Projet - Classification sur données bancaires

Objectifs: Implémenter des modèles de Classification(régression logistique et Random Forest) pour la prédiction de défauts de paiement.

Données : Jeu de données UCI_Credit_Card:

D: ID of each client
LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
SEX: Gender (1=male, 2=female)
EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE: Marital status (1=married, 2=single, 3=others)
AGE: Age in years
PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
PAY_2: Repayment status in August, 2005 (scale same as above)
PAY_3: Repayment status in July, 2005 (scale same as above)
PAY_4: Repayment status in June, 2005 (scale same as above)
PAY_5: Repayment status in May, 2005 (scale same as above)
PAY_6: Repayment status in April, 2005 (scale same as above)
BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
default.payment.next.month: Default payment (1=yes, 0=no)

A . Régression logistique

1. Charger les données et afficher les dimensions.
2. Pour plus de visibilité renommer mla variable “default.payment.next.month”.”
3. Afficher le pourcentage de défault de la base.
4. Afficher les statistiques descriptives. Commenter.
5. Données manquantes.
6. Etudiez les relations entre 2 variables explicatives au choix et la cible en utilisant des plot adéquats.
7. Analyser la corrélation. Commenter.
8. Certaines variables qualitatives ne sont pas encodées comme étant catégorielles. Corriger cela.
9. Analyser la corrélation des données.
10. Fixer la graine à “1234”.
11. Séparer le jeu de données en un échantillon d’apprentissage et un échantillon de test. Afficher leurs dimensions respectives. Rappeler l’utilité de séparer l’échantillon de base en deux.
12. Lancer un premier modèle de régression logistique avec les variables suivantes:

SEX + EDUCATION + MARRIAGE + AGE + PAY_5 + BILL_AMT1 + PAY_AMT1

13. Afficher les performances du modèle sur les deux échantillons disponibles (trian et test) en calculant l'indice de Gini et commenter.

    Gini = 2*AUC-1

14. Lancer une régression logistique en utilisant toutes les variables disponibles avec une option stepwise en utilisant la fonction stepAIC de la librairie MASS.
15. Calculer les performances du modèle résultant. Commenter. Qu'en déduisez-vous sur le modèle de régression linéaire.

B. CART et Random Forest

Il existe d'autres modèles d'apprentissage supervisé. Nous allons ici aborder les modèles CART et Random Forest. Ces modèles n'ayant pas été étudiés au préalable, je vous invite à réaliser une veille rapide sur le sujet: (ne fait pas partie de la notation)

- Qu'est ce qu'un arbre de décision?
- Comment fonctionne l'algorithme ?
- Quels sont les paramètres à faire varier?
- Qu'est ce qu'une forêt aléatoire (Random Forest)?
- Quels sont les paramètres à faire varier?
- Préciser si ces modèles sont utilisables pour la prédiction de variables qualitatives, quantitatives ou les deux.

1. Charger la librairie rpart.
2. Implémenter un modèle CART en faisant varier certains paramètres.
3. Implémenter un modèle RandomForest en faisant varier certains paramètres.