Devoir 3

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 analyse exploratoire des données et vérifiez

- s'il ne vaudrait mieux pas retirer certaines variables explicatives avec beaucoup de valeurs manquantes.
- que les variables catégorielles sont adéquatement traitées comme des facteurs (`factor`).
- s'il ne vaudrait mieux pas 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 dérivée de la variable réponse.
- que le sous-ensemble des observations employé est adéquat.
- qu'il n'y a pas d'anomalies ou de valeurs aberrantes (par ex., `999` pour valeurs manquantes) qui viendraient fausser les résultats.
- (*facultatif*) si vous ne devriez pas inclure de nouvelles variables qui sont des transformations, notamment
  - ajouter une variable binaire (0/1) pour une variable strictement positive avec excès de zéros
  - transformation monotone, par exemple $x \mapsto \ln(x+1)$, si la variable est non-négative et la relation avec la réponse est nonlinéaire.

Résumez votre analyse en une page maximum (texte et graphiques): ne soulevez que les points importants.

2. Choisissez un modèle à l'aide des méthodes de sélection couvertes en classe (sélection exhaustive, séquentielle, pénalité LASSO, etc.) Vous devez adéquatement estimer votre erreur moyenne quadratique en utilisant l'une des options suivantes :

(a) pénalisation (sélection avec critères d'information)[1]

[1]Notez que le modèle LASSO n'est pas ajusté par maximum de vraisemblance!

(b) validation croisée à cinq ou 10 plis.
(c) séparation de la base de données en échantillons d'entraînement (2/3) et de validation (1/3).

Rapportez l'erreur moyenne quadratique pour les différents modèles estimés dans un tableau et justifiez adéquatement votre choix final.

3. Écrivez un rapport résumant vos trouvailles et détaillant votre démarche.

Vous devez remettre les documents suivants:

- votre rapport au format PDF
- votre code **R** ou un fichier Rmarkdown
- une base de données en format CSV[2]

Utilisez la convention de nomenclature `d3_matricule.extension`, où `matricule` est le matricule de l'étudiant(e) qui soumet le rapport et `extension` est un de `pdf`, `R`, `csv`.

La base de données devrait contenir uniquement deux colonnes et les 25 lignes correspondant aux données manquantes pour `credm`:

- la première colonne pour les matricules (`matric`),
- la deuxième colonne pour les prédictions (`predict`) selon votre modèle préféré parmi tous ceux essayés.

Indication : vérifiez votre base de données pour vous assurer de respecter les consignes (pénalités salées pour toute personne qui déroge aux consignes).

Assurez-vous également que les prédictions sont sensées et cohérentes avec ce qui est présent dans la base de données (avez-vous des prédictions en dehors de l'étendue des données d'entraînement)? Vous ne devriez pas avoir de valeurs manquantes.

Vous serez évalués sur votre méthodologie, et non pas la performance relative de votre modèle par rapport à celles des autres étudiant(e)s : en revanche, les deux sont typiquement corrélées. Vous devez expliquer clairement votre démarche (méthodologie) dans votre rapport et décrire le modèle que vous avez retenu (méthode et nombre de variables final). Prenez garde au surajustement!

[2]Disponible à l'aide de la commande `write.csv(..., file = "d3_matricule.csv", row.names = FALSE)`