Université Cergy Paris

Diplôme universitaire : Data Analyst

Examen :
Machine Learning supervisé

11 décembre 2022

Haury Fabien

Examen Machine Learning supervisé

Haury Fabien

11 décembre 2022

Table des matières

Chapitre 1

Arbre de décision : concept de base

1.1 Expliquer comme l'indice de Gini ou l'entropie est utilisé (sans entrer dans les formules mathématiques, juste le principe), lors de la création d'un arbre de décision

L'indice de diversité de Gini mesure avec quelle fréquence un élément aléatoire de l'ensemble serait mal classé si son étiquette était choisie aléatoirement selon la distribution des étiquettes dans le sous-ensemble. L'indice de diversité de Gini peut être calculé en sommant la probabilité pour chaque élément d'être choisi, multipliée par la probabilité qu'il soit mal classé. Il atteint sa valeur minimum (zéro) lorsque tous les éléments de l'ensemble sont dans une même classe de la variable-cible. Lors de la conception de l'arbre de décision, les caractéristiques possédant la plus petite valeur de l'indice de Gini seront privilégiées.

Le gain d'information est basé sur le concept d'entropie de Shannon en théorie de l'information. L'entropie permet de mesurer le désordre dans un ensemble de données et est utilisée pour choisir la valeur permettant de maximiser le gain d'information. Le gain d'information est appliqué pour quantifier quelle caractéristique fournit l'information maximale sur la classification basée sur la notion d'entropie, c'est-à-dire en quantifiant la taille de l'incertitude, du désordre ou de l'impureté, en général, avec l'intention de diminuer la quantité d'entropie en partant du haut (nœud racine) vers le bas (nœuds feuilles).

1.2 Expliquer le concept de bootstrap

Le rééchantillonnage ou Bootstrap a été inventé à l'origine comme une méthode d'approximation de la distribution d'échantillonnage des statistiques dont les propriétés théoriques sont intraitables. Son utilisation pour estimer la performance des modèles est une application secondaire de la méthode.

Un échantillon bootstrap de l'ensemble d'apprentissage est un échantillon de la même taille que l'ensemble d'apprentissage mais qui est tiré avec remplacement, comme illustré en Figure 1. Cela signifie que certains points de données de l'ensemble de formation sont sélectionnés plusieurs fois pour l'ensemble d'analyse. Chaque point de données a 63,2 % de chances de figurer au moins une fois dans l'ensemble d'apprentissage. L'ensemble d'évaluation contient tous les échantillons de l'ensemble d'entraînement qui n'ont pas été sélectionnés pour l'ensemble d'analyse (en moyenne, 36,8 % de l'ensemble d'entraînement). Lors du bootstrapping, l'ensemble d'évaluation est souvent appelé l'échantillon hors-sac (out-of-bag sample).

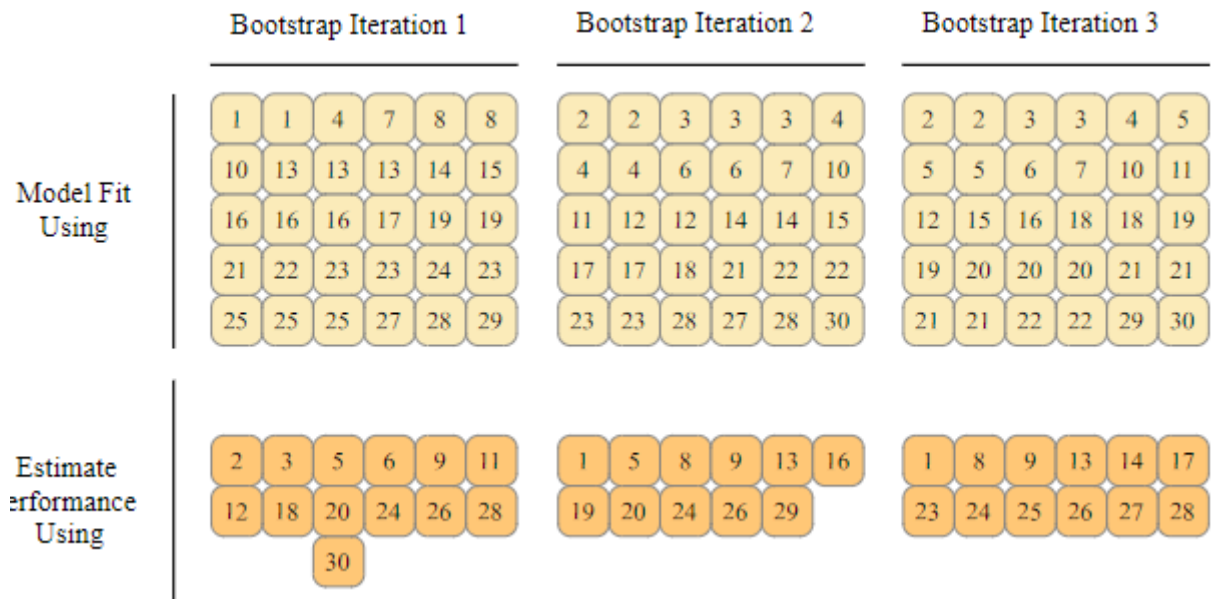

FIGURE 1 – Illustration du concept de bootstrap

1.3 Expliquer le concept de bagging

Le mot bagging est la contraction de « Bootstrap Aggregating ». Il permet de combiner les prédictions réalisées à partir de plusieurs modèles, en utilisant le même algorithme pour différents échantillons des données d'apprentissage. On utilise également le bagging pour apporter des solutions aux problèmes liés à l'instabilité des résultats quand des modèles complexes sont appliqués à des jeux de données de faible volume.

Il fonctionne en trois étapes comme le montre la Figure 2 :

1. Bootstrapping : Le Bagging utilise une technique d'échantillonnage bootstrapping pour créer des échantillons diversifiés. Cette méthode de rééchantillonnage génère différents sous-ensembles de l'ensemble de données d'entraînement en sélectionnant des points de données au hasard et avec remplacement. Cela signifie qu'à chaque fois que vous sélectionnez un point de données dans l'ensemble de données d'apprentissage, vous êtes en mesure de sélectionner la même instance plusieurs fois. Par conséquent, une valeur/instance se répète deux fois (ou plus) dans un échantillon.

2. Entraînement parallèle : Ces échantillons bootstrap sont ensuite entraînés indépendamment et en parallèle les uns des autres à l'aide d'apprenants faibles ou de base.

3. Agrégation : Enfin, en fonction de la tâche (c'est-à-dire régression ou classification), une moyenne ou une majorité des prédictions sont prises pour calculer une estimation plus précise. Dans le cas de la régression, on prend la moyenne de toutes les sorties prédites par les classificateurs individuels ; c'est ce qu'on appelle le vote doux. Pour les problèmes de classification, la classe ayant la plus grande majorité de votes est acceptée ; c'est ce qu'on appelle le vote dur ou le vote majoritaire.

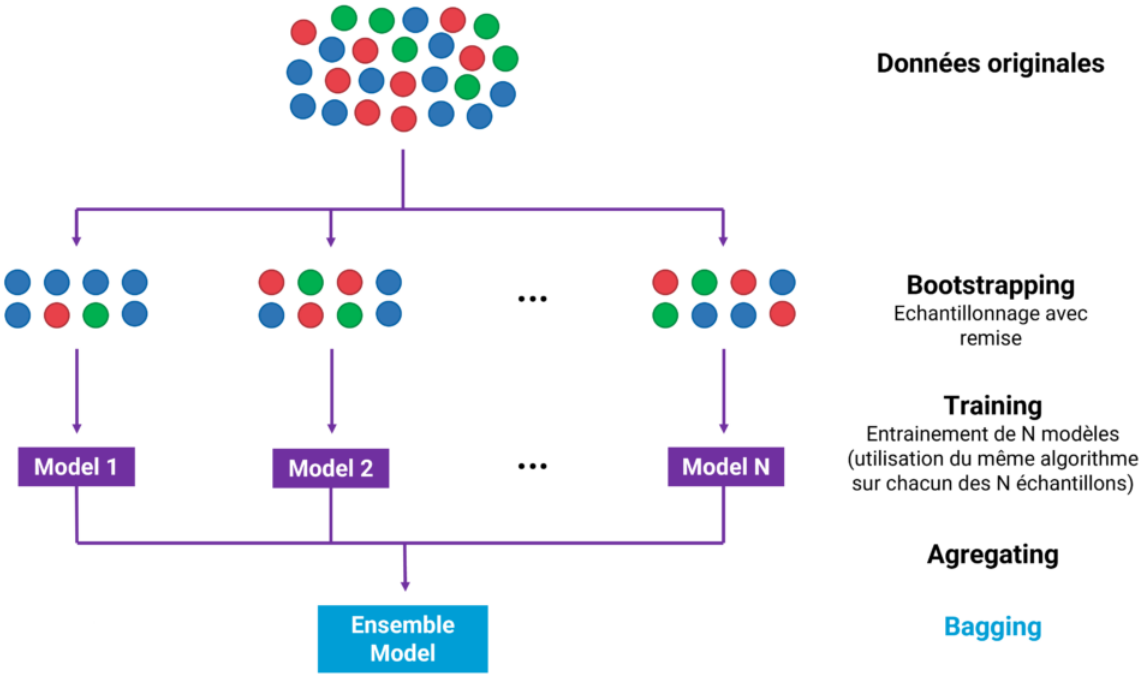

FIGURE 2 – Illustration du concept de bagging

1.4 Quelle est la différence entre une approche de type bagging et une forêt aléatoire ?

Le concept de l'échantillonnage bootstrap (bagging) consiste à former un groupe d'arbres de décision non élagués sur différents sous-ensembles aléatoires des données de formation, en échantillonnant avec remplacement, afin de réduire la variance des arbres de décision. L'idée est de combiner les prédictions de plusieurs apprenants de base pour créer un résultat plus précis. Avec les forêts aléatoires, une variation aléatoire supplémentaire est ajoutée à la procédure de mise en sac afin de créer une plus grande diversité parmi les modèles résultants. L'idée derrière les forêts aléatoires est de construire plusieurs arbres de décision et de les agréger pour obtenir un résultat précis.

Chapitre 2

Métriques de performance

2.1 Expliquer la différence entre les métriques RMSE et MAE. Dans quel contexte les emploie-t-on ? Quel est "l'avantage" du RMSE par rapport au MAE ?

L'écart quadratique moyen (REQM), root-mean-square error (RMSE) est l'écart-type des résidus (erreurs de prévision). Les résidus sont la mesure de l'écart entre les points de données et la ligne de régression. La métrique REQM est la mesure de la ventilation de ces résidus. Elle indique la concentration des données autour de la ligne du meilleur ajustement. La REQM est toujours positive et une valeur de 0 (presque jamais atteinte en pratique) indiquerait un ajustement parfait aux données. Une valeur de REQM plus petite indique une meilleur précision qu'une valeur de REQM plus élevée.

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

L'erreur absolue moyenne (EAM), mean absolute error (MAE) est la moyenne de la différence absolue entre la prévision du modèle et la valeur cible.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Ces deux métriques sont utilisés sur des problèmes de régression (régression linéaire, régression arbre de décision, etc.).
MAE : Si le nombre d'outliers dans le jeu de données n'est par un problème, c'est-à-dire que l'influence des outliers est ignorée, alors la MAE est adaptée.
RMSE : Si il y a un grand nombre d'outliers dans vos données et que vous souhaitez les prendre en compte lors de l'ajustement de votre modèle.

L'avantage de la métrique REQM par rapport à la MAE est que l'effet de chacune des erreurs sur la REQM est proportionnel à la taille de l'erreur quadratique ; ainsi, des erreurs plus importantes ont un effet disproportionné sur la REQM. Par conséquent, la REQM est sensible aux valeurs aberrantes ou anomalies

2.2 Expliquer les concepts de faux positif, de faux négatif, de sensibilité et de spécificité

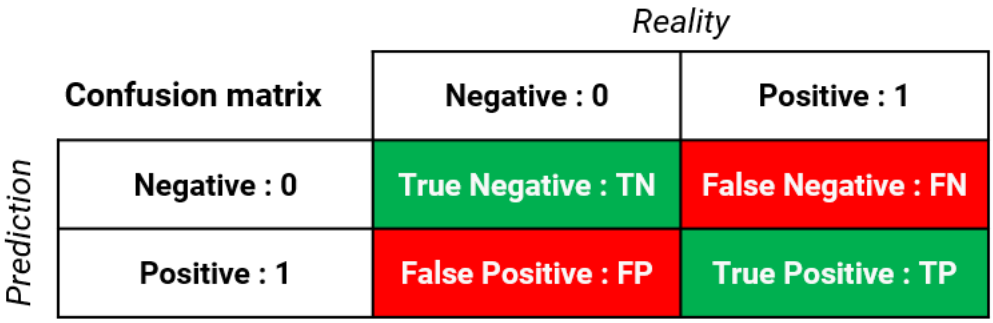

FIGURE 3 – Matrice de confusion

En prenant comme référence la Figure 3 :
les faux positifs ou false positive (FP) indiquent les valeurs classées positives par erreur. Par exemple, une personne testée positive est en réalité négative.
Les faux négatifs ou false negative (FN) indiquent les valeurs classées négatives par erreur. Par exemple, une personne aurait été testée négative, mais il est en réalité atteint par la maladie.
La sensibilité est le taux d'individus positifs correctement prédits par le modèle :

$$\frac{TP}{TP + FN}$$

Elle répond à la question : combien d'individus positifs sont prédits correctement ? Elle mesure donc la capacité du modèle à détecter l'ensemble des individus positifs. On la trouve aussi sous le nom de rappel (recall), taux de vrais positifs (True Positive Rate, TVP) ou encore taux de détection (hit rate).
La spécificité est le taux d'individus négatifs correctement prédits par le modèle :

$$\frac{TN}{TN + FP}$$

Elle répond à la question : combien d'individus négatifs sont prédits correctement ? Elle mesure donc la capacité du modèle à détecter l'ensemble des individus négatifs. On la trouve aussi sous le nom de sélectivité (selectivity) ou taux de vrais négatifs (True Negative Rate, TNR).

2.3 Comment une courbe de type ROC est-elle construite ?

La "courbe ROC" vient de l'anglais ROC pour Receiver Operating Characteristic, ou fonction d'efficacité du récepteur en français. Elle trace l'ensemble des valeurs du couple (1-Spécificité, Sensibilité) selon différents seuils de classification.

Cette courbe représente deux paramètres :

- Taux de vrais positifs ou sensibilité.
- Taux de faux positifs ou 1 - spécificité.

Le taux de vrais positifs (TVP), true positive rate (TPR) est l'équivalent du rappel. Il est donc défini comme suit :

$$TVP = \frac{TP}{TP + FN}$$

Le taux de faux positifs (TFP), false positive rate (FPR) est défini comme suit :

$$1 - \text{Specificité} = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP} = TFP$$

Une courbe ROC se construit en trois étapes :
étape n°1 : obtenir les prédictions du modèle de classification.

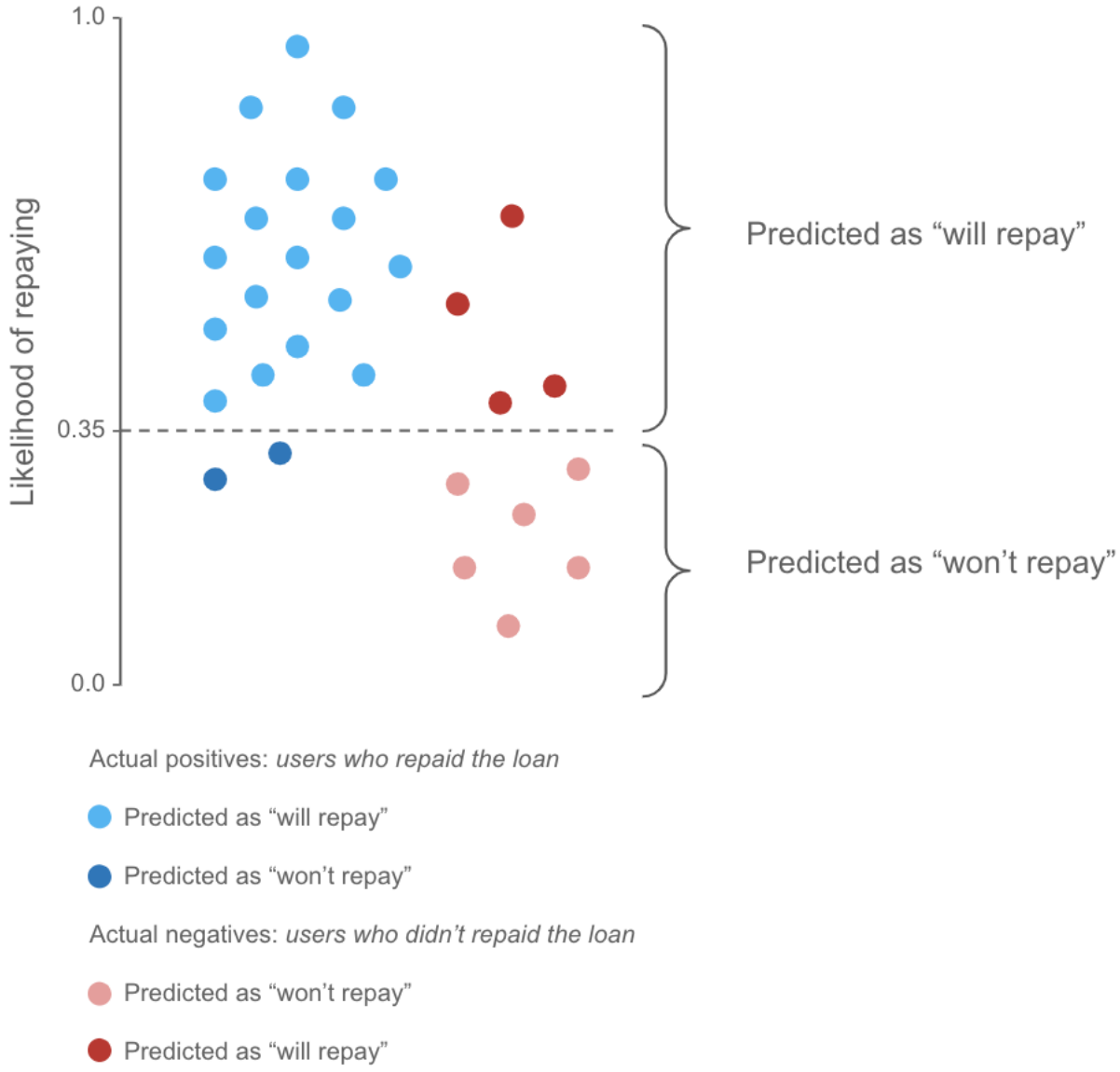

FIGURE 4 – Exemple d’un modèle de classification

Dans la Figure 4, un seuil à 0,35 est sélectionné. Toutes les prédictions égales ou supérieures à ce seuil sont classées comme "will repay". Toutes les prédictions inférieures à ce seuil sont classées comme "will not repay". À partir de ces données, nous pouvons construire une matrice de confusion telle que celle en Figure 5

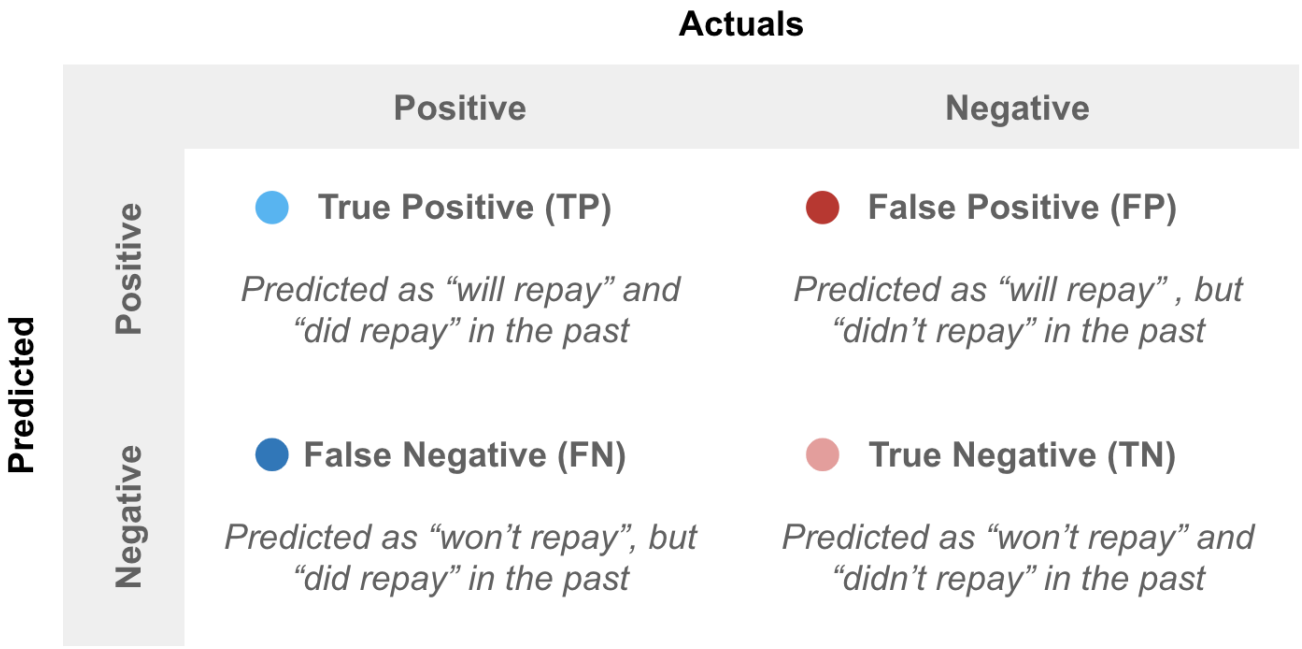

FIGURE 5 – Matrice de confusion

étape n°2 : calculez le taux de vrais positifs et le taux de faux positifs.
Maintenant que nous avons classé toutes les prédictions et nous savons si les classifications sont correctes ou non, nous pouvons calculer leurs TVP et TFP respectifs.

Les TVP et TFP pour différents seuils sont renseignés en Figure 6

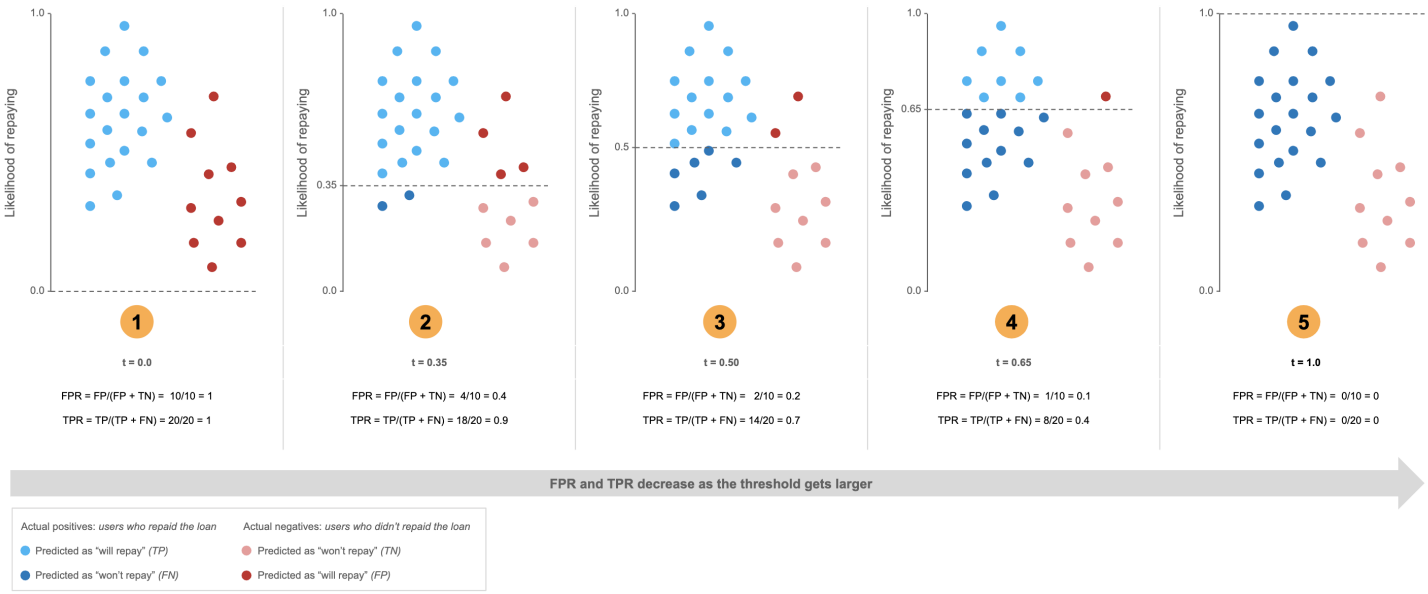

FIGURE 6 – Résultats du TVP et du TPF pour différents seuils

étape n°3 : tracez le TVP et le TPF pour chaque seuil.
Pour chaque seuil, nous traçons la valeur du TFP en abscisse et la valeur du TVP en ordonnée. Nous joignons ensuite les points avec une ligne.

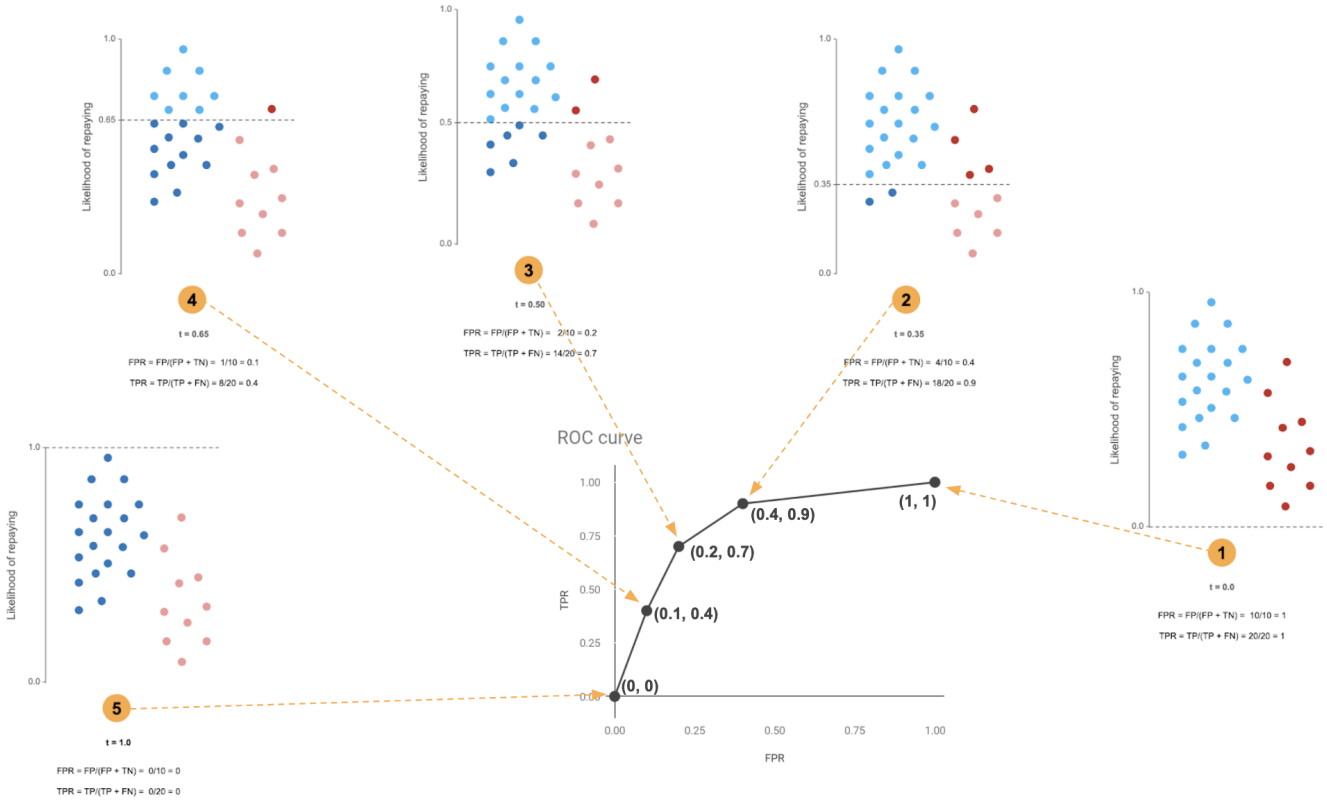

FIGURE 7 – Courbe ROC

2.4 Que signifie l'AUC ? Qu'est-ce qu'un bon AUC ? L'AUC d'un modèle peu performant ?

AUC signifie "Area under the ROC Curve"" (aire sous la courbe ROC). C'est-à-dire que l'AUC mesure toute l'aire à deux dimensions sous l'intégralité de la courbe ROC de (0,0) à (1,1). Le score de l'AUC est compris entre 0 et 1. Un modèle dont les prédictions sont correctes a un AUC qui tende vers de 1. Un modèle dont les prédictions sont totalement fausses a un AUC qui tende vers 0.

Lorsque l'AUC est de 1, le modèle est parfaitement capable de distinguer la classe positive de la classe négative. Lorsque l'AUC est d'environ 0,5, le modèle n'a aucune capacité de discrimination pour distinguer la classe positive de la classe négative. Lorsque l'AUC est de 0, le modèle est en fait en train de réciproquer les classes. Cela signifie que le modèle prédit une classe négative comme une classe positive et vice versa.

2.5 Expliquer le concept de recall, ou rappel

En termes de classification, le recall ou rappel correspond au pourcentage d'exemples positifs qu'un modèle a automatiquement classé parmi tous les exemples positifs. Le calcul du rappel se fait par la division de la valeur TP par la valeur FN combiné au total d'exemples positifs. Il peut aussi être appelé taux de réussite.

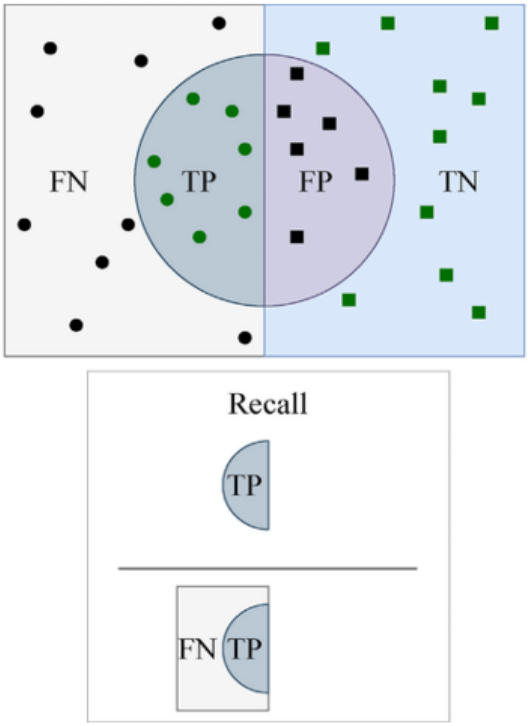

FIGURE 8 – Illustration du concept de recall

2.6 Expliquer le concept de précision

Comme le rappel, la précision correspond à un pourcentage des exemples positifs. Toutefois, ce paramètre s'intéresse aux données étiquetées positives par le modèle. Autrement dit, ici, le nombre total d'exemples positifs est divisé par la somme des TP et des FP.

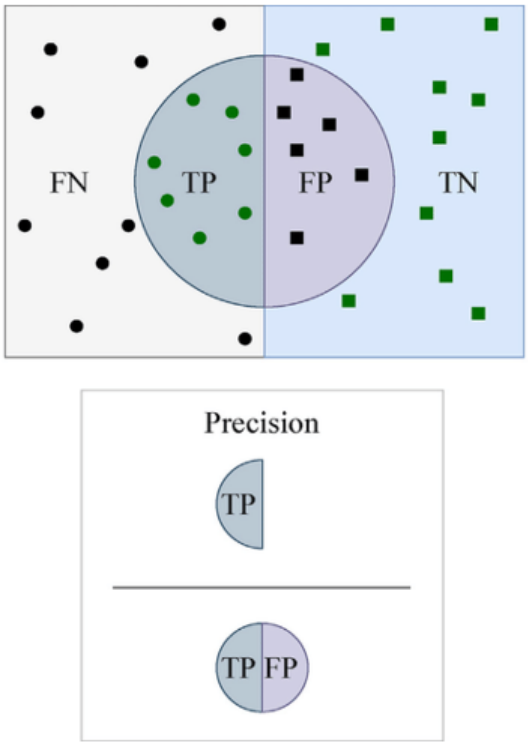

FIGURE 9 – Illustration du concept de précision

Chapitre 3

Tuning d'un modèle

3.1 A quoi sert une grid search ?

Afin d'utiliser un modèle pour de la prédiction, les paramètres de ce modèle doivent être estimés. Certains de ces paramètres peuvent être estimés directement à partir des données d'apprentissage, mais d'autres paramètres, appelés paramètres de réglage ou hyperparamètres, doivent être spécifiés à l'avance et ne peuvent pas être trouvés directement à partir des données d'apprentissage. Il s'agit de valeurs structurelles inconnues ou d'autres types de valeurs qui ont un impact significatif sur le modèle mais qui ne peuvent pas être estimées directement à partir des données. L'optimisation des paramètres de réglage relève généralement de l'une des deux catégories suivantes : recherche sur grille (grid search) et recherche itérative (iterative search).

La recherche par grille est une méthode d'optimisation des hyperparamètres (hyperparameter optimization) qui permet de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage. Pour chaque paramètre, on détermine un ensemble de valeurs que l'on souhaite tester.

3.2 Citer plusieurs hyperparamètres que l'on peut faire varier avec un arbre de décision

Les hyperparamètres possible pour un arbre de décision sont les suivants :

- Profondeur maximale de l'arbre : cet hyperparamètre détermine la profondeur maximale de l'arbre, c'est-à-dire le nombre maximum de niveaux de branches qui peuvent être créés à partir de la racine de l'arbre. Une profondeur maximale élevée peut permettre à l'arbre de capturer des modèles plus complexes dans les données, mais peut également entraîner un sur-apprentissage si la profondeur est trop élevée.
- Nombre minimum d'échantillons requis pour diviser un noeud : cet hyperparamètre détermine le nombre minimum d'échantillons d'entraînement qui doivent être présents dans un noeud pour qu'il puisse être divisé en sous-noeuds. Un nombre minimum élevé peut empêcher l'arbre de trop sur-apprendre en évitant de diviser des noeuds avec un petit nombre d'échantillons, mais peut également empêcher l'arbre d'apprendre des modèles complexes dans les données.
- fraction de poids min. de la feuille : Fraction minimale de la somme totale des poids requise pour être à un nœud feuille
- max noeuds de feuilles : Nombre maximum de nœuds feuilles que peut avoir un arbre de décision.
- caractéristiques maximales : Nombre maximum de caractéristiques qui sont prises en compte pour le découpage de chaque nœud.

Table des figures

Liste des tableaux