SAE Expliquer ou prédire une variable quantitative à partir de plusieurs facteurs

L'objectif du projet est de construire le meilleur modèle de régression multiple pour prédire le prix des maisons dans la ville d'Ames grâce aux caractéristiques données dans le dataset et de le soumettre à la plateforme kaggle qui en évaluera la performance :

House Prices - Advanced Regression Techniques | Kaggle

C'est une compétition contre tous les autres datascientists du monde !

Ce projet donnera lieu à un rendu (rapport pdf et script R correspondant) à rendre au plus tard le dimanche soir 2 avril à minuit.

Vous utiliserez les données du fichier simplifiedAmesHousing.csv

Vous devrez commencer par :

- faire une analyse exploratoire des données (EDA : Explorarory Data Analysis) c'est-à-dire étudier chaque variable du dataset séparément (graphique approprié + résultats de la fonction skim() -> commentaire si et seulement si vous remarquez quelque chose d'intéressant, de particulier)
- nettoyer les données (corriger les types de variable, traiter les données manquantes soit par suppression du prédicteur s'il est trop peu renseigné et jugé non influent sur la cible en regardant par exemple un boxplot de la cible en fonction de ce prédicteur, avec l'option notch=TRUE, s'il est qualitatif ou la corrélation s'il est quantitatif soit par suppression d'individus si et seulement leur ligne est presque vide soit par une imputation pertinente par la moyenne, la médiane ou les kNN)
- examiner les éventuelles multicolinéarité et les corriger.
- agréger en une seule variable certaines variables éventuellement redondantes
- proposer un premier modèle de régression multiple pour expliquer la variable SalePrice (cible) à l'aide des autres variables pertinentes.
- effectuer une sélection pertinente des variables pour proposer votre meilleur modèle après en avoir comparé plusieurs.
- Soumettre le résultat à la plateforme Kaggle, en faire une capture que vous intègrerez à votre rapport