Projet

UE Modèle linéaire

Groupe DS

Objectif du projet : prédire les prix immobiliers

Données

Les données sont celles du `California Housing Dataset` disponible dans la librairie `sklearn` de Python. Il s'agit d'un jeu de données contenant des informations sur les logements en Californie. Les variables du jeu de données sont :

- `MedInc` : revenu médian du bloc
- `HouseAge` : âge médian des logements du bloc
- `AveRooms` : nombre moyen de pièces par ménage
- `AveBedrms` : nombre moyen de chambres par ménage
- `Population` : population du bloc
- `AveOccup` : nombre moyen de personnes par ménage
- `Latitude` : latitude du bloc
- `Longitude` : longitude du bloc

Ce jeu de données a été obtenu à partir du dépôt StatLib. https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

La variable cible est la valeur médiane des logements pour les districts de Californie, exprimée en centaines de milliers de dollars (100 000 $).

Ce jeu de données a été dérivé du recensement américain de 1990, en utilisant une ligne par groupe de blocs de recensement. Un groupe de blocs est la plus petite unité géographique pour laquelle le Bureau du recensement des États-Unis publie des données d'échantillon (un groupe de blocs a généralement une population de 600 à 3 000 personnes).

Un ménage est un groupe de personnes résidant dans un logement. Étant donné que le nombre moyen de pièces et de chambres dans ce jeu de données est fourni par ménage, ces colonnes peuvent prendre des valeurs étonnamment élevées pour les groupes de blocs avec peu de ménages et de nombreux logements vides, comme les stations de vacances.

Problématique

Il s'agit de prédire la valeur médiane des logements.

Questions pour vous guidez

Après avoir importé puis pris en main les données, vous pouvez vous poser les questions suivantes pour vous guider dans votre analyse :

- Prédire en sélection les covariables par critère BIC ou AIC
- Estimer R^2 par validation croisée et comparer avec la valeur de R^2 brute et du R^2-ajusté
- Remplacer longitude et latitude par une variable de moyenne des prix médian dans un rayon de x km autour de chaque bloc (ce bloc exclu)
- Utiliser des méthodes de sélection de variables ou de pénalisation pour prédire en utilisant les covariables du jeu et plusieurs covariables dérivées du point précédent pour différentes valeurs de x.
- Interpréter les résultats obtenus, proposer des pistes d'amélioration et de nouvelles analyses possibles

Livrables

- Codes commentés (code avec commentaires, ou notebook)
  - Codage de l'import des données aux méthodologies statistiques
- Rapport de 5 pages maximum (hors annexes et figures)
  - Résumé de l'objectif du projet, des données de départ
  - Résumé de la méthodologie utilisée
  - Présentation des résultats numériques

Objectifs des livrables :

- Le code doit être compréhensible par un tiers pour être adapté, appliqué à un autre jeu de données
- Le rapport doit être compréhensible en 10 minutes par un lecteur cherchant les résultats numériques