

note méthodologique P7

Parcours Data Scientist « Implémentez un modèle de scoring »

**Table des matières**

[I. Contexte 2](#_Toc90919113)

[II. La méthodologie d'entraînement du modèle 2](#_Toc90919114)

[A. Split jeu d’entrainement et jeu de test 2](#_Toc90919115)

[B. Classification binaire avec déséquilibre des classes 2](#_Toc90919116)

[C. Test de différents modèles. 3](#_Toc90919117)

[III. La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation 3](#_Toc90919118)

[A. Choix de la métrique 3](#_Toc90919119)

[B. Algorithme d’optimisation 4](#_Toc90919120)

[IV. L’interprétabilité globale et locale du modèle 4](#_Toc90919121)

[V. Les limites et les améliorations possibles 5](#_Toc90919122)

# Contexte

Cette note méthodologique est un des livrables du projet « Implémentez un modèle de scoring » du parcours Data Scientist d’Openclassroom. Son objet est de présenter le processus de modélisation et d’interprétabilité du modèle mis en place dans le cadre du projet 7.

Le projet consiste à développer pour vous, l’entreprise, « Prêt à Dépenser », société de crédit de consommation, un modèle de scoring de la probabilité de défaut de paiement d’un client avec pas ou peu d’historique de prêt.

La base de données que vous nous avez confiée est constituée de fichiers relatifs à 307 000 clients comportant 121 caractéristiques (âge, sexe, emploi, logement, revenus, informations surle crédit, notation externe, etc.)

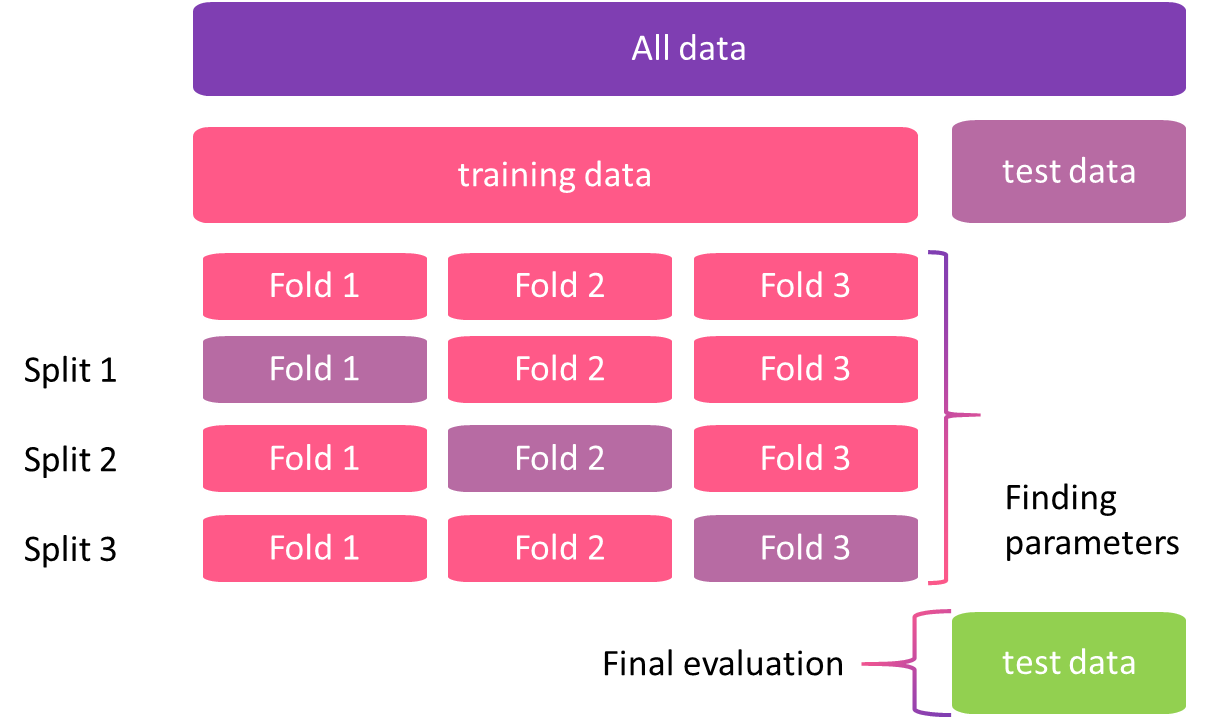
# La méthodologie d'entraînement du modèle

## Split jeu d’entrainement et jeu de test

Nous avons entraîné notre modèle sur la base du jeu de données, après analyse exploratoire et création de nouvelles features. Le notebook utilisé est consultable sur github.

Le jeu de données initial a été séparé en plusieurs parties de façon à disposer :

* d’un jeu de training (80% des individus) qui a été séparé en plusieurs folds pour entraîner les différents modèles et optimiser les paramètres (cross validation) sans overfitting.
* d’un jeu de test (20 % des individus) pour l’évaluation finale du modèle



## Classification binaire avec déséquilibre des classes

Nous sommes face à un problème de classification binaire et une problématique de déséquilibre des classes. En effet, nous avons une classe sous représentée (9 % de clients en défaut contre 91 % de clients sans défaut). Nous avons dû prendre cela en compte dans l’entraînement des modèles.

Nous avons appliqué un poids sur classe 1 (= 10) pour contrebalancer ce déséquilibre. Puis nous avons procédé à une validation croisée (3 folds stratifiés).

Pour chaque approche, nous avons testé différents modèles avec recherche d’hyperparamètres et cross validation (3 folds).

## Test de différents modèles.

Voici les différents modèles qu’il nous a paru intéressant de tester :

* Light GBM Classifier
* LightGBM with prior class balancing
* t-SVD+LGBM

Seul LightGBM a fourni des temps CPU suffisamment bas.

* Utilisation du score AUC-ROC car c'était le score de la compétition Kaggle
* Utilisation de Random CV plutôt que Grid CV en raison du grand nombre d'hyperparamètres.

Le travail sur jeu réduit a montré que les 2 approches présentaient des résultats d’ordre de grandeur comparable sur le jeu de validation. Elles ont donc été conservées pour un entraînement sur le jeu complet. Le processus a ensuite été réalisé sur l’intégralité du jeu de données.

Le choix du meilleur modèle a été effectué en retenant le modèle avec le meilleur score Random CV sur le jeu de validation.

Logigramme de méthodologie d’entraînement du modèle

# La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation

## Choix de la métrique

La matrice de confusion est la suivante :

|  |  |  |
| --- | --- | --- |
|  | Clients prédits en défaut | Clients prédits sans défaut |
| Clients sans défaut | Faux positifs | Vrais négatifs |
| Clients réellement en défaut | Vrais positifs | **Faux négatifs** |

Dans votre métier, la banque, mieux vaut éviter de mal catégoriser un client avec un fort risque de défaut.

Nous avons donc cherché à minimiser le pourcentage de faux négatifs et à maximiser le pourcentage de vrais positifs. Cela revient à maximiser le « RECALL » de 𝑣𝑟𝑎𝑖𝑠 𝑝𝑜𝑠𝑖𝑡𝑖𝑓𝑠, soit le nombre de *vrais positifs* par rapport à l’ensemble (v𝑟𝑎𝑖𝑠 𝑝𝑜𝑠𝑖𝑡𝑖𝑓𝑠+𝑓𝑎𝑢𝑥 𝑛é𝑔𝑎𝑡𝑖𝑓𝑠).

Par ailleurs, pour étendre votre base de clientèle, nous voulons éviter de tous les classer en défaut. On cherche donc à éviter d'avoir un trop grand nombre de faux positifs.

Nous avons donc cherché à maximiser la PRECISION, soit le soit le nombre de vrais positifs par rapport à l’ensemble (v𝑟𝑎𝑖𝑠 𝑝𝑜𝑠𝑖𝑡𝑖𝑓𝑠+𝑓𝑎𝑢𝑥 posi𝑡𝑖𝑓𝑠)

Pour votre problématique métier, le RECALL est plus important que la PRECISION car on préfèrera vraisemblablement limiter un risque de perte financière plutôt qu’un risque de perte de client potentiel.

En une phrase : « Plus le recall est haut, plus le modèle repère de positifs, plus la précision est haute, moins le modèle se trompe sur les positifs »

On cherche donc une fonction qui optimise les 2 critères en donnant plus d'importance au recall. C’est ce que fait la Fonction F1 Score. Elle nous permet d’effectuer une bonne évaluation de la performance de notre modèle. Voici comment elle se calcule (moyenne harmonique) :

F1 Score = (2 x recall x précision) / (recall + précision)

Comme vous l’aurez compris, le plus votre F1 Score est élevé, le plus votre modèle est performant. <https://inside-machinelearning.com/recall-precision-f1-score/>

Il faut pour cela estimer :

* le coût moyen d'un défaut
* le coût d'opportunité d'un client potentiel accidentellement écarté

Nous avons pris des hypothèses, que nous pourrons ajuster selon vos directives métiers :

* que chaque défaut entraîne la dépense de 30 % du montant du crédit en pertes et frais de recouvrement
* qu'un client a 10 % de chance de souscrire au crédit quand il en fait la demande à un conseiller, et donc que le coût d'opportunité pour un client potentiel accidentellement écarté est de 10 % du montant du crédit.

Nous obtenons les coefficient suivants :

* Coefficient RECALL : de 69 % \* le montant moyen de crédit des personnes en défaut
* Coefficient PRECISION : 16 % \* le montant moyen de crédit des personnes sans défaut

Ce qui nous permet d’établir la valeur de F1 Score à 26%.

## Algorithme d’optimisation

Nous avons retenu la meilleure combinaison d’hyperparamètres pour chaque algorithme. Le modèle ayant le meilleur score en cross validation sur le jeu de training a été retenu : il s’agit du modèle light GBM.

* C’est un framework d'amplification de gradient qui utilise un algorithme d'apprentissage basé sur des arbres.
* C’est un algorithme Leaf-wise (fait croître l'arbre verticalement)
* tandis qu'un algorithme level-wise fait croître les arbres horizontalement
* Lors de la croissance de la même feuille, l'algorithme Leaf-wise peut réduire davantage les pertes qu'un algorithme level-wise.

# L’interprétabilité globale et locale du modèle

L’interprétation de modèles de Machine Learning (ML) complexes, encore appelés modèles ”black box”, est aujourd’hui un enjeu important dans le domaine de la Data Science. Le modèle que nous avons développé pour vous permet de prédire la probabilité de défaut de paiement d’un client.

Cependant, cette prédiction n’apporte pas d’informations sur les paramètres qui ont influencés chaque valeur prédite. Par ailleurs, vos conseillers doivent être capables d’expliquer les prédictions que ce soit pour des raisons légales (article 22 du RGPD), pour des raisons de validation métier enfin pour des raisons de transparence vis-à-vis de vos clients.

Ainsi, notre objectif a été de leur donner les moyens d’expliquer les décisions de l’algorithme aux clients réels. Pour cela nous avons utilisé la librairie SHAP, librairie adaptable à n’importe quel modèle de machine learning et permettant de comprendre quelles caractéristiques expliquent la décision d’octroi du crédit.

Pour extraire des informations du modèle, la première étape est l’approche globale qui consiste à définir l’importance des variables du modèle de manière globale. La seconde étape consiste à changer d’échelle afin d’extraire des informations locales pour des exemples spécifiques de notre dataset.

La valeur de Shap proposée par Lundberg et al. est la valeur SHapley Additive exPlanation. L’idée est de calculer la valeur de Shapley pour toutes les variables à chaque exemple du dataset, puis de moyenner l’impact qu’une variable a pour toutes les combinaisons de variables possibles.

Cette approche combine respecte les trois propriétés suivantes :

* Additivité : la somme des effets des variables est égale à la prédiction du modèle pour tous les exemples,
* Variables nulles sans effet : si une variable de l’exemple considéré est à zéro, alors la variable ne doit pas avoir d’impact pour cet exemple,
* Cohérence : si un modèle change, tel que l’effet d’une variable est plus important sur le modèle, l’attribution assignée à cette variable ne doit pas baisser.

La valeur de Shap est très coûteuse à calculer et c’est un désavantage de cette approche. Sauf lorsqu’on la combine avec Tree explainer comme nous l’avons fait, cette approche utilise la construction des arbres afin de réduire les coûts de calcul. L’algorithme récursif garde en mémoire tous les ensembles possibles de variables qui descendent dans chaque feuille de l’arbre, et pour chaque exemple les valeurs des feuilles, la proportion d’exemples et les chemins des « hot » et « cold child » sont utilisés pour l’estimation des valeurs de Shap.

Pour plus de détail vous pouvez consulter cet article très didactique : <https://www.aquiladata.fr/insights/shap-mieux-comprendre-linterpretation-de-modeles/>

# Les limites et les améliorations possibles

Nous avons effectué notre modélisation sur la base d’une hypothèse forte : la définition d’une métrique d’évaluation : le F Beta Score avec Beta fixé suivant certaines hypothèses non confirmées par le métier.

L’axe principal d’amélioration serait de définir plus finement la métrique d’évaluation en collaboration avec les équipes métier.

L’interprétabilité du modèle pourrait être étoffée en considérant les variables issues du one hot encoding comme une seule et même variable dans la perturbation (un client ne pouvant cumuler plusieurs caractéristiques dans la logique du jeu de données initial.

Par ailleurs, la partie de traitement préalable du jeu de données a été abordée de façon superficielle en réutilisant un notebook issu de Kaggle qui se base uniquement sur une table du jeu de données. Il y a très probablement l’opportunité d’améliorer la modélisation en utilisant d’autres features des données fournies, ainsi qu’en créant de nouvelles features en collaboration avec les équipes métier.