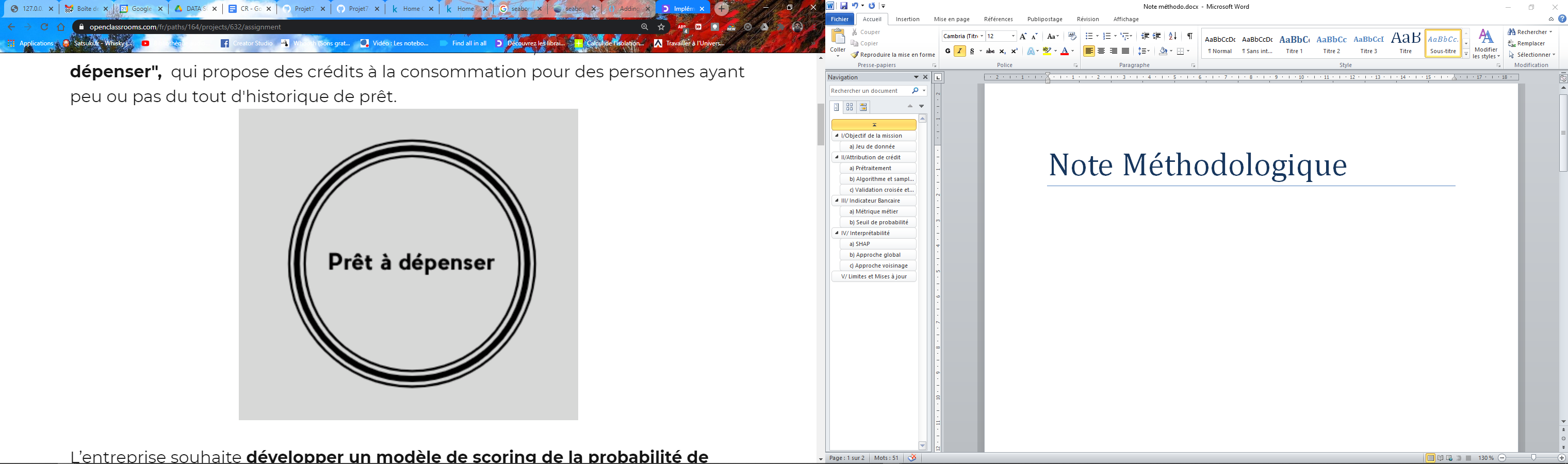
Note Méthodologique

Attribution de crédit à la consommation



Février 2021

**Projet 7 – OpenClassroom**

*Implémentez un modèle de scoring* DENIS Jean-Benoît

Table des matières

[I/Objectif de la mission 3](#_Toc62812810)

[a) Contexte 3](#_Toc62812811)

[b) Jeu de données 3](#_Toc62812812)

[II/Attribution de crédit 3](#_Toc62812813)

[a) Prétraitement 3](#_Toc62812814)

[b) Validation croisée et optimisation Bayésienne 4](#_Toc62812815)

[c) Algorithme et Sampling 4](#_Toc62812816)

[III/ Indicateur Bancaire 5](#_Toc62812817)

[a) Métrique métier 5](#_Toc62812818)

[b) Seuil de probabilité 6](#_Toc62812819)

[IV/ Interprétabilité 6](#_Toc62812820)

[a) SHAP 6](#_Toc62812821)

[b) Approche global 7](#_Toc62812822)

[c) Approche voisinage 8](#_Toc62812823)

[V/ Limites et Mises à jour 8](#_Toc62812824)

# I/Objectif de la mission

## Contexte

La société financière «***Prêt à Dépenser*** », société de crédit de consommation pour des personnes ayant peu ou pas du tout d'historique de prêt, souhaite développer un outil de scoring permettant d’obtenir la probabilité de défaut de paiement de ses clients.

Dans une volonté de transparence auprès de ses clients, la société souhaite également mettre en place un dashboard interactif de présentation afin que les chargés de relation clients puissent exposer les décisions d’octroi ou non de prêt.

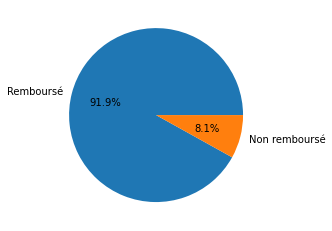
## Jeu de données

La base de données fournit pour cette mission rassemble des informations pour 301 511 clients selon 122 indicateurs. Ces informations sont du type générique comme l’âge, le sexe, les revenus, l’emploi, le logement, les informations de crédit en cours, des notations externes, etc.

En se penchant de plus près sur la base de données, il apparaît un fort déséquilibre entre les personnes solvables et les personnes non solvables. Ce constat est représenté sur la figure ci-dessous.

**Non Solvable**

**Solvable**



# II/Attribution de crédit

## Prétraitement

Afin d’exploiter au mieux le set de données, des étapes de prétraitement doivent être appliquées. Il s’agit d’étapes de nettoyage des données, d’encodage numérique des variables catégorielles, ainsi que la création de nouveaux indicateurs génériques comme le ratio du montant de crédit sur les revenu ou encore le terme du crédit. Enfin une imputation des valeurs manquantes est réalisée et une mise à l’échelle est appliquée.

## Validation croisée et optimisation Bayésienne

La méthodologie de modélisation employée est la suivante. D’abord, le jeu de données est divisé en deux sets, le premier pour entrainer le modèle (Training) et le second pour l’évaluer (Test). Le set de training est ensuite séparé en plusieurs sous-ensembles (**validation croisée**) afin d’optimiser les hyper paramètres du modèle. Et pour améliorer cette optimisation, une **approche bayésienne** est mise en place. Cette méthode probabilistique permet de sélectionner les valeurs des hyper paramètres à tester dans le but de minimiser une fonction de perte, ce qui permet un gain de temps dans le training du modèle.



## Algorithme et Sampling

Le modèle de classification choisit pour cette étude est le modèle ***Light Gradient Boosting Machine*** ou Light-GBM. Il s’agit d’un modèle robuste, originellement développé par Microsoft, et qui repose sur le principe d’arbres de décisions.

Comme mentionné précédemment la base de données utilisée comporte un fort déséquilibre entre les clients solvables ou non. Cette différence du nombre d’entrée va avoir un impact sur la performance de l’algorithme et faussera les prédictions attendues, attribuant beaucoup plus fréquemment la classe la plus représentée aux clients. Pour pallier à ce problème de déséquilibrage, trois approches de sampling, appliquées uniquement sur les sous-ensembles de la validation croisée, ont été testées :

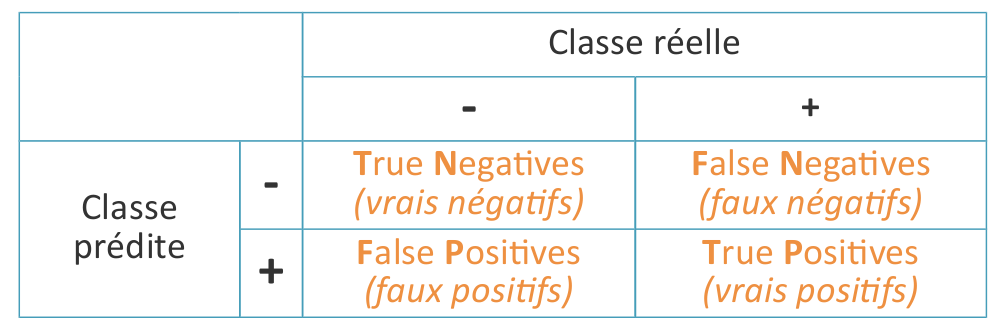
* Le **Sample weights** est une méthode directement gérée par le modèle Light-GBM et qui permet de pénaliser les poids associés aux observations sur représentées.
* L’**Over-sampling** ou **SMOTE** est une méthode qui va synthétiser de nouvelles entrées de la classe sous représentée à partir des données existantes pour équilibrer la base de données.
* L’**Under-sampling** est une méthode qui va sélectionner une partie des observations majoritairement représentées pour équilibrer la base de données.

# III/ Indicateur Bancaire

## Métrique métier

Afin de rendre le modèle le plus cohérent possible une nouvelle **métrique spécifique** à la mission est implémentée. Il s’agit d’un indicateur bancaire qui va permettre au modèle de pénaliser ou non un client en fonction de ses informations.

L’objectif pour la société bancaire étant de gagner de l’argent, le gain sera alors calculé sur les prédictions faites par l’algorithme. La matrice de confusion en sortie de prédiction est représentée dans le tableau suivant :



Les ‘**-**‘ représentent les clients solvables, c’est-à-dire **négatif au refus de prêt** et les ’**+**’ les clients non solvables, c’est-à-dire **positif au refus de prêt**.

Dans ce calcul de gain, si un prêt est accordé à une personne non solvable, la banque perdra de l’argent. Il s’agit des « **Faux Négatif** » dans la matrice. De même si un prêt est refusé à une personne solvable, les « **Faux Positif** », la banque aura perdu un gain d’argent potentiel. Au contraire si l’attribution, **Vrai Négatifs** » ou le refus, « **Vrais Positifs** », de prêt est correct la banque gagnera de l’argent ou à défaut n’en perdra pas.

La fonction métrique se construit alors en pénalisant d’une valeur arbitraire, ces prédictions erronées et en gratifiant les prédictions correctes.

Les scores arbitrairement choisis sont les suivantes :

* FN 🡺 Perte d’argent pour la banque : **- 100**
* TP 🡺 Refus de prêt, la banque ne perd pas d’argent : **0**
* TN 🡺 Prêt accordé, gain d’argent pour la banque : **+ 10**
* FP 🡺 Client potentiel perdu, perte d’argent pour la banque : **- 1**

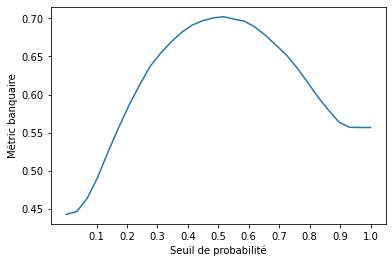
A partir de ces scores, le **gain total** va être calculé et normalisé selon les équations suivantes :

Avec :

C’est donc cette fonction gain que le modèle va chercher à **maximiser** durant l’optimisation de ses hyper paramètres.

## Seuil de probabilité

Le **seuil de probabilité** est la valeur de prédiction à partir de laquelle le modèle va considérer qu’un client est solvable ou non compte tenu de la fonction gain employée. Les résultats montrent un seuil à environ **52%**.



Un client avec une prédiction au-dessus de cette valeur sera catégorisé par le modèle comme non solvable et un client avec une prédiction en dessous de cette valeur sera considéré comme solvable.

Pour plus de précision, l’optimisation de cette valeur seuil est considérée comme un hyper paramètre dans l’approche bayésienne.

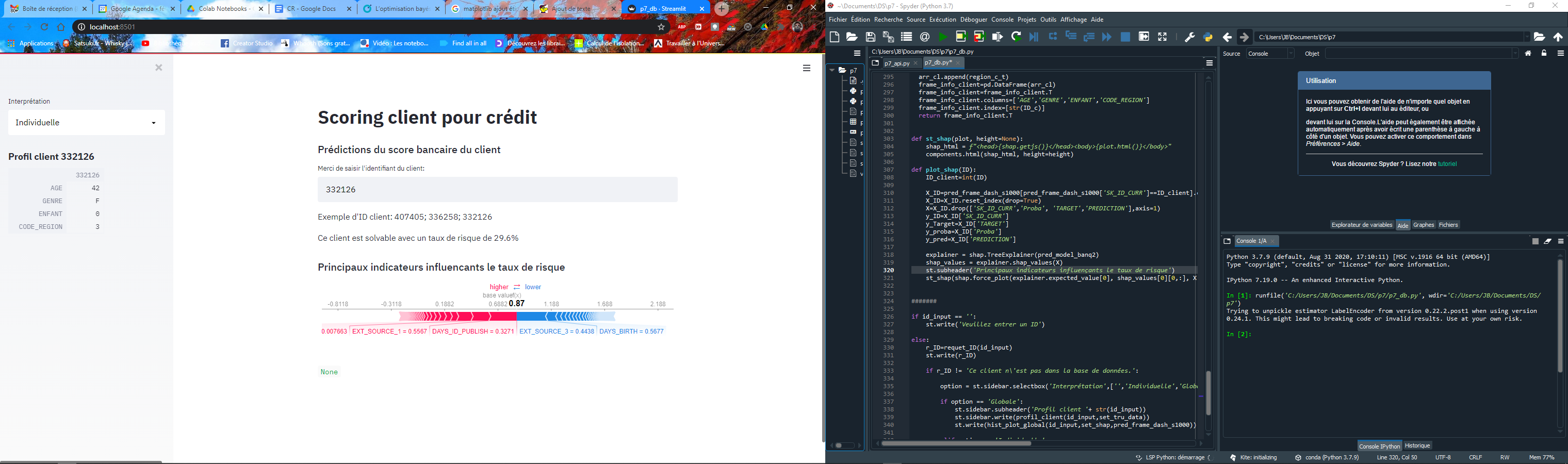
# IV/ Interprétabilité

Pour garantir la transparence auprès de ses clients, un **dashboard interactif** est réalisé afin de comprendre comment la décision d’attribution ou de refus de prêt est prise.

**URL du dashboard** : <https://creditscorep7.herokuapp.com/>

## SHAP

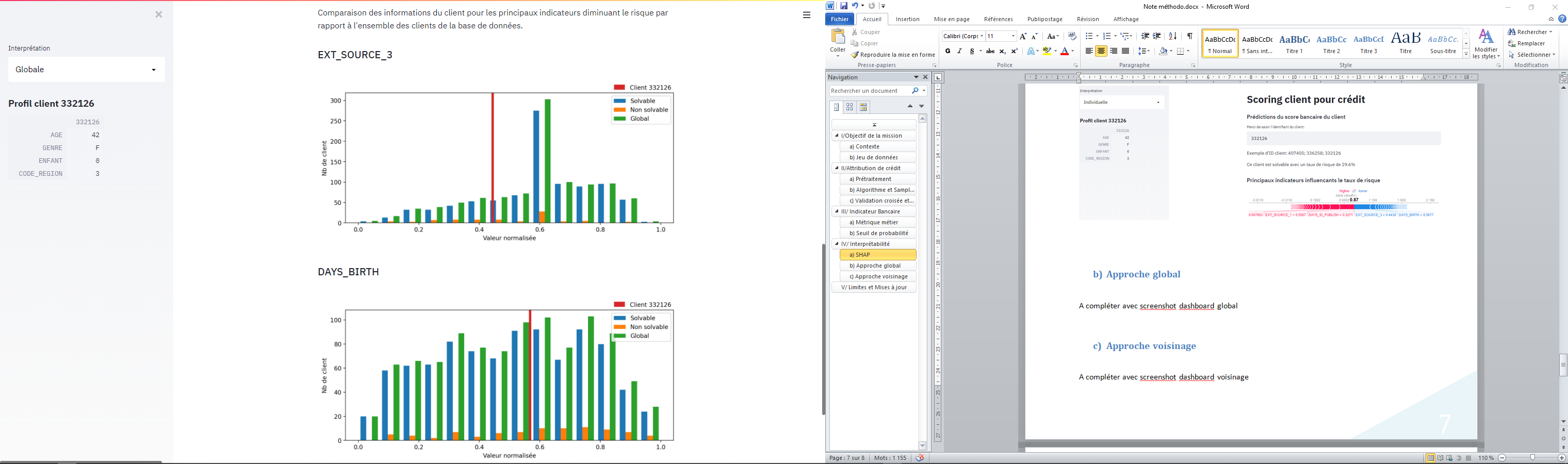
La librairie **SHAP** permet d’observer l’évolution des prédictions en fonction de l’importance de certains critères. Par exemple, pour un client solvable, le dashboard va montrer un graphe qui représente les critères les plus importants retenus par le modèle pour réaliser l’attribution du prêt.



Pour ce client, les indicateurs « ***Days\_ID\_PUBLISH*** » et « ***EXT\_SOURCE\_1*** » vont faire augmenter le risque de défaut de remboursement. Alors que les indicateurs « ***EXT\_SOURCE\_3*** » et « ***DAY\_BIRTH*** » vont faire tendre le modèle vers l’attribution du prêt bancaire.

## Approche global

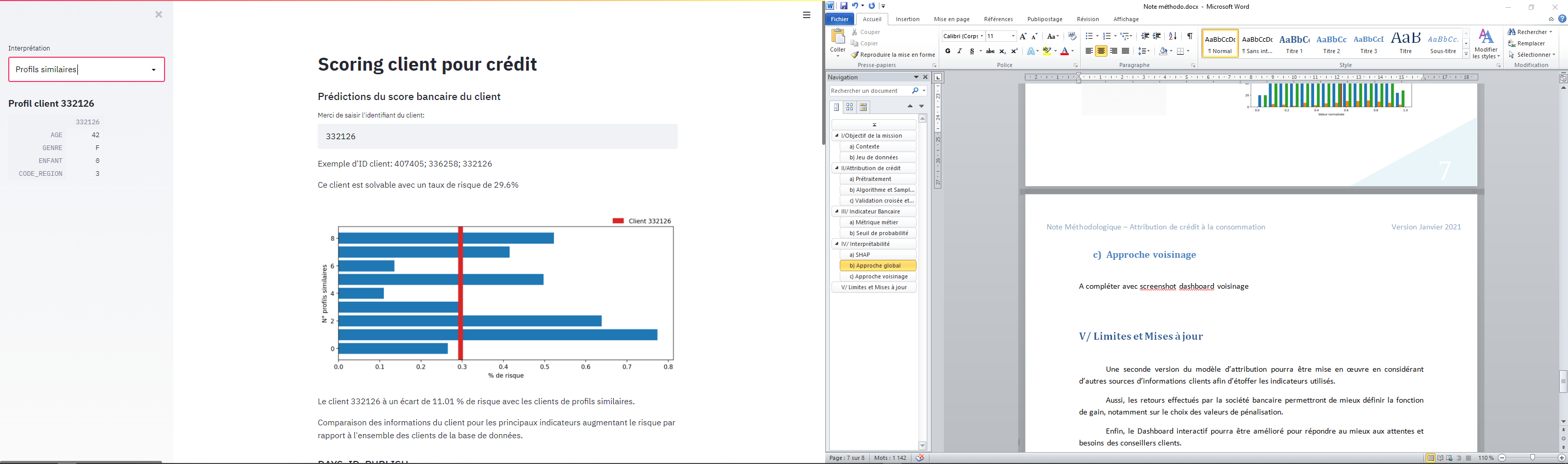
A partir des indicateurs identifiés par la librairie SHAP, il est alors possible de comparer l’information d’un client pour ces indicateurs par rapport aux données des autres clients de la base de données. Les distributions de la totalité des clients ainsi que celles des clients solvables et non solvables sont affichées sur le dashboard.



Pour ce client, les graphiques pour ses indicateurs montrent qu’il se situe plutôt dans la moyenne de la totalité de la base de données ainsi que celle des clients solvable.

## Approche voisinage

Enfin, il est également possible d’afficher ou se situe un client par rapport aux clients de même profil. Le profil est définit selon des critères simples comme l’âge, le sexe, le nombre d’enfant ou encore le code région d’habitation.



Le graphique montre que le risque de ce client se situe encore une fois dans la moyenne avec les scores des autres clients de même profils avec un écart moyen de 11%. Le dashboard permet également d’afficher où se situe le client par rapport aux distributions des indicateurs les plus influents des clients de mêmes profils.

# V/ Limites et Mises à jour

Une seconde version du modèle d’attribution pourra être mise en œuvre en considérant d’autres sources d’informations clients afin d’étoffer les indicateurs utilisés. Egalement une attention particulière devra être posée sur les méthodes de gestion de déséquilibrage des données, notamment l’over sampling, afin d’éviter le sur apprentissage du modèle sur des données fictives qui peuvent être erronées.

Aussi, les retours effectués par la société bancaire permettront de mieux définir la fonction de gain, notamment sur le choix des valeurs de pénalisation. Il serait également intéressant de mettre en place une pondération de la métrique selon le montant du prêt demandé.

Enfin, le Dashboard interactif pourra être amélioré pour répondre au mieux aux attentes et besoins des conseillers clients.