Dossier de RShiny:

Master I ESA
Projet individuel et données au choix

à rendre 25/04

Vous choisissez une des deux bases de données ci-dessous pour construire une application qui permet de prédire la probabilité de défaut par une régression logistique. Votre application doit permettre à un utilisateur de prédire la probabilité de défaut d'un nouveau client à partir de ses caractéristiques.

1 Description du projet

Votre application doit comprendre deux parties :

1.1 Modélisation

L'utilisateur doit choisir entre les trois rubriques suivantes :

1. Données : quelques statistiques descriptives, lien entre la variable cible et les variables explcatives
2. Modèle : Table de régression
3. Performances : score AUC, courbe de ROC, Taux d'erreur

1.2 Prédiction de la probabilité de défaut

L'utilisateur doit renseigner le nom et les caractéristiques de l'emprunteur et obtenir automatiquement son score (variable latente linéairement reliée aux explicatives) et son classement. Vous pouvez classer les emprunteurs en 5 catégories suivant les notations : A+++, A++, A+, A, A-, A- -

1. Excellent A+++ : Si la probabilité de défaut est inférieure à 0.1
2. Très bon client A++ : Si la probabilité de défaut est entre 0.1 et 0.3
3. Bon client A+ : Si la probabilité de défaut est entre 0.3 et 0.5
4. Mauvais payeur A- : Si la probabilité de défaut est entre 0.5 et 0.7
5. Très mauvais payeur A- - : Si la probabilité de défaut est supérieure 0.7

Exemple : Le Score de Prénom Nom emprunteur est de $E(y^*|x) = x_i\widehat{\beta}$ et sa probabilité de défaut est .... Vous pouvez lui faire confiance c'est un très bon client ....

2 Les données de Kaggle

Vous choisissez une de ces deux bases de données disponibles sur le site `https://www.kaggle.com`. Vous nettoyez votre base choisie et limitez le nombre de variables explicatives. Vous construisez un modèle performant (avec un bon pouvoir prédictif) et simple (avec seulement les variables explicatives pertinentes pour l'analyse). Pour sélectionner les variables, vous pouvez considérer celles qui ont le plus d'impacts sur la variable cible et supprimer la colinéariténtre les variables afin d' éviter une répétition de l'information.

2.1 Give Me Some Credit

Vous pouvez trouver toutes les informations de la base kaggle (kaggle.txt), "Give Me Some Credit" sur le site suivant : `https://www.kaggle.com/c/GiveMeSomeCredit/data`. Les informations sur les variables de la bases sont dans le fichier Data Dictionary.xls.

Description des variables

Variable	Type	Description
SeriousDlqin2yrs	Binary	The person experienced 90 days past due delinquency or worse (Yes/No)
RevolvingUtilizationOfUnsecuredLines	Percentage	Total balance on credit cards and personal lines of credit except real estate and no instalment debt such as car loans divided by the sum of credit limits
Age	Interval	Age of the borrower (in years)
NumberOfTime30-59DaysPastDueNotWorse	Interval	Number of times a borrower has been between 30 and 59 days past due but not worse in the last 2 years
DebtRatio	Percentage	Monthly debt payments, alimony and living costs over the monthly gross income
MonthlyIncome	Interval	Monthly Income
NumberOfOpenCreditLinesAndLoans	Interval	Number of open loans (like car loan or mortgage) and credit lines (credit cards)
NumberOfTimes90DaysLate	Interval	Number of times a borrower has been 90 days or more past due
NumberRealEstateLoansOrLines	Interval	Number of mortgage and real estate loans including home equity lines of credit
NumberOfTimes60-89DaysPastDueNotWorse	Interval	Number of times a borrower has been between 60 and 89 days past due but not worse in the last 2 years
NumberOfDependents	Interval	Number of dependents in family excluding themselves (spouse, children, etc.)

2.2 Default Payments of Credit Card Taiwan

Vous pouvez trouver toutes les informations de la base taiwan (taiwan_credit.xls) sur les sites suivant :

1. `https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset?select=UCI_Credit_Card.csv`
2. `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

Il s'agit de la même base de données exploitées différemment.

Description des variables

Variable	Type	Description
Y	Binary	default payment (Yes = 1, No = 0)
X1	Quantitative	Amount of the given credit (NT dollar)
X2	Binary	Gender (1 = male; 2 = female)
X3	Nominal	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
X4	Nominal	Marital status (1 = married; 2 = single; 3 = others)
X5	Quantitative	Age (year)
X6-X11	Nominal	X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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
X12-X17	Quantitative	Amount of bill statement (NT dollars). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005
X18-X23	Quantitative	Amount of previous payment (NT dollars). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005

La variable de réponse

— Y est binaire : paiement par défaut (Oui = 1, Non = 0)

Les variables explicatives

— X1 : Montant du crédit accordé (dollar NT) : il comprend à la fois le crédit à la consommation individuel et le crédit familial (supplémentaire).

— X2 : Sexe (1 = homme ; 2 = femme).

— X3 : Niveau d'études (1 = études supérieures ; 2 = université ; 3 = lycée ; 4 = autres).

— X4 : état civil (1 = marié ; 2 = célibataire ; 3 = autres).

— X5 : Age (année).

— X6 - X11 : Historique des paiements antérieurs. Nous avons suivi l'historique des paiements mensuels passés (d'avril à septembre 2005) comme suit : - X6 = le statut de remboursement en septembre 2005 ; X7 = le statut de remboursement en août 2005 ; ... ; X11 = le statut de remboursement en avril 2005. L'échelle de mesure de l'état de remboursement

est la suivante : -1 = paiement en bonne et due forme ; 1 = retard de paiement d'un mois ; 2 = retard de paiement de deux mois ; . . . ; 8 = retard de paiement de huit mois ; 9 = retard de paiement de neuf mois et plus.

— X12-X17 : Montant de la facture (en dollars NT). X12 = montant de la facture de septembre 2005 ; X13 : montant de la facture d'août 2005 ; . . . ; X17 = montant de la facture d'avril 2005.

— X18-X23 : Montant du paiement précédent (dollar NT). X18 = montant payé en septembre 2005 ; X19 = montant payé en août 2005 ; . . . ; X23 = montant payé en avril 2005.