Ce travail est à réaliser en équipe (**minimum deux, maximum quatre** personnes). On cherche à estimer le total du solde de la carte de crédit VISA Première (`credm`, en francs) à l'aide d'autres variables explicatives présentes dans la base de données `visacredm21a`; 25 données sont intentionnellement manquantes pour permettre d'évaluer vos modèles finaux et faire un classement des équipes.

1. Faites une brève analyse exploratoire des données : vérifiez
   - — que les variables catégorielles sont déclarées comme telles avec `class`
   - — s'il vaudrait mieux fusionner des modalités de variables catégorielles si le nombre d'observation par modalité est trop faible
   - — qu'il n'y a pas de variable explicative qui soit un dérivé de la variable réponse
   - — que le sous-ensemble des observations employé est adéquat
   - — qu'il n'y a pas des variables explicatives qui ont une relation nonlinéaire avec la réponse. Le cas échéant, ça peut être payant de faire des transformations $\log(x+1)$, ou créer des indicateurs binaires pour ajouter en plus.
   - — s'il n'y a pas d'anomalies ou des valeurs aberrantes (e.g., 999 pour valeurs manquantes) qui viendraient fausser les résultats.

   Utilisez des graphiques pour déterminer si vous devriez inclure des transformations des variables explicatives. Résumez votre analyse en une page maximum (texte et graphiques).
2. **Sélection de variables** : Essayez les méthodes de sélection suivantes (`choose`) pour effectuer votre sélection pour ces données avec les trois options suivantes :
   (a) pénalisation (sélection avec critères d'information), **sauf pour le LASSO** [1]
   (b) validation croisée à cinq plis
   (c) séparation de la base de données en échantillons d'entraînement (2/3) et de validation (1/3)

   Pour chacune des méthodes de sélection, utilisez les procédures suivantes pour la recherche de modèles :
   - — Sélection séquentielle
   - — Régression avec pénalité q_1 (LASSO)
   - — Moyenne de modèles

   Rapportez l'erreur moyenne quadratique pour les différents modèles avec les méthodes (b) et (c) dans un tableau pour chaque modèle estimé.
3. Écrivez un rapport résumant vos trouvailles et détaillant votre démarche. Fournissez votre code SAS et une base de données au format `d4_matricule.sas7bdat` avec deux colonnes et les 25 lignes correspondant aux données manquantes pour `credm` : la première colonne contiendra les matricules (`matric`), la deuxième colonnes les prédictions (`predict`) selon votre modèle préféré parmi tous ceux essayés. Justifiez votre choix adéquatement.

Vous serez évalués sur votre méthodologie, et non pas la performance relative de votre modèle par rapport à celles des autres étudiant(e)s. Vous devez expliquer clairement votre démarche (méthodologie) dans votre rapport et décrire le modèle que vous avez retenu (les variables utilisées).

1. Le LASSO n'est pas ajusté par maximum de vraisemblance.

Compilé le 09/11/2021 à 10:44