Implémentez un modèle de scoring

Formation Data scientist
Projet n°7
Bayram DONAT

Sommaire

- Problématique
- Présentation du jeu de données
- Approche de modélisation
- Présentation du dashboard

Problématique

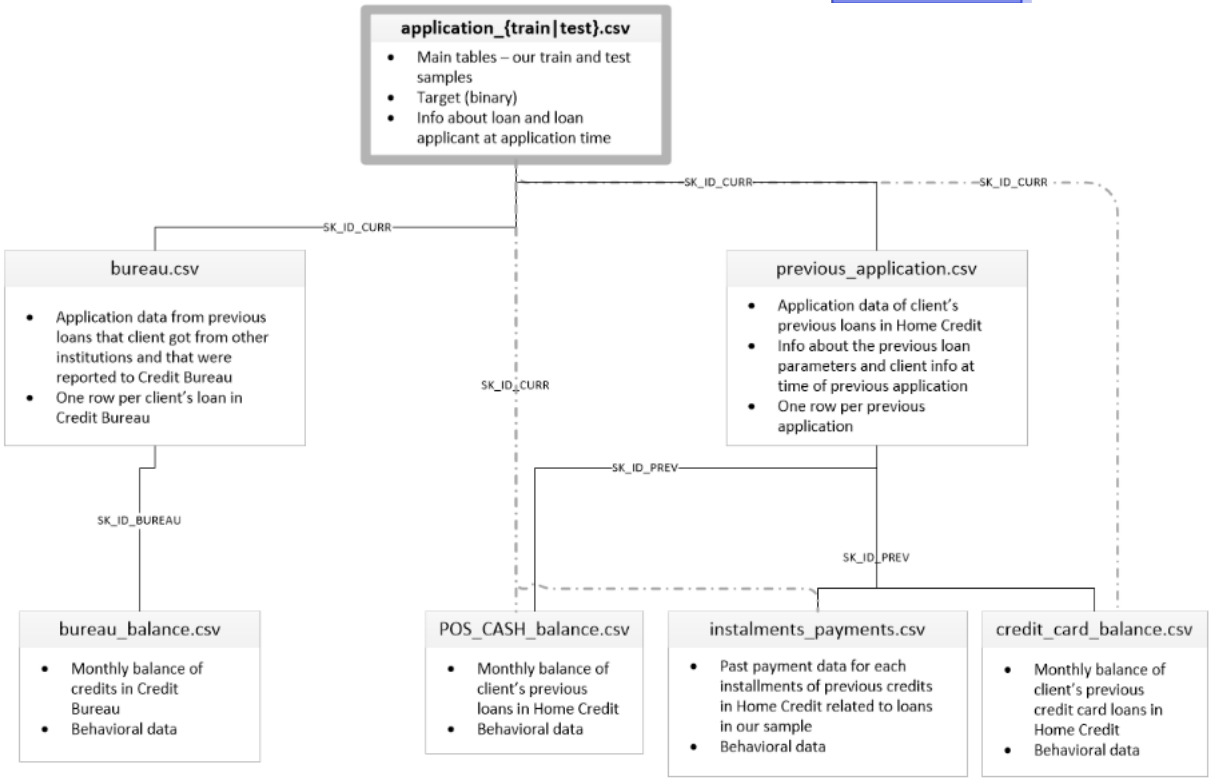
- Contexte
  - Entreprise : **"Prêt à dépenser"**, société financière
  - Activité : proposer des **crédits à la consommation** pour des personnes ayant peu ou pas du tout d'historique de prêt.
  - Besoins :
    - à partir de sources de données variées, réaliser un **outil de “scoring crédit”** pour :
      - calculer la probabilité qu’un client rembourse son crédit,
      - classifier la demande en crédit accordé ou refusé.
      - transparence vis-à-vis des décisions d’octroi de crédit auprès des clients (valeurs de l’entreprise)
    - développer un **dashboard interactif** pour les chargés de clientèle et clients

Problématique

- Mission
  - Construire un modèle de scoring avec **prédiction sur la probabilité de faillite** d'un client de façon automatique à partir d'une BDD kaggle.
  - Construire un dashboard interactif :
    - Visualiser le score et l'interprétation de ce score pour chaque client
    - Visualiser des informations d'un client (via un système de filtre).
    - Comparer les informations descriptives relatives à un client à l'ensemble des clients ou à un groupe de clients similaires

Jeux de données

- Une base de données client disp
  Kaggle :
  - 8 fichiers CSV
- Travail sur le fichier application_t



Approche de modélisation

- Analyse exploratoire
  - Définition des colonnes
  - Nombre de ligne colonnes
  - Valeurs manquantes
  - Valeurs uniques
  - Analyse univariée
  - Analyse bivariée
- Traitement
  - Passage des variables catégorielles en numérique
  - Suppression des colonnes à variance nulle
  - Traitement des données aberrantes et remplissage des données manquantes
  - Création de variables fonctionnelles et polynomiales
  - Normalisation

Approche de modélisation

- Traitement
  - Séparation données entrainement et test
  - Déséquilibre de la variable TARGET
    - 0 : 169611 valeurs
    - 1 : 14895 valeurs
  - Equilibrage TARGET par SMOTE*
- Modélisation
  - Entrainement différents modèles
  - Optimisation du meilleur modèle
- Interprétabilité SHAP**
  - Globale
  - Locale
- Recherche des plus proches voisins

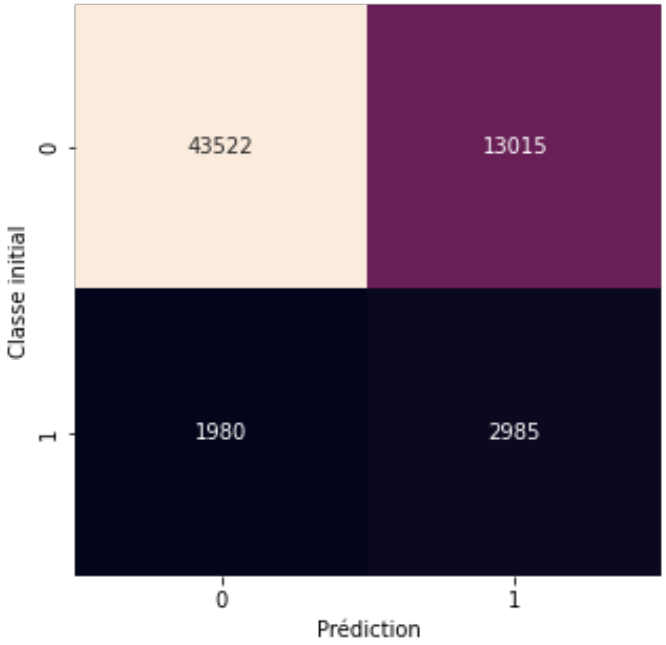
* Synthetic Minority Oversampling Technique
** SHapley Additive exPlanations

Approche de modélisation

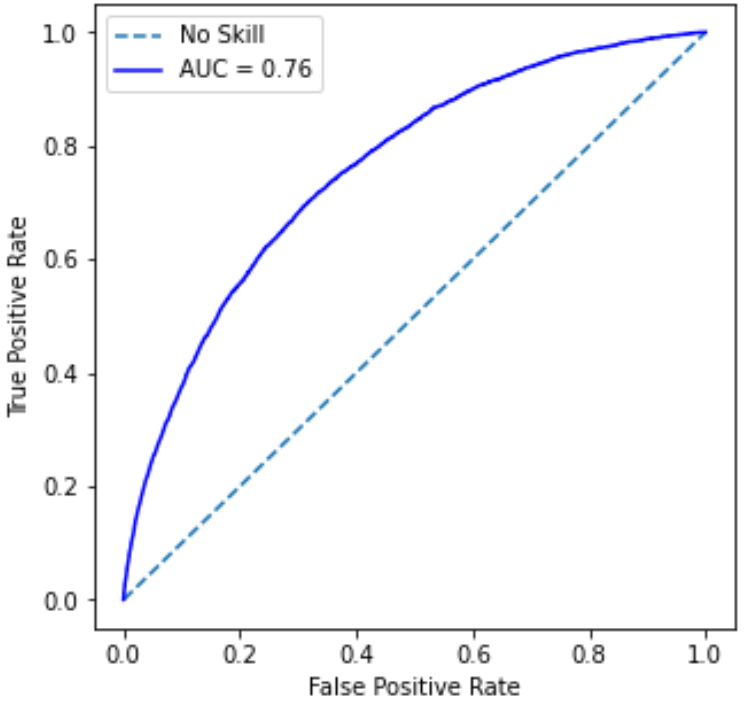
	Model Type	F1-Score	FBeta-Score	Recall_score	Precision_score	Roc_AUC_score
0	logistic regression classifier	0.000000	0.000000	0.000000	0.000000	0.578986
1	decision tree classifier	0.151854	0.162898	0.171198	0.136437	0.546818
2	random forest classifier	0.006400	0.004021	0.003223	0.457143	0.719823
3	gradient boosting classifier	0.026646	0.017001	0.013696	0.489209	0.748325
4	light gradient boosting machine classifier	0.024385	0.015516	0.012487	0.516667	0.758652

Approche de modélisation

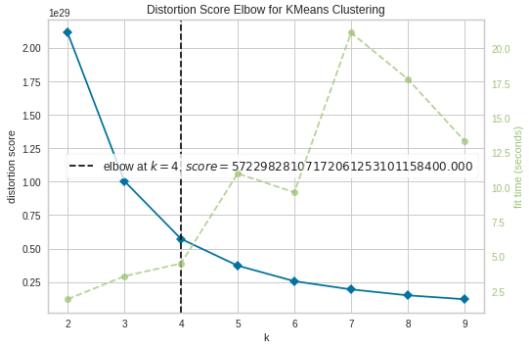
Matrice de confusion : threshold RFE_light gradient boosting machine classifier

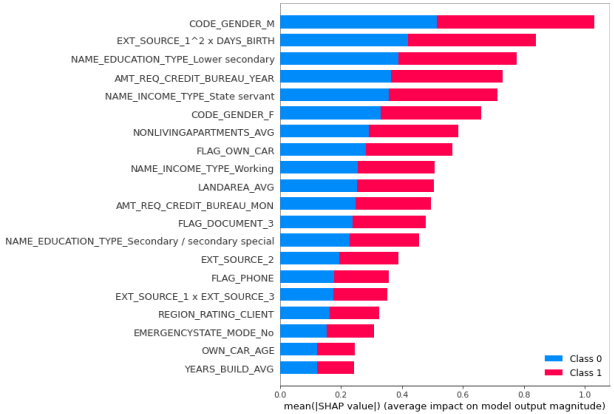


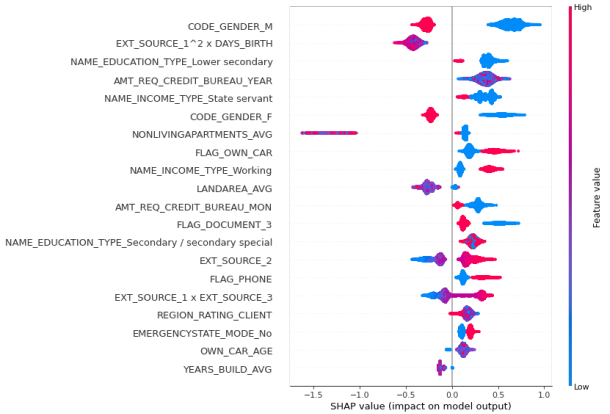
ROC Curve : threshold RFE_light gradient boosting machine classifier



Approche de modélisation







Présentation du dashboard

Solution	Description
SHAP	Explication / interprétabilité de la prédiction
fastapi	API permettant d'appeler la prédiction à partir de l'identifiant du client
streamlit	Tableau de bord
Heroku	Déploiement sur le cloud
Git hub	Versionning

https://api-oc-p7-mbd.herokuapp.com/docs

https://dash-oc-p7-mbd.herokuapp.com/

https://github.com/m3hm3tb4yr4m/OC-P7-Backend

https://github.com/m3hm3tb4yr4m/OC-P7-Frontend

Présentation du dashboard

Présentation du dashboard

	NAME_CONTRACT_T…	CODE_GEN…	FLAG_OWN_…	FLAG_OWN_REA…	CNT_CHILDREN	AMT_INCOME_TO…	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PR…	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_T…	NAME_FAMILY_STATUS	NAME_HOUSING_T…	REGION_POP
152,591.0000	Cash loans	M	N	Y	0.0000	171,000.0000	840,285.0000	43,033.5000	738,000.0000	Unaccompanied	State servant	Higher education	Married	House / apartment	
347,498.0000	Cash loans	F	N	Y	0.0000	301,500.0000	995,562.0000	55,719.0000	922,500.0000	Unaccompanied	Working	Higher education	Married	House / apartment	
229,001.0000	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	
300,223.0000	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	<NA>	
272,937.0000	Cash loans	M	N	Y	0.0000	147,775.5000	1,093,500.0000	45,126.0000	1,093,500.0000	Unaccompanied	Commercial associate	Higher education	Civil marriage	House / apartment	

Conclusion

- Projet très intéressant
- Modélisation
  - Perfectionnement du features engineering
    - Choix des métriques d’évaluation important
    - Choix des hyperparamètres important
- Outils de présentation
  - Visualisation et présentation du travail compréhensible
  - Travail de versionning essentiel en programmation
  - Problèmes de compatibilités des bibliothèques python