Machine Learning
TD - Classification et Régression

M2 Informatique - BI&A (2022-2023)

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Abstract

Ce TD a pour but de vous faire mettre en application différents algorithmes de classification mais aussi de régression que nous avons pu voir ou revoir en cours. On pourra traiter du problème de classification et de régression à travers deux jeux de données différents.

1 Un problème de Classification

Nous allons travailler sur une base de données décrivant des clients d'une banque et leurs comportements sur le plan bancaire (mouvements, soldes des différents comptes). L'objectif est l'estimation d'un score d'appétence à la carte VISA Premier. C'est une carte de paiement haut de gamme qui cherche à renforcer le lien de proximité avec la banque en vue de fidéliser une clientèle aisée.

Après échantillonnage et une première sélection de variables, la base de données (VisaPremier.txt)

Jeu de données VisaPremier.

est composée de 1,073 clients décrits avec les variables suivantes :

Nom	Description	Nom	Description
Identif	Libellé	matricul	Matricule (clé primaire)
departem	département de résidence	ptvente	point de vente
sexe	sexe	age	age en année
sitfamil	situation familiale	anciente	durée de connaissance du client en mois
csp	catégorie socio-profiessionnelle	codeqlt	qualité du client
nbimpaye	nombre d'impayés	mtrejet	montant total des rejets
nboguic	nombre opérations guichet par mois	moycred3	moyenne des mouvements créditeurs sur les 3 derniers mois
aveparmo	total épargne	endette	taux endettement
endette	taux endettement	engagemt	total engagements
engagemc	total engagements court terme	engagemm	total engagements moyen terme
nbcptvue	nombre de compte à vue	moysold3	solde moyen sur 3 mois
moycredi	moyenne mouvements créditeurs	agemvt	age du dernier mouvement
nbop	nombre d'opérations à M-1	mtfactur	montant des facteurs sur l'année
engageml	total engagements long terme	nbvie	nombre contrats assurances vie
mtvie	montant contrats assurances vie	nbeparmo	nombre de produits épargne monétaire
mteparmo	montant des produits d'épargne monétaire	nbeparlo	nombre de produits épargne logement

mteparlo	montant des produits épargne logement	nblivret	nombre de livret
mtlivret	montant des livrets	nbeparlt	nombre de produits épargnes long terme
mteparlt	montant des produits épargne long terme	nbeparte	nombre de produits épargne fermés
mteparte	montant des produits épargne fermés	nbbon	nombe de produits bons
mtbon	montant des produits bons	nbpaiecb	nombre de paiement par CB à M-1
nbcb	nombre de CB	nbcbptar	nombre de carte point argent
avtscpte	total des avoir sur le compte	aveparfi	total des avoirs financiers
cartevp	possession de la carte Visa Premier	sexer	sexe codé en 0-1
artevpr	possession de la carte VisaPremier codé en 0-1	nbjdebit	Nombre de jours de débit

Comparer différents modèles de classification pour prédire la probabilité qu'un client dispose de la carte Visa Premier, en vous basant :

- sur les variables quantitatives uniquement,
- sur les variables catégorielles uniquement,
- sur toutes les variables.

Les modèles seront comparés sur la base de l'aire sous la courbe ROC évaluée sur un échantillon test de 200 clients tirés aléatoirement dans la base (seed fixée à 1). Notez également les temps de calculs des différents modèles. Présentez l'ensemble des résultats dans un tableau récapitulatif.

Quelques remarques : il faudra effectuer les transformations nécessaires sur les variables, préalablement à la construction ou à l'apprentissage des paramètres des différents modèles. En outre, il faudra vérifier que certaines variables ne sont pas redondantes, procéder à un encodage si nécessaire, traiter les valeurs manquantes, etc.

2 Un problème de Régression

Nous allons travailler dans ce TP sur une base de données relative à la criminologie aux Etats-Unis :

Jeux de données Crimes

Votre objectif sera de trouver le meilleur modèle de régression pour prédire le taux de crimes violents aux Etats-Unis :

ViolentCrimesPerPop

Les modèles seront comparés sur la base de l'erreur quadratique moyenne évaluée par validation croisée 10-fold.