|  |
| --- |
| **PROJET 7 :** NOTE MÉTHODOLOGIQUE |

1. **MÉTHODOLOGIE D’ENTRAÎNEMENT**

Le jeu de données initial a été nettoyé des lignes et colonnes pour lesquelles le taux de remplissage était inférieur à 80%.

Dans le jeu de données initial (train), nous avons séparé :

* X : matrice des variables
* y : vecteur des cibles

Nous avons ensuite séparé en deux :

* X\_fit, y\_fit : jeu de données servant à la sélection du modèle et des hyperparamètres
* X\_eval, y\_eval : jeu de données servant à l’évaluation finale du modèle

Les valeurs manquantes résiduelles ont été imputées « par la moyenne » (moyennes calculées sur les données X\_fit, pour éviter la fuite des données).

Pour faciliter l’apprentissage du modèle, nous avons procédé à l’équilibrage des classes par sur-échantillonnage (*oversampling*) de la classe minoritaire.

Nous avons prélevé un aléatoire échantillon de 5'000 demandes de crédit dans le jeu (X\_fit, y\_fit) pour accélérer la recherche d’optimisation des modèles et hyperparamètres.

Les autres étapes de prétraitement (normalisation, standardisation), la nature du modèle et les hyperparamètres ont été sélectionnés en utilisant l’algorithme d’optimisation discrète *Tree of Parzen Estimators* (TPE) implémentée dans la librairie Hyperopt.

Enfin nous avons comparé le résultat à une prédiction utilisant Lightgbm.

**La liste de modèles utilisés :**

* Bayésien naïf :

Notre baseline. Point 0 sur lequel on va pouvoir comparer les autres modèles.

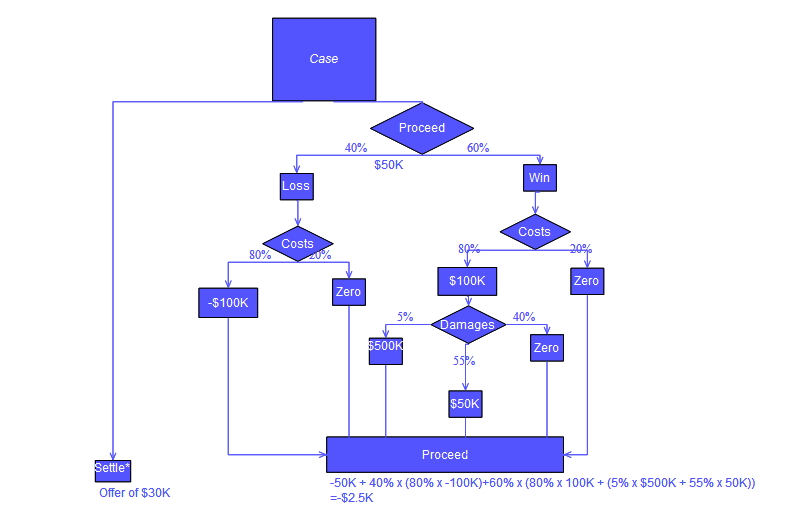
La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le [théorème de Bayes](https://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_Bayes) avec une forte indépendance (dite naïve) des hypothèses. Elle met en œuvre un classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des [classifieurs linéaires](https://fr.wikipedia.org/wiki/Classifieur_lin%C3%A9aire).

Un terme plus approprié pour le modèle probabiliste sous-jacent pourrait être « modèle à caractéristiques statistiquement indépendantes ».

En termes simples, un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Un fruit peut être considéré comme une pomme s'il est rouge, arrondi, et fait une dizaine de centimètres. Même si ces caractéristiques sont liées dans la réalité, un classifieur bayésien naïf déterminera que le fruit est une pomme en considérant indépendamment ces caractéristiques de couleur, de forme et de taille.

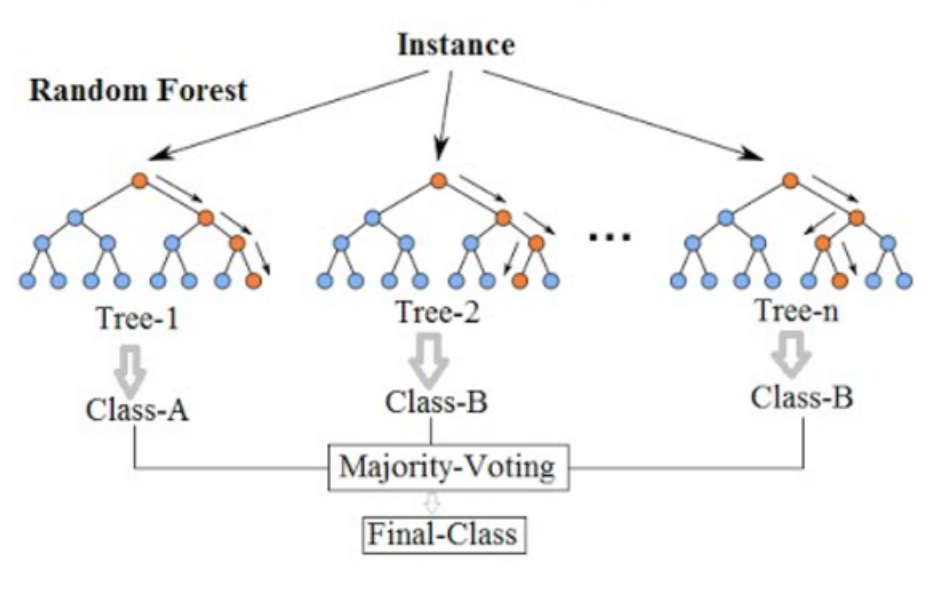
* Arbre de décision

Un arbre de décision est un outil d'aide à la [décision](https://fr.wikipedia.org/wiki/D%C3%A9cision) représentant un ensemble de choix sous la forme [graphique](https://fr.wikipedia.org/wiki/Repr%C3%A9sentation_graphique) d'un [arbre](https://fr.wikipedia.org/wiki/Arbre_(graphe)). Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteintes en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans des domaines variés tels que la [sécurité](https://fr.wikipedia.org/wiki/S%C3%A9curit%C3%A9), la [fouille de données](https://fr.wikipedia.org/wiki/Fouille_de_donn%C3%A9es), la [médecine](https://fr.wikipedia.org/wiki/M%C3%A9decine), etc. Il a l'avantage d'être lisible et rapide à exécuter. Il s'agit de plus d'une représentation calculable automatiquement par des algorithmes d'apprentissage supervisé.



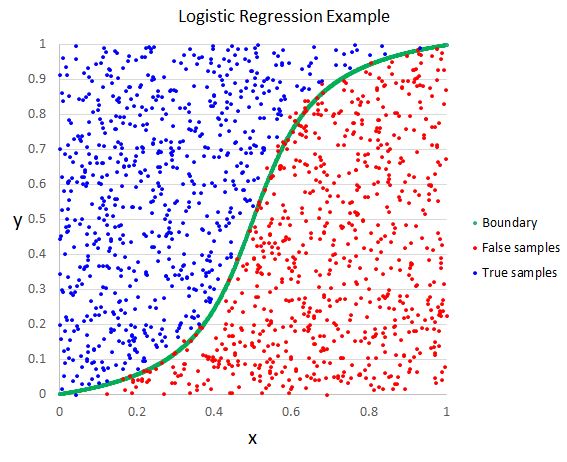
* Forêt aléatoire

Les forêts d'arbres décisionnels[1](https://fr.wikipedia.org/wiki/For%C3%AAt_d%27arbres_d%C3%A9cisionnels#cite_note-1) (ou forêts aléatoires de l'anglais random forest classifier) font partie des techniques d'[apprentissage automatique](https://fr.wikipedia.org/wiki/Apprentissage_automatique). Cet algorithme combine les concepts de sous-espaces aléatoires et de [bagging](https://fr.wikipedia.org/wiki/Bagging). L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.



* Régression logistique

La régression logistique ou modèle logit est un modèle de [régression](https://fr.wikipedia.org/wiki/R%C3%A9gression_(statistiques)) binomiale. Comme pour tous les modèles de régression binomiale, il s'agit de modéliser au mieux un modèle mathématique simple à des observations réelles nombreuses. En d'autres termes d'associer à un vecteur de variables aléatoires une variable aléatoire. La régression logistique constitue un cas particulier de [modèle linéaire généralisé](https://fr.wikipedia.org/wiki/Mod%C3%A8le_lin%C3%A9aire_g%C3%A9n%C3%A9ralis%C3%A9). Elle est largement utilisée en [apprentissage automatique](https://fr.wikipedia.org/wiki/Apprentissage_automatique).



* LightGBM

Light Gradient Boosting Machine, libre et opensource, developpé par Microsoft. Basé sur un arbre de décision LightGBM adopte deux nouvelles techniques GOSS (Gradient-based One-Side Sampling) et EFB (Exclusive Feature Bundling). Avec GOSS, LightGBM peut entraîner chaque arbre avec seulement une petite fraction du jeu de données complet. Avec EFB, LightGBM gère beaucoup plus efficacement les fonctionnalités clairsemées de haute dimension. LightGBM prend également en charge l’entrainement distribué avec un faible coût ce qui réduit les besoins sur les GPU.

1. **FONCTION COÛT, ALGORITHME D’OPTIMISATION, MÉTRIQUE D’ÉVALUATION**
2. **Fonctions de pertes**

Les fonctions de pertes utilisées par chacun des modèles listés ci-dessus sont celles implémentées sous scikit-learn.

Certains modèles scikit-learn permettent de modifier la fonction de perte en modifiant un paramètre d’entrée. Lorsque cela était le cas, nous l’avons spécifié comme hyperparamètre dans l’espace de recherche, par exemple pour la forêt aléatoire :

'criterion': hp.choice('rf\_criterion', ["gini", "entropy"])

La sélection du modèle est faite en se basant sur l’évaluation, par validation croisée à 3 plis, à l’aide de la métrique personnalisée décrite ci-dessous.

1. **Algorithmes d’optimisation**

Les algorithmes d’optimisation utilisées par chacun des modèles listés ci-dessus sont ceux implémentés sous scikit-learn.

Certains modèles scikit-learn permettent de modifier l’algorithme d’optimisation en modifiant un paramètre d’entrée. Lorsque cela était le cas, nous l’avons spécifié comme hyperparamètre dans l’espace de recherche, par exemple pour la régression logistique :

'solver' : hp.choice('logit\_solver', ['newton-cg', 'lbfgs', 'liblinear']),

Lightgbm est optimisé par randomsearch, préféré à gridsearch pour sa plus grande rapidité.

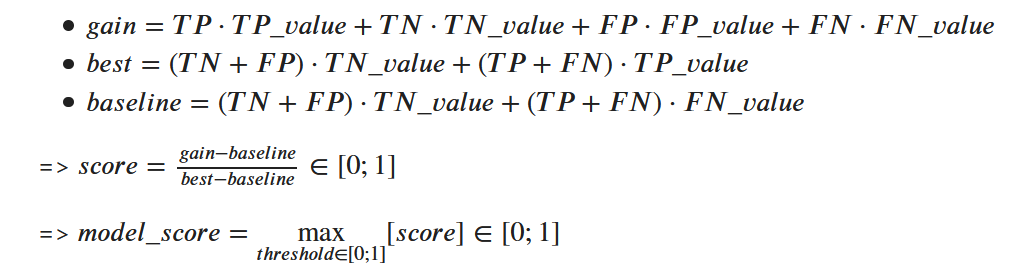
1. **Métrique d’évaluation**

Les classes cibles du jeu de données initial sont très déséquilibrées (plus de 90% des crédits sont remboursés sans défaut). Cela rend la métrique *accuracy* peu pertinente.

De plus, on peut supposer que l’impact pour l’entreprise d’un faux négatif et ceux d’un faux positif ne sont pas les mêmes, ce qui rend une *AUC* peu pertinente.

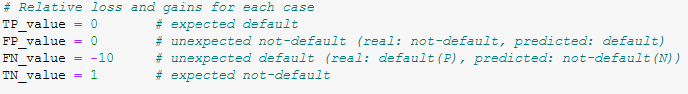
* FN, erreur de 1ère espèce :  client à qui l’on refuse un prêt, mais qui l’aurait honoré
* FP, erreur de 2ème espèce : client à qui l’on accorde un prêt, mais qui fait défaut

Nous avons donc construit une métrique d’évaluation « métier » qui a servi à la sélection des modèles et des hyperparamètres lors de l’exécution de l’algorithme TPE (hyperopt).



Ce métrique se base sur 4 paramètres TP\_value, TN\_value, FN\_value et FP\_value qui sont respectivement les coûts associés aux vrais positifs (TP), vrais négatifs (TN), faux négatifs (FN) et faux positifs (FP). Voici les valeurs de ces paramètres que nous avons utilisés pour la construction du modèle. Ces valeurs peuvent facilement être modifiés pour tenir compte de la « réalité métier ».

Les FN présentent un risque important et ont donc un coefficient de 10.



La valeur de la perte ou de gain pour l’entreprise est fonction du seuil de probabilité de défaut à partir duquel l’entreprise refusera le prêt. C’est donc un hyperparamètre qui est également à optimiser, dans une optique d’aide à la décision, et c’est pourquoi le score du modèle correspond à un maximum lorsque le seuil varie entre 0 et 100%.

Enfin, ce métrique a été normalisée afin que :

* 0 : représente le score obtenu par un modèle naïf (prédiction systématique de la classe majoritaire N : pas de défaut de paiement).
* 1 : représente le score maximal, obtenu lorsque toutes les prédictions sont correctes.

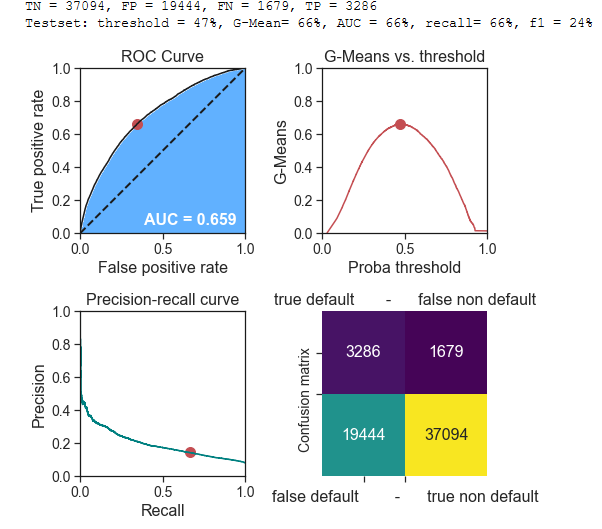
1. **INTERPRÉTABILITÉ DU MODÈLE**

Le modèle étant sélectionné sur la seule base de ses performances, il n’est pas certain qu’il s’agisse d’un modèle interprétable.

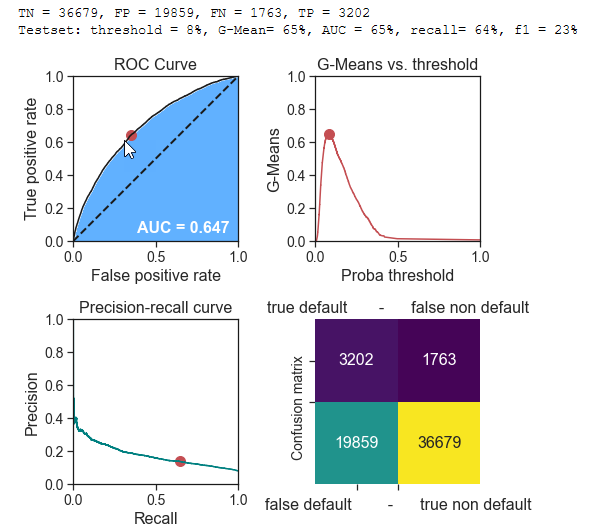
Pour répondre au besoin d’interprétabilité et au besoin de rapidité de calcul pour l’API, nous avons donc privilégié des méthodes « indépendantes du modèle » (*model-agnostic*). Nous avons utilisé la méthode du « *surrogate model* » (modèle de substitution) qui consiste à entraîner un modèle interprétable sur les prédictions du modèle à interpréter en forçant le sur-apprentissage. Ce modèle (arbre) est alors interprétable : globalement, avec les « *features importances* », et localement en utilisant la librairie treeinterpreter.

Le surrogate model a été testé sur le Random Forest et sur Lightgbm, a été sélectionné Lightgbm car ses performances sont supérieures sur le faux non défaut, coefficient de 10 dans la fonction de coût.

LGBM optimisé avec randomizedsearch



RF optimisé avec gridsearch



1. **LIMITES & AMÉLIORATIONS**
2. **Modèles et hyperparamètres**
3. Un premier axe d’amélioration est d’intégrer dans l’espace de recherche d’autres familles de modèles et notamment :

* Boosting : XGBoost, CatBoost…
* Réseaux neuronaux : fully-connected avec Keras/TensorFlow

1. On peut aussi essayer d’intégrer davantage d’étapes de pré-traitement (pour, par exemple comparer les méthodes d’équilibrage des classes, ou les méthodes d’imputation).
2. **Interprétabilité**

La prédiction de solvabilité repose sur le modèle entier.  
L’interprétabilité repose sur le modèle surrogate: treeinterpreter sur arbre de décision. Il faut alors faire une mise à jour régulière du modèle surrogate, au fur et à mesure que le nombre de nouveaux clients augmente. Ceci afin que les features globales de la population ne diffèrent pas du modèle entier.