Machine Learning S8

RAVOAVISON Fenomanana Irénée Stevenson
MAFI

18 Décembre 2025

Le jeu de donnée est un dataframe contenant des information sur des information sur des personnes ayant preter de l'argent .Notre but est de trouver des bons modèles de machine learning adapté aux données afin de pouvoir predidire les personnes qui ont rendu ou pas l'argent qu'ils ont preté.Donc il s'agit d'un probléme de classification.
Nous allons faire l' évaluation de 8 modèles puis trouver les 3 meilleurs en utilisant les données de validation ,les optimiser et retourner leurs performances sur les données de test .

1 Les étapes

1.1 Preparation des données

- Chargement du jeu de donée
- Voir les informations général sur le données dans et la distribution de la variable cible "loan paid back" (qui est trés deséquilibré panchant vers la classe 1)
- On voit aussi une corrélation entre le statut d'emploi des personnes et la variable cible.
- la separation des données en train,validation ,test
- Séparation des variables numériques et categorielles

- Création de encodeur pour les variables categorielles adapté à la version de Scikit learn pour garantir que OneHotEncoder fonctionne correctement quel que soit la version de scikit-learn. (j'a eu ce probléme sur ma machine)
- Creation de Pipeline :"categorical pipeline dense" et "preprocessor dense" adaptée à la modele GaussianNB, car il exige des entrées denses.

1.2 Trouver la performance des modèles

Voici les modèles utilisés

- "$LogisticRegression$"
- "$DecisionTree$"
- "$RandomForest$"
- "$GradientBoosting$"
- "$AdaBoost$"
- "$LinearSVC$"
- "$GaussianNB$"
- "$LogRegL1$"

On va faire la validation afin de voir les performances de chaque modèle correspondant à chaque metrique.On prenant la moyemme pondérée ,comme les cas positifs sont très grands par rapport aux cas négatifs :

- Accuracy : peut être trompeuse, car la classe majoritaire domine.
- Precision : proportion de vrais positifs parmi les prédits positifs.
- Recall : proportion de vrais positifs parmi les réels positifs.
- F1 : moyenne harmonique de précision et rappel, bonne métrique globale pour déséquilibre.
- ROC AUC : mesure de performance globale, indépendante du seuil, tr

Donc l vaut mieux donner plus de poids à F1 et ROC AUC, car elles capturent mieux la performance sur la classe minoritaire et l'équilibre global.Accuracy peut être moins importante ici.Les top 3 sont 'GradientBoosting', 'DecisionTree', 'AdaBoost'.
L'etape suivante est d'optimiser ces 3 modèles en cherchant les meilleur paramètres .Après évaluez les modèles optimisés avec les données de test.