Sujet :Projet M2 2022-23

Modèles pour la régression

Katia Meziani

Les données

Le jeu de données concerne la vente de 5891 logements et de 29 variables explicatives décrivant (presque) tous les aspects de ce logement et de son environnement en Corée du Sud dans la ville de Daegu, sur une période de 10 ans.

But de l'étude

L'objectif est de prédire le prix de vente et d'obtenir le modèle qui aura le plus petit **RMSE**. Vous devez contruire un (ou 2 modèles) qui permet de calculer le prix de vente du logement (Y=**SalePrice**) en fonction de toutes ou une partie des variables (regresseurs/facteurs) contenues dans le jeu de données. Il y a 2 jeux de données :

- Le "train" avec 4189 observations, à partir duquel vous devez contruire votre modèle final
- Le "test" avec 1702 observations, que vous ne devez pas utiliser pour contruire votre modèle.

```
train=read.csv2("train.csv",header=TRUE,sep=",")
test=read.csv2("test.csv",header=TRUE,sep=",")
```

Obligatoire

- Vous devez mettre la seed à 2022 pour toute l'étude.

```
set.seed(2022)
```

- Vous ne devez utiliser que les modèles vus en cours (modèles linéaires, modèles linéaires généralisés et les estimateurs biaisés. Pas de XGboost,....

Vous devez fournir:

- Un rapport de 5 pages maximum au format .pdf. C'est un résumé de vos recherches. Ce rapport doit être compris par un **non initié. AUCUN CODE!** Il comportera les parties suivantes

Part A. Vous y présenterez une étude descriptives des données: les variables explicatives que vous avez conservées dans l'étude, la raison pour laquelle vous les avez conservées. On attachera une attention particulière à la variable cible. Vous donnerez quelques représentations numériques et/ou graphiques qui résument/expliquent vos choix/données (boxplot, table, plot....). La qualité des graphiques et des tableaux sera autant appréciée que leur pertinence!! On commentera soigneusement cette étude.

Part B. Vous présenterez votre modèle final et les démarches effectuées pour arriver à ce dernier. Vous mettrez en évidence par des représentations numériques et/ou graphiques la pertinence de votre modèle. Vous présenterez vos scores : le **RMSE** à la fois sur le train et le test dans un tableau soigné. Vous commenterez vos résultats.

Part C. Vous ferez une conclusion qui résumera votre étude. Quelles sont vos conclusions finales ? Mentionnez les limites de votre analyse ou les orientations futures possibles de la recherche...

- Une Annexe (pas de limite) au **format .rmd et au format .pdf** où **toutes les commandes auront été compilées**. Devront appararaitre:
  - Votre prise en main des données: après avoir téléchargé les 2 jeux de données, faites les vérifications usuelles (la bonne déclaration des données, suppression de colonnes inutiles...)
  - Faire les représentations numériques et/ou graphiques qui résument/expliquent vos données (boxplot, table, plot....) (*Plus nombreuses que celles du rapport*). La qualité des graphiques et des tableaux sera autant appréciée que leur pertinence!! On commentera soigneusement cette étude. On attachera une attention particulière à la variable cible.
  - Les différents modèles que vous avez testés, leurs scores seront résumés dans un tableau unique et clair. Vous proposerez un modèle final en expliquant pourquoi ce choix.
  - Vous mettrez en évidence votre recherche d'outliers ET la "validation" de votre modèle préféré.
  - Vous mettrez en évidence par des représentations numériques et/ou graphiques la pertinence de votre modèle préféré.
  - **Le score**: Vous devez évaluer vos scores finaux (**RMSE**) à la fois sur le train et le test. Vous présenterez vos résultats dans un tableau soigné.Attention le score final doit être calculé à l'échelle (pas sur une transformation de la target Y !!. Vous commenterez vos résultats.
  - Vous ferez une conclusion qui résumera votre étude. Quelles sont vos conclusions finales ? Mentionnez les limites de votre analyse ou les orientations futures possibles de la recherche.
  - Mettre une table des matières grace à la commande ' \tableofcontents'.

Notation

Sont pris en compte dans la notation 3 points importants:

- La qualité de la présentation et la compréhension du *Rapport* par un non initié.
- La variété, la qualité et la pertinence des représentations numériques et/ou graphiques.
- La qualité des codes et le soin apporté à leur explication.
- Le Score par rapport à ma baseline et les autres groupes.
- La conclusion (critique, ouverture. . .)

Les fichiers doivent être envoyés VIA TEAMS par message au plus tard le 27 janvier 2023, chaque jour entrainera une pénalité d'1 point.

Les Packages vus en cours

Vous pouvez bien entendu (il est même souhaitable) enrichir la librairie!

```
library(MASS)
library(knitr)
library(ggplot2)
library(cowplot)
library(reshape2)
library(dplyr)
library(GGally)
library(corrplot)
library(carData)
library(car)
library(questionr)
library(multcomp)
library(dplyr)
library(tidyverse)
library(forestmodel)
library(effects)
library(pscl)
library(ResourceSelection)
library(survey)
library(caret)
library(pROC)
library(ROCR)
library(mlr)
library(randomForest)
library(party)
library(rpart)
```

```
library(rpart.plot)
library(caret)
library(nnet)
library(ResourceSelection)
```