Ce travail est à réaliser en équipe (minimum deux, maximum quatre personnes).
Vous devez remettre les documents suivants :

— votre rapport au format PDF
— votre code **R** ou un fichier Quarto
— une base de données en format CSV [1]

Utilisez la convention de nomenclature `d1_matricule.extension`, où `matricule` est le matricule de l'étudiant(e) qui soumet le rapport et `extension` est un de `pdf`, `qmd`, `R`, `csv`.
On cherche à prédire si une personne s'est prévalue d'une offre promotionnelle pour une nouvelle carte de crédit (`promo`, en francs) à l'aide d'autres variables explicatives présentes dans la base de données `visapromo`; 50 données sont intentionnellement manquantes pour permettre d'évaluer vos modèles finaux et faire un classement des équipes.

1. Faites une analyse exploratoire pour vous assurer de la conformité de la base de données.
   Les étapes suivantes sont obligatoires :
   (a) Fusion de catégories : ré-étiquetez les situations familiales (`famiq`) selon que la personne est seule (`seu`) ou en couple (`cou`).
   (b) Transformez les valeurs manquantes (`inc` pour inconnu) en `NA` (voir `dplyr::na_if`).
   (c) Que représentent les variables manquantes résiduelles de `zocnb`? Remplacez ces valeurs manquantes par des valeurs numériques adéquates.
   (d) Produisez et rapportez un histogramme de la variable ancienneté du compte (`relat`) et un nuage de point de `relat` et `age`. Que remarquez vous? Commentez sur la relation entre les deux variables.
2. Choisissez un modèle à l'aide des méthodes de sélection couvertes en classe (sélection exhaustive, séquentielle, pénalité LASSO, etc.) pour prédire si les personnes se prévalent de l'offre promotionnelle ou pas. Le critère employé pour juger sera le taux de bonne classification.

Parmi les méthodes couvertes figurent notamment.

(a) pénalisation (sélection avec critères d'information)
(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 de classification, la sensibilité et la spécificité pour les différents modèles estimés dans un tableau et justifiez adéquatement votre choix final de modèle.
Écrivez un rapport détaillant votre démarche et vos résultats. Résumez votre analyse exploratoire de manière succinte : ne soulevez que les points importants.
La base de données devrait contenir uniquement deux colonnes et les 50 lignes correspondant aux données manquantes pour `promo` :

— la première colonne pour les matricules (`matric`),
— la deuxième colonne pour les prédictions (`predict`) selon votre modèle préféré parmi tous ceux essayés (prédictions binaires avec valeur '0' pour non et '1' pour oui).

1. Disponible à l'aide de la commande `write.csv(..., file = "d1_matricule.csv", row.names = FALSE)`

Compilé le 20/08/2023 à 15:12

Indication : vérifiez votre base de données pour vous assurer de respecter les consignes (pénalités 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 binaires) ? 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 !

Compilé le 20/08/2023 à 15:12