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 `visacredm`; 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. Quelles variables seraient d'après-vous le plus utile pour modéliser `credm`? Utilisez des graphiques pour déterminer si vous devriez inclure des transformations (logarithmiques, indicateurs de points de ruptures, interactions, polynômes) des variables explicatives. Résumez votre analyse en une page maximum de texte et des graphiques choisis.
2. **Sélection de variables :**
   (a) Essayez les méthodes de sélection suivantes et discutez de leur performance relative pour ces données :
      — Sélection séquentielle avec critères d'information (plusieurs options possibles pour les critères d'entrée et de sortie).
      — Sélection séquentielle avec critères d'information, suivie d'une recherche exhaustive sur un nombre réduit de modèles.
      — Régression avec pénalité q_1 (LASSO) — choisissez la valeur du paramètres de pénalité λ par validation croisée avec votre échantillon d'apprentissage.
      — Mélange de modèles avec sélection séquentielle avec critères d'information..
   (b) Séparez votre échantillon 50/50 ou utilisez la validation croisée pour calculer l'erreur moyenne quadratique de vos différents modèles et effectuer votre sélection.
   (c) Pour chaque modèle, estimez l'erreur moyenne quadratique sur un échantillon de validation (ou par validation croisée) et rapportez cette valeur dans votre rapport PDF.
3. Variations : refaites les questions précédentes, mais en imposant les changements suivants :
   (a) Vous pourriez imposer des contraintes pour n'inclure des interactions que si les effets principaux sont présents (en retirant `hier=none`). Essayez pour chaque méthode d'imposer cette contrainte; est-ce que vous perdez beaucoup de performance ce faisant?
   (b) Considérez une transformation logarithmique de la variable réponse $y \mapsto \ln(y+1)$ pour corriger la forte asymmétrie. Est-ce que ça améliorerait votre prédiction? (attention à calculer l'erreur moyenne quadratique à l'échelle des données originales.
4. 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`).

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).

Compilé le 17/02/2021 à 21:47