Projet Data Mining
Classement et ciblage

Master 1 Informatique – Université Lyon 2

Description de la base

La base « data_avec_etiquettes.txt » comporte 200 variables (V1…V200) et 494 021 observations.

V200 est la variable cible, elle comporte 23 modalités.

V1…V199 sont les variables explicatives potentielles. Certaines sont qualitatives (V160, V161, V162), les autres sont quantitatives.

Objectifs de l'étude

1. Produire un système de classement :
   a. Qui permet de prédire le plus précisément possible les valeurs de la variable cible V200.
   b. Annoncer le nombre d'erreurs de classement si on appliquait votre modèle sur une base de déploiement comportant 4 898 424 obs.
2. Produire un système de « scoring » :
   a. Qui permet de cibler la modalité « m16 » (les positifs) de V200
   b. Annoncer le nombre de positifs parmi les 10 000 observations qui présentent les scores les plus élevés dans une base de 4 898 424 obs.
3. Construire une variable cible modifiée « V200_Prim » via le regroupement des modalités de « V200 » en argumentant votre démarche (nombre de groupes, constitution des groupes). Construire un modèle qui permet de prédire le plus précisément possible cette nouvelle variable cible « V200_Prim ».

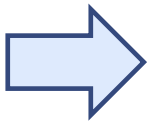
Choix des méthodes et des outils

(1)

Choix des méthodes à utiliser libre pourvu que :

a. Le modèle définitif peut s'exprimer à l'aide d'un système à base de règles s'appliquant sur les variables initiales du dataset

b. (et/ou) Le modèle peut s'exprimer à l'aide d'une combinaison linéaire basée sur les variables initiales (ou codée 0/1 pour ce qui est des explicatives qualitatives).



Dans tous les cas (objectifs 1, 2 et 3), je m'attends à ce que opériez une sélection de variables : seules sont conservées les variables pertinentes, et on dispose d'indications sur leur importance dans les modèles à déployer.

(2)

Choix des outils (logiciels) à utiliser libre pourvu qu'ils soient accessibles gratuitement et que je sois en capacité de reproduire vos calculs.

Livrables (1)

Un rapport retraçant votre démarche pour chaque objectif (1, 2 et 3). Un plan type serait d'adopter le mode de présentation préconisée par la méthodologie CRISP-DM, voir en particulier « The CRISP-DM outputs ». On doit bien identifier notamment pour chaque objectif :

a. quelles sont les approches que vous avez testées, comment vous avez identifié le modèle définitif ;
b. comment s'exprime le modèle définitif ;
c. quelles sont les variables intégrées dans le modèle définitif, avec un classement selon leur importance ;
d. comment avez-vous estimé les performances sur la base de déploiement ;
e. pour l'objectif (3), quelle est la stratégie adoptée pour procéder au regroupement des classes.

Livrables (2)

Un programme (R, Python ou autre) pour chaque objectif (1, 2 et 3) que l'on peut appliquer sur une base de déploiement (DB) au format texte avec séparateur tabulation de 4 898 424 obs. décrite exclusivement par les variables (V1… V199 ; les noms de variables sont en en-tête de chaque champ). Ils doivent :

1. Prendre en entrée DB et produire une prédiction stockée dans un fichier texte nommé « predictions.txt » du répertoire courant.
2. Prendre en entrée DB et produire le score d'appartenance à la classe « m16 » dans un fichier texte nommé « scores.txt » du répertoire courant.
3. Prendre en entrée DB et un fichier « classes.txt » contenant les classes d'appartenance originelles (nom de variable : V200). Il doit effectuer le regroupement selon votre stratégie, effectuer la prédiction sur DB, et sauvegarder dans un fichier « sorties.txt » les prédictions et les classes regroupées.

Critères d'évaluation

- Travail à faire en groupe de 3 étudiants max.

- Performances prédictives
- Conformité des performances annoncées avec les performances effectivement mesurées durant la correction.

- Qualité et fiabilité des programmes de déploiement (attention, taille du fichier de déploiement $\approx 2\mathrm{GB}$, faites des essais en dupliquant la base à votre disposition)

- Argumentation des choix, positionnement des différentes alternatives, pertinence de la sélection de variables
- Lisibilité des modèles prédictifs, identification des variables pertinentes
- Qualité de rédaction du rapport (texte, tableaux, graphiques). Rédigez correctement (disons une 15-aine de pages max pour donner un ordre d'idées).

Calendrier

Diffusion du sujet et des données : lundi 07 décembre 2020

Retour du travail des étudiants : vendredi 18 décembre 2020 au soir

Envoyer un e-mail à ricco.rakotomalala@univ-lyon2.fr, mettez-vous en copie :

- Sujet de l'e-mail : [M1 INFO – Data Mining] Noms des étudiants
- Dans le corps de l'e-mail, indiquez les coordonnées d'un drive où je pourrai récupérer votre travail (ne mettez pas votre travail en fichier attaché, ma boîte mail n'y survivrait pas)
- Je m'attends à trouver sur votre drive : le rapport ; les 3 programmes de déploiement ; tout autre fichier me permettant de retracer et reproduire votre travail (documentez vos programmes).