Université Paris-Dauphine
Machine Learning pour la finance
Examen Blanc — 28 juin 2023

Aucun document, calculatrice ou autres objets électroniques ne sont autorisés.

Prédiction du risque de crédit

Vous travaillez dans une banque au sein de l'équipe chargée de l'octroi de crédit immobilier. Nouvellement arrivé, vous êtes chargé de construire un modèle de Machine Learning pour calculer la probabilité de défaut du demandeur de crédit.
Pour cela, vous avez à votre disposition un jeu de donnée historique labellisé de dix milles lignes, dont 10% représente un demandeur de crédit qui présente un risque de crédit. Vous avez à votre disposition les informations suivante :

- Informations sur le demandeur de crédit : âge, revenu mensuel, statut d'emploi : (employé, indépendant ou sans emploi), situation familiale (célibataire ou marié), statut de logement (locataire ou propriétaire)
- Informations sur le prêt : montant, durée (en année) et ratio prêt-valeur (ratio entre le montant du prêt hypothécaire par rapport à la valeur de la propriété)
- Risque de crédit : la variable binaire cible où 1 représente un risque

Nous noterons `df` le dataframe Python correspondant à ce jeu de donnée, X la matrice correspondante aux informations présentent dans `df` et `y` la variable cible.

1. Dans quel cadre sommes-nous : apprentissage supervisé ou non-supervisé ? Et quel sous-type ? Quelles sont, dans ce cadre, les métriques de performances que vous allez considérer ? Pourquoi ?

   Votre data engineer vous affirme que `df` ne contient pas de valeur manquante ni de valeur aberrante. Vous commencez donc la partie d'exploration des données, en prenant soin de vérifier ces informations tout de même.

2. Lors de votre exploration, vous remarquez la tendance suivante :

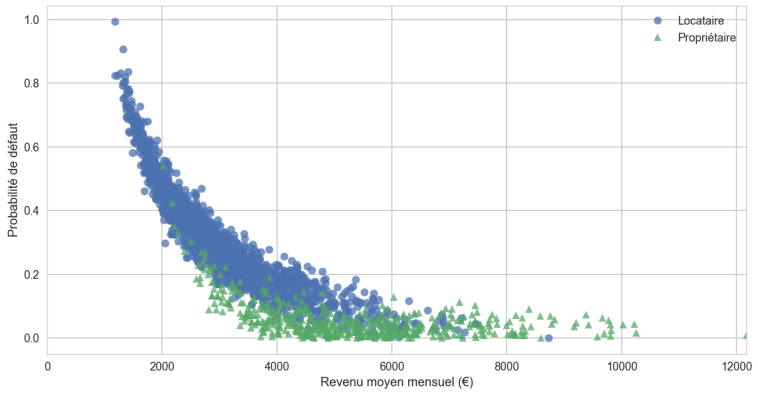

FIGURE 1 – Probabilité de défaut en fonction du revenu mensuel moyen pour deux populations

Quelle forme de fonction semble suivre la tendance pour les deux populations ?

A. $f(x) = e^{-\alpha x}$, avec $\alpha > 0$ un paramètre
B. $f(x) = \dfrac{1}{x^{\alpha}}$, avec $\alpha > 0$ un paramètre
C. $f(x) = (x - \alpha)^2$ avec $\alpha > 0$ un paramètre

3. Jusqu'à présent, nous n'avons pas traité les données catégorielles. Comment les traitez-vous ?

   Un premier travail simple de feature engineering terminé, vous commencez le travail de modélisation du risque de crédit. Puisqu'il vous a été demandé de concevoir un algorithme très explicable, vous décidez de privilégier la régression logistique, l'arbre de décision et les forêts aléatoires.

4. Quelle est la commande Python que vous allez utiliser pour séparer votre base de données en une base de données d'entraînement et une base de données de test, et pourquoi ?

   A. `X_train, X_test, y_train, y_test = train_test_split(X, y)`

   B. `X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y)`

   C. `X_train, X_test, y_train, y_test = KFold(X, y, k=5)`

5. Sans réglage et sans sélection de variable, vous calculez les performances des trois algorithmes :

Algorithme	Précision (%)	Recall (%)	F1-Score (%)
Régression Logistique	60	55	57
Arbre de décision	65	80	71
Random Forest	75	75	75

   Commentez.

6. Quelle est l'intérêt d'utiliser une Random Forest plutôt qu'un arbre ? Le tableau ci-dessus permet-il de voir cela ? Si non, expliquer la démarche à suivre pour le mesurer.

7. Pour avoir de meilleures performances, vous décidez de trouver les meilleurs paramètres pour l'arbre et la Random Forest. Comment vous-y prenez-vous ?

8. Lors de votre recherche des meilleurs paramètres, vous vous rendez compte que vous arrivez à obtenir un arbre performant avec 7 comme profondeur. Vous souhaitez construire une forêt aléatoire ayant une complexité (au sens du nombre de coupures) inférieure ou égale à cet arbre avec uniquement des arbres de profondeurs 4. Combien d'arbres peut-on mettre dans la forêt ?

9. Après avoir trouvé les meilleurs paramètres, vous inspectez les variables les plus importantes pour le modèle (nous ne présentons que les trois premières) :

   - Régression Linéaire : Montant, Âge, Ratio prêt-valeur
   - Arbre de décision : Revenu mensuel, Ratio prêt-valeur, Montant
   - Random Forest : Revenu mensuel, Ratio prêt-valeur, Âge

   Vous trouvez étonnant que l'arbre et la Random Forest utilisent autant la variable Revenu mensuel et aussi peu la régression logistique. Proposer une piste d'explication en exploitant les questions précédentes.

   Vous souhaitez approfondir cette question et décidez de linéariser la variable Revenu mensuel. On se concentre d'abord sur la sous-population des demandeurs de crédit qui sont locataires de leurs logement.

10. Justifier pourquoi on peut modéliser la relation entre le risque de crédit et le revenu mensuel par :

$$y = \frac{\theta_0}{x + \theta_1}$$

    Peut-on apprendre les paramètres θ_0 et θ_1 avec une régression linéaire ? Si non, trouvez une solution pour se placer dans le cadre d'une régression linéaire.

11. Vous obtenez un $R^2 = 0.95$. Expliquez à quoi correspond cette métrique et commentez sa valeur.

    Après avoir fait les ajustements nécessaires, les performances sont meilleures et homogènes. La Random Forest gagne toujours le concours pour la métrique $F1$. Votre référent IT se pose la question du gain que pourrait apporter une méthode de Boosting par rapport aux modèles que vous avez choisis.

12. Quelle est la différence principale entre une méthode de Boosting et une Random Forest ? Quelle démarche allez-vous suivre pour répondre à son interrogation ?

    Finalement, la demande est dépriorisée au profit d'une demande du marketing. L'objectif est d'identifier à partir de votre base de données des potentiels clients pour une offre de déménagement personnalisé, afin de sur-vendre un crédit consommation.
    Vous filtrez donc votre base de données pour ne considérer que les personnes qui ne représentent pas de risque de crédit.

13. Dans quel cadre sommes-nous : apprentissage supervisé ou non-supervisé ? Qu'est-ce que cela change par rapport au travail précédent ?

14. Présentez succinctement deux méthodes que vous pourriez utiliser pour répondre à ce problème.

15. Comment allez-vous mesurer la performance de votre modèle ?

Cosinus decay

Votre manager vous parle avec enthousiasme d'un papier de recherche qu'il vient de lire. Il propose une variante de la descente de gradient sous la forme (différente de l'originale) :

$$\begin{cases} \theta_{t+1} & = \theta_t - \eta_t \nabla f_{\theta_t}(x) \\ \eta_t & = \eta^i_{\min} + \frac{1}{2}\left(\eta^i_{\max} - \eta^i_{\min}\right)\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right) \end{cases}$$

Il explique que la descente de gradient va se faire par plusieurs redémarrages i qui auront lieu après T_i itérations. A chaque redémarrage, la valeur de T_{cur} est initialisée à 0 et est incrémenté à chaque itération. Voici un exemple avec 2 itérations, puis 4, puis 8 :

- Valeur de t : 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,12, 13
- Valeur de T_i : 1, 1, 3, 3, 3, 3, 7, 7, 7, 7, 7, 7, 7, 7
- Valeur de T_{cur} : 0, 1, 0, 1, 2, 3, 0, 1, 2, 3, 4, 5, 6, 7

Il vous informe également que c'est particulièrement utile lorsque la fonction que l'on cherche à optimiser n'est pas convexe. Sans vous en dire plus, il vous laisse seul lui préparer une note de synthèse sur cette technique.

1. Rappeler le problème que résout une descente de gradient, l'équation de la descente de gradient simple et expliquer en quoi la version proposée est bien une variante.

2. En fixant $\eta^i_{\min} = 0.1$ et $\eta^i_{\max} = 1$, dessiner l'allure de η_t si on suit l'exemple de l'introduction.

3. Pour $i \in \mathbb{N}^*$ fixé, combien y-a-t-il d'hyper paramètres à spécifier avant de commencer la descente de gradient ?

4. Dans l'objectif de réduire le nombre d'hyper paramètres, vous proposez de fixer $\eta^i_{\min}$ à une unique valeur et de ne spécifier que la valeur $\eta^0_{\max}$. Pour les valeurs suivantes de $\eta^i_{\max}$ vous proposez de fixer la valeur de l'itération $i+1$ de sorte à réduire de 25% l'écart entre $\eta^i_{\max}$ et $\eta_{\min}$. Quelle est la formule de $\eta^i_{\max}$?