



PROJET 7

IMPLÉMENTEZ UN MODÈLE DE SCORING

SOMMAIRE

- ❑ PROBLÉMATIQUE
- ❑ DONNÉES
- ❑ MODÉLISATION
- ❑ DASHBOARD
- ❑ CONCLUSION



- ❑ PROBLÉMATIQUE
- ❑ DONNÉES
- ❑ MODÉLISATION
- ❑ DASHBOARD
- ❑ CONCLUSION



❑ Prêt à dépenser :

Société financière d'offre de crédit à la consommation pour la clientèle ayant peu ou pas d'historique de prêt.

❑ Mission :

- ✓ Construire un **modèle de scoring** prédisant automatiquement la **probabilité** de **défaut de paiement** d'un client.
- ✓ Développer un **dashboard interactif**.

❑ Objectifs :

- ✓ **Étayer** la décision d'accorder ou non un prêt.
- ✓ Améliorer la relation avec le client en faisant preuve de **transparence**.
- ✓ Montrer au client les **informations** le concernant grâce à **l'interactivité**.

PROBLÉMATIQUE

EDA

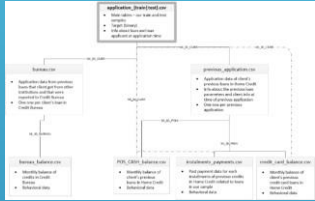
PRÉ
PROCESSING

FEATURES
SELECTION

MODELISATION

DASHBOARD

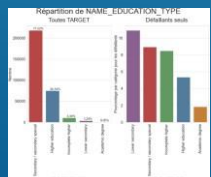
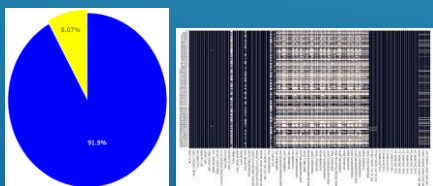
Kaggle, 8 fichiers CSV



Kernel Rishabhrao



Stats, types données,
NaN, unique



Nettoyage

- Type de données (bool, mémoire..)
- Val. aberrantes
- Imputation

Feature
Engineering

- Création variables métiers
- Création variables automatiques (min, max, mean, sum, count..)
- Encodage
- Suppression des colinéarités fortes
- Assemblage (merge)

1 train set
1 test set

LightGbm

- Plusieurs itérations
- Features importances

Boruta

BorutaShap

Permutation
importance

- Sklearn, eli5

RFECV

→ Conserve les variables les plus répétées pour toutes ces méthodes

Pycaret

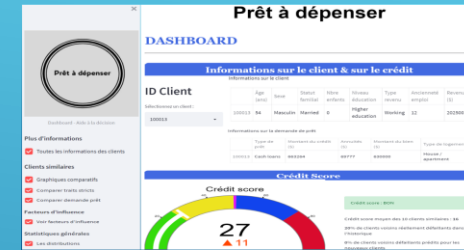
- Première idée
- Choix du jeux de données
- Choix du modèle

LightGbm

- Split du jeux de données, rééquilibrage
- Choix des métriques
- Optimisation du modèles
- Seuil de probabilité optimal

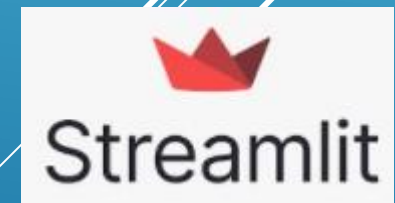
Modèle Final

Dév. Local



- Prédictions
- Visualisations

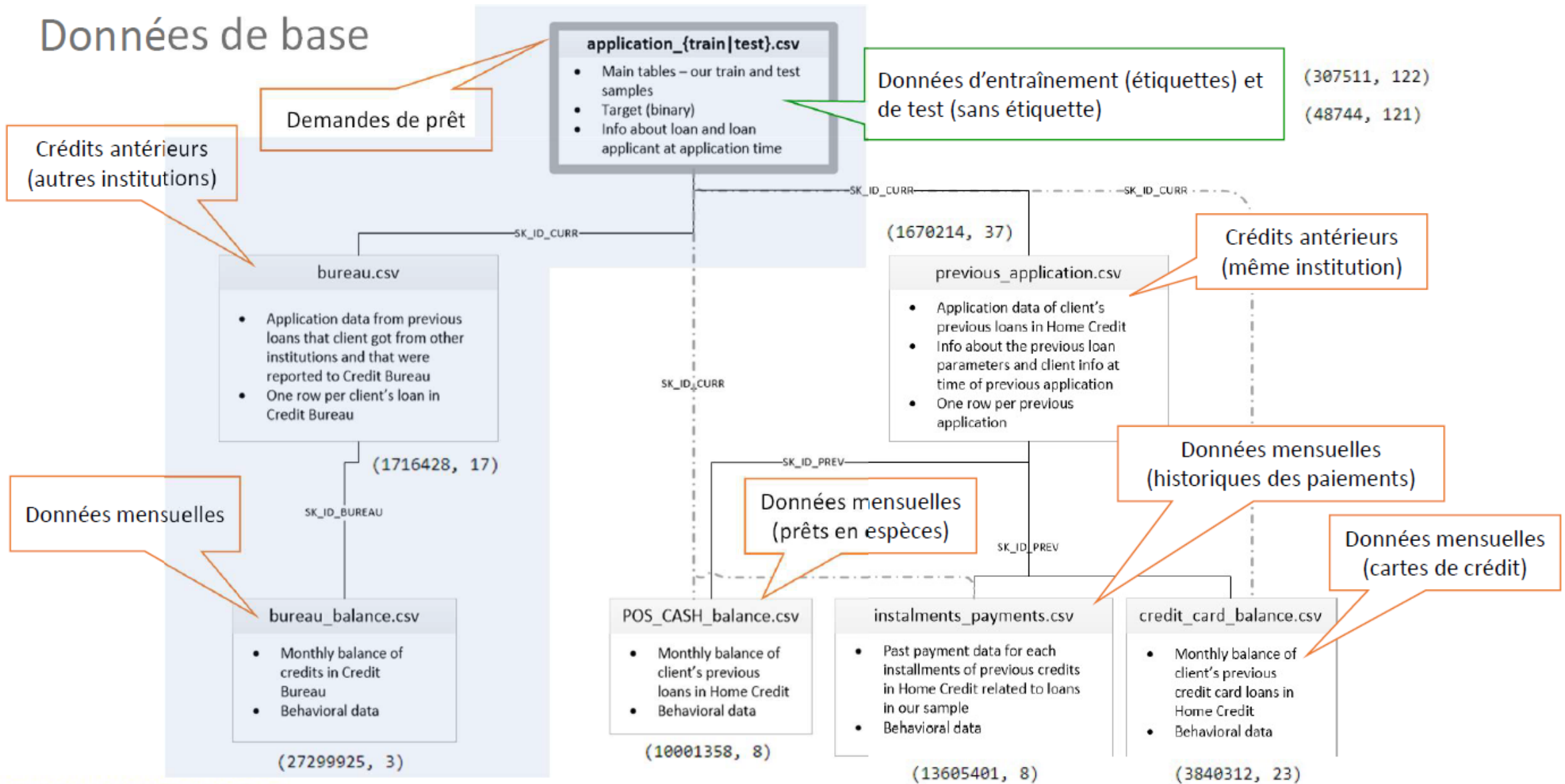
Déploiement



- ❑ PROBLÉMATIQUE
- ❑ **DONNÉES**
- ❑ MODÉLISATION
- ❑ DASHBOARD
- ❑ CONCLUSION

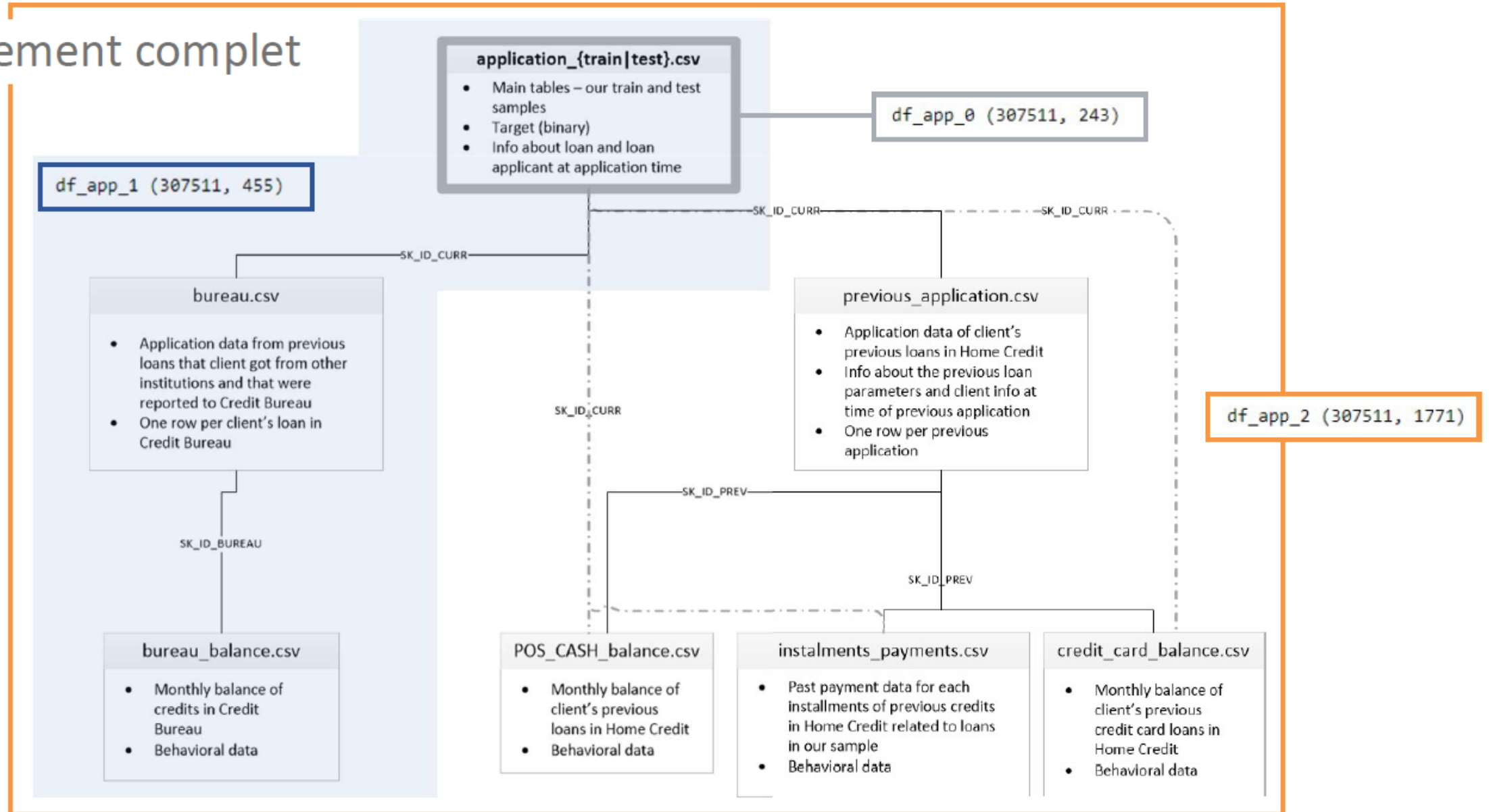


Données de base



DONNÉES – ANALYSE EXPLORATOIRE

Traitement complet



DONNÉES – ANALYSE EXPLORATOIRE

Dataframe initial	Nbr lignes var. initiales	Nbr lignes var. après FE	Merge avec application_train/test et suppr var. colinéaires + > 90% nan
application_train/test	(307511, 122) (48744, 121)	(307507, 206) (48744, 205)	
credit_card_balance	(3840312, 23)	agg_ccb_cat (103558, 21) agg_ccb_num (103558, 68)	(307507, 246) (48744, 245)
installments_payments	(13605401, 8)	agg_pay_num (339587, 30)	(307507, 265) (48744, 264)
POS_CASH_balance	(10001358, 10)	agg_pos_num (337252, 27)	(307507, 285) (48744, 284)
previous_application	(1670214, 37)	agg_prev_num (338857, 114)	(307507, 552) (48744, 551)
bureau_balance	(27299925, 3)	agg_bureau_balance_par_demandeur (305811, 12)	(307507, 555) (48744, 554)
bureau	(1716428, 17)	agg_bureau_num (305811, 60)	(307507, 615) (48744, 614)



train set : 615 variables
test set : 614 variables



**FEATURE SELECTION
NÉCESSAIRE**

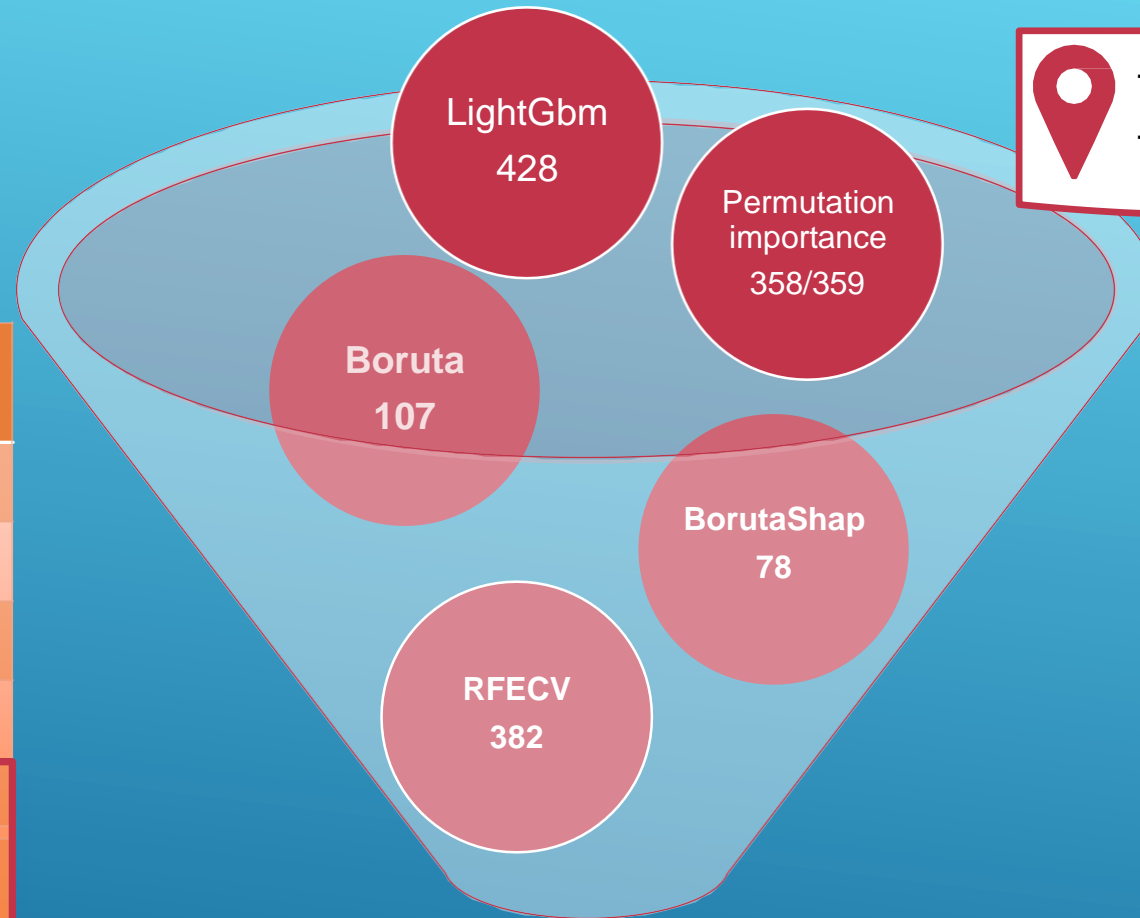



↓ ↓
+ 493


DONNÉES – FEATURES ENGINEERING

Nbr répétition	Nbr variables	%_variables/615
(0, 1]	99	19.9195
(1, 2]	53	10.6640
(2, 3]	57	11.4688
(3, 4]	182	36.6197
(4, 5]	28	5.6338
(5, 6]	78	15.6942

+ IDENTIFIANT
+ TARGET 



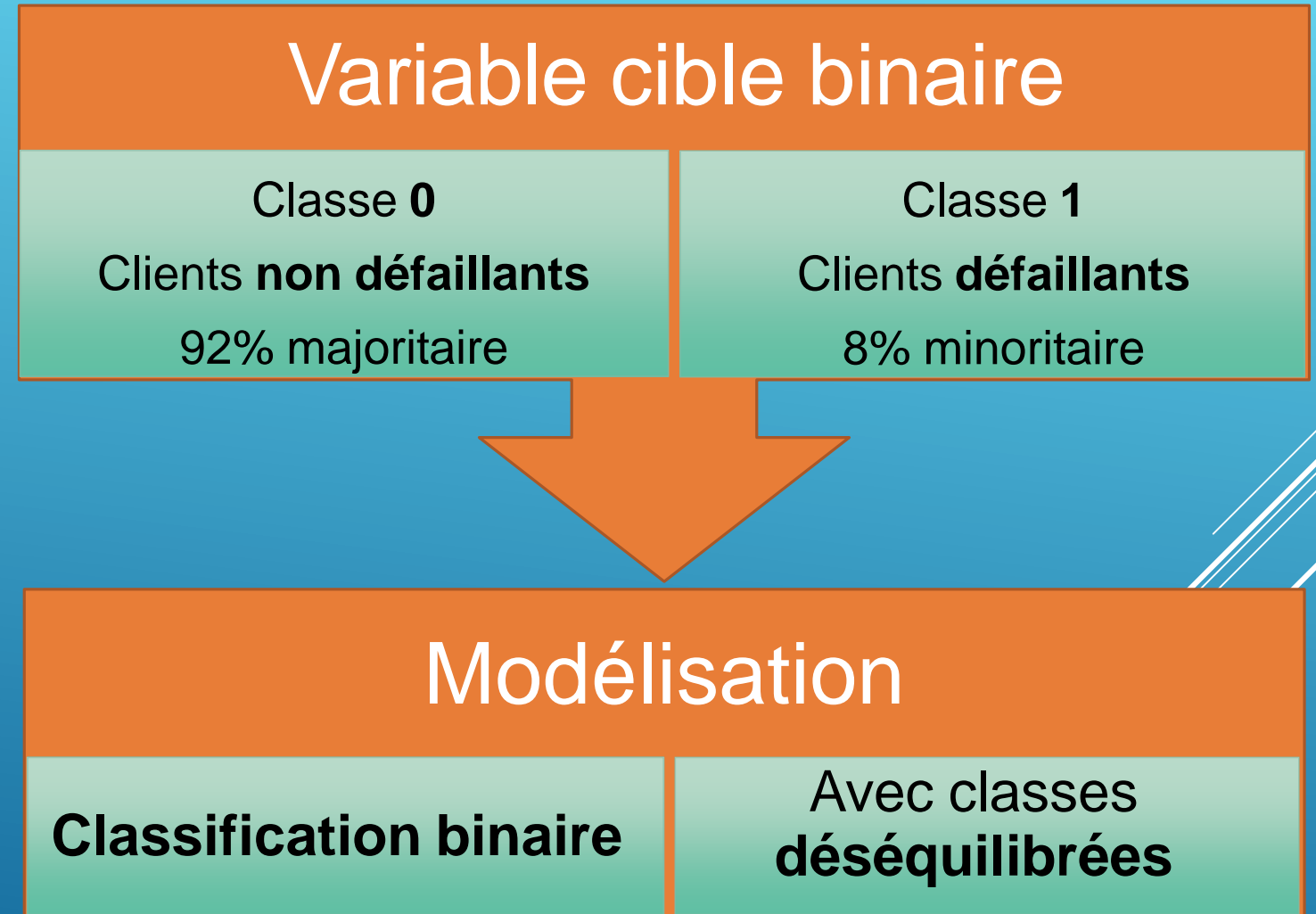
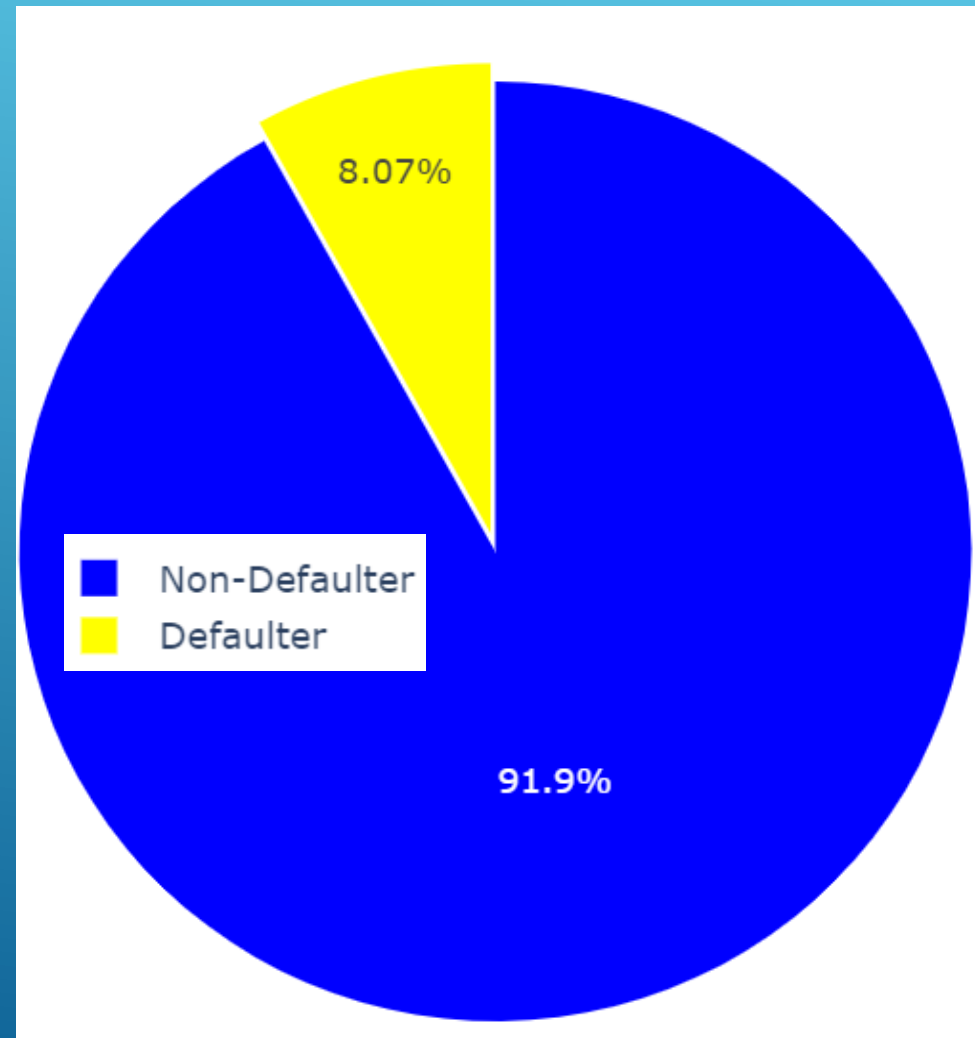
 train set : **615** variables
test set : 614 variables

 train set : **108** variables
test set : 107 variables

 **DONNÉES PRÊTES
POUR MODELISATION**

- ❑ PROBLÉMATIQUE
- ❑ DONNÉES
- ❑ **MODÉLISATION**
- ❑ DASHBOARD
- ❑ CONCLUSION







MODÉLISATION – CHOIX DES MÉTRIQUES

Rappel :

Classe 0 négative = non défaillant

Classe 1 positive = défaillant

Matrice de confusion :

Classe réelle	+	 TP Vrais positifs	FN Faux négatifs
	-	FP Faux positifs	 TN Vrais négatifs
		+	-
		Classe prédite	

Minimiser les pertes argent :


Prédiction	Réalité	PERTE
+ défaillant	- non-défaillant	Intérêt du prêt non accordé
- non-défaillant	+ défaillant	Somme empruntée en partie ou totale



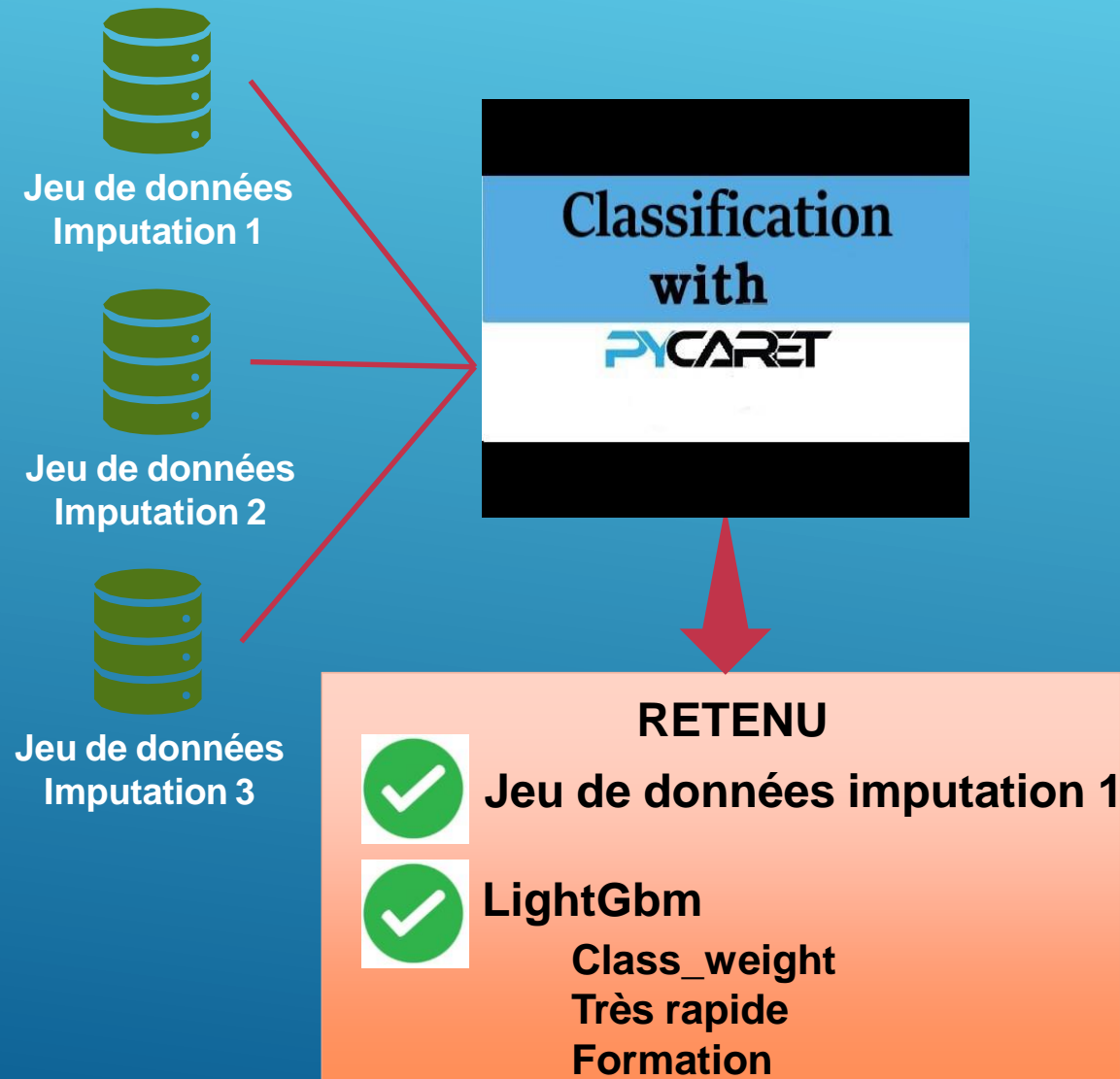
minimiser le nombre de **faux positifs**
maximiser la métrique **Précision**



minimiser le nombre de **faux négatifs**
maximiser les métriques **Recall** ou **Fbeta 10**

Perte > pour FN que FP : privilégier Recall/F10
Compromis FN/FP (si FN  FP  et vis versa)
Tester métrique métier / fonction coût pour jouer sur le taux de FN/FP et privilégier les bons prêts

MODÉLISATION – PYCARET



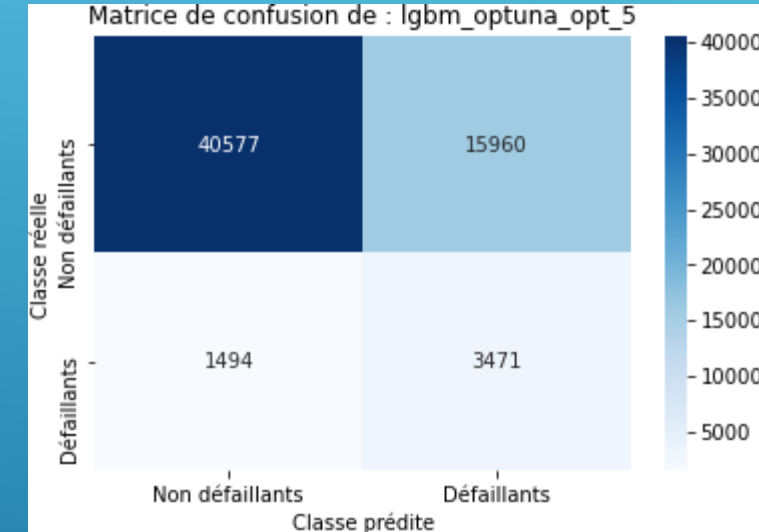
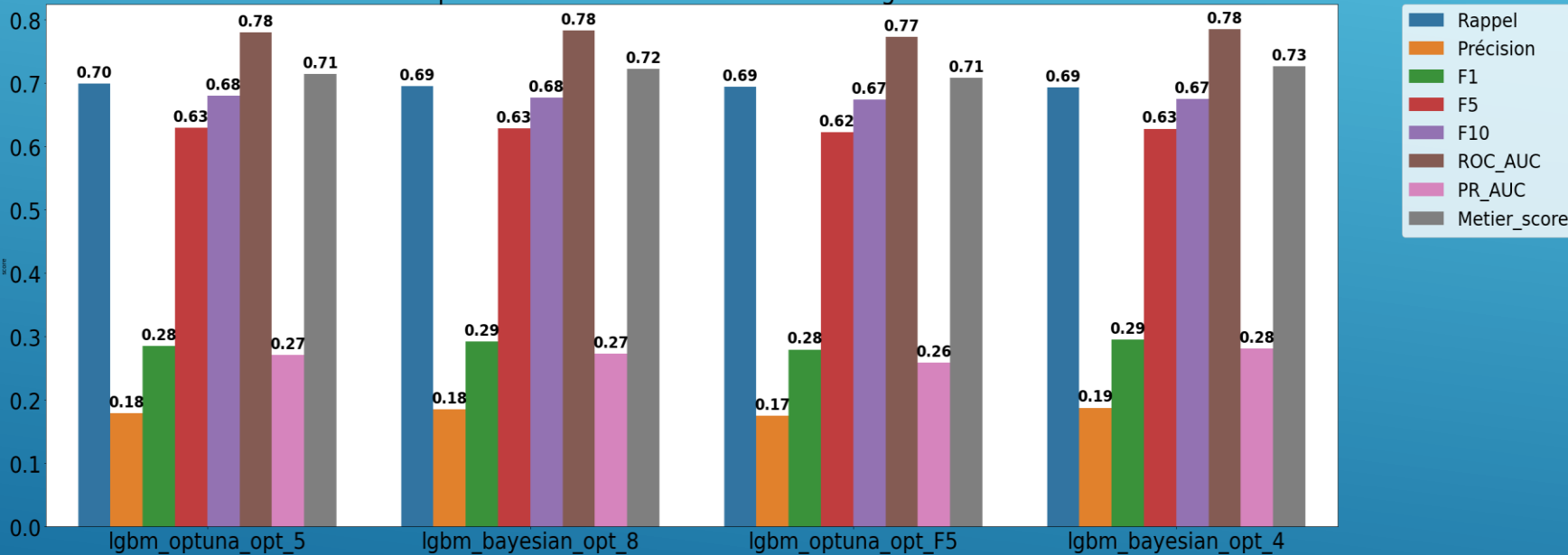
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.9179	0.7685	0.0683	0.4452	0.1184	0.0991	0.1498	65.995
xgboost	Extreme Gradient Boosting	0.9158	0.7562	0.0800	0.3930	0.1329	0.1086	0.1481	51.939
lightgbm	Light Gradient Boosting Machine	0.9162	0.7550	0.0503	0.3640	0.0883	0.0701	0.1104	10.222
rf	Random Forest Classifier	0.9124	0.7342	0.0586	0.2897	0.0974	0.0722	0.0987	49.923
gbc	Gradient Boosting Classifier	0.9019	0.7168	0.1046	0.2466	0.1468	0.1037	0.1146	125.717
et	Extra Trees Classifier	0.9018	0.7124	0.0947	0.2331	0.1347	0.0924	0.1030	42.072
ada	Ada Boost Classifier	0.8723	0.6979	0.1959	0.2010	0.1984	0.1290	0.1291	32.810
lda	Linear Discriminant Analysis	0.7316	0.6739	0.4904	0.1484	0.2278	0.1185	0.1498	7.882
knn	K Neighbors Classifier	0.6839	0.5661	0.3829	0.1040	0.1636	0.0419	0.0556	116.716
nb	Naive Bayes	0.1113	0.5592	0.9762	0.0816	0.1506	0.0019	0.0174	5.304
qda	Quadratic Discriminant Analysis	0.1459	0.5553	0.9502	0.0828	0.1523	0.0044	0.0266	10.489
dt	Decision Tree Classifier	0.8286	0.5473	0.2119	0.1370	0.1664	0.0758	0.0780	10.933
lr	Logistic Regression	0.6327	0.4720	0.1981	0.0948	0.1270	0.0455	0.0500	22.488
svm	SVM - Linear Kernel	0.5116	0.0000	0.5890	0.0956	0.1613	0.0286	0.0541	9.513
ridge	Ridge Classifier	0.7315	0.0000	0.4898	0.1482	0.2275	0.1182	0.1493	5.319

OPTIMISATION

MODÉLISATION – LES 4 MEILLEURS MODÈLES

Modèle	Jeu_donnees	FN	FP	TP	TN	Metrique	Optimisation	Class_weight	Rappel	Précision	F1	F5	F10	ROC_AUC	PR_AUC	Metier_score	D
lgbm_optuna_opt_5	train	1494	15960	3471	40577	F10	optuna	oui	0.6991	0.1786	0.2846	0.6286	0.6795	0.7795	0.2702	0.7135	
lgbm_bayesian_opt_8	train	1515	15278	3450	41259	F10	bayes_opt	oui	0.6949	0.1842	0.2912	0.6279	0.6763	0.7826	0.2728	0.7219	
lgbm_optuna_opt_F5	train	1522	16272	3443	40265	F5	optuna	non	0.6935	0.1746	0.2790	0.6223	0.6736	0.7727	0.2579	0.7079	
lgbm_bayesian_opt_4	train	1526	14937	3439	41600	roc_auc	bayes_opt	oui	0.6926	0.1871	0.2947	0.6275	0.6746	0.7843	0.2801	0.7260	

Comparaison des scores des 4 meilleurs algorithmes

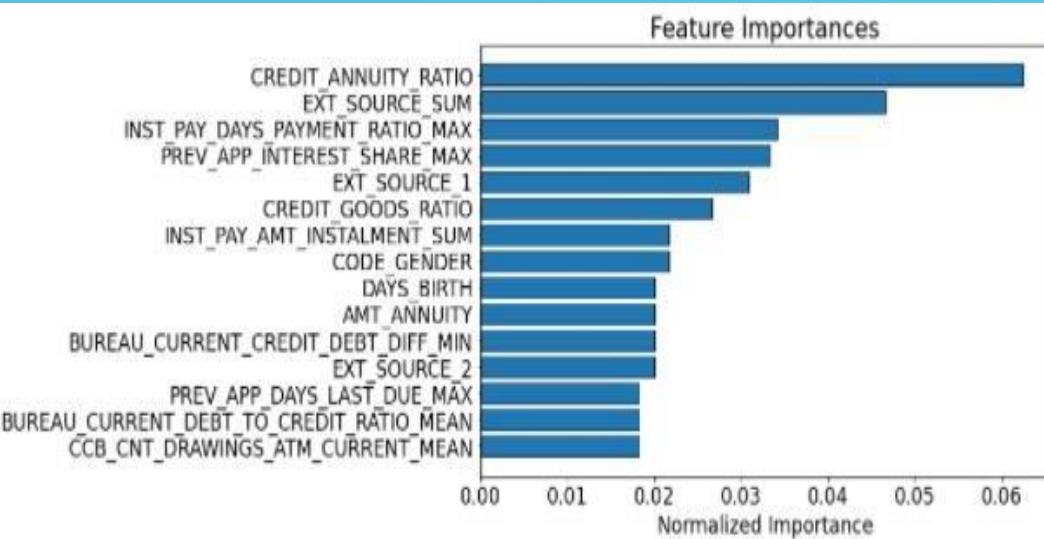


RETENU pour Dashboard

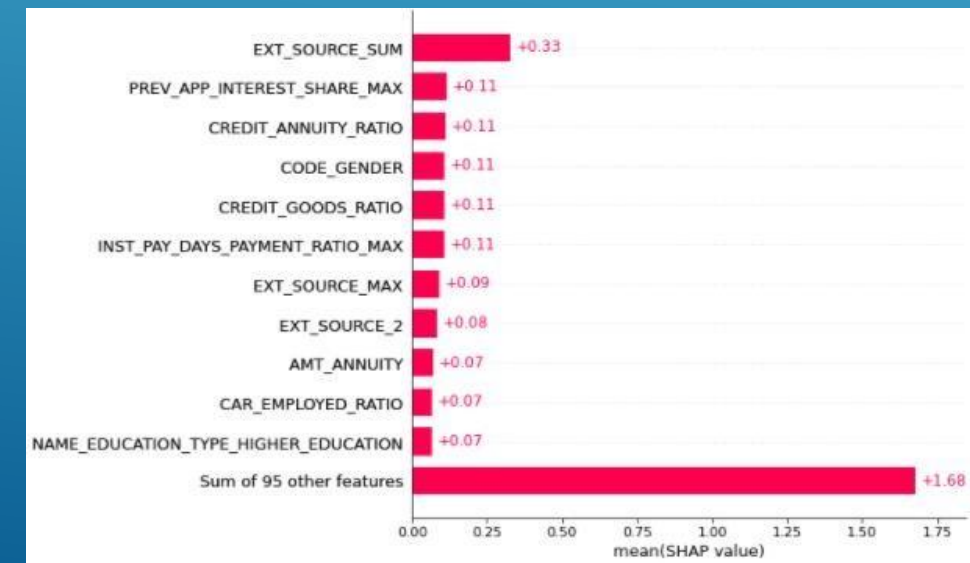
lgbm_otuna_opt_5

MODÉLISATION – INTERPRÉTABILITÉ

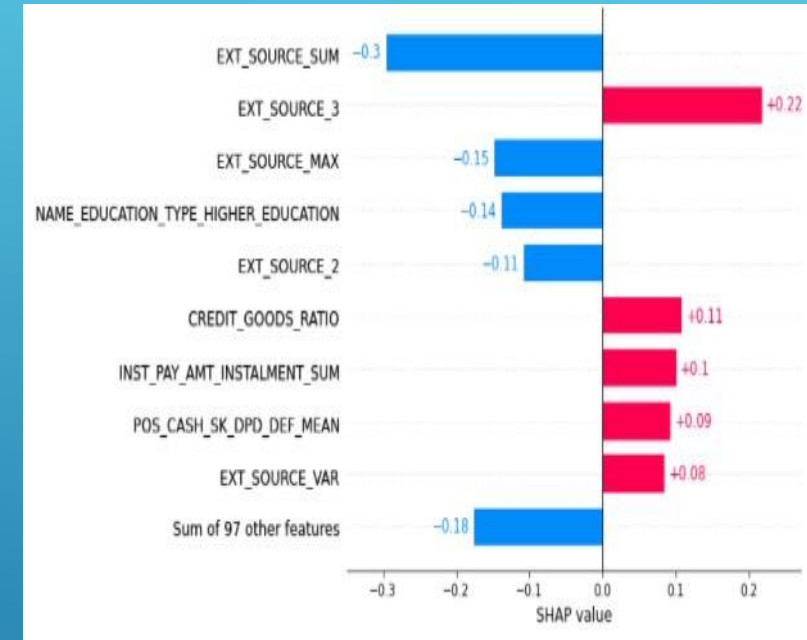
GLOBAL



Feature importance LightGBM



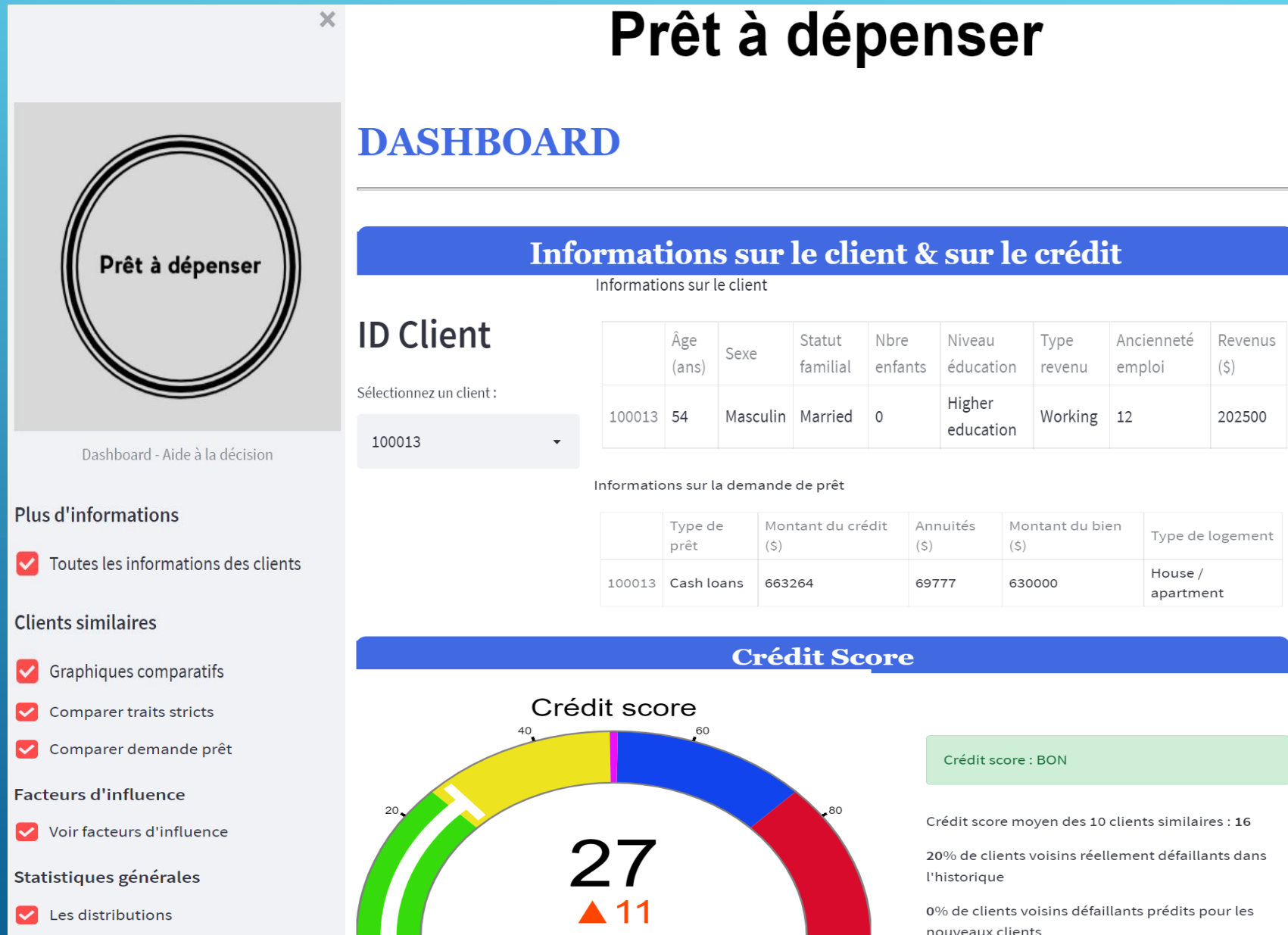
Shap values global



POUR UN CLIENT

- ❑ PROBLÉMATIQUE
- ❑ DONNÉES
- ❑ MODÉLISATION
- ❑ **DASHBOARD**
- ❑ CONCLUSION





- ❑ PROBLÉMATIQUE
- ❑ DONNÉES
- ❑ MODÉLISATION
- ❑ DASHBOARD
- ❑ CONCLUSION





Problématique de classification binaire avec classes déséquilibrées



Modèle final LightGBM optimisé avec optuna sur la métrique F10 interprétable



Autre modèle XGBoost (recall)



Optimisation des hyperparamètres de LightGBM



Échange avec notre client :

- variables métiers ajoutées adéquates?
- compromis FN/FP (seuil? Taux?)
- revoir métrique métier bancaire inefficace lors des essais
- source externe très influente mais ininterprétable



Dashboard : fonctions avancées d'optimisation de Streamlit (cache...) – phase de réentraînement du modèle



Éthique : sexe, âge, transparence, automatisation

**MERCI POUR VOTRE
ATTENTION !**