Programme d'Ingénierie en Intelligence Artificielle – OC IA P4

- Projet 4 – Construire un modèle de scoring

Les défauts de paiement des prêts peuvent entraîner des pertes considérables pour les banques.

C'est pourquoi elles consacrent une attention particulière à ce problème et mettent en œuvre diverses méthodes pour détecter et prédire les comportements de défaut de leurs clients.

Dans cette présentation, je vais aborder le processus fondamental de la prédiction des défauts de paiement des prêts en utilisant des algorithmes d'apprentissage automatique.

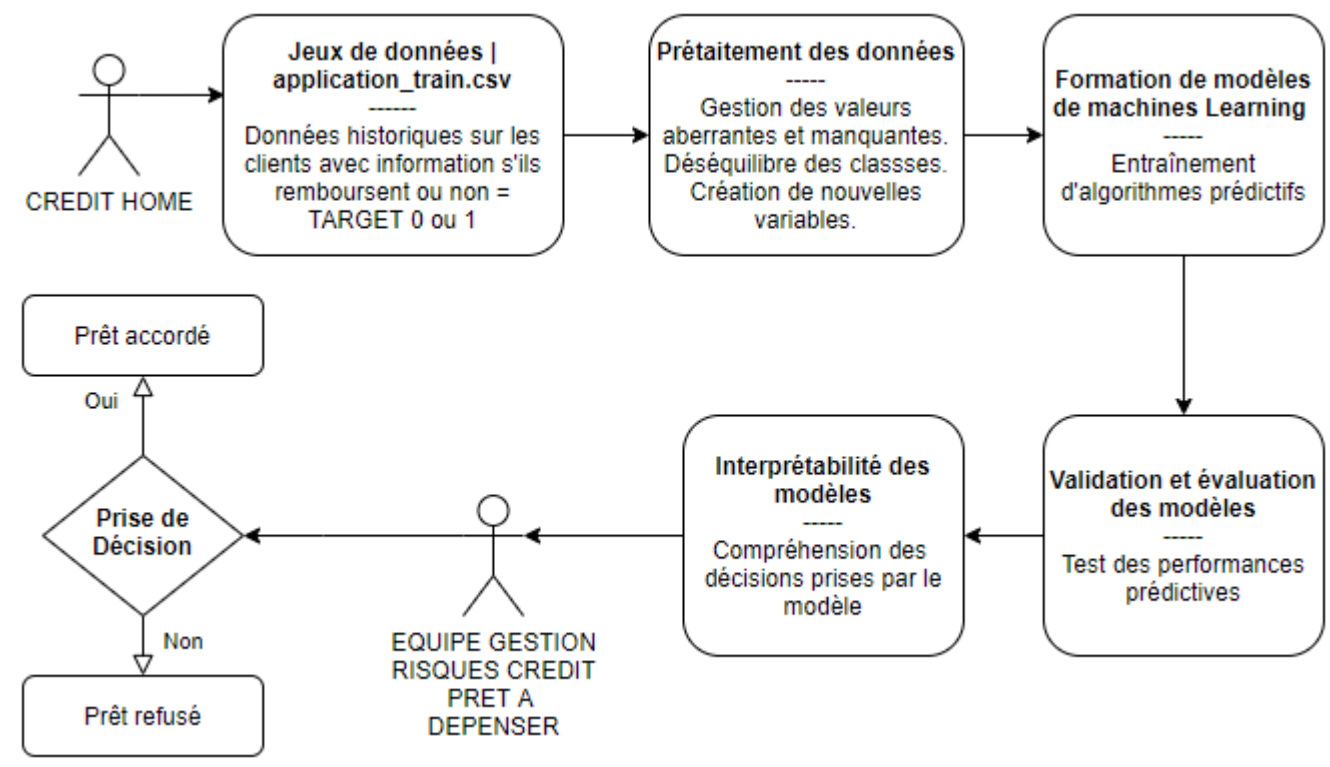
Sommaire

Préambule

- Cette présentation s'adresse à la fois aux professionnels du métier du Credit Scoring et aux techniciens de la Data Science.
- Un compromis a été fait pour inclure des explications claires et accessibles tout en fournissant suffisamment de détails techniques.
- Aucun code n'est fourni dans cette présentation, mais les techniciens souhaitant plus d'informations peuvent consulter les 2 notebooks pour des détails approfondis avec la synthèse des résultats.
- Un notebook est comme un cahier numérique où l'on peut écrire du texte et du code informatique au même endroit. Cela permet de voir les explications et les résultats des calculs en même temps.
- Pour les non datascientists, les questions lors de la présentation sont les bienvenues pour expliquer les concepts techniques.

1.Processus de mise en place de l'Outil de Scoring de Crédit.

Le processus de scoring de crédit comprend plusieurs étapes clés. Ce schéma illustre les principales phases de mise en œuvre de l'outil, garantissant ainsi la robustesse et la précision des prédictions.



2.Compréhension de la problématique métier

- La société financière **Prêt à dépenser** propose des **prêts à la consommation** aux particuliers ayant peu ou pas d'antécédents de crédit.

- Pour octroyer un crédit à la consommation de manière responsable et éviter les risques de non-remboursement, l'entreprise souhaite mettre en place un outil de **scoring de crédit**. Cet outil est essentiellement un **modèle mathématique** qui évalue la probabilité qu'un client rembourse ou non le prêt. Basé sur une série de critères—tels que les **revenus**, l'**emploi**, les **dépenses passées**, et d'autres **facteurs financiers**—ce modèle permet de prendre une décision éclairée sur l'octroi ou le refus du crédit.

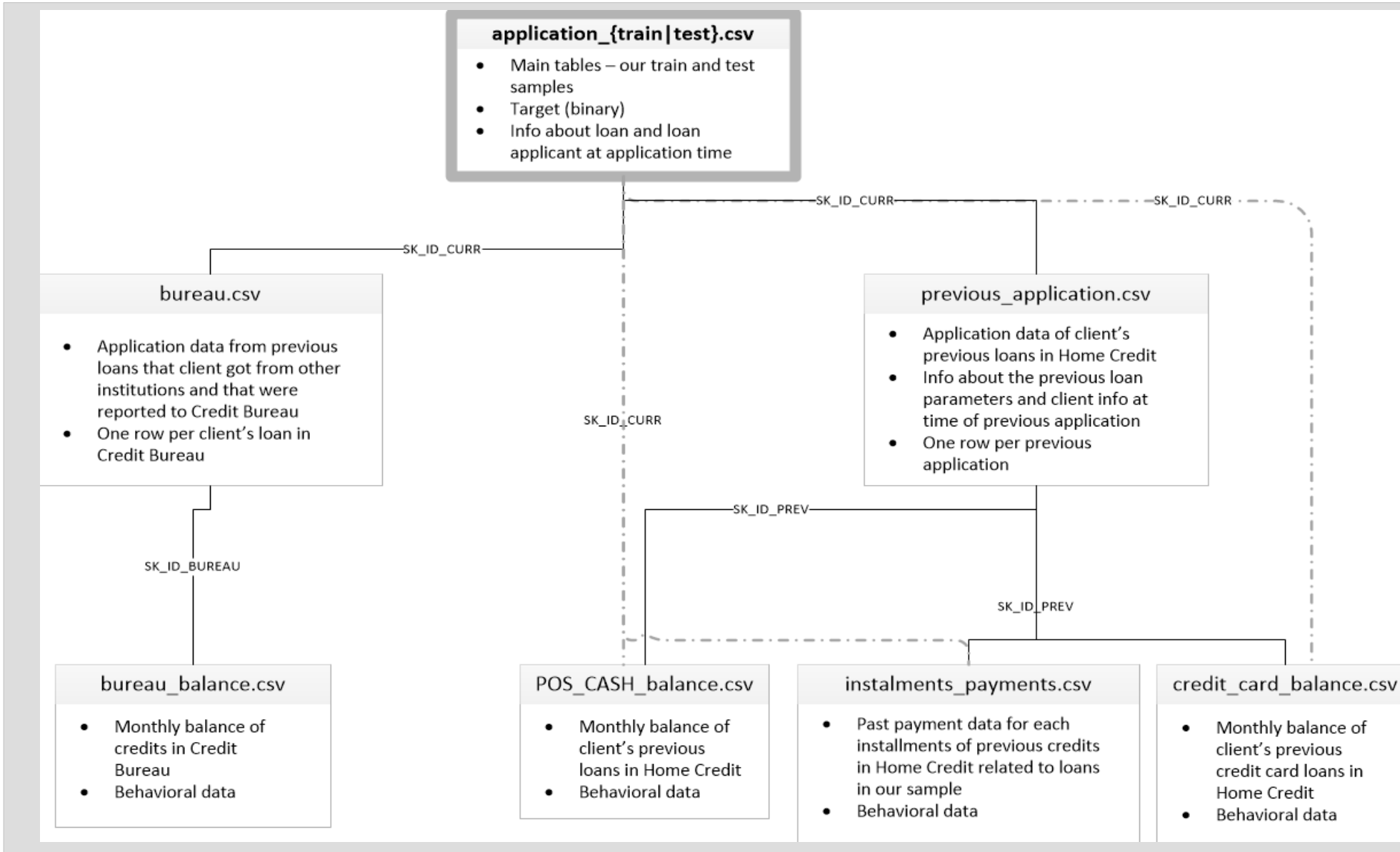
- Ce qui est attendu :
- - **Former un modèle de classification** pour aider à décider si un prêt peut être accordé à un client.
- - **Analyser les facteurs** qui justifient la décision du modèle.
- Dans le contexte du scoring, il est crucial de minimiser les erreurs. Par exemple, une **fausse alerte** (faux positif, **FP**) où l'on pense qu'un client ne remboursera pas alors qu'il le fera, est 10 fois moins coûteuse q**u'une erreur de non-détection** (faux négatif, **FN**), où l'on pense qu'un client remboursera alors qu'il ne le fera pas. Cela aide à mieux gérer les **risques financiers.**

- **Cet outil de scoring contribuera à réduire les risques financiers pour l'entreprise et offrira une opportunité de crédit équitable aux clients, renforçant ainsi la confiance et la stabilité financière de l'entreprise.**

3.Description du jeu de données.

- Nous disposons d'un ensemble de données contenant :
- Un **historique des prêts**
- Un **historique des informations financières**
- Des **informations sur le comportement des emprunteurs** (s'ils ont fait défaut ou non)
- Il y a plusieurs fichiers CSV
- Pour le fichier **application_train qui est le plus important** :
- - Ce fichier contient les données de **307 511 prêts**.
- - Chaque prêt est décrit par **122 caractéristiques (ou features)**, telles que les **revenus, l'âge, la situation familiale,** etc. **Ces caractéristiques sont utilisées pour évaluer le profil de l'emprunteur.**
- La colonne **TARGET** indique si le client a remboursé le prêt **(TARGET=0)** ou s'il ne l'a pas remboursé **(TARGET=1)**. Cette information est cruciale pour entraîner notre modèle de scoring, car elle permet de différencier les bons payeurs des mauvais payeurs.
- En résumé, le fichier **application_train** est utilisé pour **entraîner** notre **modèle de scoring** en lui fournissant des exemples de prêts **remboursés et non remboursés**. Les caractéristiques permettent de définir le profil de chaque emprunteur, et la colonne TARGET nous indique l'issue du prêt, ce qui est essentiel pour que le modèle apprenne à faire des prédictions précises. **C'est un problème de classification supervisée.**

3.Description du jeu de données.

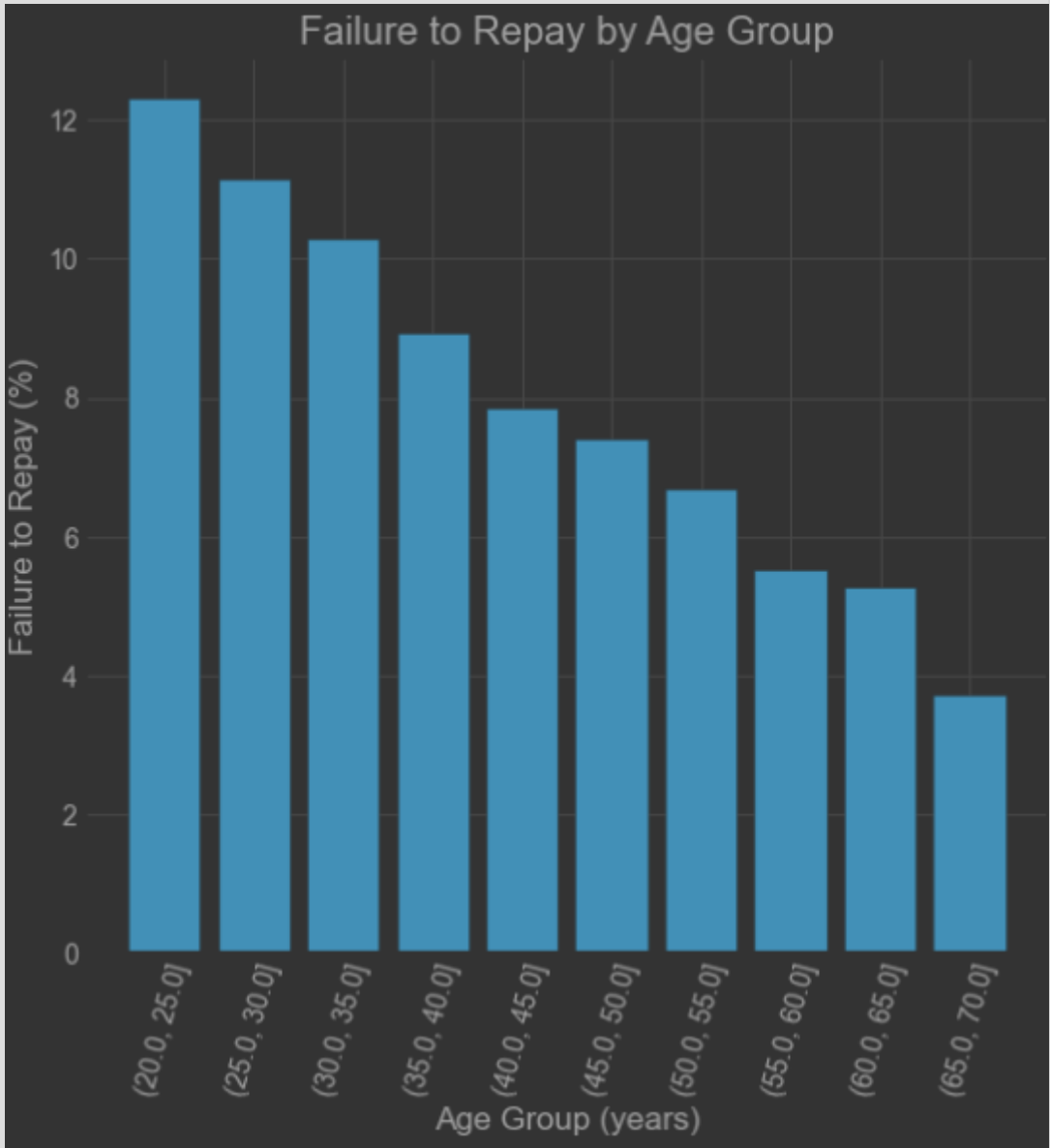


4.Transformation du jeu de données (Analyse exploratoire des données).

Tendance claire : Les **jeunes emprunteurs** ont plus de **probabilités de ne pas rembourser le prêt**, avec un taux de non-remboursement supérieur à **10 % pour les trois groupes d'âge** les plus jeunes et inférieur à **5** % pour le **groupe d'âge le plus âgé.**

Notre outil de scoring permet à la banque de **mieux comprendre les risques** liés à l'âge.

En identifiant **les jeunes clients** à risque, la banque peut **adapter ses stratégies** de gestion des risques en offrant des **programmes de soutien** ou des **conseils financiers** spécifiques, réduisant ainsi les risques et améliorant la **satisfaction des clients.**



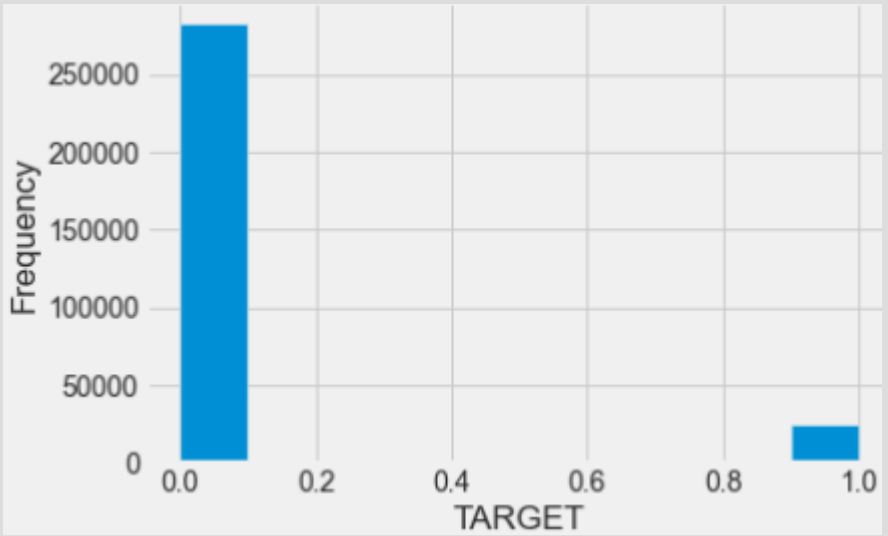
4.Transformation du jeu de données (Déséquilibre des classes).

Déséquilibre des Classes

Le graphique montre un déséquilibre important entre les classes dans notre variable TARGET :

- **Classe 0 (clients qui remboursent leur prêt) : Environ 90% des données.**
- **Classe 1 (clients qui ne remboursent pas leur prêt) : Environ 10% des données**.

Ce déséquilibre peut biaiser les modèles de machine learning en faveur des clients qui remboursent.



Solution Proposée

Pour résoudre ce problème, **nous équilibrons les classes pour que les deux groupes soient mieux représentés lors de l'entraînement du modèle.**

Valeur Ajoutée de Notre Application

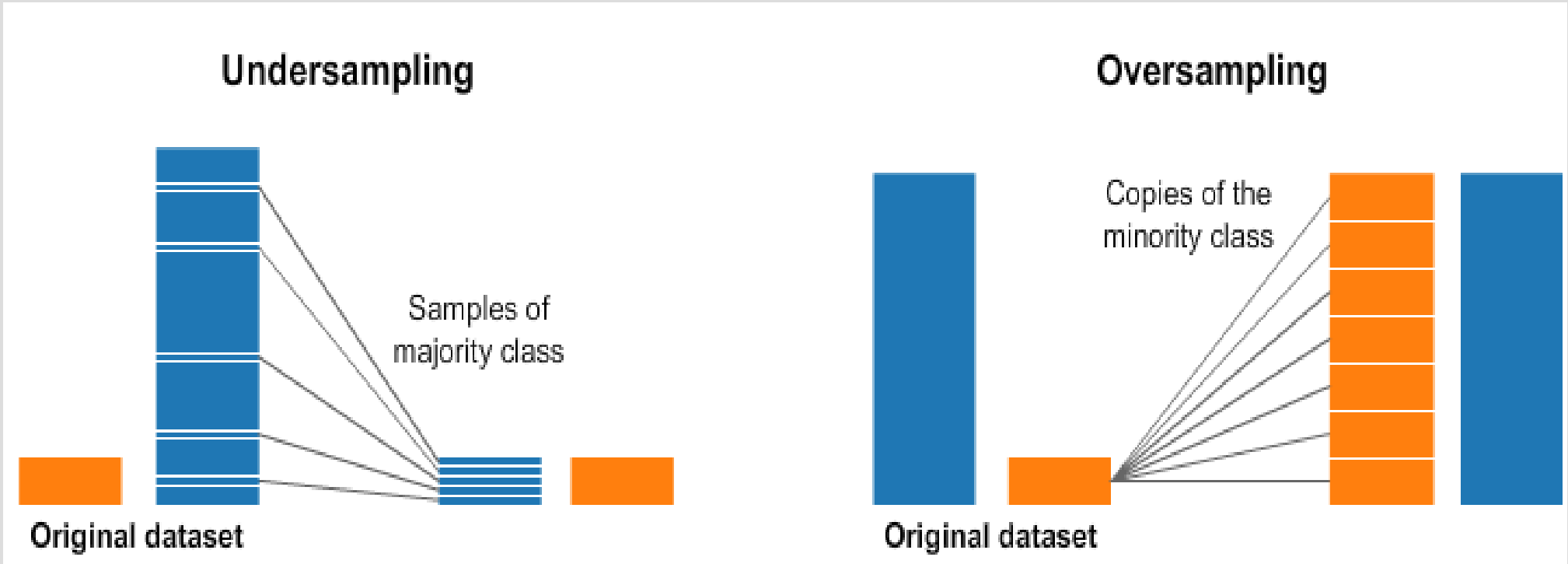
En rééquilibrant les classes, notre modèle de scoring de crédit pourra :

- **Mieux détecter les clients à risque** (classe 1), aidant à réduire les risques de défaut de paiement.
- **Fournir des évaluations de prêt plus justes** en tenant compte équitablement de tous les clients.

4.Transformation du jeu de données (Réquilibrage des classes TARGET).

Comme nous l'avons vu, nos **classes de TARGET sont largement déséquilibrées**, ce qui peut introduire un fort biais dans l'apprentissage de nos modèles. Pour remédier à ce problème, nous allons rééquilibrer ces classes dans nos jeux d'apprentissage.

Il existe **deux approches principales pour rééquilibrer les classes :**



4.Transformation du jeu de données (Valeurs Manquantes).

Valeurs Manquantes

- Notre jeu de données contient 122 colonnes :
- **67 colonnes ont des valeurs manquantes.**
- Nous conserverons toutes les colonnes.

Cependant, nous devrons **imputer les valeurs manquantes avant d'entraîner nos modèles supervisés.**
Imputer les valeurs manquantes signifie les remplacer par une valeur calculée. Par exemple, **nous pouvons utiliser la médiane** pour remplacer les valeurs manquantes car elle est moins influencée par les valeurs extrêmes que la moyenne.

4.Transformation du jeu de données (Imputation des données manquantes).

- Exemple simple pour illustrer l'imputation

Avant Imputation

Client ID	Revenu Annuel	Âge
1	50,000	35
2	60,000	NaN
3	NaN	45

Après Imputation

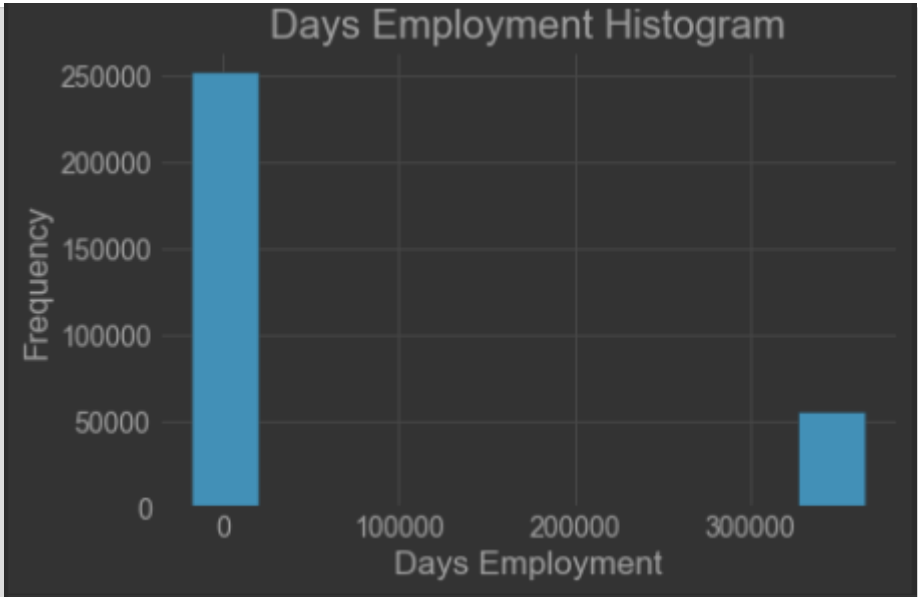
Client ID	Revenu Annuel	Âge
1	50,000	35
2	60,000	40
3	55,000	45

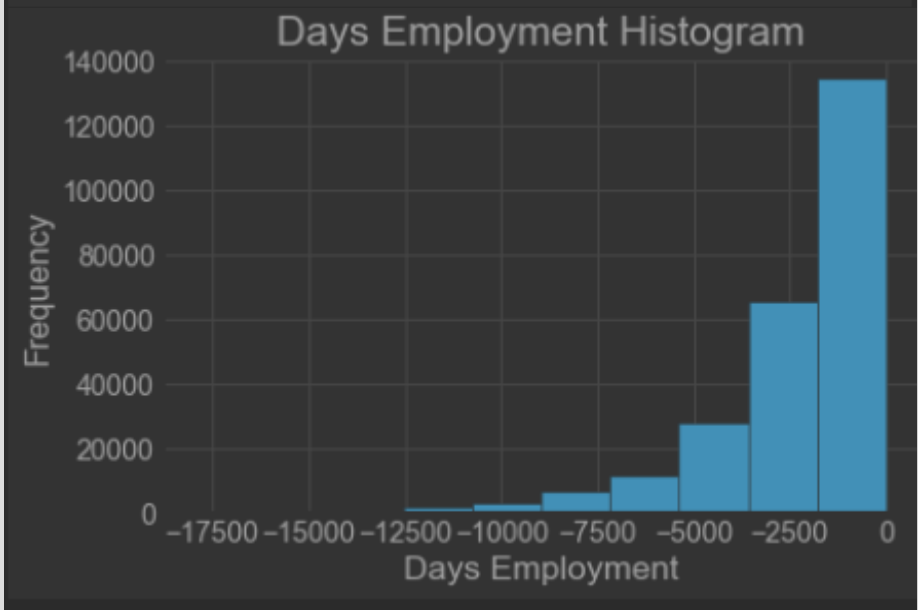
Explications

- Pour le revenu annuel, la médiane des valeurs non manquantes (50,000, 60,000, 55,000) est utilisée pour remplacer les valeurs manquantes (55,000).
- Pour l'âge, la médiane des valeurs non manquantes (35, 45, 40) est 40.

4.Transformation du jeu de données (Valeurs aberrantes -).

- **Cela ne semble pas correct !** La valeur maximale représente environ **700 ans pour les jours d'emplois cumulés**, ce qui est **irréaliste.**
- Cela indique des **erreurs** dans les données, probablement dues à des valeurs **mal enregistrées** ou **incorrectes.**
- Dans un projet de **scoring de crédit**, il est crucial de **nettoyer** ces anomalies pour assurer la **fiabilité** des modèles.
- En corrigeant les **valeurs incorrectes**, Cela permet à notre modèle de faire des **prédictions plus précises** en se basant sur des informations **réalistes**. Enfin, cela contribue à une **meilleure interprétabilité** des résultats, renforçant ainsi la **confiance** dans les décisions prises par l'outil de scoring.
- Après correction, nous obtenons une **distribution plus normale** des données, ce qui est essentiel pour l'**efficacité** des algorithmes de **machine learning**.





4.Transformation du jeu de données (Création de nouvelles colonnes – Feature engineering).

- Le **feature engineering** consiste à **transformer et créer des variables** (features) à partir des **données brutes** pour améliorer les **performances** des **modèles de machine learning**. Par exemple, à partir des informations de **revenus** et d'**emprunts**, nous pouvons créer des **variables supplémentaires** comme le **ratio d'endettement**. Cela permet au modèle de mieux comprendre **les relations cachées** dans les données et de faire des **prédictions plus précises**. Cette étape est **cruciale** pour obtenir des résultats **fiables** et **pertinents**.

- **Nouvelles Caractéristiques Créées sur la Connaissance Métier :**

- **Pour rendre notre ensemble de données plus compréhensible, nous avons créer de nouvelles caractéristiques à partir de celles existantes :**

- **Pourcentage du Crédit sur le Revenu :** Cette caractéristique représente la **proportion du montant du crédit par rapport au revenu total**. Elle aide à comprendre combien du revenu d'une personne est pris par son crédit.

- **Pourcentage de l'Annuité sur le Revenu :** Cette caractéristique indique la **proportion du montant de l'annuité (paiement régulier) par rapport au revenu total.** Elle montre combien du revenu d'une personne est consacré aux paiements réguliers.

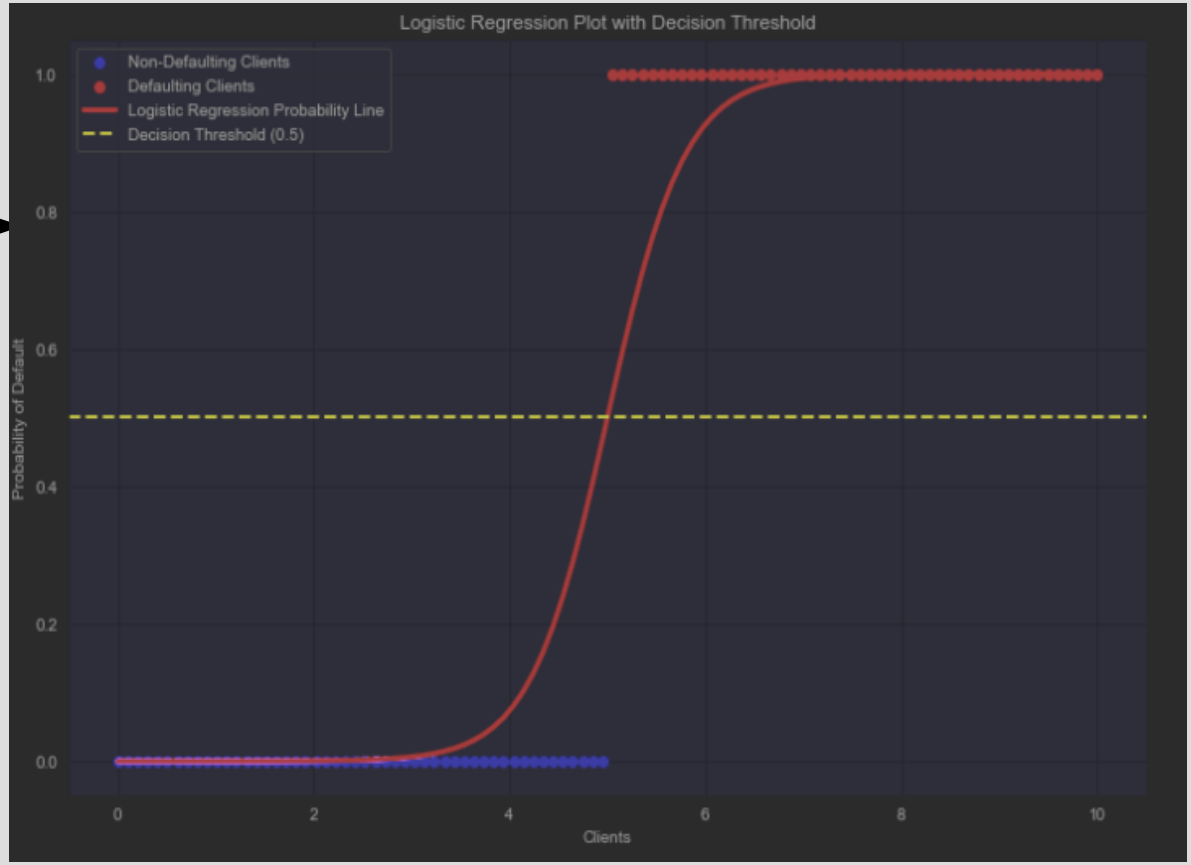
- **Durée du Crédit :** Cette caractéristique est le **ratio du montant de l'annuité sur le montant du crédit**, montrant essentiellement la durée du crédit. Elle aide à comprendre la durée sur laquelle le crédit doit être remboursé.

- **Pourcentage de Jours Employés :** Cette caractéristique montre la **proportion du nombre de jours employés par rapport au nombre de jours depuis la naissance**. Elle donne une idée de la stabilité de l'emploi d'une personne au cours de sa vie.

5.Les modèles utilisés

Logistic Regression

Modèle simple et interprétable



Dans ce graphique :

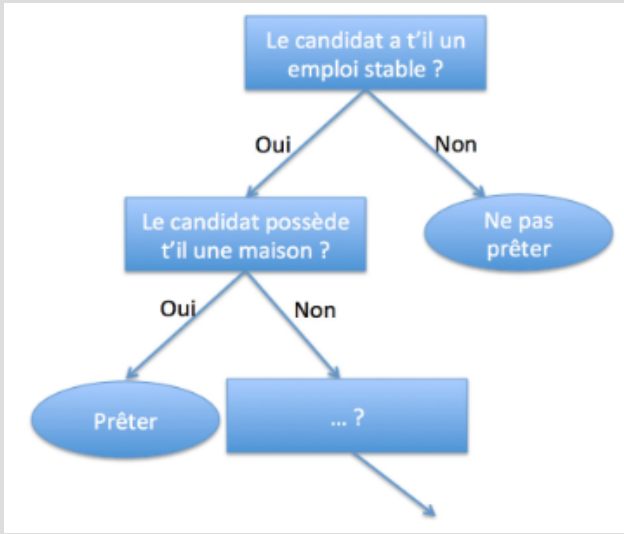
- L'axe horizontal représente les **clients ou une caractéristique des clients.**
- L'axe vertical représente la **probabilité de défaut**.

Les **points rouges et bleus indiquent les clients défaillants et non-défaillants**. La courbe rouge prédit la probabilité de défaut en fonction des Clients. La ligne jaune est le **seuil de décision** : **au-delà de ce seuil, un client est classé comme défaillant.**

5. Les modèles utilisés

Random Forest

Ensemble de nombreux arbres de décision qui améliore la précision et réduit le surapprentissage.

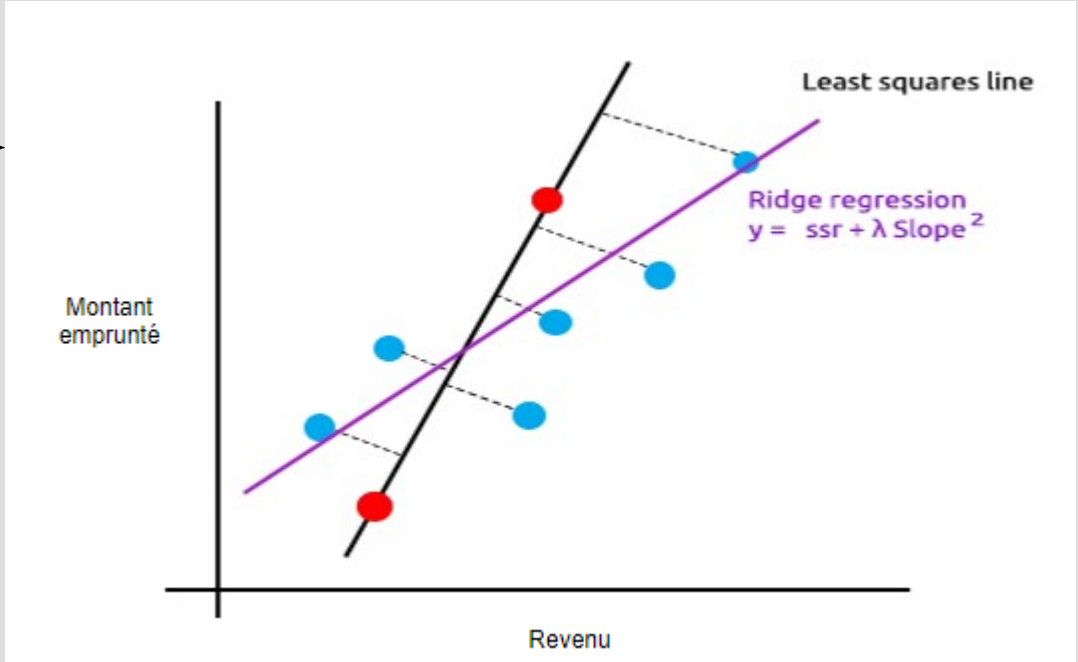


Une **forêt aléatoire** est composée de plusieurs **arbres de décision**, comme celui illustré. Par exemple, **on évalue si un candidat** a un **emploi stable**. **S'il n'en a pas**, la décision est de ne **pas prêter**. S'il en a un, on vérifie s'il possède **une maison**. S'il en a une, **on prête**. Sinon, d'autres critères peuvent être considérés. Chaque arbre dans la forêt aléatoire **donne une prédiction**, et la décision finale est basée sur la **majorité des prédictions**

5.Les modèles utilisés

Ridge Regression

La régression Ridge est une technique avancée qui améliore la régression linéaire en ajoutant une pénalité aux coefficients pour éviter qu'ils deviennent trop grands.

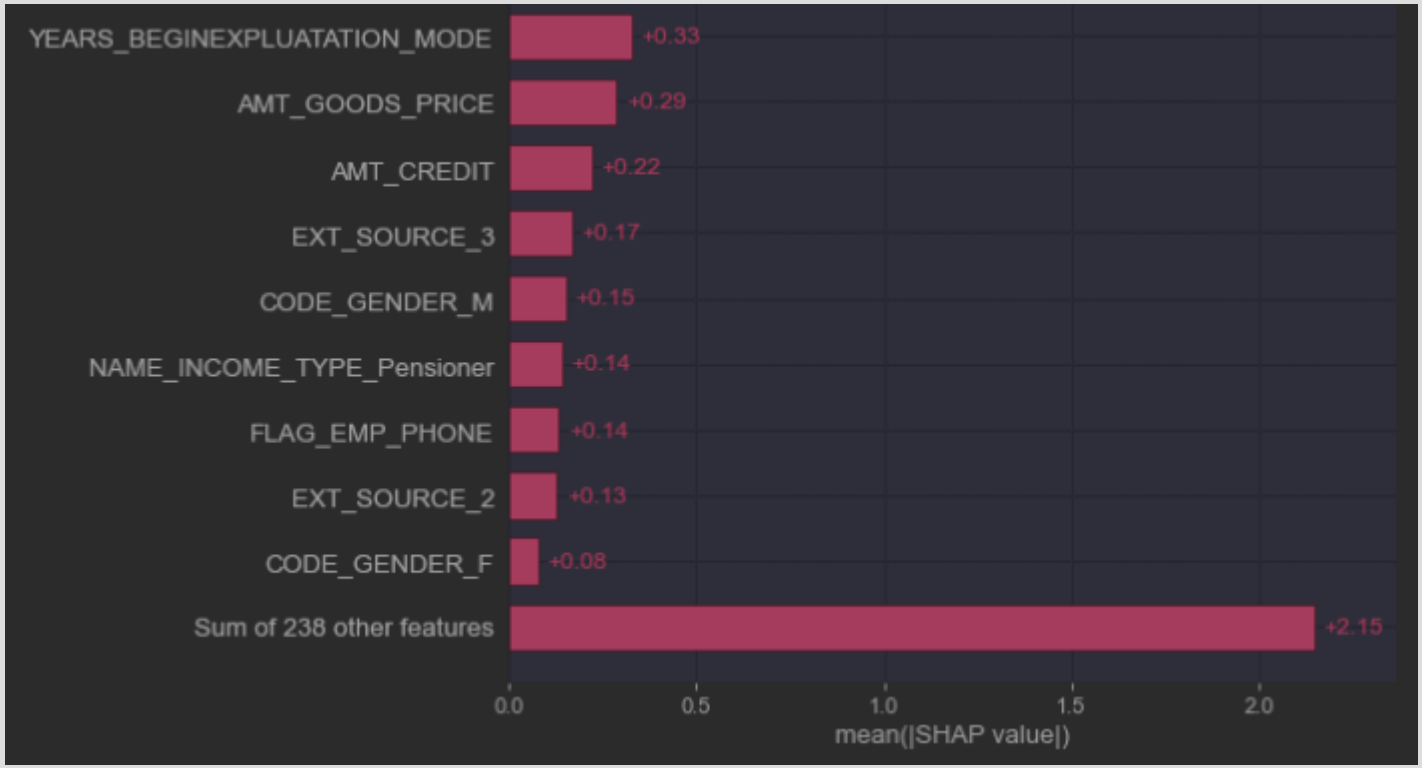


Imaginez que **vous évaluez la capacité d'un client** à rembourser un prêt en fonction de son **revenu et du montant emprunté**. Les **coefficients** sont les nombres qui multiplient ces variables (revenu, montant emprunté) pour faire la prédiction. Une régression classique peut donner des coefficients très élevés, ce qui peut conduire à des erreurs sur de nouvelles données. **La régression Ridge réduit ces coefficients** pour créer un modèle **plus équilibré et fiable pour de futures prédictions**.

6.Interprétabilité du modèle

Importance Globale des Variables (Global Feature Importance) :

- **Objectif** : Expliquer l'**impact général de chaque variable sur les prédictions du modèle.**
- **Méthode :** Utilisation de l'analyse SHAP pour visualiser l'importance des variables.
- **Exemple :** Les variables comme **AMT_GOODS_PRICE**, **AMT_CREDIT, et EXT_SOURCE_3** sont cruciales.
-

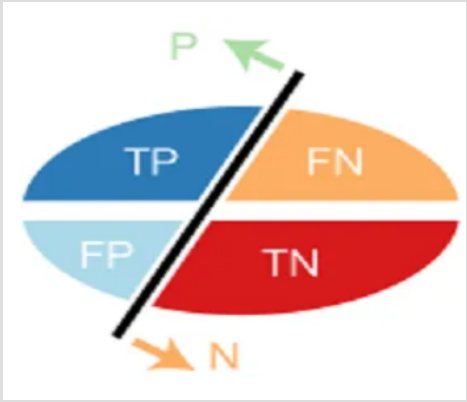


6.Interprétabilité du modèle

Importance Locale des Variables (Local Feature Importance) :

- **Objectif :** Analyser les **prédictions individuelles pour comprendre les décisions spécifiques.**
- **Méthode :** Utilisation de l'analyse SHAP pour des instances spécifiques.
- **Exemple : Pour un client donné,** des variables telles que **EXT_SOURCE_3** et **EXT_SOURCE_2** peuvent expliquer une prédiction de risque élevé, tandis que d'autres variables comme **CODE_GENDER_F** réduisent cette probabilité.

7.Conclusion



- En réalité, la plupart des **modèles de classification** binaire prédisent d'abord une **probabilité** avant de l'assigner à 1 ou 0 **en se basant sur un seuil par défaut de 0,5**.

- Pour améliorer **le rappel du modèle**, nous pouvons utiliser les probabilités prédites par le modèle et définir nous-mêmes le **seuil**. Le seuil est défini en fonction de plusieurs facteurs tels que les **objectifs commerciaux**. Dans la **prédiction du comportement de remboursement des prêts**, par exemple, la banque veut **contrôler les pertes à un niveau acceptable**. Elle peut donc utiliser un seuil relativement bas. Cela signifie que davantage de clients seront classés comme **"clients à risque potentiel"** et leurs profils seront examinés plus attentivement par l'**équipe de gestion des risques de crédit**. De cette manière, la banque peut **détecter les comportements de défaut de paiement à un stade précoce** et prendre les mesures correspondantes pour **réduire les pertes possibles.**

- Dans ce contexte, il est important de comprendre les concepts suivants, illustrés par le schéma :

- **TP (Vrai Positif)** : Les clients à risque de défaut correctement identifiés.
- **FP (Faux Positif)** : Les clients identifiés à tort comme à risque de défaut.
- **TN (Vrai Négatif)** : Les clients non à risque correctement identifiés.
- **FN (Faux Négatif)** : Les clients à risque de défaut non identifiés.

- **En ajustant le seuil de classification, nous pouvons influencer le taux de TP, FP, TN et FN, et ainsi mieux aligner le modèle avec les objectifs de gestion des risques de la banque.**

7.Conclusion

En analysant les performances des différents modèles, nous constatons que **Random Forest avec Class Weight** et **Ridge avec Class Weight** ont chacun leurs points forts :

- **Random Forest avec Class Weight** : **Meilleur pour détecter les défauts**
- **Ridge avec Class Weight** : **Performance plus équilibrée**

Recommandation : Selon la priorité de l'entreprise **(minimiser le risque financier ou avoir un modèle équilibré**), l'un ou l'autre modèle peut être sélectionné.

Il est important de noter qu'il s'agit d'un **MVP (Produit Minimum Viable)** et que nous allons **continuer les cycles d'analyses avec les équipes de gestion du crédit scoring pour identifier les caractéristiques actuellement utilisées pour les décisions de prêt (analyse de l'existant) et trouver les meilleurs modèles.**