**NOTE METHODOLOGIQUE**

**Projet 7 :** **Implémentez un modèle de scoring**

**;kloàlol)koù$  
$**

**Introduction :**

Le projet 7 a pour objectif de mettre en œuvre un outil de « scoring crédit » pour calculer la probabilité qu’un client rembourse ou non un crédit.

Il s’agit alors de développer un algorithme de classification binaire avec une variable cible « TARGET » ayant 2 classes « 0 : ceux qui rembourse » et « 1 : ceux qui auront des difficultés à rembourser ».

**Plan de la méthodologie :**

Nous allons dans ce document :

* Décrire la **méthodologie d’entraînement** du modèle,
* Voir la **fonction coût métier** qui nous servira de métrique pour choisir le meilleur modèle, **l’algorithme d’optimisation** de notre modèle choisi et la **métrique d’évaluation**,
* Puis à l’aide de Shap, nous **interpréterons de manière globale et locale** notre modèle,
* Et enfin, nous proposerons **les limites et les améliorations** que nous pourrions avoir pour ce modèle.

1. **Méthodologie d’entraînement du modèle :**

Cette partie donne suite à la préparation de notre base de données finale (fusion des bases de données en s’inspirant du kaggle mentionné dans le projet, ainsi que feature engineering et cleaning de la base fusionnée).

Nous avons ensuite séparé notre base de données en une partie entrainement (80%) et une partie test (20%).

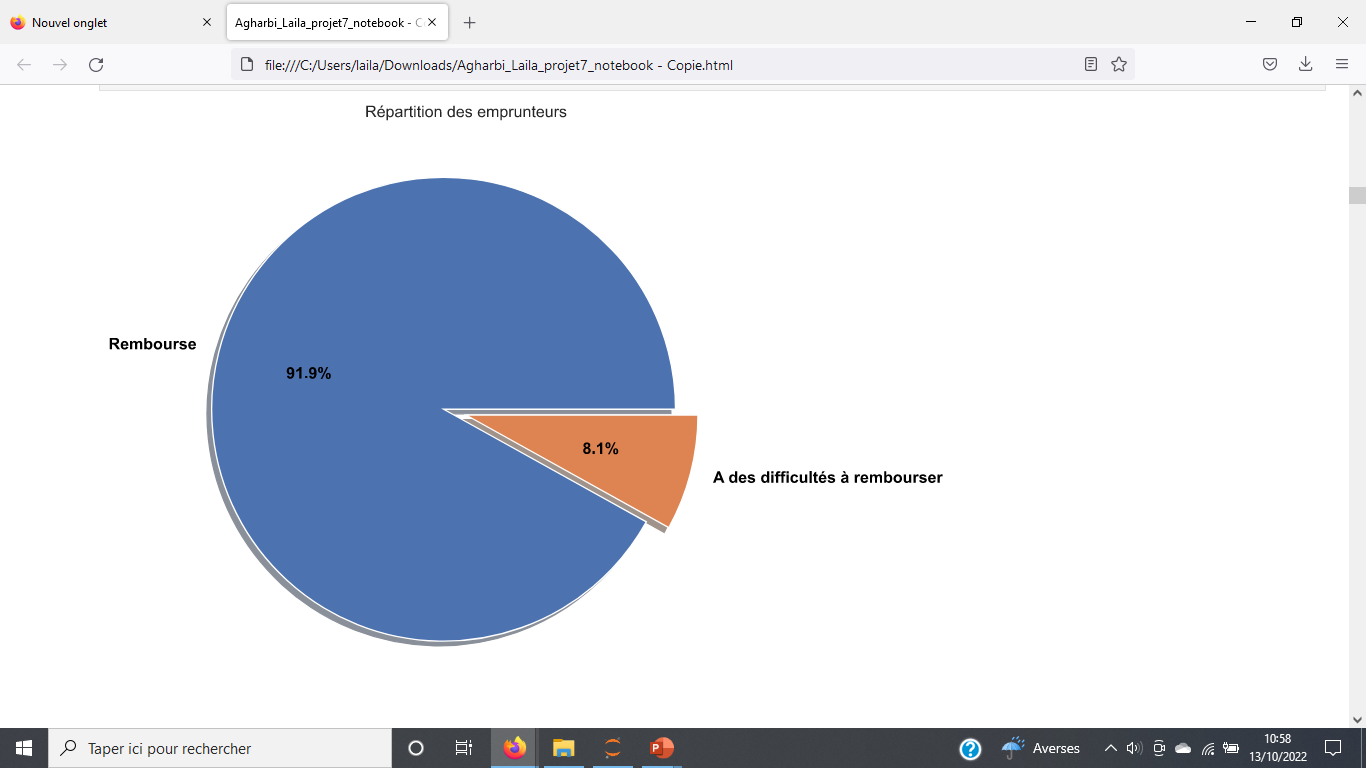
Puis, nous avons imputé la médiane aux valeurs manquantes des variables et standardisé les données des variables non binaires.

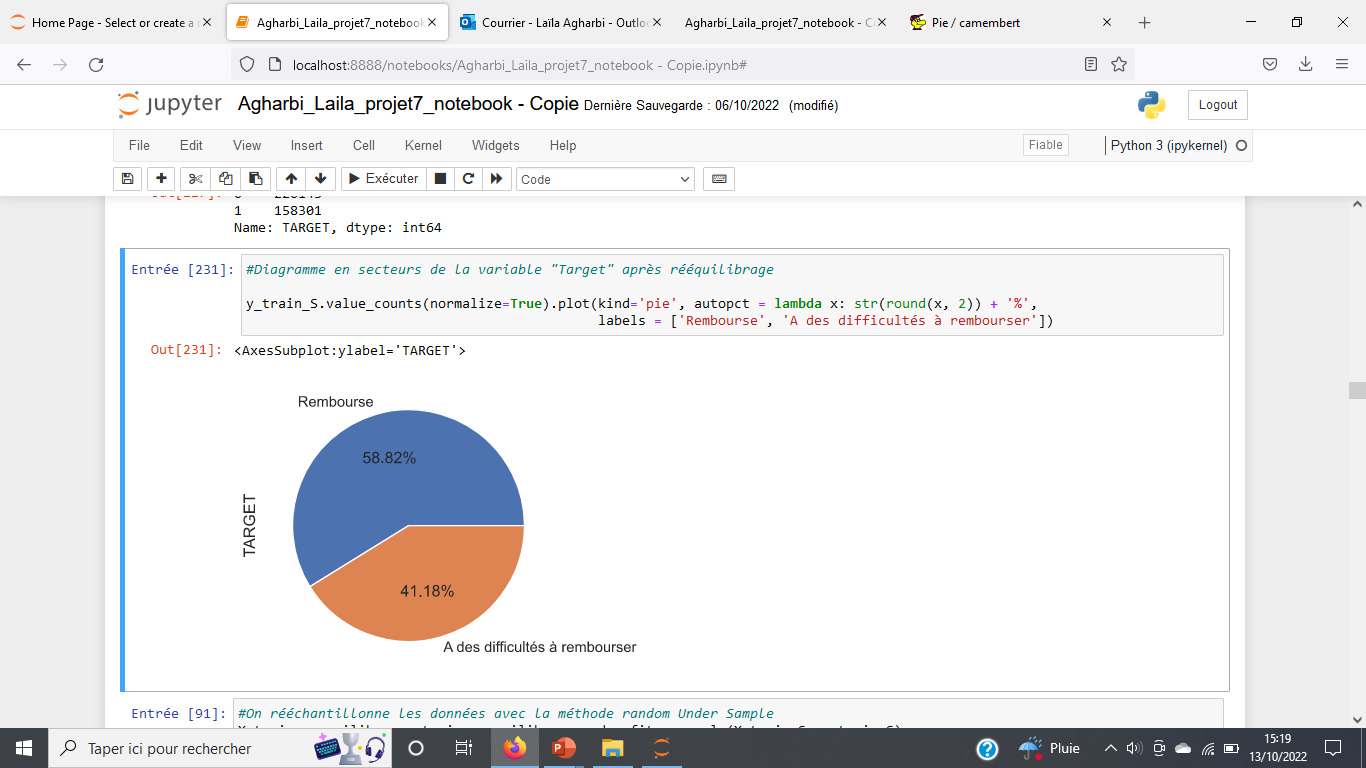
Nous avons ensuite choisi de tester et de comparer les métriques des classifieurs suivants :

* 1. Dummy Classifier (notre baseline)
  2. Light GBM Classifier
  3. Random Forest
  4. Regression logistique

Avant de faire ces tests, nous allons résoudre le problème de déséquilibrage des classes.

En effet, la variable « TARGET » a la répartition suivante :



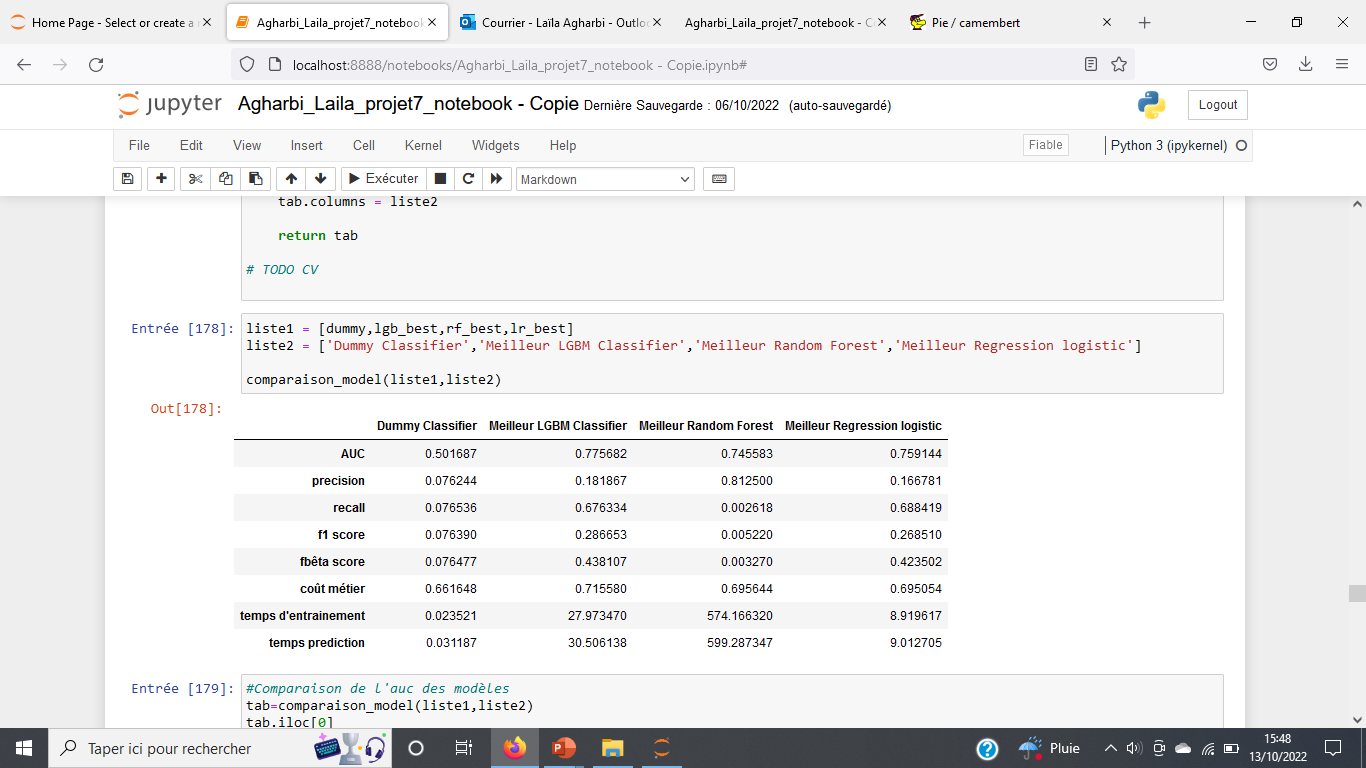
Pour rééquilibrer les classes, nous allons utiliser la librairie SMOTE. Celle-ci va répéter les individus minoritaires ùùùjusqu’à obtenir l’équilibre souhaité.

Ce qui nous permet d’obtenir la répartition des classes suivante :

Elle sera utilisée pour répartir les classes dans le modèle dummy, mais on va se servir des paramètres des modèles telles que « class\_weight » pour permettre de palier à ce problème.

Pour sélectionner le meilleur modèle, nous allons entrainer chacun de ses modèles (sauf le dummy qui sert de baseline) en cherchant à optimiser les paramètres à l’aide du gridsearchCV, et procédant à une validation croisée grâce à la méthode K-Folds (séparation en 5 folds).

On compare ensuite les métriques des meilleurs modèles (ceux obtenus après avoir lancer le gridSearchCV), et on obtient le tableau comparateur suivant :



Le modèle LGBM est celui qui a les meilleurs rapports temps / métriques. C’est donc celui que l’on choisira.

1. **La fonction coût métier, l’algorithme d’optimisation et la métrique d’évaluation :**

Pour calculer le cout métier, nous avons besoin desavoir l’importance d’un faux positif ou d’un faux négatif pour notre banque. Le point de vue métier nous permet de dire qu'un faux positif coutera moins cher qu'un faux négatif. En effet, un individu qui sera prévu comme pouvant remboursant son prêt (donc un faux négatif) mais qui en réalité, rencontrera des difficultés à le faire, coûtera 10 fois plus d'argent que celui à qui on n’a pas fait de prêt (parce que prévu comme étant mauvais emprunteur, alors qu'il aurait pu être un bon emprunteur en réalité).

A l’aide de la matrice de confusion, nous pouvons déterminer nos faux positifs, faux négatifs, vrais positifs et vrais négatifs.

Nous allons accorder à chacun d’eux un poids en fonction de ce qu’ils apportent ou non à la banque dans le calcul de ce coût métier :

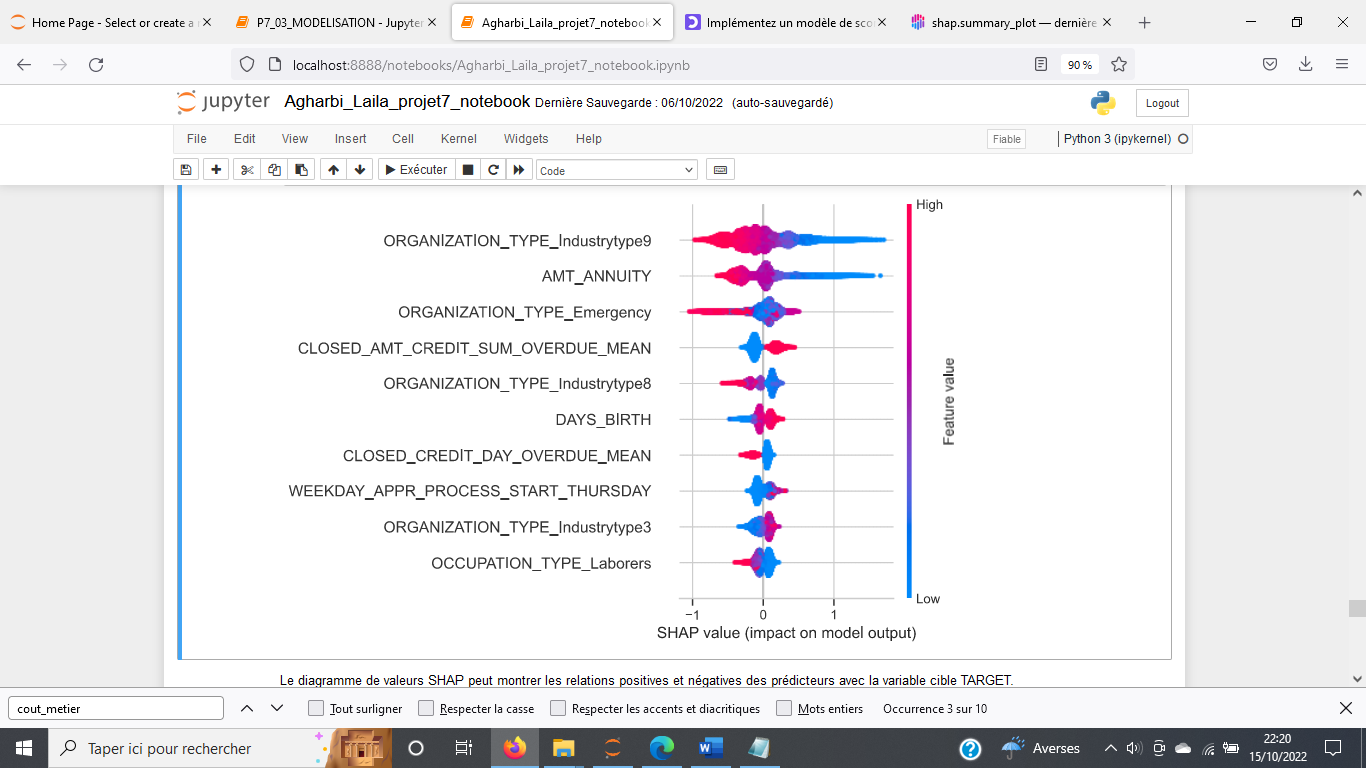
*-* ***Faux négatifs : -10.*** *Le prêt est accordé mais le client a difficulté à rembourser : la banque perd de l'argent.  
-* ***Faux positifs : -1.*** *Le prêt est refusé par erreur : la banque aurait pu gagner de l'argent, mais ne perd pas réellement d'argent.  
-* ***Vrais positifs : 0.*** *Le prêt est (à juste titre) refusé : la banque ne gagne ni ne perd d'argent.  
-* ***Vrais négatifs : 1.*** *Le prêt est remboursé : la banque gagne de l'argent.*

A partir d’HyperOpt, on va rechercher les meilleurs paramètres, en prenant pour métrique ce coût métier, de notre modèle choisie LGBM. Cela nous permettra d’optimiser celui-ci de façon à maximiser le gain pour notre banque.

Les métriques permettant d’évaluer notre modèle sont donc la fonction coût métier, l’AUC score ainsi que le f1score.

1. **Interprétation globale et locale du modèle :**

* **Global :**



Les features sont classés de haut en bas par ordre d’importance. Ainsi, dans notre modèle, la variable « type d’organisation où le client travaille ; Industrie » est la plus influente. Il semble que la majorité des individus de cette catégorie sont susceptibles de pouvoir rembourser. Elle est suivie du montant de l’annuité ; ¨Plus le montant de l’annuité est faible et plus, on pourra rembourser.

* Une image contenant texte

  Description générée automatiquement**Local :**

Il s’agit ici du premier individu de notre liste de clients test.

Sa valeur de prédiction moyenne est de -2.75, elle sera ensuite convertie en une valeur de 0 ou 1. Shap a tracé les features qui ont le plus influencé pour la prédiction pour cet individu. Les features en rouge influencent positivement et ceux en bleu négativement.

Ainsi le client a demandé le prêt un jeudi, et cela a influencé positivement.

En revanche, ce qui a eu plutôt un impact négatif est le montant du prêt ainsi que sa date de naissance.

1. **Limites et améliorations possibles :**

Le nombre important d’individus et de variables demandent un temps machine assez important lors du déploiement du dashboard.

Avec une meilleure compréhension des variables et de leur impact pour la demande de crédit, ceci grâce à l’aide d’un expert, nous aurions peut-être pu diminuer ce nombre, et avoir un plus grand impact sur nos prédictions.

Les scores de notre métrique créer par la fonction coût métier ont été donné de manière arbitraire. Ceux-ci auraient pu être affinés par le point de vue d’un éventuel expert.