Manon BONNAUD-DUBOIS Alexandre BESNIER

**SAÉ - Régression sur des données réelles**

Pour élaborer un modèle de prédiction du prix de vente des logements en 2021 dans les Deux-Sèvres et la Charente-Maritime, nous avons utilisé le jeu de données de train qui contient le prix de vente et des variables communiquant des informations concernant chaque logement. Nous avons également utilisé le fichier de test qui contient les mêmes variables que train, mais pas le prix de vente. Notre objectif était de déterminer le meilleur modèle de prédiction du prix de vente et de l'appliquer pour prédire le prix de vente des logements présents dans le fichier test. Nous allons vous présenter les démarches faites et le choix du modèle de prédiction que nous avons trouvé le plus pertinent.

Nous avons commencé par examiner les données et vérifier s'il y avait des valeurs manquantes ou des données aberrantes dans les deux fichiers csv. Nous avons d’abord remarqué des valeurs aberrantes dans les valeurs foncières au niveau du premier et du dernier quartile grâce à une visualisation graphique. Nous avons ainsi nettoyé les données en supprimant les observations extrêmes pour éviter toute perturbation dans nos analyses.

Une image contenant graphique

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement

Nous avons également construit des modèles de régression linéaire, logarithmique, puissance et exponentielle pour identifier le meilleur modèle pour prédire le prix de vente. Nous avons décidé de travailler principalement sur la surface réelle bâti car nous avons estimés, après analyse des SCR entre les différentes variables et la valeur foncière, que c’était celle qui donnait le SCR le moins élevé et donc la plus corrélée (un SCR de 4,32e+14 avec le modèle linéaire simple environ).

Nous avons comparé les performances de chaque modèle à l'aide de la mesure de la somme des carrés des résidus. le modèle linéaire avait ainsi la meilleure performance en termes de prédiction du prix de vente car c’était celui qui possédait la somme des carrés des résidus la plus faible.

Nous avons utilisé des graphiques pour visualiser les distributions et les corrélations entre différentes variables telles que le nombre de pièce du logement en faisant des classements : inférieure à 6 pièces et supérieure à 6 pièces, la superficie du bâtiment ou encore le département : 79 ou 17.

Une image contenant graphique

Description générée automatiquementLe modèle qui nous a donné le SCR le plus faible était celui entre les département et plus précisément le 79 (avec un SCR de 1.21e+13 environ).

Pour finir, nous avons donc utilisé le modèle linéaire pour prédire le prix de vente des logements dans le fichier test, en utilisant la superficie d’un bâtiment en fonction de son département. Nous avons utilisé la formule “y=a\*x+b” pour prédire les prix de vente des logements dans le fichier test.

Nous avons ensuite créé un fichier nommé "prediction.csv" au format CSV contenant les prédictions du prix de vente pour les logements dans le fichier test. Ce fichier contient deux variables : la variable id du jeu test et une variable “Valeur. Foncière” contenant nos prédictions.

En conclusion, nous avons élaboré un modèle de prédiction du prix de vente des logements en 2021 dans les Deux-Sèvres et la Charente-Maritime en utilisant le jeu de données de train. Nous avons utilisé le modèle linéaire en fonction du département pour prédire les prix de vente des logements dans le fichier test et avons créé un fichier de prédictions au format CSV.