusion des datasets, imputation nan et var. colinéaires
application_train/test	(307 511, 122) (48744,121)	(307507, 259) (48744, 258)	
cc_balance	(3840312, 23)	credit_balance_fe (103558, 389)	(307507, 294) (48744, 292)
installments_payments	(13605401,8)	installment_fe (339587, 56)	(307507, 301) (48744, 292)
POS_CASH_balance	(10001358, 8)	pos_cash_fe (337252, 38)	(307507, 322) (48744, 314)
previous_application	(1670214, 37)	previous_application_fe (338857, 258)	(307507, 499) (48744, 488)
bureau_balance	(27299925, 3)	bureau_bb_fe (817395, 12)	(307507, 495) (48744, 485)
bureau	(1716428, 17)	bureau_fe (305811, 60)	(307507, **530**) (48744, **521**)

train set: 530 variables
test set: 521 variables

FEATURE SELECTION OBLIGATOIRE

2 Traitement des données – Feature Engineering

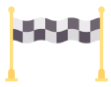
Plage	nb_données	%_données
(0, 1]	125	23.674242
(1, 2]	187	35.416667
(2, 3]	216	40.909091
(3, 4]	0	0.000000

+ ID CLIENT
+ VARIABLE TARGET

LightGBM:2
40 Features

Permutation
Importance:
344
Features

train set: 530 variables
test set: 521 variables

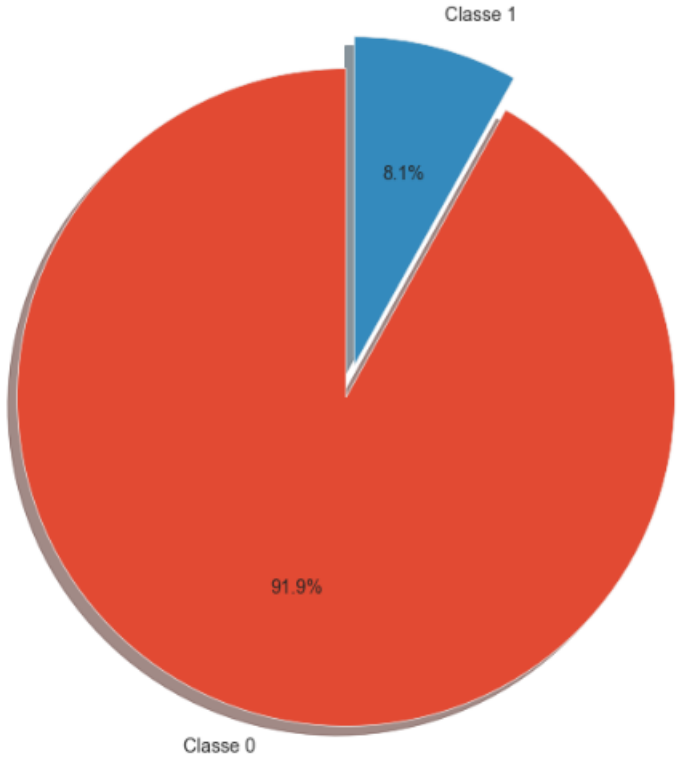


train set: 218 variables
test set: 217 variables

DONNEES FINALES
MODELISATION

Sommaire

3 Modélisation – variable cible



Constat variable cible

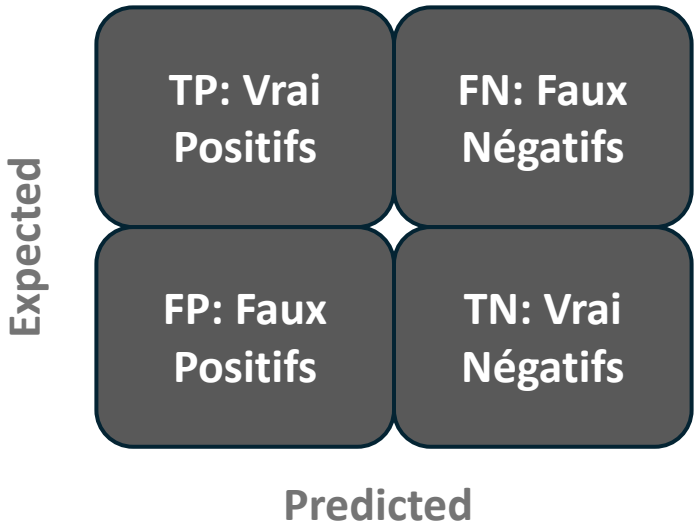
Fort déséquilibre entre les clients à risque (Target =1) ou sain (Target = 0)

Modélisation

Rééchantillonnage des classes déséquilibrées: SMOTE, ADASYN

3 Modélisation – les métriques

Utilisation de plusieurs métriques standards (minimisation du déséquilibre des classes):
Accuracy, ROC, F1 Score, F2 Score, FBeta, Custom metric



Précision: - défaillants = classe **positive**
- non-défaillants = classe **négative**

Minimisation des pertes de bénéfices (prédiction):

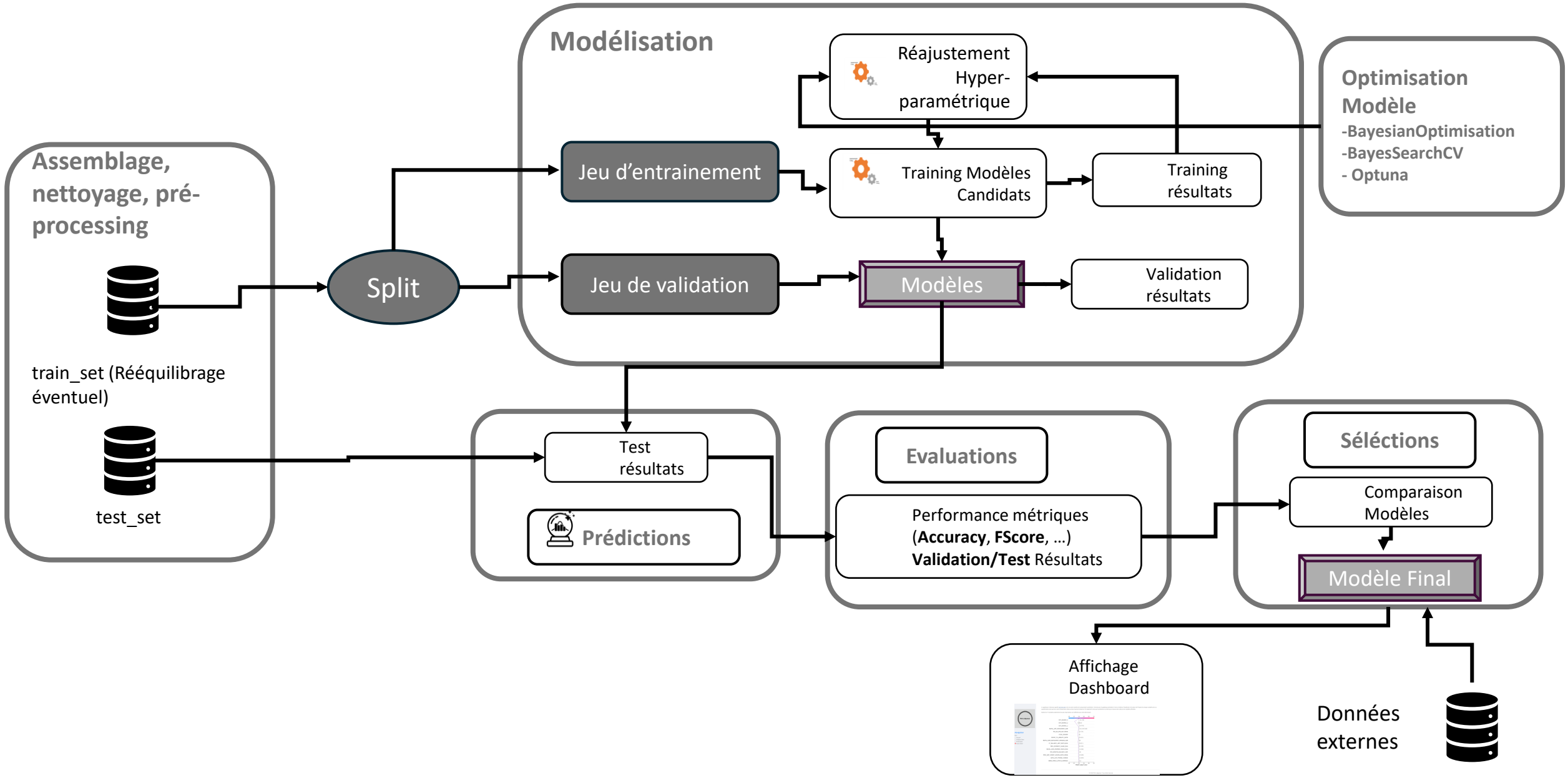
- un client non-défaillant s'il est défaillant ==> **minimiser le nombre de faux négatifs (erreur de type II)** (prédit non-défaillant mais client défaillant) ==> **maximiser** le Recall

- le client n'est pas défaillant ==> **minimiser les faux positifs (erreur de type I)** (classe 1 défaillant alors que non-défaillant) ==> **maximiser** la Précision

- **Compromis FN/FP** (si **FN** - **P** + et inversement)
- **Test de minimisation fonction cout** pour différentes métriques

3 Modélisation – déroulement

Prêt à dépenser



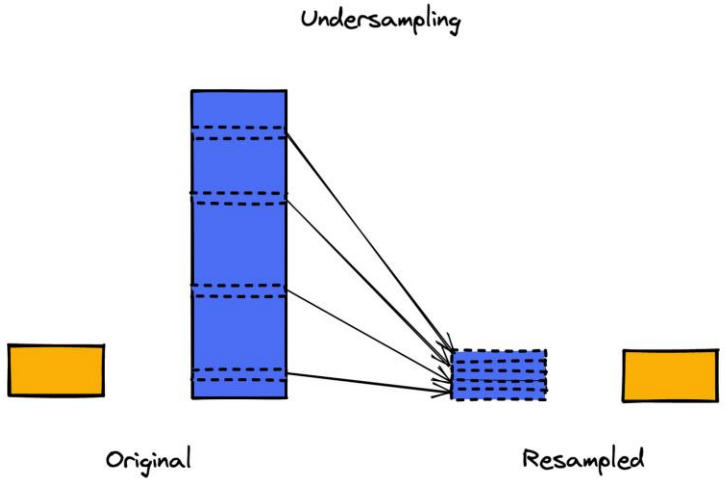
Modélisation – Oversampling methods (désequilibre des classes)

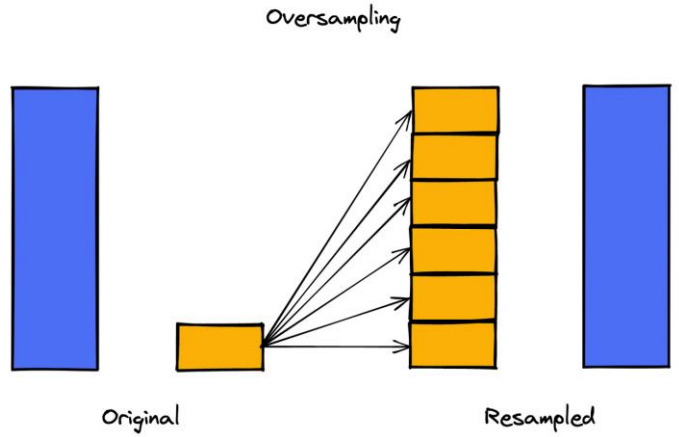
Méthodes utilisées

UNDERSAMPLING - ***SMOTE***

OVERSAMLPING - ***ADASYN***

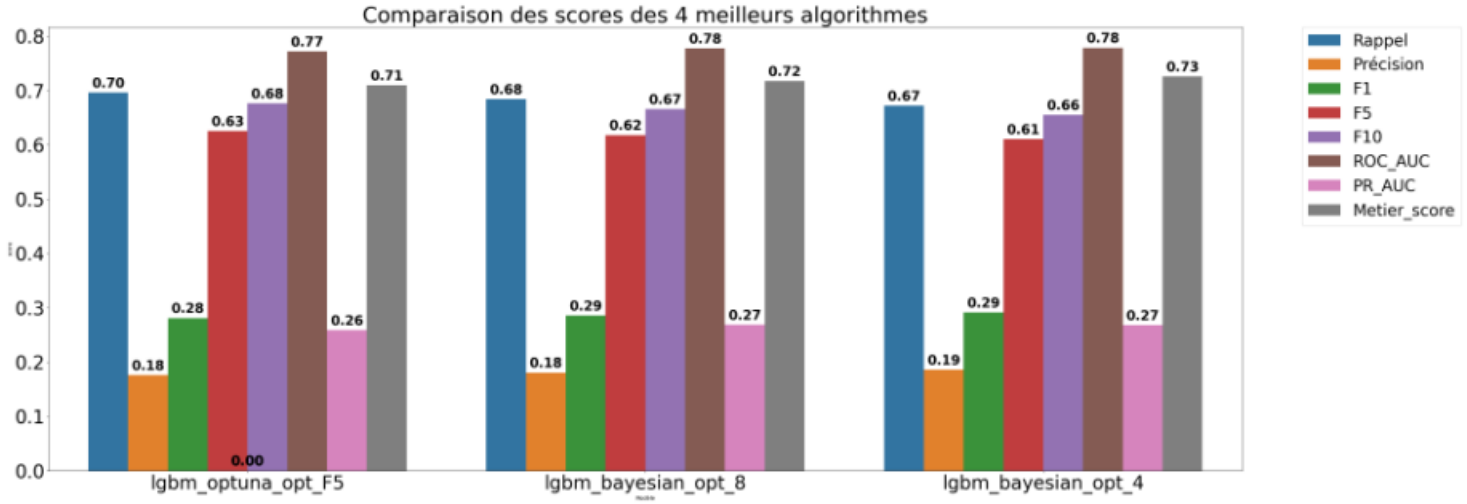
COMBINAISON OVERSAMLPING et UNDERSAMPLING - ***SMOTE***

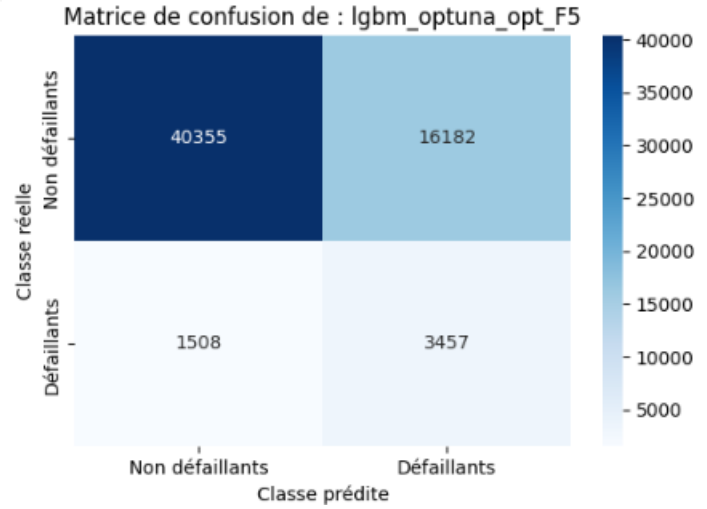




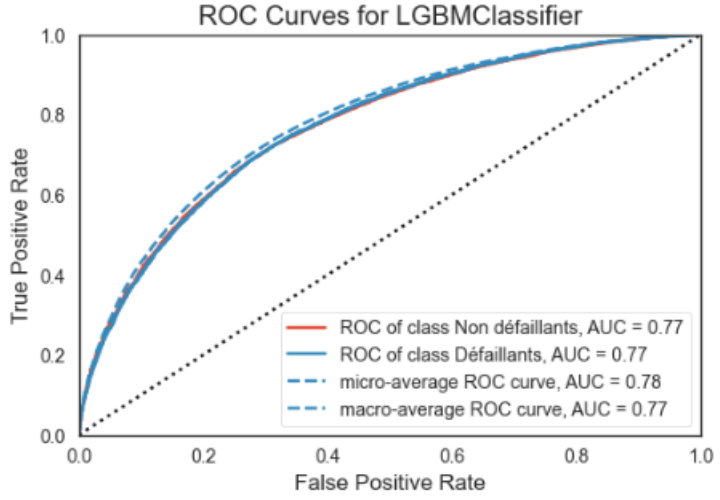
3 Modélisation – Conclusion meilleurs modèles

Modèle	Jeu_donnees	FN	FP	TP	TN	Metrique	Optimisation	Class_weight	Rappel	Précision	F1	F5	F10	ROC_AUC
lgbm_optuna_opt_F1	train	1500	16369	3465	40168	F1	optuna	oui	0.697885	0.174700	0.279447	0.625803	0.677788	0.770713
lgbm_optuna_opt_F5	train	1508	16182	3457	40355	F5	optuna	non	0.696274	0.176027	0.281011	0.625205	0.676479	0.771812
lgbm_optuna_opt_F10	train	1514	15871	3451	40666	F10	optuna	non	0.695065	0.178605	0.284185	0.625499	0.675720	0.776303
lgbm_optuna_opt_10_train	train	1544	15089	3421	41448	F10	optuna	oui	0.689023	0.184819	0.291459	0.623592	0.670902	0.780011





MODELE RETENU: LGBM_OPTUNA_OPT_F5



3 Modélisation – Hyperparamètres retenus

Prêt à dépenser

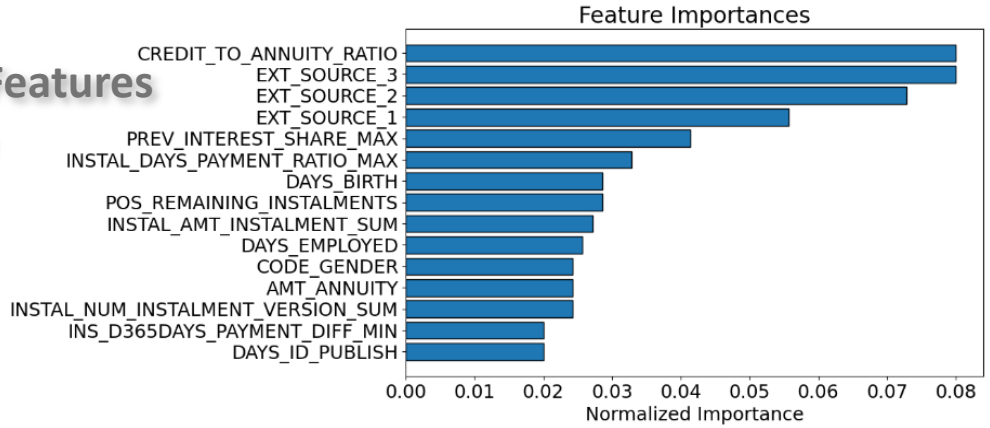
MODELE LIGHTGBM:
Meilleurs Paramètres

Hyperparamètre	Meilleurs paramètres
n_estimators	**100 (par défaut)**
learning_rate	**0,1 (par défaut)**
objective	**binary**
boosting_type	**gbdt**
class_weight	**balanced**
colsample_tree	**0.883696173865355**
max_depth	**4**
min_child_samples	**37**
min_child_weight	**0.9053832802852111**
num_leaves	**8**
reg_alpha	**0.0013343227256418153**
reg_lambda	**1.1168060057563535e-06**
subsample	**0.876335534267455**
subsample_freq	**4**

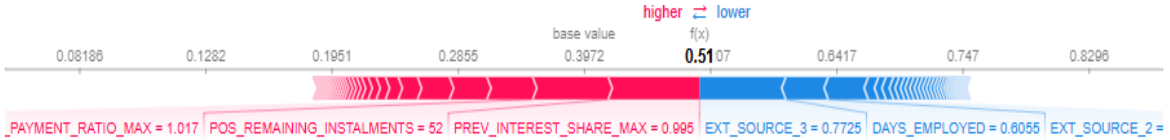
Modélisation – Graphes d'interprétation

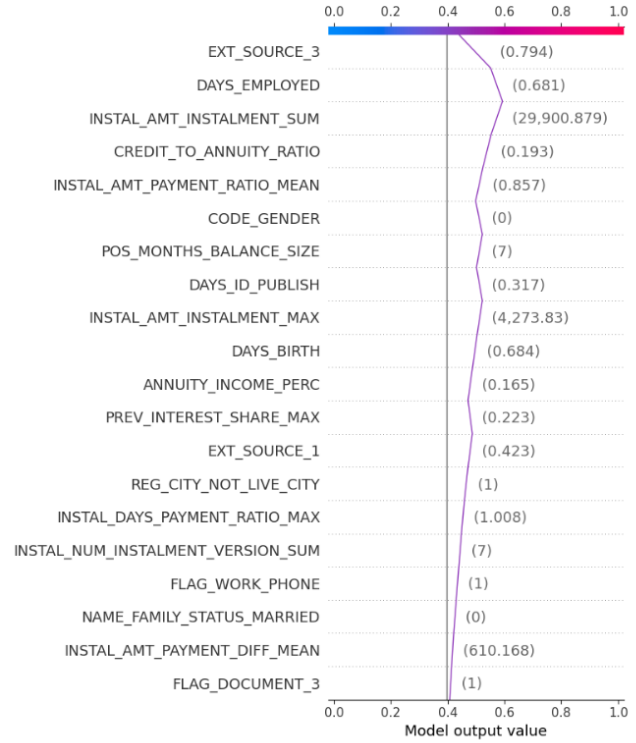
VUE GLOBALE

LightGBM: Features Importance



VUE CLIENT





Sommaire

4 Dashboard

Prêt à dépenser

Deploy

Prêt à dépenser

Navigation

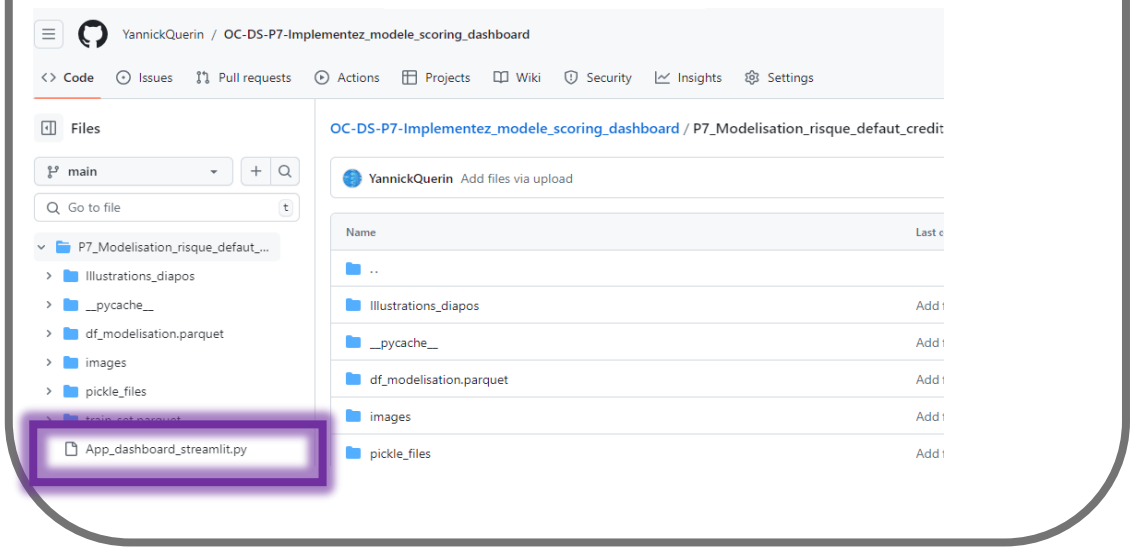
Menu

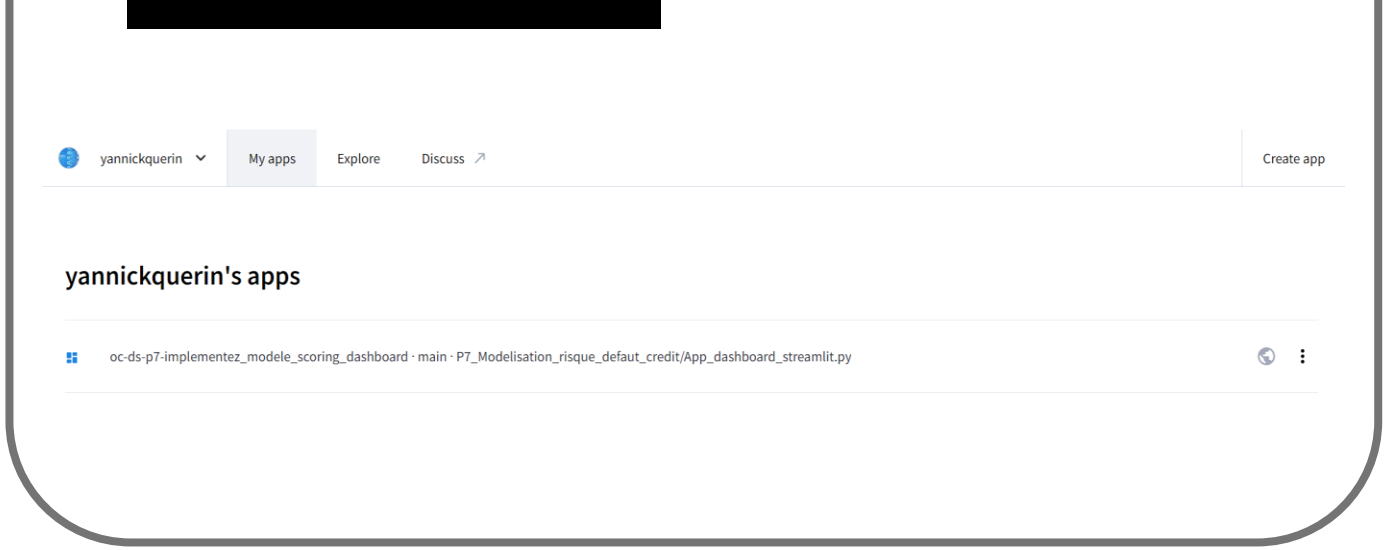
- Accueil
- Analyse Client
- Profil Client
- Score Client

Bienvenue sur l'application de Crédit 🏠

Cette application vous permet d'analyser le risque de défaut de paiement de vos clients, de visualiser le profil de vos clients et de comprendre comment les scores de risque sont calculés. Utilisez le menu à gauche pour naviguer entre les différentes sections.

4 Dashboard





Lien Github:
https://github.com/YannickQuerin/OC-DS-P7-Implementez_modele_scoring_dashboard/blob/main/P7_Modelisation_risque_defaut_credit/App_dashboard_streamlit.py

Lien local: App_dashboard_streamlit.py

Lien Streamlit:
https://yannickquerin-p07-dashboard.streamlit.app

4 Dashboard

Prêt à dépenser

Boîte de réception - yannickque | YannGwad/OC-DS-P7-Impleme | yanngwad's apps · Streamlit | Application de Crédit · Streamlit | Mes tableaux de bord - Session

appdashboardapppy-xjlkesdmhuhvtrgsbgeasz.streamlit.app

Page Web non disp... | Gmail | YouTube | Maps | Meet – dwp-frmb-n... | Data blue - Google... | IMDB_DATASET_1 |... | Meet – rrx-uihj-jxe | Content-Based Mov... | Slack | data-blue-09... | Tous les favoris

Share

Prêt à dépenser

Navigation

Menu

- Accueil
- Analyse Client
- Profil Client
- Score Client

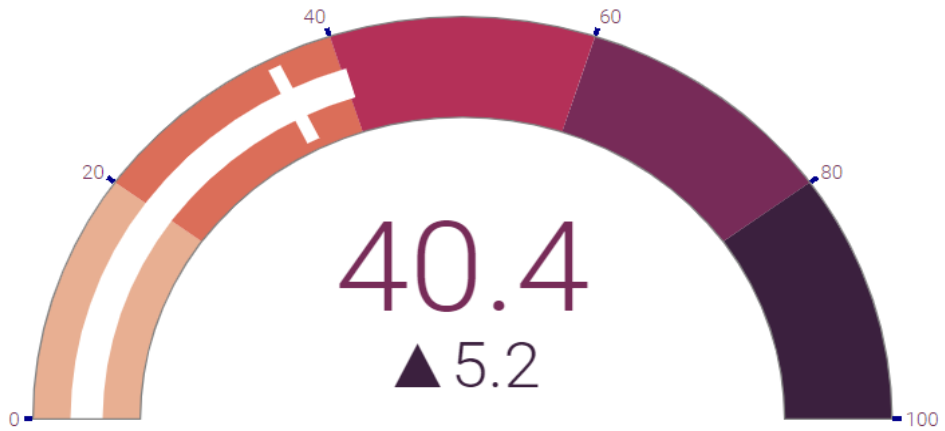
Score Client

Veuillez sélectionner le numéro de votre client 👇

100001

Vous avez sélectionné l'identifiant n : 100001

Score du client



Le client dont l'identifiant est **100001** a obtenu le score de **40.4%**.

Il y a donc un risque de 40.4% que le client ait des difficultés de paiement.

Le client est donc considéré par *'Prêt à dépenser'* comme **solvable** et décide de lui **accorder** le crédit.

Manage app

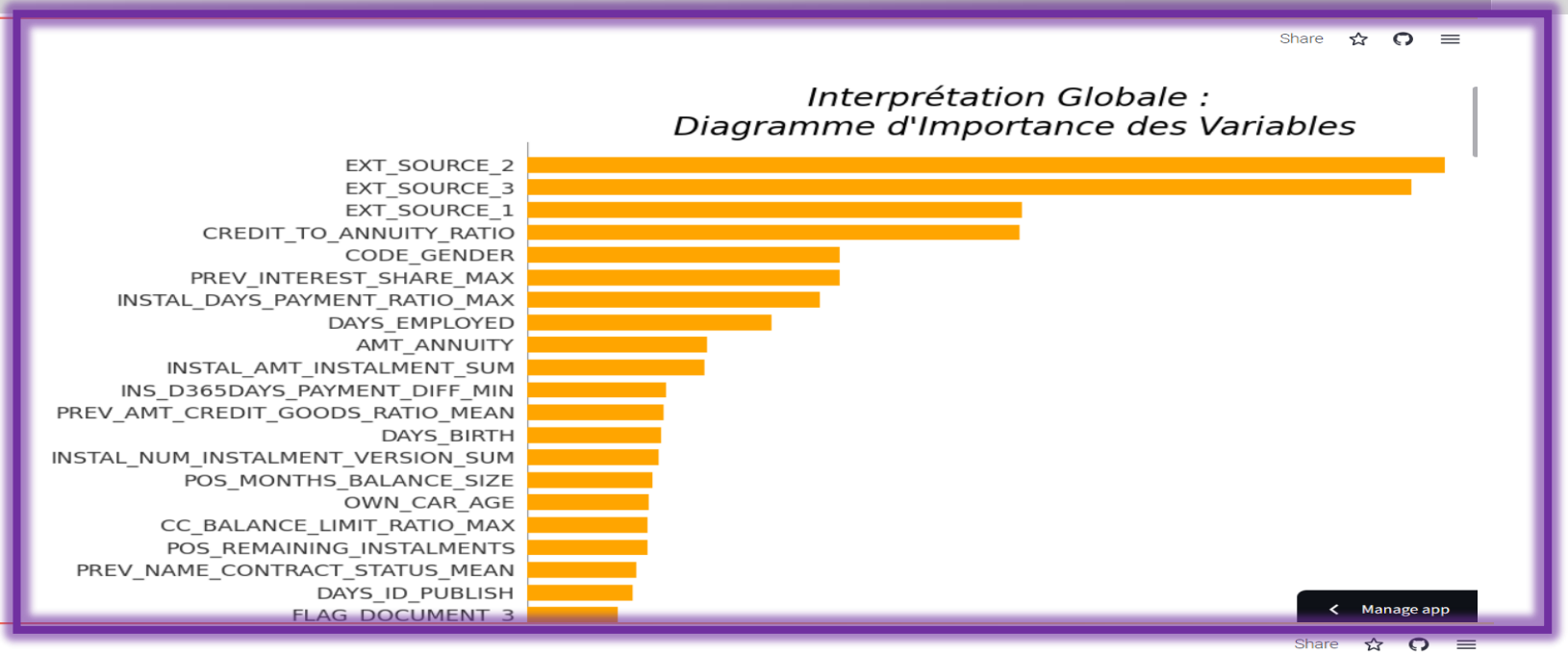
4 Dashboard

Prêt à dépenser

Prêt à dépenser

Navigation

Menu
- Accueil
- Analyse Client
- Profil Client
- Score Client



Share

2. Quel est l'Impact de chaque caractéristique sur la prédiction ?

Le diagramme des valeurs SHAP ci-dessous indique également comment chaque caractéristique impacte la prédiction. Les valeurs de Shap sont représentées pour chaque variable dans leur ordre d'importance. Chaque point représente une valeur de Shap (pour un client).

Les points rouges représentent des valeurs élevées de la variable et les points beiges des valeurs basses de la variable.

Prêt à dépenser

Navigation

Menu
- Accueil
- Analyse Client
- Profil Client
- Score Client

4 Dashboard

Prêt à dépenser

Share

Profil Client

Prêt à dépenser

Navigation

Menu

- Accueil
- Analyse Client
- Profil Client
- Score Client

1. Quel est le profil de votre client ?

Veuillez sélectionner le numéro de votre client à l'aide du menu déroulant

100001

Vous avez sélectionné l'identifiant n° : `100001`

Le client dont l'identifiant est **100001** a obtenu le score de **40.4%**.

Il y a donc un risque de 40.4% que le client ait des difficultés de paiement.

Profil socio-économique

Sexe de l'individu : **Féminin**

Situation familiale : **Married**

Nbre enfants : **0**

Niveau éducation : **Higher education**

Type revenu : **Working**

Ancienneté emploi : **6 ANS**

Revenus: **135000.0 $**

Profil emprunteur

Type de prêt : **Cash loans**

Montant du crédit: **568800.0 $**

Annuités: **20560.5$**

Montant du bien: **450000.0 $**

Type de logement : **House / apartment**

Manage app

Sommaire

5 Conclusions

- Manque de connaissance précise du **secteur bancaire**: vérification au niveau **métier** des **variables retenus** lors du pré-process

- Définir de façon plus détaillée la **métrique d' évaluation** et de **fonction de cout** avec les équipes métiers

- Développer un dashboard avec une page **Données bancaires** coté décideur et une page **Données Clients** de façon à séparer, et sécuriser certaines données sensibles détenus uniquement du côté de la banque, sans pour autant certaines données au client.

- Entrevoir une section de r**ecommandation** qui permettrait au client de voir quelle variable aurait pu influencer sur son obtention ou pas du prêt en question.