

P7 - IMPLÉMENTEZ UN MODÈLE DE SCORING

08/10/2021

Etudiant : Luc Rogers
Mentor : Etienne Sanchez

Sommaire

2

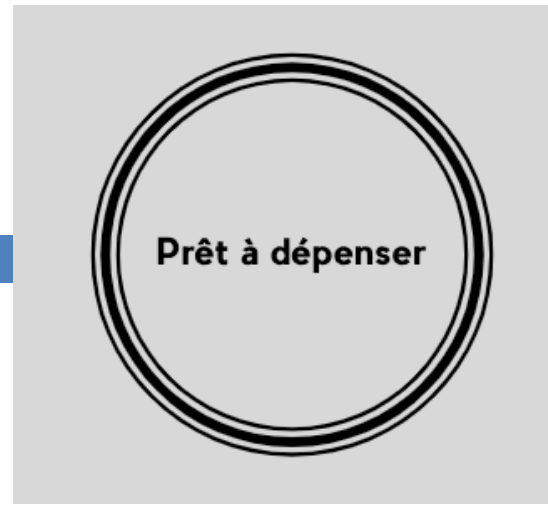
- 1. Problématique
- 2. Modélisation
- 3. Scoring et fonction coût
- 4. Présentation du dashboard interactif
- 5. Limites du modèle et améliorations

3

Problématique

Problématique

4

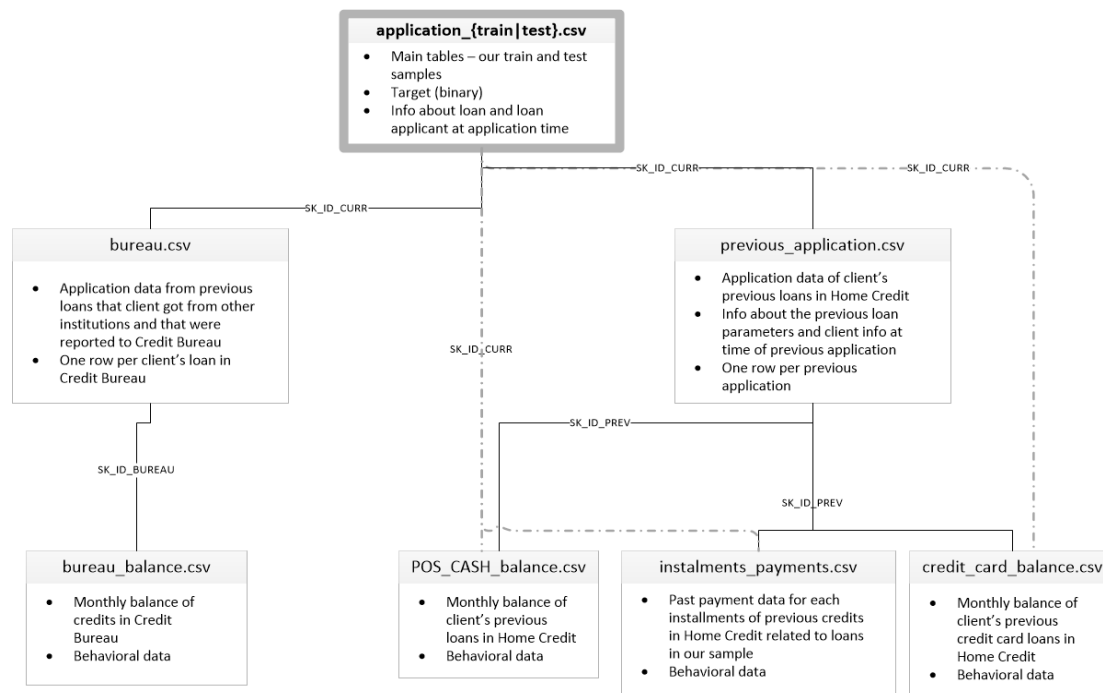


- Contexte: société de crédits à la consommation « Prêt à dépenser »
- Objectifs:
 - ▣ Développer un modèle de scoring de la probabilité de défaut de paiement client
 - ▣ Déployer un dashboard interactif en ligne répondant à des critères de transparence sur l'explicabilité du modèle
- Liens du projet:
 - ▣ Données : <https://www.kaggle.com/c/home-credit-default-risk/data>
 - ▣ Livrables : <https://github.com/lucrogers/OpenClassrooms-Projet-7-Data-Scientist-Implementez-un-modele-de-scoring>
 - ▣ Dashboard interactif : <https://loan-management-dash.herokuapp.com>

Problématique

5

- Base de données:
 - ▣ ~ 300 000 clients provenant de diverses sources
 - ▣ 121 features : salaire, métier, nombre d'enfants, âge, informations relatives aux crédits antérieurs, ...
 - ▣ Etiquette cible:
 - 0 pour un client sans incident de paiement
 - 1 pour un client ayant connu des difficultés de paiement



6

Modélisation

Modélisation

7

- Reprise d'un [kernel Kaggle](#)
 - Pre-processing:
 - ▣ Valeurs aberrantes supprimées
 - ▣ Valeurs manquantes remplacées par la moyenne ou par le mode selon leur nature quantitative ou catégorielle
 - ▣ Les variables catégorielles sont ensuite encodées avec un encodage de type one hot encoding
 - ▣ Création de features métiers
 - ▣ Aggrégation des différents datasets par client
- ➔ Obtention d'un jeu de données de 657 features

Modélisation

8

- Méthodologie type:
 - ▣ Essais de plusieurs modèles de classification
 - ▣ Recherche des hyperparamètres optimaux
 - ▣ Validation croisée
 - ▣ Choix d'une métrique appropriée: score AUC
- Modèle retenu:
 - ▣ Classifieur LightGBM
 - ▣ Score AUC: 0,80

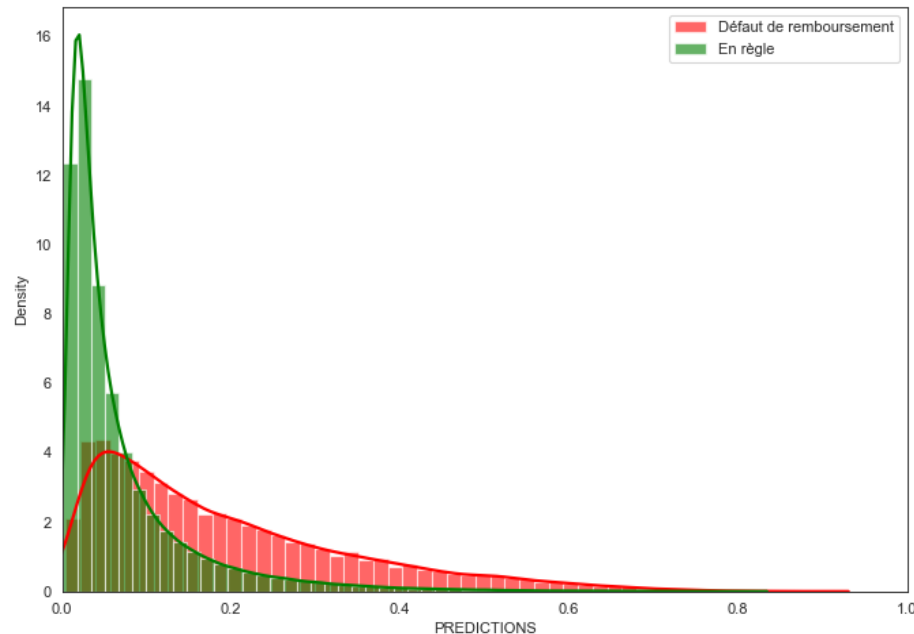
9

Scoring et fonction coût

Scoring et fonction coût

10

- Score client = probabilité de défaut de paiement (entre 0 et 1)



- ➔ Une distribution des défauts de paiement légèrement plus étalée vers la droite
- ➔ Quel seuil choisir ? Compromis à trouver en fonction des coûts et profits

Scoring et fonction coût

11

□ Fonction coût = bénéfice réalisé par l'entreprise:

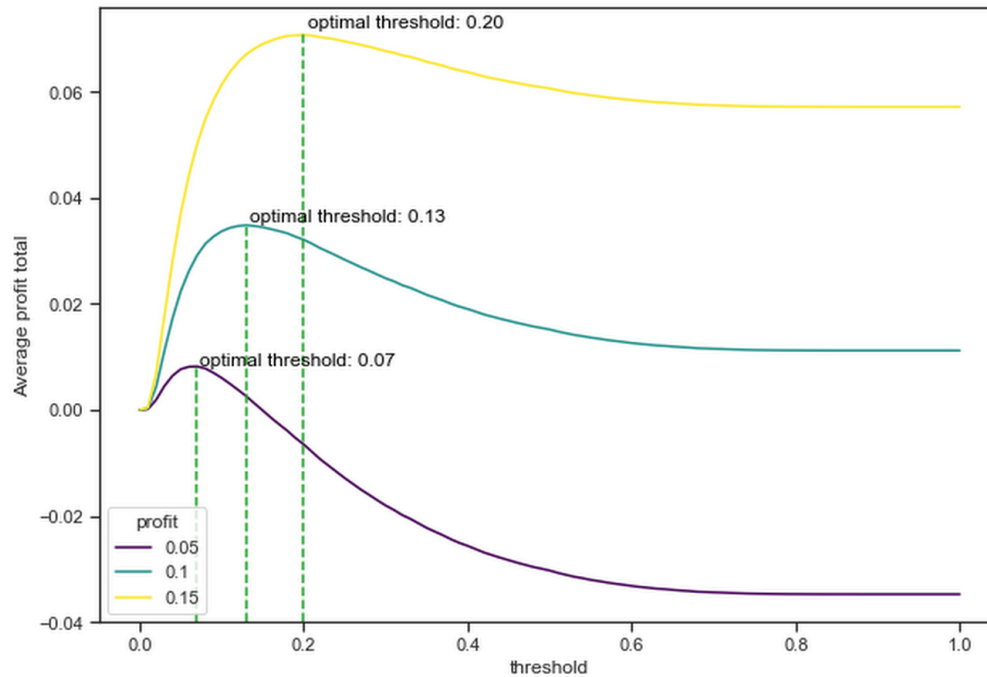
$$\text{Bénéfice} = [(1 - p(\text{seuil})) * \text{profit} + p(\text{seuil}) * \text{coût}] * n(\text{seuil})$$

Avec:

- $p(\text{seuil})$: la proportion de clients en défaut de paiement
- profit : le profit moyen réalisé par un prêt en règle (en %)
- coût : le coût moyen associé à un défaut de paiement (en %)
- $n(\text{seuil})$: le nombre de prêts accordés

Scoring et fonction coût

12



➔ Seuil optimal trouvé en fonction des coûts et profits générés par les pertes et le gains

13

Présentation du dashboard

Présentation du dashboard

14

- ❑ Dashboard interactif à l'attention des gestionnaires clients
- ❑ Cahier des charges:
 - ▣ Permettre de visualiser le score et l'interprétation de ce score pour chaque client de façon intelligible pour une personne non experte en data science.
 - ▣ Permettre de visualiser des informations descriptives relatives à un client (via un système de filtre).
 - ▣ Permettre de comparer les informations descriptives relatives à un client à l'ensemble des clients ou à un groupe de clients similaires.
 - ▣ Déploiement en ligne.
- ❑ Lien du dashboard: <https://loan-management-dash.herokuapp.com/>
 - ▣ Réalisé avec Dash
 - ▣ Déployé sur Heroku

Présentation du dashboard

15

Client Loan Management

Client sélectionné

100002

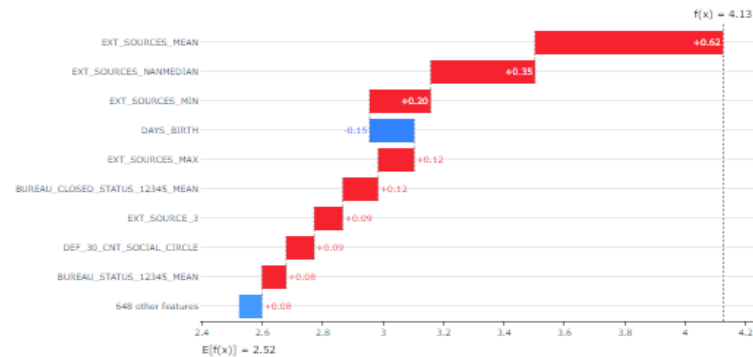
Infos	Client
SK_ID_CURR	100002
TARGET	1
NAME_CONTRACT_TYPE	Cash loans
CODE_GENDER	M
FLAG_OWN_CAR	N
FLAG_OWN_REALTY	Y
CNT_CHILDREN	0
AMT_INCOME_TOTAL	202500
AMT_CREDIT	406597.5
AMT_ANNUITY	24700.5
AMT_GOODS_PRICE	351000
NAME_TYPE_SUITE	Unaccompanied
NAME_INCOME_TYPE	Working
NAME_EDUCATION_TYPE	Secondary / secondary special

P prédiction
Crédit à risque

Score client (cible = 0.2)
0.23

Montant crédit
406597.5

Client Group



Observations

notes de sessions client

Explicabilité du modèle (définitions des variables ciblées)

feature	Définition
EXT_SOURCES_MEAN	
EXT_SOURCES_NANMEDIAN	
EXT_SOURCES_MIN	
DAYS_BIRTH	
EXT_SOURCES_MAX	

Présentation du dashboard

16

Client Loan Management

Client sélectionné

100002

Prediction
Crédit à risque

Score client (cible = 0.2)
0.23

Montant crédit
406597.5

Infos	Client
SK_ID_CURR	100002
TARGET	1
NAME_CONTRACT_TYPE	Cash loans
CODE_GENDER	M
FLAG_OWN_CAR	N
FLAG_OWN_REALTY	Y
CNT_CHILDREN	0
AMT_INCOME_TOTAL	202500
AMT_CREDIT	406597.5
AMT_ANNUITY	24700.5
AMT_GOODS_PRICE	351000
NAME_TYPE_SUITE	Unaccompanied
NAME_INCOME_TYPE	Working
NAME_EDUCATION_TYPE	Secondary / secondary special



Observations

notes de sessions client

Explicabilité du modèle (définissons des variables ciblées)

feature	Définition
EXT_SOURCES_MEAN	
EXT_SOURCES_NANMEDIAN	
EXT_SOURCES_MIN	
DAYS_BIRTH	
EXT_SOURCES_MAX	

17

Limites du modèle et améliorations

Limites du modèle et améliorations

18

□ Limites du modèle:

- Population des demandeurs de prêts \neq population d'entraînement du modèle → modèle qui ne peut se suffire à lui-même
- Données provenant de différentes sources → règles d'acceptation de crédit différentes
- Fonction coût qui repose sur l'observation empirique du nombre d'individus en difficulté de paiement dans notre jeu de données
- Fonction coût calculée sur l'approximation d'un montant moyen par client
- Fonction coût = rentabilité uniquement

□ Amélioration du dashboard:

- Prévoir espace de simulation qui permettrait de jouer avec les features client pour modifier dynamiquement le score

Merci de votre attention

Annexe