Note Méthodologique -

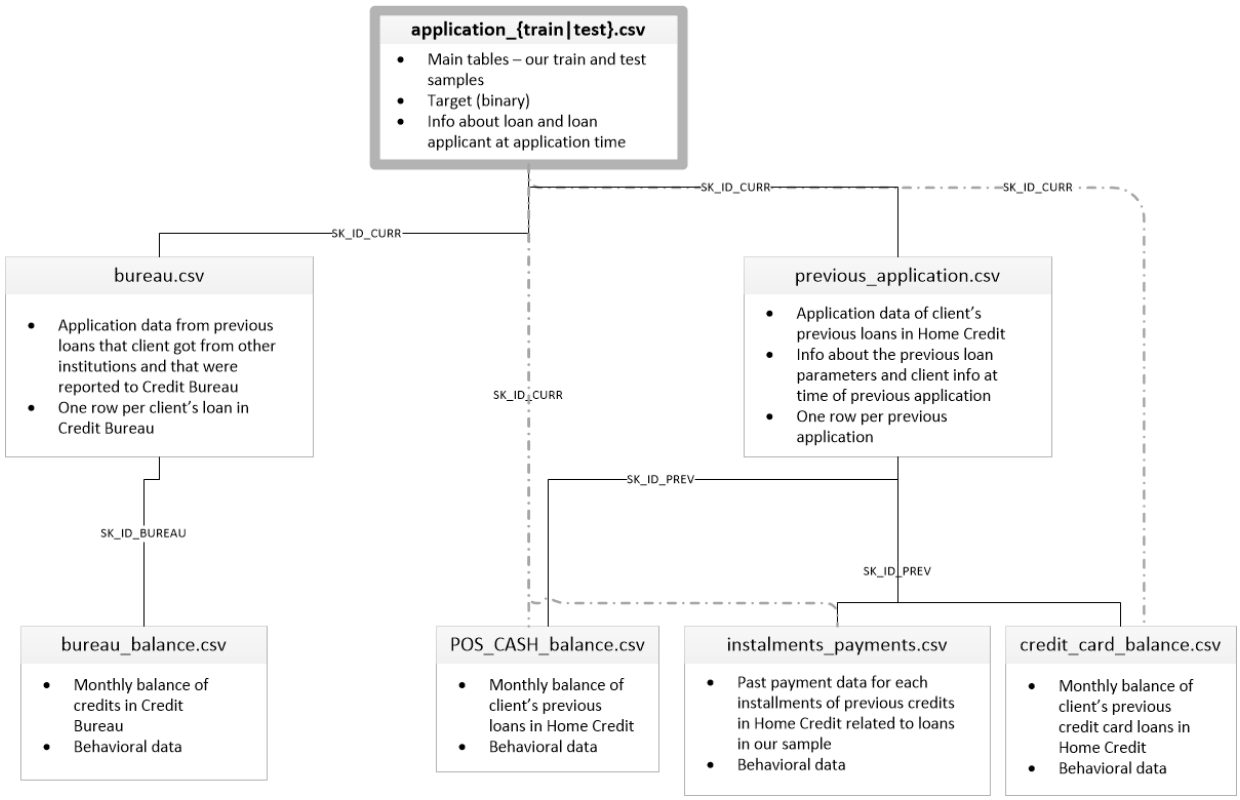
Modèle de “Home Credit Scoring”

Objectif : mettre œuvre un outil de “home credit scoring” pour modéliser la probabilité qu’un client rembourse son crédit.

1. Entrainement du Modèle

1.1 Le Jeu des Données

Le modèle de “Home Credit Scoring” est basé sur un jeu de données fournies par le Home Credit Group[1] et est reparti sur 7 fichiers distinctes comme le montre la figure ci-dessous.



[1] Disponible ici : [Home Credit Default Risk | Kaggle]

Dispoin

1.2 Prétraitement

Le nettoyage, prétraitement et analyse des données tirent parti de précédent travail de AGUIAR[2] et WILL KOEHRSEN[3] .

Première nettoyage, prétraitement et agrégation

Le principales étapes sont:

- remplacement des valeurs aberrantes par une valeur manquante (NaN)
- regroupement cohérente des valeurs catégorielles pour des valeurs peu fréquentes
- transformation des valeurs catégorielles en one-hot variables (encodage binaire) à l'exceptions des catégories pour lesquelles une hiérarchie ou séquence est évidente (e.g. niveau d'éducation, de rendement, jours de semaine) qui sont transformées en variables ordinales.
- création des nouvelles variables métier potentiellement plus pertinentes
- agrégations statistiques des données liées à un individu venant des fichiers autre que l'application avec des différents fonctions (e.g. minimum, maximum, moyenne, déviations, ...)

→ 773 variables par individu

Réduction de nombre de variables

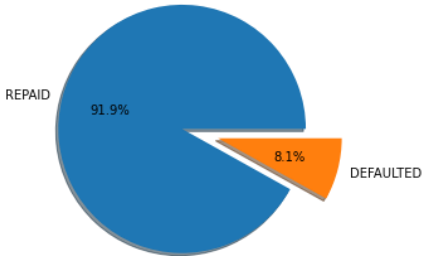
Pour réduire le nombre des variables, on a supprimé les variables avec

- plus de 70% de valeurs manquantes
- une corrélation (positive ou négative) plus élevée que 90% avec une autre variable (seulement un des deux variables est supprimée)
- une déviance de moins de 10%, s'ils ont des valeurs entre 0 et 1.

→ 377 variables par individu

Division du jeu de données en jeux d'entrainement et de test

Le jeu de données contient 307507 individus dont que 8.1% étaient défaillants. La variable cible est donc très déséquilibrée et le partage en jeux d'entrainement et de test est fait d'une façon stratifiée pour garantir la même distribution de la variable cible dans les deux jeux de données.



→ 70% jeu d'entrainement et 30% jeu de test (échantillonnage stratifié)

[2] Kaggle Notebook: LightGBM with Simple Features | Kaggle

[3] Kaggle Notebook : Start Here: A Gentle Introduction | Kaggle (et ses références à des notebooks suivantes)

1.3 La Sélection et Optimisation du Modèle

Les Modèles testés

Un modèle entier consiste en générale de trois étapes, dont les deux premiers ne sont pas nécessairement obligatoires:

1. une mise à l'échelle (e.g. MinMaxScaler(), un scaler customisé qui prend en compte l'asymétrie de la distribution des valeurs (skew) pour éventuellement faire une mise à l'échelle logarithmique)
2. une imputation des valeurs manquantes (e.g. avec la valeur médiane)
3. un modèle de classification (e.g. LogisticRegresstion, RandomForestClassifier, RUSBoostClassifier, LightGBM)

Méthode d'Optimisation des Hyperparamètres

Les paramètres des différents modèles sont optimisés avec une recherche de grille qui est validé par une validation croisée de 5 folds (sous-groupes) stratifiés. Le score de validation est la moyenne de ces 5 évaluations. [4]

Pour accélérer le processus, la recherche des bons hyperparamètres se fait avec un sous-échantillonnage aléatoire de la classe majoritaire (i.e. le crédit est remboursé) de sorte qu'on a une distribution équilibrée entre les deux classes cibles. Après le modèle est entraîné et validé sur tout le jeu d'entrainement et si un meilleur score est obtenu, ce modèle est retenu.

Métrique d'Evaluation

Donnant la distribution déséquilibrée entre les deux classes cibles, on a utilisé la métrique 'ROC-AUC' (receiver operating characteristic curve – area under the curve) principale pour optimiser les hyperparamètres et choisir le modèle. Cette métrique synthèse la relation entre le rappel, ou taux de vrais positifs (i.e. la capacité d'identifier des mauvais crédits), et le taux de faux positifs (i.e. le risque de mal identifier des bons crédits). De plus, on regarde les métriques la précision-moyenne et le F1-score.

L'évaluation finale d'un modèle est faite avec le jeu de données de test.

Résultats

Model	Mise à l'échelle	Imputation	Sous-échant.	AUC-ROC valid.	AUC-ROC test	Précision-moyenne test	F1-score test	Temps (s)
LightGBM	--	--	non	0.7807	0.7780	0.1525	0.2935	20.5
LogisticRegr.	Cust.	médiane	non	0.7704	0.7694	0.1442	0.2754	45.2
RUSBoost	--	médiane	non	0.7701	0.7694	0.1452	0.2780	114.6
RandomForest	--	médiane	oui	0.7538	0.7516	0.1369	0.2627	31.7

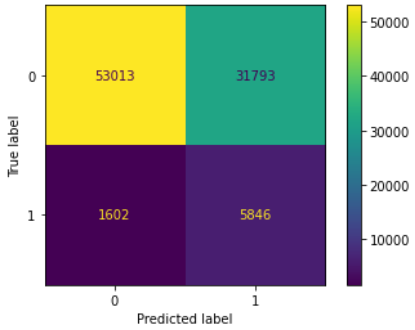
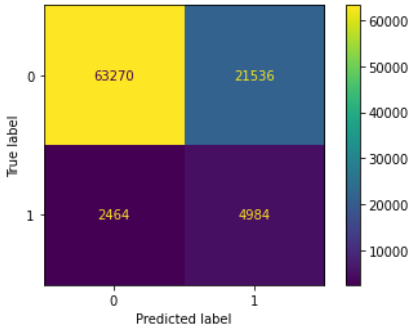
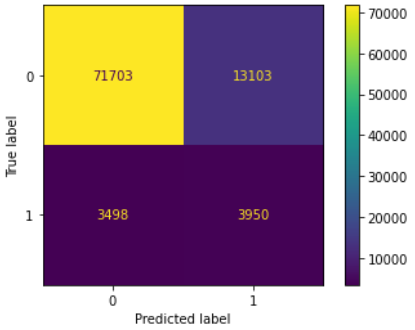
[4] Pour le modèle LightGBM on prend simplement les hyperparamètres optimisée dans le Kaggle notebook : Start Here: A Gentle Introduction | Kaggle

2. Le Modèle - LightGBM

Le modèle qui est mis en œuvre est LightGBM. Ce modèle non seulement donne les meilleurs scores mais a aussi un temps d'entrainement plus court.

2.1 Matrice de Confusion

La décision de donner un crédit ou pas peut être adapté avec un seuil de prédiction en accord avec une analyse coûts-avantages, entre mauvais clients acceptés et bons clients rejetés. En bas des matrics de confusion avec des seuils différents.

Seuil de Prédiction 0.4	Seuil de Prédiction 0.5	Seuil de Prédiction 0.6
		

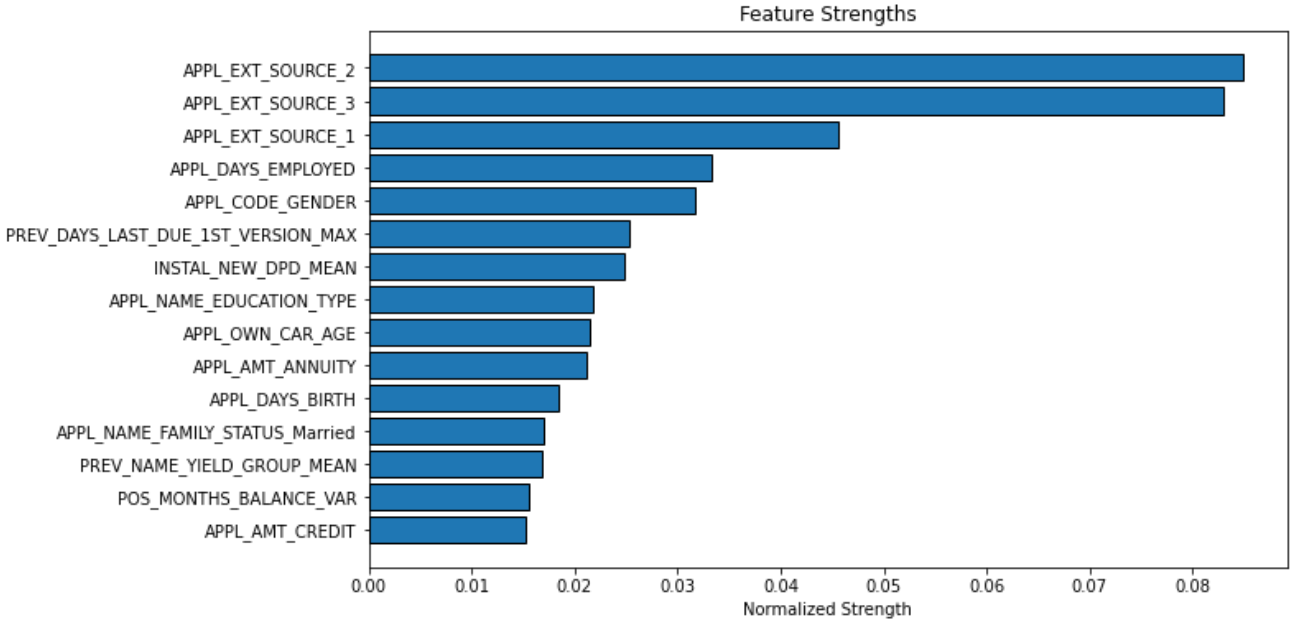
2.2 Interprétation du Modèle

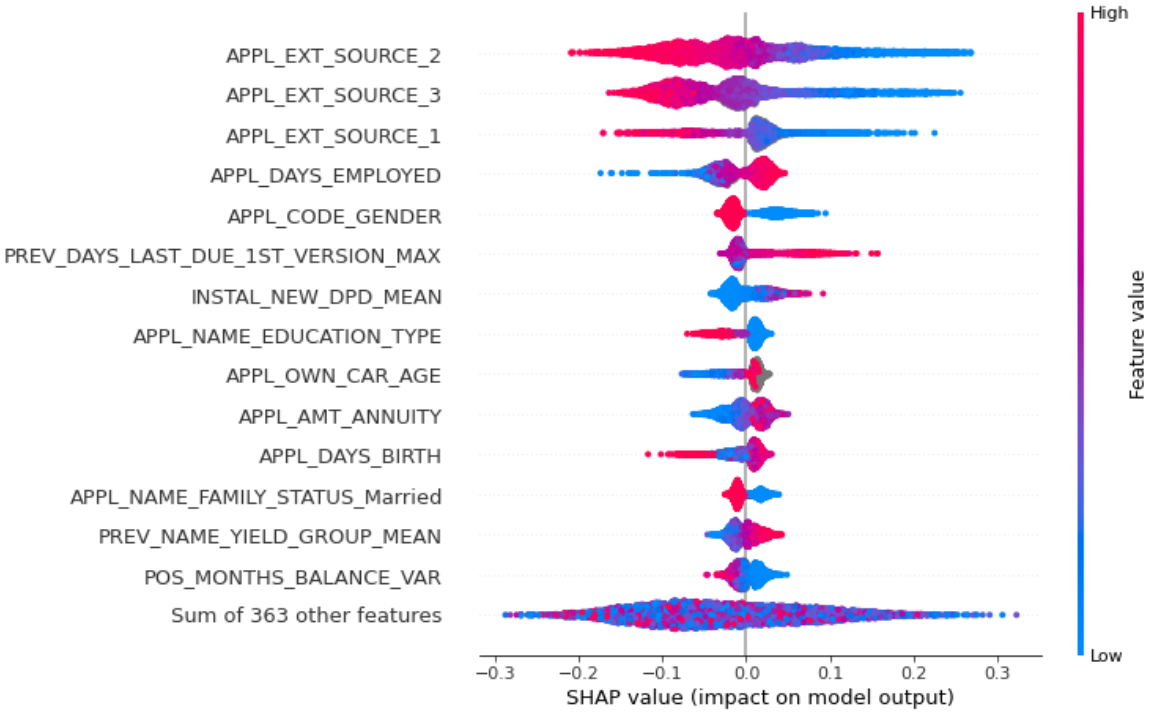
SHAP - (SHapley Additive exPlanations)

L'interprétation du modèle est faite avec SHAP, un algorithme de la théorie des jeux coopératifs qui explique un modèle en déterminant les contributions des différents features à les observations. Pour obtenir un point de vue globale les contributions des features sont moyennées. Pour des raisons de temps de calcul, on ne fait qu'une estimation des valeurs SHAP sur un échantillon des clients.

Interprétation globale

Le point de vue globale montre que les informations venant des sources externes (APPL_EXT_SOURCE) ont un impact significatif sur le modèle. De plus, le nombre de jours employé, le genre, le niveau d'éducation et l'âge ont un impact sur la prédiction du modèle.



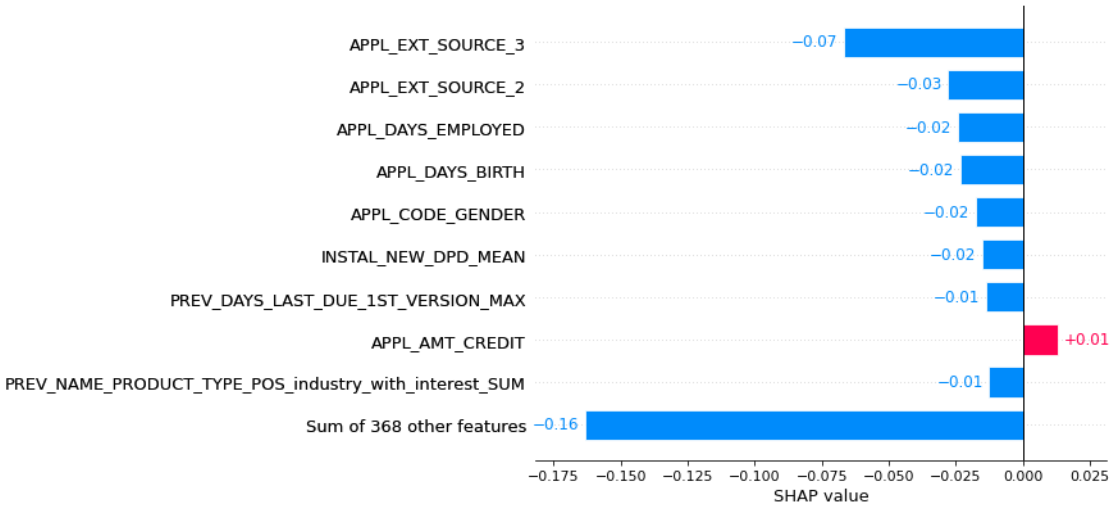


Dispoin

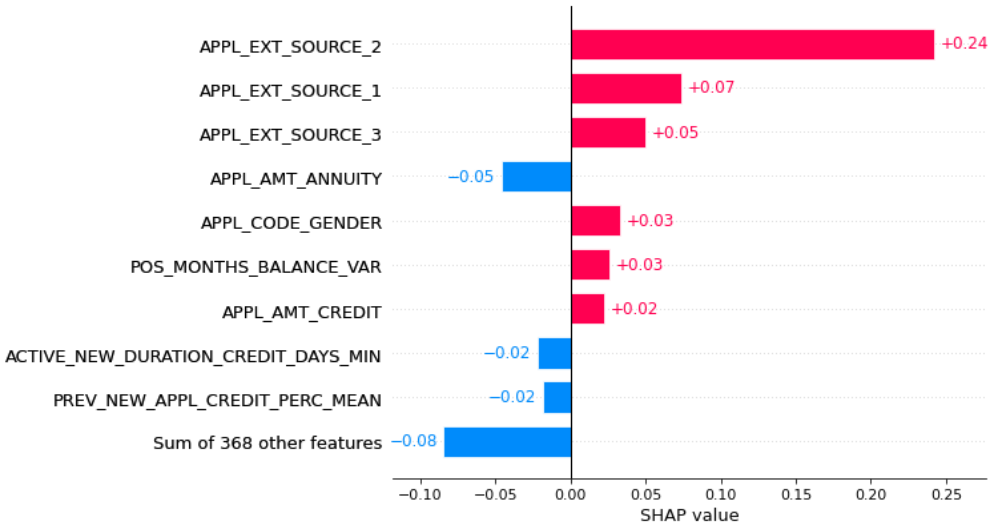
Les valeurs de SHAP pour un individu

Les valeurs de SHAP peuvent aussi donner une indication quelle features contribuent à quel taux à la prédiction du modèle pour un individu, i.e. que sont les caractéristiques particulières de l'individu conduisant à la prédiction du modèle.

Les valeurs de SHAP pour un individu qui a remboursé son crédit :



Les valeurs de SHAP pour un individu qui n'a pas remboursé son crédit :



Les features les plus importantes varient pour les individus, mais font souvent parties des features qui sont globalement important, même si elles n'ont pas forcément le même ordre ou ni la même magnitude d'impact.

3. Points d'Améliorations

3.1 Préparation et Sélection des Données

- regroupement des features catégorielles dans groupes plus larges et pertinents pour évider des classes très peu peuplées
- réduction de nombre des features utilisés, éventuelle en regardant les SHAP values, pour réduire le temps de calcul et faciliter l'interprétation du modèle
- création des autres features métiers ou des features polynomiales avec les features le plus important

3.2 Le Modèle - LGBM

- évaluation du modèle avec un mise en échelle customisé
- re-évaluation des hyperparamètres, pour assurer que le modèle est parfaitement optimisé

3.3 La fonction de coût

- analyse coûts-avantages, entre mauvais clients acceptés et bons clients rejetés
- prendre en compte le montant du crédit dans l'analyse coûts-avantages

3.4 Evaluation du biais dans le modèle

- le genre est un feature le plus important dans le modèle, ce que problématique éthiquement