



22 Juillet 2022

SOUTENANCE DU PROJET 7: IMPLÉMENTER UN MODÈLE DE SCORING

Check KOUTAME

PLAN

I/ Projet et donnés

- Mission & Description du projet
- Observations des données

II/ Traitement des données

- Processus de traitements des données

III/ Modélisation

- Choix des métriques
- Entraînement et optimisation
- Analyse des résultats
- Interprétabilité

IV/ Dashboard

Conclusion

I. PROJET ET DONNÉES

MISSION & DESCRIPTION DU PROJET:



- **Problématique :**

- Société financière d'offre de crédit à la consommation pour la clientèle ayant peu ou pas d'historique de prêt
- Possession d'une base de données avec plusieurs variables concernant les clients
 - Attribution de prêt ou non: client défaillant ou non
 - Créer un algorithme pour ces attributions.
- Comment implémenter un modèle de scoring et le présenter avec un Dashboard pour la clientèle?

- **Mission :**

- Développer un modèle de Scoring de la probabilité de défaut de paiement du client pour étayer la décision d'accorder ou non un prêt à un client potentiel.
- Développement d'un Dashboard interactif pour que les chargés de relation client
 - Améliorer la relation avec le client en faisant preuve de transparence.
 - Montrer au client les informations le concernant grâce à l'interactivité.

- **Contrainte:** utiliser un Kernel Kaggle pour l'analyse exploratoire et du pre-processing

I. PROJET ET DONNÉES: OBSERVATION DES DONNÉES

application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

Informations principales
concernant les clients

Pas de Target pour le test set!

bureau.csv

- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

SK_ID_BUREAU

bureau_balance.csv

- Monthly balance of credits in Credit Bureau
- Behavioral data

previous_application.csv

- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

SK_ID_CURR

SK_ID_PREV

SK_ID_PREV

POS_CASH_balance.csv

- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv

- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

credit_card_balance.csv

- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

- Informations anonymes
- 307511 clients
- 218 variables

Informations d'autres organismes

Informations recueillies auprès de Homme Crédit group

I. PROJET ET DONNÉES: OBSERVATION DES DONNÉES

| | Données | Dimension | Nombre de variables | Nombre observations | Nombres de types de variables | Nombre de cellules manquantes | % de cellules manquantes | Nombre de lignes dupliquées | % de lignes dupliquées |
|---|-------------------------------|---------------|---------------------|---------------------|--|-------------------------------|--------------------------|-----------------------------|------------------------|
| 0 | [application_train, data] | (307511, 122) | 122 | 307511 | float64 65 int64 41 object 16 dtype: ... | 9152465 | 0.243959 | 0 | 0.0 |
| 1 | [application_test, data] | (48744, 121) | 121 | 48744 | float64 65 int64 40 object 16 dtype: ... | 1404419 | 0.238117 | 0 | 0.0 |
| 2 | [bureau, data] | (1716428, 17) | 17 | 1716428 | float64 8 int64 6 object 3 dtype: ... | 3939947 | 0.135026 | 0 | 0.0 |
| 3 | [bureau_balance, data] | (27299925, 3) | 3 | 27299925 | int64 2 object 1 dtype: int64 | 0 | 0.000000 | 0 | 0.0 |
| 4 | [cc_balance, data] | (3840312, 23) | 23 | 3840312 | float64 15 int64 7 object 1 dtype: ... | 5877356 | 0.066541 | 0 | 0.0 |
| 5 | [installments_payments, data] | (13605401, 8) | 8 | 13605401 | float64 5 int64 3 dtype: int64 | 5810 | 0.000053 | 0 | 0.0 |
| 6 | [POS_CASH_balance, data] | (10001358, 8) | 8 | 10001358 | int64 5 float64 2 object 1 dtype: ... | 52158 | 0.000652 | 0 | 0.0 |
| 7 | [previous_application, data] | (1670214, 37) | 37 | 1670214 | object 16 float64 15 int64 6 dtype: ... | 11109336 | 0.179769 | 0 | 0.0 |

II.TRAITEMENT DES DONNÉES: PROCESSUS

- Utilisation du Kernel Kaggle de [Rishabh RAO](#)

Process des 8 fichiers

Analyse exploratoire

- Sur l'ensemble des fichiers
- Utilisation du kernel Kaggle

Pré-traitement

- Changement du type de données(Yes/Non→ 1/0)
- Réduction de la mémoire des données
- Traitement des valeurs aberrantes trouvées pendant l'EDA
- Imputations

Feature Engineering

- Création de variables statistique et métier
- Encodage + merging
- Suppression des variables corrélées et > 90% de V.manquantes
- Feature selection LighGBM, RFECV, Boruta...
- Selection des variables les plus fréquences

Modélisation

- Utilisation de Pycarest
- Choix de la métrique
- Optimisation du modèle
- Seuil de probabilité optimale

Dashboard

- Développement & déploiement sur Streamlit
- API utilisant mlflow

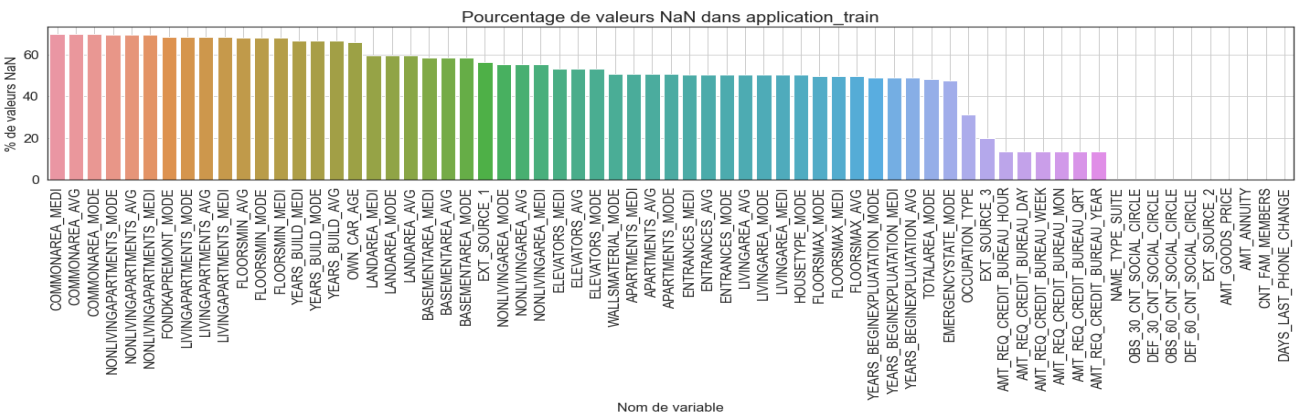
II.TRAITEMENT DES DONNÉES: ANALYSE EXPLORATOIRE

- Analyse exploratoire sur l'ensemble **des variables de tous les 8 fichiers**
 - L'objectif est de voir quelles sont les variables qui présentent une grande variabilité

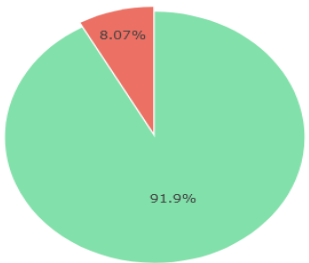
Exemple du fichier application_train

| | SK_ID_CURR | TARGET | NAME_CONTRACT_TYPE | CODE_GENDER | FLAG_OWN_CAR | FLAG_OWN_REALTY | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | |
|---|------------|--------|--------------------|-------------|--------------|-----------------|--------------|------------------|------------|--|
| 0 | 100002 | 1 | Cash loans | M | N | Y | 0 | 202500.0 | 406597.5 | |
| 1 | 100003 | 0 | Cash loans | F | N | N | 0 | 270000.0 | 1293502.5 | |
| 2 | 100004 | 0 | Revolving loans | M | Y | Y | 0 | 67500.0 | 135000.0 | |

3 rows × 122 columns



| | Nombre par type de variable | % des types de variable |
|---------|-----------------------------|-------------------------|
| float64 | 65 | 53.280000 |
| int64 | 41 | 33.610000 |
| object | 16 | 13.110000 |



Target: 8% défaillant – 92% Non-défaillant

67 variables sur 122 contiennent des valeurs manquantes

II. TRAITEMENT DES DONNÉES: PRE-PROCESSING

- Utilisation du Kernel Kaggle de [Rishabh RAO](#)
- Sur seulement les fichiers `application_train` & `application_test`
- **Changement du type de données**
 - Homme/Femme ou Yes/Non → I/O
 - Réduction de la taille en mémoire des données (par exemple int32 en int8)
- **Valeurs aberrantes**
 - Correction des valeurs aberrantes trouvées pendant l'EDA
 - Correction/Suppression des valeurs uniques
- **Valeurs manquantes & imputation**
 - Suppression des Nan des variables ayant plus de 67%
 - Imputation en faisant 3 tests
 - Test I: imputation par la médiane sur les V. quantitative, et par la mode pour les V. catégorielles
 - Test II: imputation par 0 sur les V. quantitatives, et par la XNA pour les V. catégorielles
 - Test III: imputation par un algorithme NaNimputer pour les V. quantitatives et par XNA pour les V. catégorielles

II. TRAITEMENT DES DONNÉES: FEATURE ENGINEERING

- Utilisation du Kernel Kaggle de [Rishabh RAO](#)
- Sur l'ensemble des fichiers

➤ Ajout de variables métiers

- Revenu, de rente et de crédit : ratio/différence
- Jours en années, changement de jours : ratio
- Âge de la voiture, ancienneté d'emploi : ratio/différence
- Flag sur les téléphones : ratio/différence
- Membres de la famille : ratio/différence
- Note de la région où vit le client : ratio/différence
- Données externes : ratio, moyenne, max, min
- Informations sur le bâtiment : somme, multiplication
- Défauts de paiements et les défauts observables : somme/ratio
- Flag sur les documents : somme, moyenne, variance, écart-type
- Modification du demandeur : somme/ratio

➤ Ajout de variables statistiques

- Quantitatives: min, max, sum,
- Qualitatives : sum, mean, count

II. TRAITEMENT DES DONNÉES: FEATURE ENGINEERING: MERGING

| Dataframe initial | Nbr lignes var. initiales | Nbr lignes var. après suppression des variables corrélées | 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) |

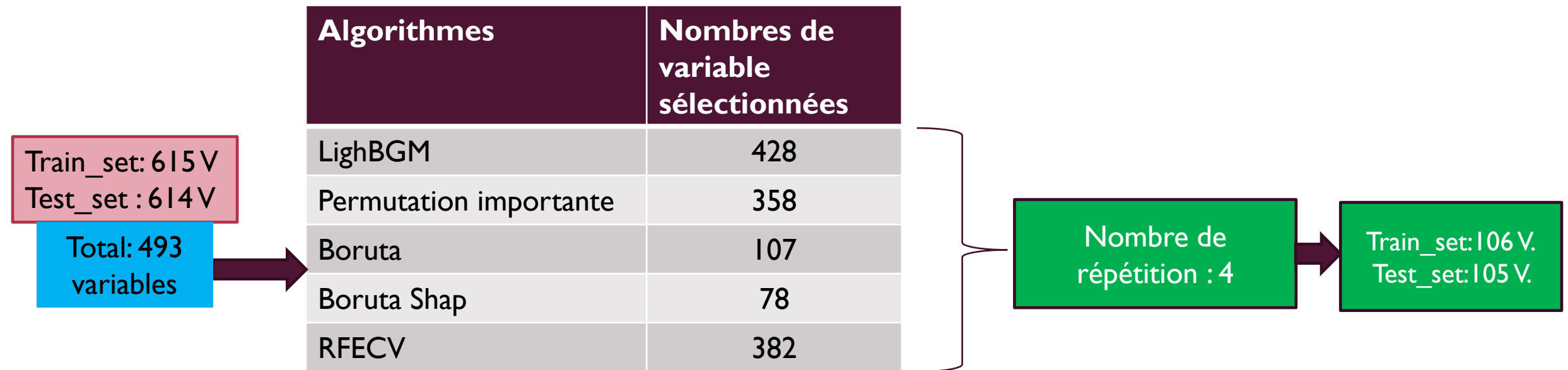
Cellule finale: train_set: (307507, **615**)
test_set (48744, **614**)

Total: 493 variables

Nécessité d'une feature selection

II. TRAITEMENT DES DONNÉES: FEATURE ENGINEERING: FEATURE SELECTION

- Utilisation de plusieurs algorithmes de sélection de variables importantes
- Sélection finale basée sur le nombre de répétition d'une variable dans tous les algorithmes



III. MODÉLISATION MÉTHODOLOGIE

- La variable cible: binaire
 - Client défaillants: 8%
 - Client non défaillants: 92%

Modélisation

- Classification binaire
 - Client défaillants: classe 1
 - Client non défaillants: classe 0

Méthodologie

- **Pycaret:** Idée des différents algorithmes de classification les plus performants
- **Equilibrer les données:** SMOTE
- **Choix des métriques:** Précision, Recall, Fbeta, métrique métier, etc...
- **Optimisation des modèles :** modèles bayésiens
- **Choix du modèle final & Seuil de probabilité optimale**

III. MODÉLISATION

CHOIX DES MÉTRIQUES

Matrice de confusion

| Réelles | + | TP : Vrais Positifs | FN : Faux Négatifs |
|---------|---|-------------------------------|-------------------------------|
| | - | FP : Faux positifs | TN : Vrais Négatifs |
| | | + | - |
| | | Prédictions | |

- De ne pas prédire « défaillant » un client non défaillant : minimiser les faux positifs (erreur de type I: Il convient donc de maximiser la métrique Précision
- De ne pas prédire un client non-défaillant s'il est défaillant : minimiser le nombre de faux négatifs (erreur de type II): Dans notre cas, il convient donc de maximiser les métriques Recall ou FBeta 10.

$$\text{Précision} = \frac{TP}{TP+FP} \quad \text{Recall} = \frac{TP}{TP+FN}$$

$$F_{\beta}\text{-score} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Fonction Coût

$$J = TP * TP_{value} + TN * TN_{value} + FP * FP_{value} + FN * FN_{value}$$

- Ces valeurs de coefficients signifient que les Faux Négatifs engendrent des pertes 10 fois plus importantes que les gains des Vrai Négatifs

| | |
|-----------------|------|
| TP_value | :0 |
| FN_value | :-10 |
| TN_value | :1 |
| FP_value | :0 |

III. MODÉLISATION

MODÉLISATION: PYCARET

- Objectif: Avoir une idée de plusieurs algorithmes simultanément
- Possibilité de rééquilibrer les variables cibles

| | 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 | 65.9950 |
| xgboost | Extreme Gradient Boosting | 0.9158 | 0.7562 | 0.0800 | 0.3930 | 0.1329 | 0.1086 | 0.1481 | 51.9390 |
| lightgbm | Light Gradient Boosting Machine | 0.9162 | 0.7550 | 0.0503 | 0.3640 | 0.0883 | 0.0701 | 0.1104 | 10.2220 |
| rf | Random Forest Classifier | 0.9124 | 0.7342 | 0.0586 | 0.2897 | 0.0974 | 0.0722 | 0.0987 | 49.9230 |
| gbc | Gradient Boosting Classifier | 0.9019 | 0.7168 | 0.1046 | 0.2466 | 0.1468 | 0.1037 | 0.1146 | 125.7170 |
| et | Extra Trees Classifier | 0.9018 | 0.7124 | 0.0947 | 0.2331 | 0.1347 | 0.0924 | 0.1030 | 42.0720 |
| ada | Ada Boost Classifier | 0.8723 | 0.6979 | 0.1959 | 0.2010 | 0.1984 | 0.1290 | 0.1291 | 32.8100 |
| lda | Linear Discriminant Analysis | 0.7316 | 0.6739 | 0.4904 | 0.1484 | 0.2278 | 0.1185 | 0.1498 | 7.8820 |
| knn | K Neighbors Classifier | 0.6839 | 0.5661 | 0.3829 | 0.1040 | 0.1636 | 0.0419 | 0.0556 | 116.7160 |
| nb | Naive Bayes | 0.1113 | 0.5592 | 0.9762 | 0.0816 | 0.1506 | 0.0019 | 0.0174 | 5.3040 |
| qda | Quadratic Discriminant Analysis | 0.1459 | 0.5553 | 0.9502 | 0.0828 | 0.1523 | 0.0044 | 0.0266 | 10.4890 |

- ✓ Choix des données du Test I
- ✓ Choix du les 3 modèles ensemblistes
 - ✓ Choix porté sur LightGBM
 - ✓ Plus rapide...

Rappel:

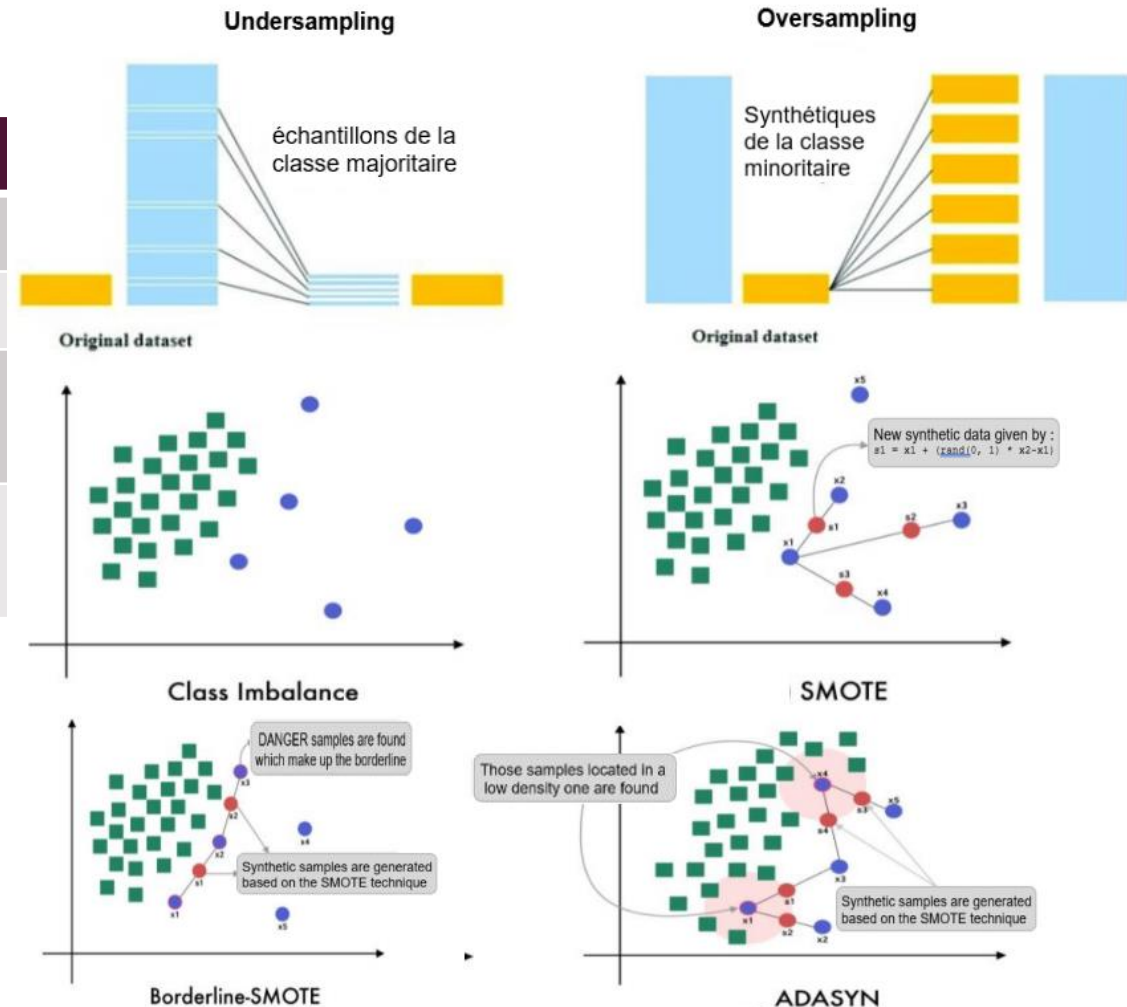
- Test I: imputation par la médiane sur les V. quantitative, et par la mode pour les V. catégorielles
- Test II: imputation par 0 sur les V. quantitatives, et par la XNA pour les V. catégorielles
- Test III: imputation par un algorithme NaNimputer pour les V. quantitatives et par XNA pour les V. catégorielles

III. MODÉLISATION

MODÉLISATION: RÉÉQUILIBRAGE - SMOTE

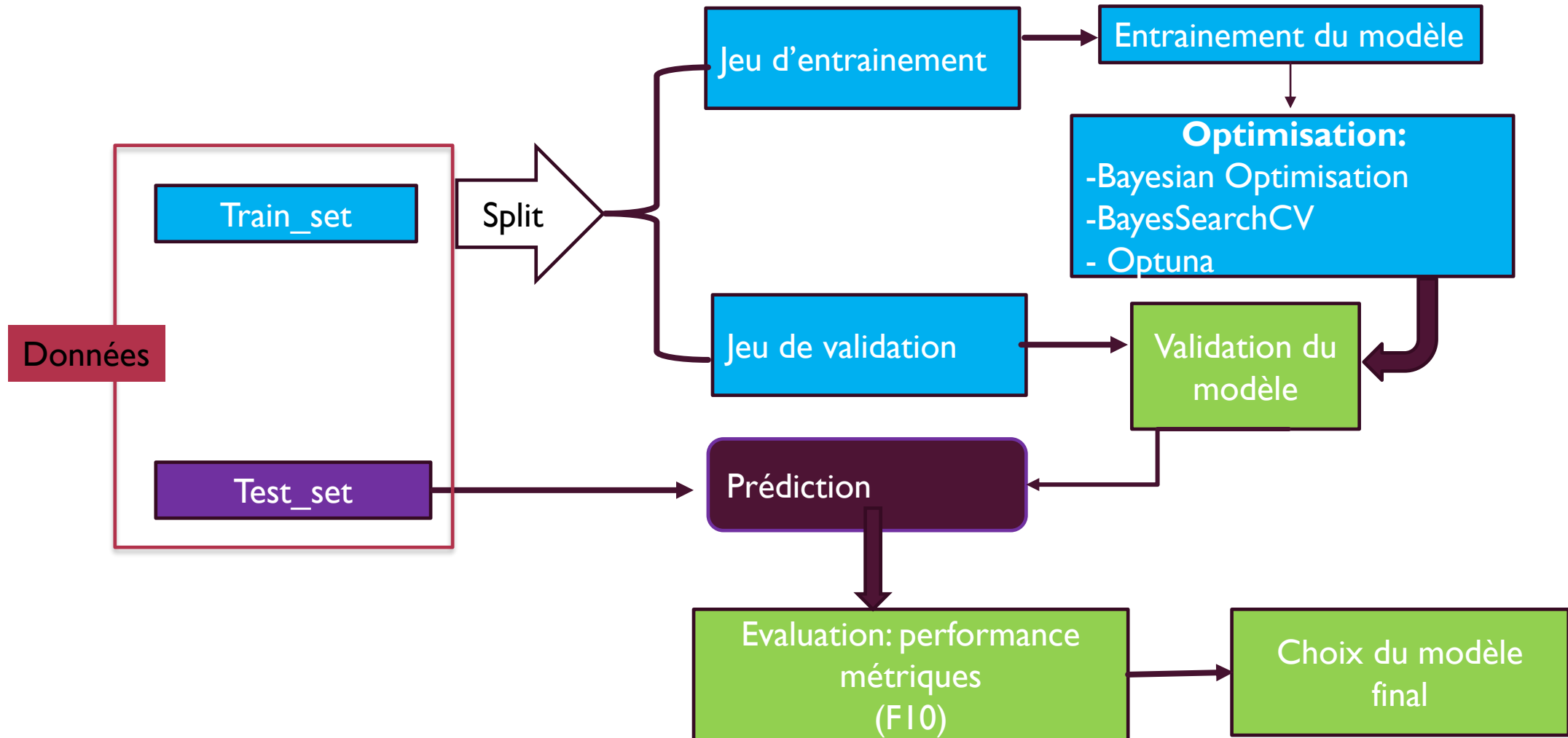
Techniques de rééquilibrage

| | |
|--------------------------------|-------------------------------|
| LightGBM class_weight | class_weight |
| Undersampling | SMOTE |
| Oversampling | SMOTE, BordelineSMOTE, ADASYN |
| Oversampling + d'Undersempling | SMOTE |



III. MODÉLISATION

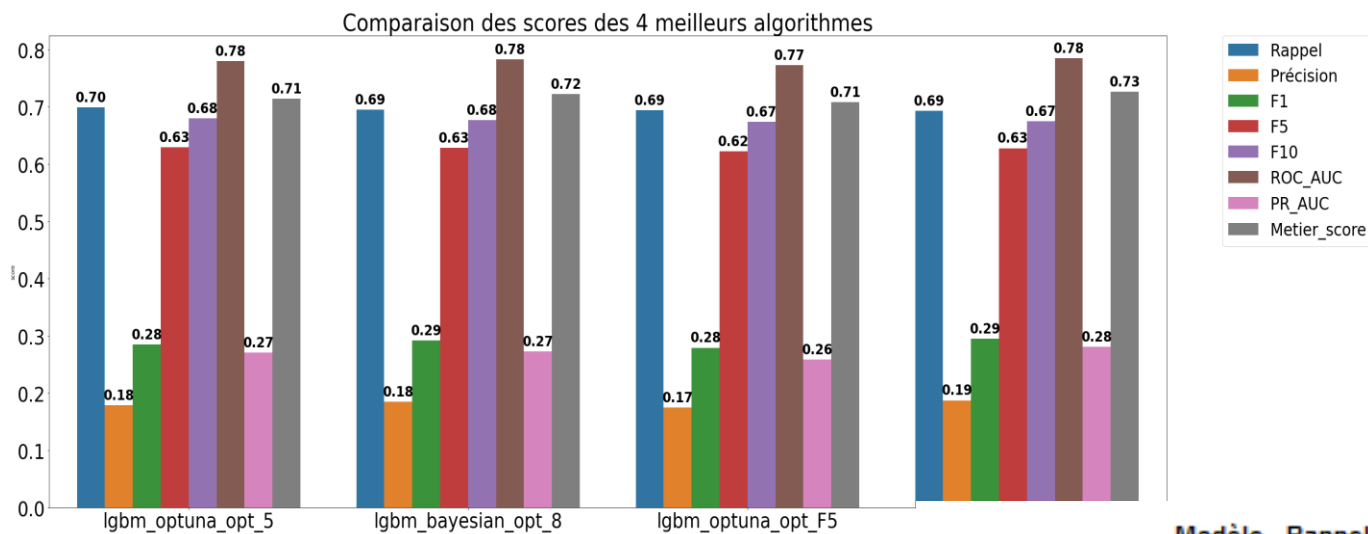
MODÉLISATION: OPTIMISATION



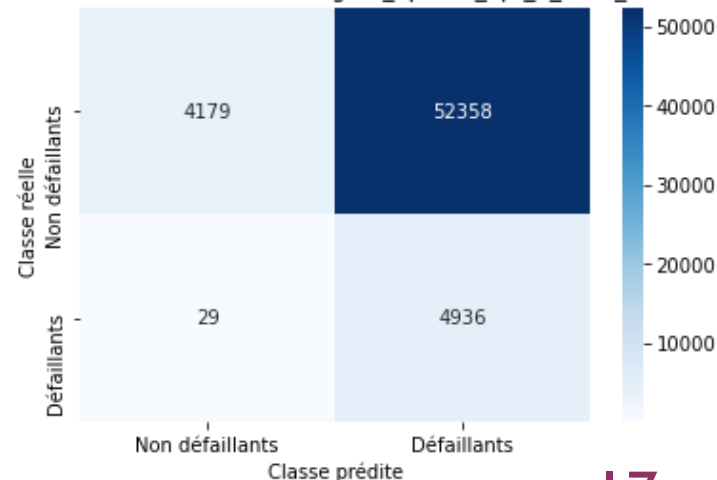
III. MODÉLISATION

MODÉLISATION: BILAN DES MEILLEURS MODÈLES

| Modèle | Jeu_donnees | FN | FP | TP | TN | Metrique | Optimisation | Class_weight | Rappel | Précision | F1 | F5 | F10 | ROC_AUC | PR_AUC | Metier_score |
|---------------------|-------------|------|-------|------|-------|----------|--------------|--------------|--------|-----------|--------|--------|--------|---------|--------|--------------|
| 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 |



Matrice de confusion de : lgbm_optuna_opt_5_seuil_1



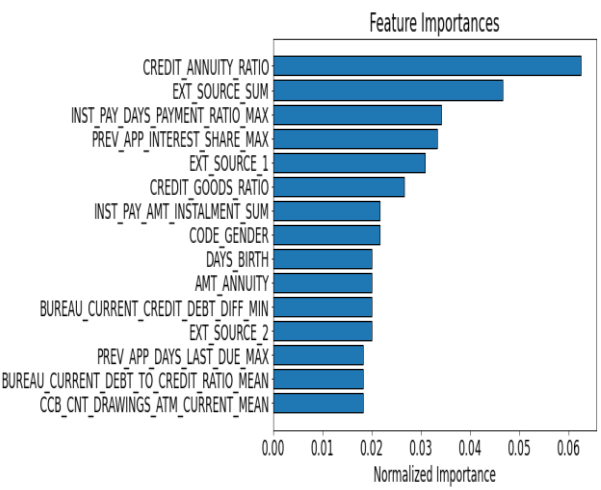
✓ **Modèle retenu: lgbm_optuna_5** ➡

| Modèle | Rappel | Précision | F1 | F5 | F10 | ROC_AUC | PR_AUC | Metier_score | Durée_train |
|---------------------------|--------|-----------|--------|--------|--------|---------|--------|--------------|-------------|
| lgbm_optuna_opt_5_seuil_1 | 0.9942 | 0.0862 | 0.1586 | 0.7074 | 0.9002 | 0.7795 | 0.2702 | 0.2823 | 8.4321 |

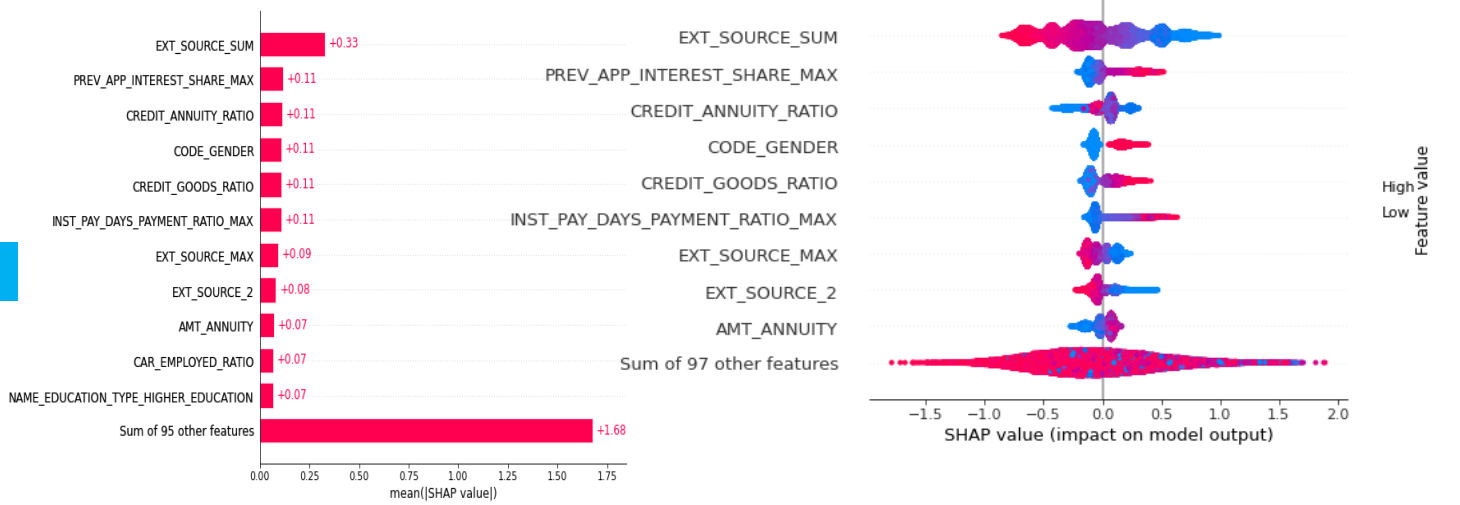
III. MODÉLISATION

MODÉLISATION: INTERPRÉTABILITÉ

Modèle lightGBM

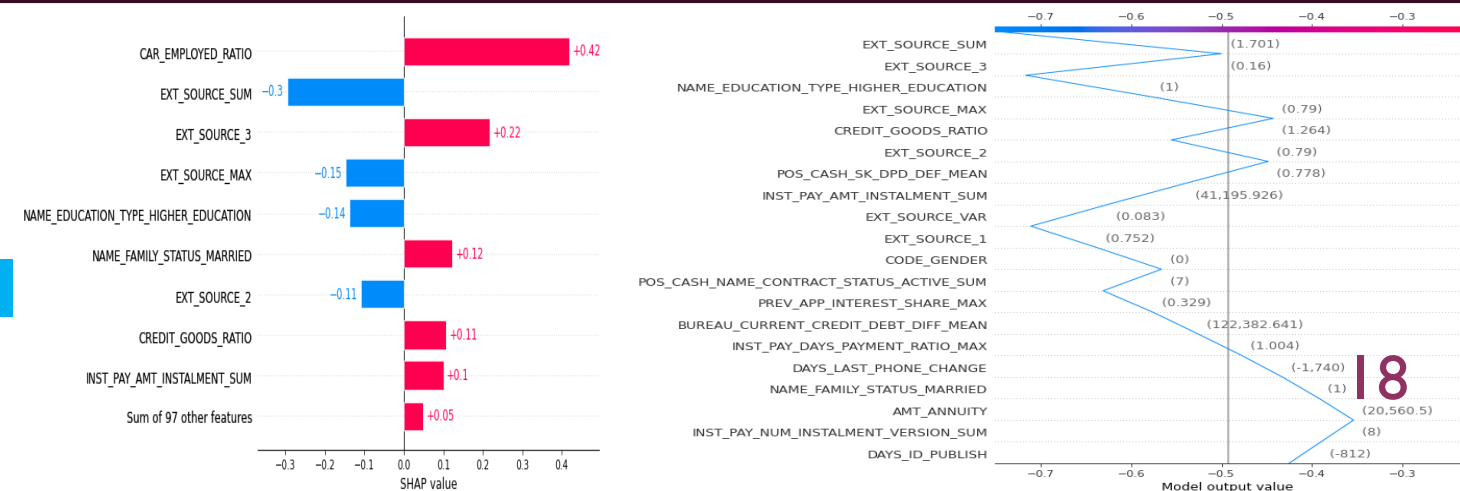


Interprétabilité globale



SHAP

Interprétabilité locale



Client: I00001

IV. DASHBOARD STREAMLIT + MLFLOW

Streamlit

Streamlit: DashBoard

Requête (input: variables importantes)

mlflow 1.27.0

API: mlflow

Réponse (Output: Prédiction:)

mlflow 1.27.0 Experiments Models GitHub Docs

Experiments + < Default

Search Experiments

Default

Track machine learning training runs in experiments. [Learn more](#)

Experiment ID: 0

Description [Edit](#)


Refresh Compare Delete Download CSV Source All time

Columns Only show differences metrics.rmse < 1 and params.model = "tree" Search Filter Clear

Showing 8 matching runs

| | Start Time | Duration | User | Run Name | Source | Version | Tags | Linked Models | Parameters |
|--------------------------|---------------------|----------|---------|----------|--------|---------|------|---------------|------------|
| <input type="checkbox"/> | 2022-07-19 10:21:27 | 5.6s | koutame | | API.py | | | | |
| <input type="checkbox"/> | 2022-07-18 09:18:21 | 5.8s | koutame | | API.py | | | | |

IV. DASHBOARD STREAMLIT



Prêt à dépenser

Dashboard - Aide à la décision

Plus infos

- ☒ Voir toutes infos clients ?

Clients similaires

- ☒ Graphiques comparatifs
- ☒ Comparer traits stricts ?
- ☒ Comparer demande prêt ?

Facteurs d'influence

- ☒ Voir facteurs d'influence

Stats générales

- ☒ Voir les distributions

Prêt à dépenser

DASHBOARD

Informations sur le client / demande de prêt

ID Client

Sélectionnez un client :

100001

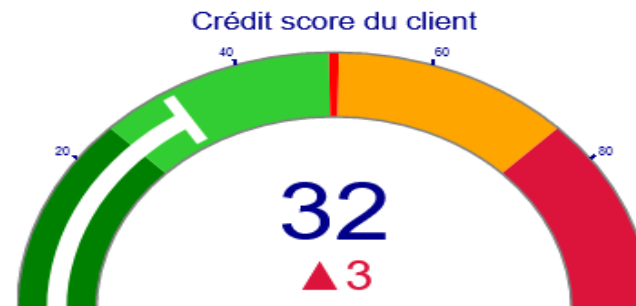
| | Âge (ans) | Sexe | Statut familial | Nbre enfants | Niveau éducation | Type revenu | Ancienneté emploi | Revenus (\$) |
|--------|-----------|---------|-----------------|--------------|------------------|-------------|-------------------|--------------|
| 100001 | 52 | Féminin | Married | 0 | Higher education | Working | 6 | 135000 |

| | Type de prêt | Montant du crédit (\$) | Annuités (\$) | Montant du bien (\$) | Type de logement |
|--------|--------------|------------------------|---------------|----------------------|-------------------|
| 100001 | Cash loans | 568800 | 20560.5 | 450000 | House / apartment |

Crédit Score

Prédire

Non défaillant




Crédit score : BON

Crédit score moyen des 10 clients similaires : 29

20% de clients voisins réellement défaillants dans l'historique

IV. DASHBOARD STREAMLIT



Dashboard - Aide à la décision

Plus infos

☒ Voir toutes infos clients ?

Clients similaires

☐ Graphiques comparatifs

☐ Comparer traits stricts ?

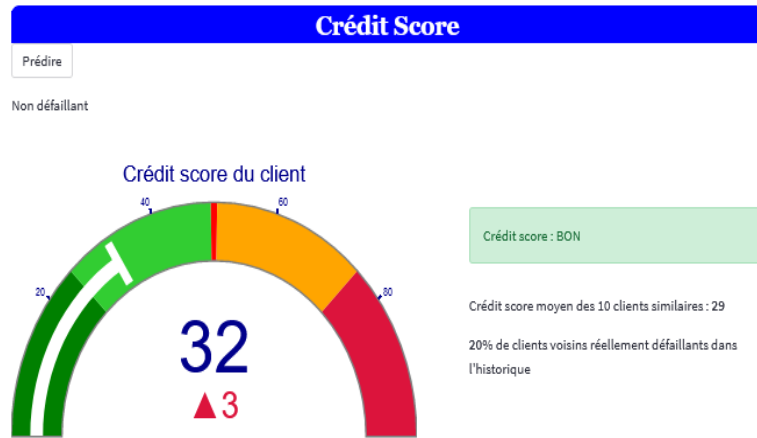
☐ Comparer demande prêt ?

Facteurs d'influence

☐ Voir facteurs d'influence

Stats générales

☐ Voir les distributions




20% de clients voisins défaillants prédits pour les nouveaux clients

Plus infos

Toutes les informations du client courant

| | SK_ID_CURR | NAME_CONTRACT_TYPE | CODE_GENDER | FLAG_OWN_CAR | FLAG_OWN_REALTY | CNT_CHILDREN | AMT_INCOME_TOTAL |
|---|------------|--------------------|-------------|--------------|-----------------|--------------|------------------|
| 0 | 100001 | Cash loans | F | N | Y | 0 | 135000 |

| | SK_ID_CURR | CREDIT_ANNUITY_RATIO | PREV_APP_AMT_ANNUITY_MEAN | PREV_APP_DAYS_FIRST_DUE_SUM | PREV_APP_AMT_ANNUITY |
|---|------------|----------------------|---------------------------|-----------------------------|----------------------|
| 0 | 100001 | 27.671875 | 3951 | -1709 | 3951 |



Dashboard - Aide à la décision

Plus infos

☒ Voir toutes infos clients ?

Clients similaires

☒ Graphiques comparatifs

☐ Comparer traits stricts ?

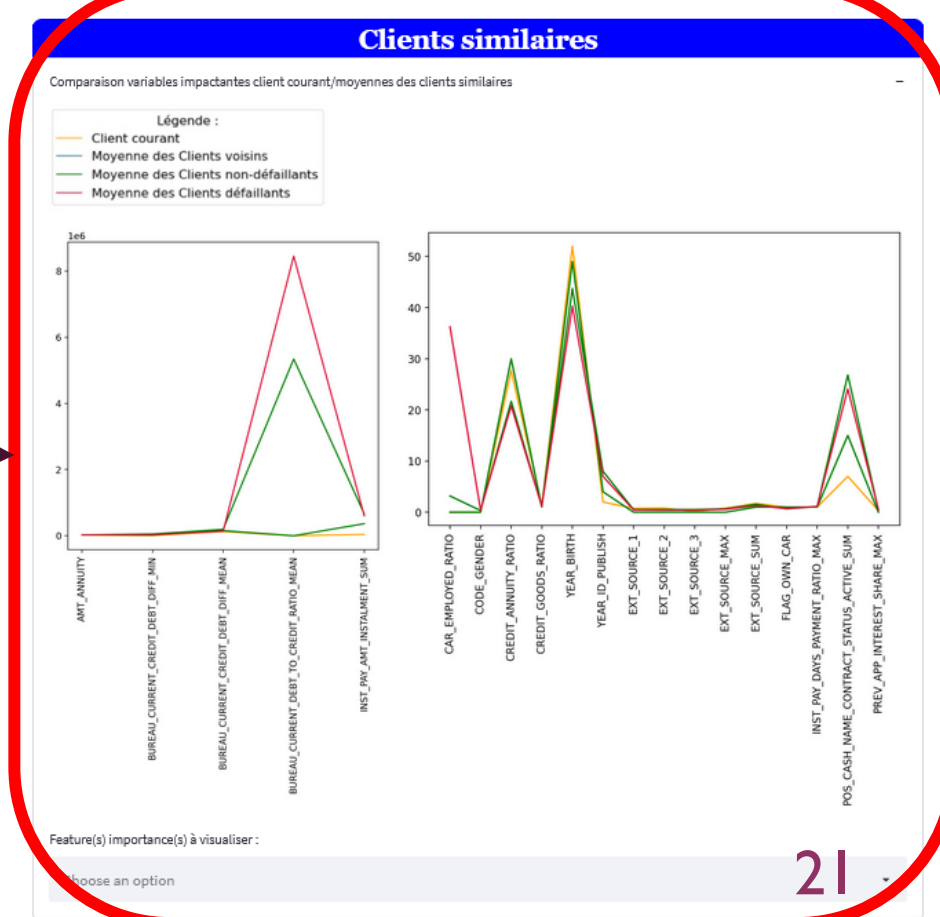
☐ Comparer demande prêt ?

Facteurs d'influence

☐ Voir facteurs d'influence

Stats générales

☐ Voir les distributions



IV. DASHBOARD STREAMLIT



Dashboard - Aide à la décision

Plus infos

☒ Voir toutes infos clients ?

Clients similaires

☒ Graphiques comparatifs

☒ Comparer traits stricts ?

☐ Comparer demande prêt ?

Facteurs d'influence

☐ Voir facteurs d'influence

Stats générales


☐ Voir les distributions

Client courant

| | Âge (ans) | Sexe | Statut familial | Nbre enfants | Niveau éducation | Type revenu |
|--------|-----------|---------|-----------------|--------------|------------------|-------------|
| 100001 | 52 | Féminin | Married | 0 | Higher education | Working |

10 clients similaires

| | Âge | Sexe | Statut familial | Nbre | Niveau éducation | Type revenu | A |
|--------|-----|---------|-----------------|------|------------------|----------------------|---|
| 77677 | 42 | Féminin | Married | 0 | Higher education | Working | 1 |
| 257447 | 47 | Féminin | Married | 0 | Higher education | State servant | 1 |
| 109458 | 48 | Féminin | Married | 0 | Higher education | Working | 1 |
| 212270 | 48 | Féminin | Married | 1 | Higher education | Commercial associate | 2 |
| 139230 | 51 | Féminin | Married | 0 | Higher education | Working | 3 |
| 97944 | 54 | Féminin | Married | 0 | Higher education | Pensioner | 0 |
| 229248 | 53 | Féminin | Married | 0 | Higher education | State servant | 5 |
| 213485 | 51 | Féminin | Married | 0 | Higher education | Working | 2 |
| 293935 | 44 | Féminin | Married | 0 | Higher education | Working | 5 |
| 181115 | 50 | Féminin | Married | 0 | Higher education | Commercial associate | 1 |



Dashboard - Aide à la décision

Plus infos

☒ Voir toutes infos clients ?

Clients similaires

☒ Graphiques comparatifs

☒ Comparer traits stricts ?

☒ Comparer demande prêt ?

Facteurs d'influence

☐ Voir facteurs d'influence

Stats générales

☐ Voir les distributions

| | | | | | | | | |
|--------|----|---------|---------|---|------------------|----------------------|---|---------------|
| 293935 | 44 | Féminin | Married | 0 | Higher education | Working | 5 | 153000.000000 |
| 181115 | 50 | Féminin | Married | 0 | Higher education | Commercial associate | 1 | 495000.000000 |

Comparaison demande de prêt

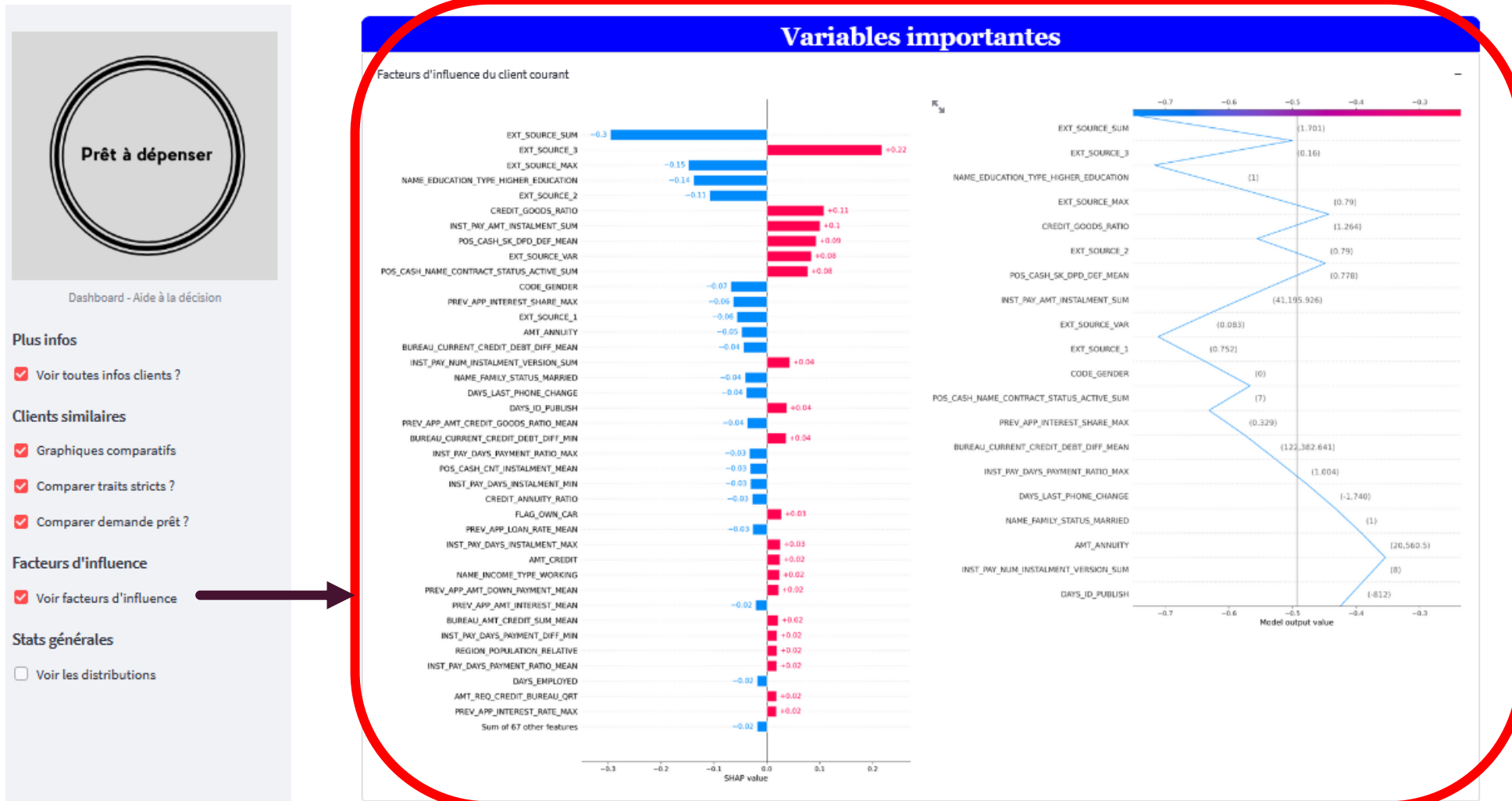
Client courant

| | Type de prêt | Montant du crédit (\$) | Annuités (\$) | Montant du bien (\$) | Type de logement |
|--------|--------------|------------------------|---------------|----------------------|-------------------|
| 100001 | Cash loans | 568800 | 20560.5 | 450000 | House / apartment |

10 clients similaires

| | Type de prêt | Montant du crédit (\$) | Annuités (\$) | Montant du bien (\$) | Type de logement |
|--------|--------------|------------------------|---------------|----------------------|-------------------|
| 77677 | Cash loans | 665892.000000 | 21609.000000 | 477000.000000 | House / apartment |
| 257447 | Cash loans | 905688.000000 | 29214.000000 | 756000.000000 | House / apartment |
| 109458 | Cash loans | 526491.000000 | 19039.500000 | 454500.000000 | House / apartment |
| 212270 | Cash loans | 1226511.000000 | 35860.500000 | 1071000.000000 | House / apartment |
| 139230 | Cash loans | 1298655.000000 | 35842.500000 | 1134000.000000 | House / apartment |
| 97944 | Cash loans | 781920.000000 | 23706.000000 | 675000.000000 | House / apartment |
| 229248 | Cash loans | 545040.000000 | 20677.500000 | 450000.000000 | House / apartment |
| 213485 | Cash loans | 840951.000000 | 33480.000000 | 679500.000000 | House / apartment |
| 293935 | Cash loans | 545040.000000 | 17712.000000 | 450000.000000 | House / apartment |
| 181115 | Cash loans | 1206954.000000 | 34717.500000 | 945000.000000 | House / apartment |

IV. DASHBOARD STREAMLIT



IV. DASHBOARD STREAMLIT



CONCLUSION

- Classification binaire avec variable cible déséquilibrée: utilisation de SMOTE
- Utilisation du modèle final LightGBM optimisé sur la métrique F10
- Utilisation de SHAP pour l'interprétabilité globale et locale
- API: utilisation mlflow
- Dashboard: utilisation de Streamlit

Pour aller plus loin...:

- Amélioration avec les feedbacks des experts + les utilisateurs
- Utilisation de métriques d'experts métiers
- Utilisation de cluster de calcul sur le cloud pour utiliser les modèles CatBoost & XGBoost

GitHub: <https://github.com/chkoutam/Implementer-un-modele-de-scoring>

Wabapp: <https://chkoutam-implementer-un-modele-de-scoring-dashboard-bis-h429gc.streamlitapp.com/>