Implémenter un modèle de scoring

Sommaire

1. Rappel de la problématique et présentation du jeu de données

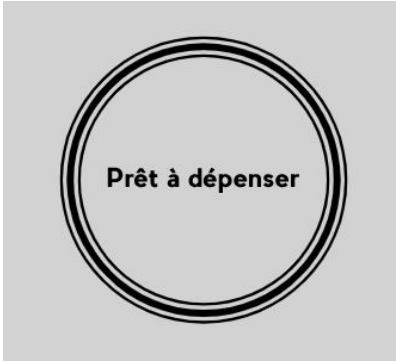
2. Explication de l’approche de modélisation

3. Présentation du dashboard métier

Associer un score à un profil client

Contexte :

Prêt à dépenser est une société de crédits à la consommation pour des personnes avec peu ou pas d'historique de prêt.



Objectifs :

Développer un modèle de scoring reflétant la probabilité de défaut de paiement du client.

Développer un dashboard interactif permettant d'interpréter le score d'un client

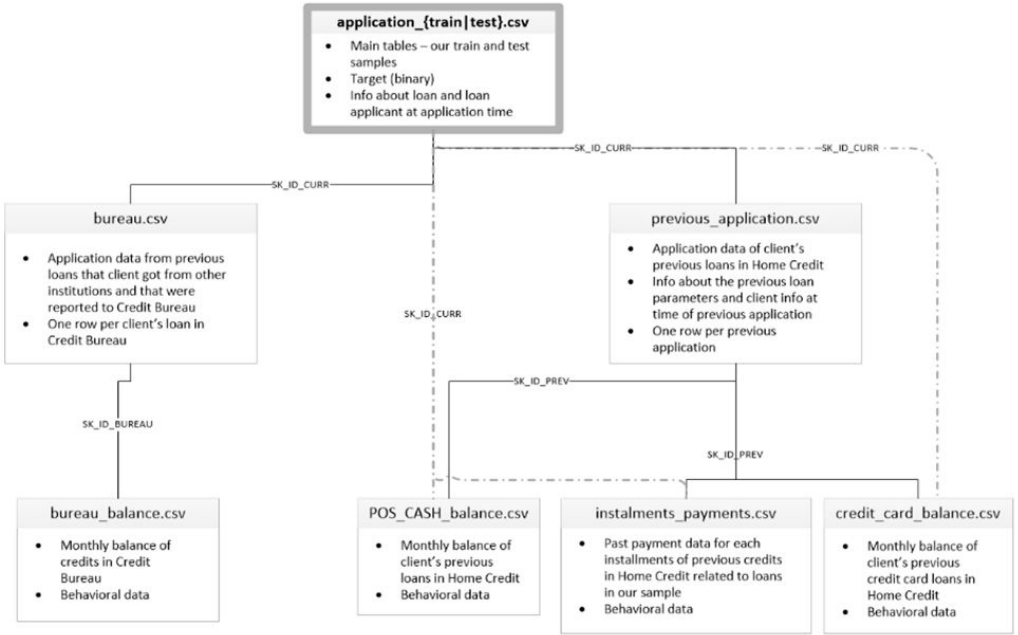
Structures des données à disposition

7 sources de données (informations relatives aux clients et la société, historique et balance de crédit, ...)

307000 clients référencés

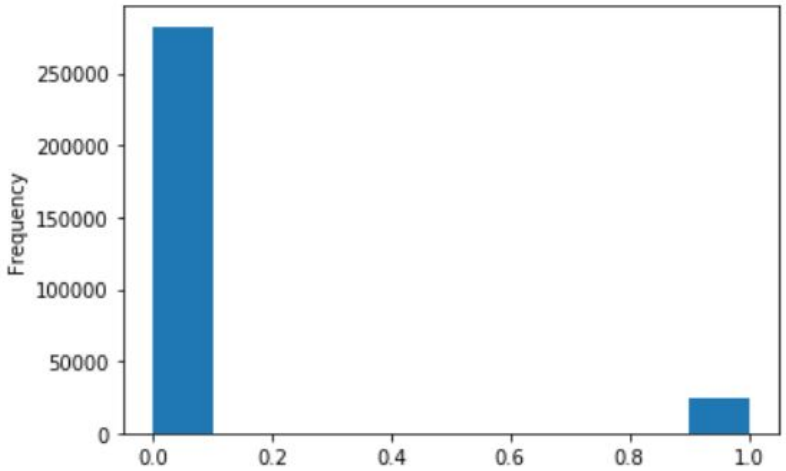
+121 variables (âge, sexe, emploi, logement, informations crédit, ...)

Information cible : profil ayant fait **défaut ou non** au recouvrement du crédit



Un jeu de données déséquilibré

répartition des clients ayant fait défaut ou non au crédit



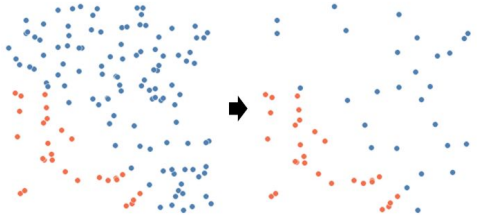
91 % des clients sans défaut de paiement

Risques :

- Modèle naïf
- Surreprésentation de la classe majoritaire dans la prédiction

Méthode de réduction du déséquilibre

Under sampling

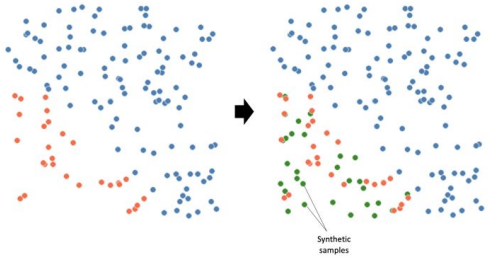


Réduction du nombre d'observations associées à la classe majoritaire

---

SMOTE

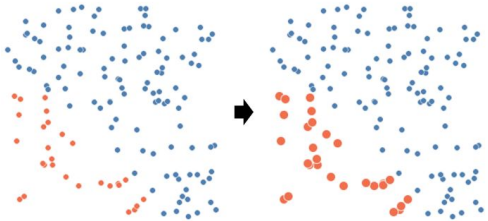
(Synthetic Minority Oversampling Technique)



Création d'observations synthétiques associées à la classe minoritaire

---

Class Weights



Pondération de chaque classe pour que le poids cumulé de chaque individu soit identique entre chaque classe

Implication métier de l'estimation

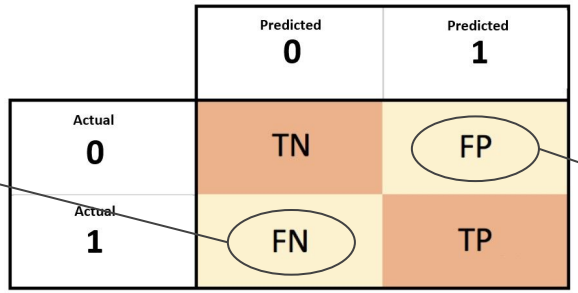
Définition d'une métrique de score adaptée

Faux Négatif:

Clients identifiés sans défaut, aboutissant à un défaut réel

Représentent un risque financier pour la société

Minimiser le taux de faux négatif revient a chercher a maximiser le recall



Faux Positif:

Clients identifiés en défaut, mais n'aboutissent à aucun défaut réel

Représentent un manque à gagner

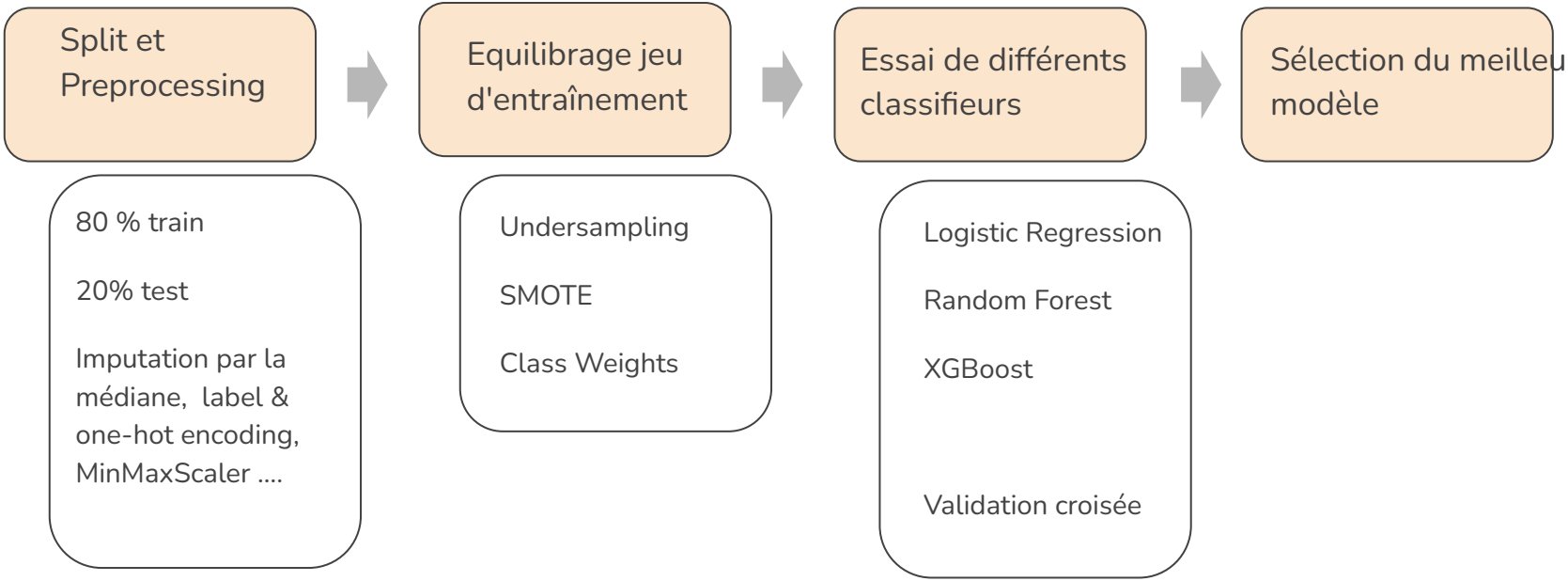
Minimiser le taux de faux positif revient a chercher a maximiser la précision

F-bêta score correspond à la moyenne harmonique du recall et de la precision

Le coefficient bêta permet de pondérer l'importance du recall par rapport à la précision

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

Méthodologie de modélisation

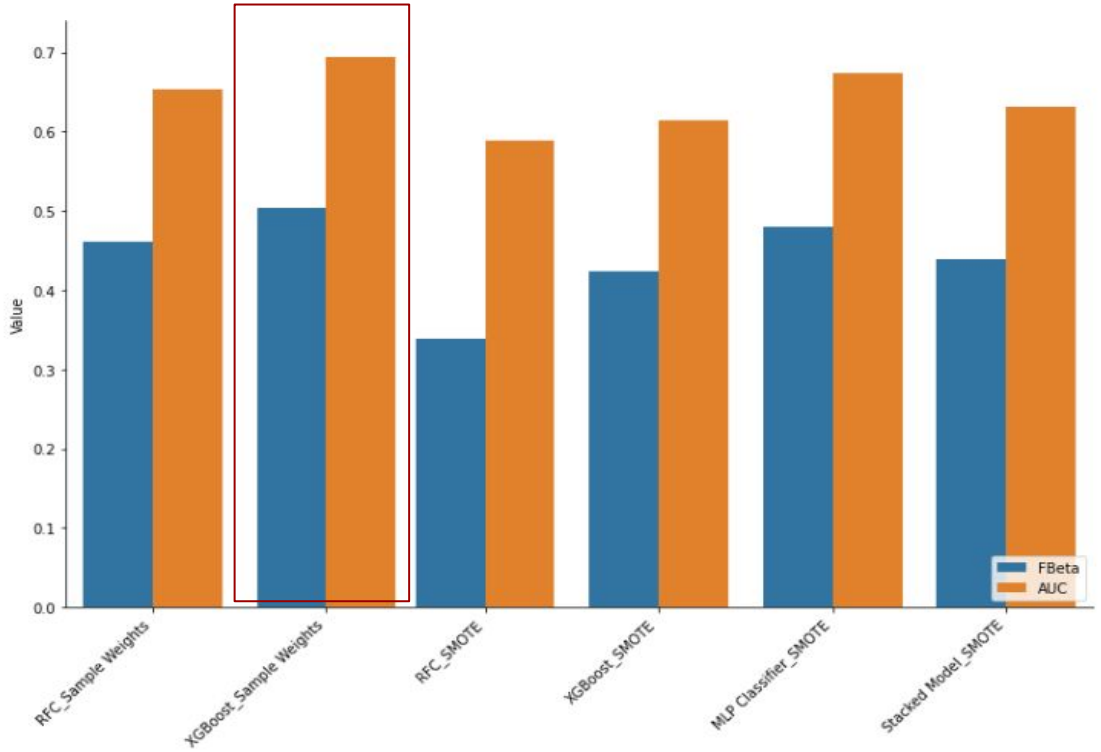


Résultats

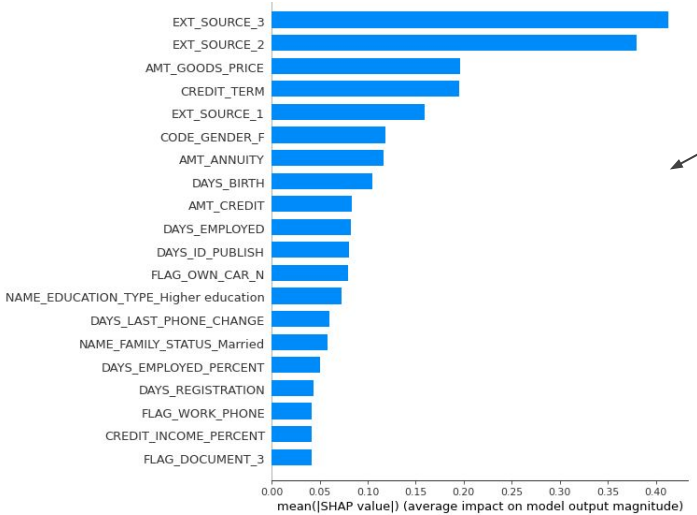
Meilleur modèle : XGBoost (class weights)

F-beta = 0.51

AUC = 0.70



Explicabilité des estimations



Les **sources extérieures**, le **montant** du bien et le **terme** du prêt sont les variables les plus importantes pour expliquer le comportement global de l'estimateur

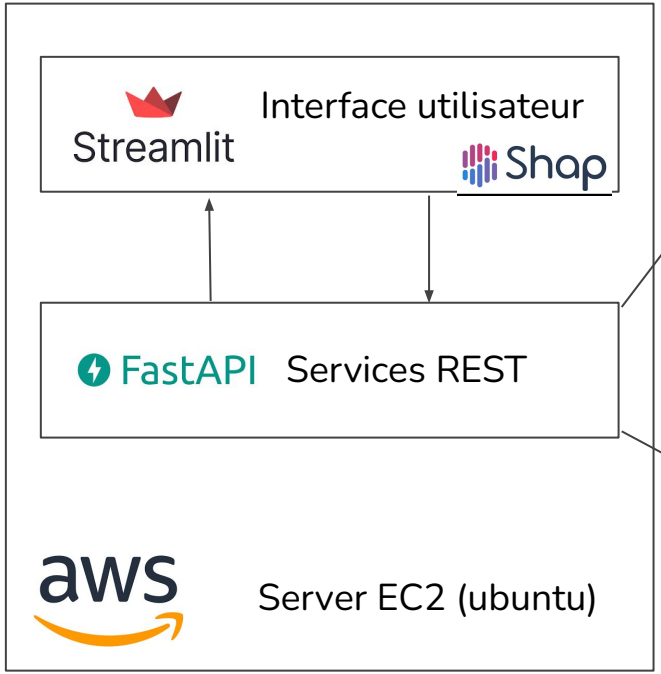
Applicabilité locale permise:

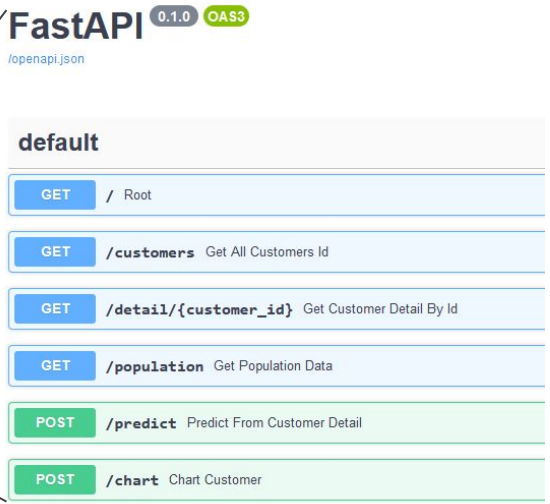
Facteurs favorables : âge, possède un véhicule personnel

Facteurs défavorable : Condition précaire, faible apport extérieur, montant du crédit

3. Présentation du Dashboard

Structure applicative





Ressources

Dépot github : https://github.com/analyst236/ocr_p7

API : http://ec2-35-181-58-38.eu-west-3.compute.amazonaws.com:8000/docs

Dashboard : http://ec2-35-181-58-38.eu-west-3.compute.amazonaws.com:8501/

Pour aller plus loin

Un score métier plus adapté:

- Définir le coefficient bêta en accords avec les objectifs fixé par la société

Un modèle plus performant:

- Amélioration du feature engineering
- Model stacking / deep-learning

Amélioration du dashboard:

- Afficher la valeur réelle des variables (shap)
- re-condenser les variables transformé par one hot encoding
- interactivité des graphes
- Possibilité de modification et saisie manuelle

Merci de votre attention