Évaluation Pratique : Modélisation Statistique du Prix des Voitures

Achille Thin
4 novembre 2025

1 Introduction

Cette évaluation vise à tester vos compétences en modélisation statistique en utilisant le langage de programmation R ou python en utilisant l'ensemble des outils vus en cours. Vous travaillerez sur ce jeu de données qui contient diverses caractéristiques des voitures et leur prix. L'objectif est de construire un modèle statistique pour prédire le prix des voitures en fonction de leurs caractéristiques. Ce travail est à rendre pour le 15 janvier 2026. Vous le rendrez sous la forme d'un dossier zip, sous le format *NOM_prenom.zip*, contenant : votre code, votre rapport en pdf, et les données que vous avez utilisées.

2 Description du Jeu de Données

3 Questions

3.1 Exploration des Données

1. Importez le jeu de données en R (ou python) et affichez les premières lignes. Faites une description statistique sommaire des variables (que représentent-elles ? sont-elles catégorielles ou quantitatives ?).
2. Visualisez la distribution du prix des voitures. Utilisez un histogramme ou un diagramme en boîte.
3. Examinez les relations entre le prix des voitures et les autres variables à l'aide de graphiques appropriés. Concluez.

3.2 Construction du Modèle Linéaire

1. Construisez un modèle linéaire initial en utilisant toutes les variables explicatives.
2. Analysez l'influence des points de données individuels sur le modèle. Identifiez et traitez les valeurs aberrantes.
3. Vérifiez les hypothèses de base du modèle linéaire (normalité des résidus, homoscédasticité, etc.). Qu'en concluez-vous ? Si les hypothèses ne sont pas vérifiées, que faut-il faire ? Le cas échéant, recommencer l'analyse.

3.3 Sélection de Variables et Affinement du Modèle

1. Testez les effets des différentes variables dans le modèle. Qu'en concluez-vous ?
2. Interprétez les coefficients obtenus.
3. Comparez les performances de différents modèles en utilisant les différents critères vus en cours.

4 Validation du Modèle

1. Évaluez la performance du modèle final sur un ensemble de test. Calculez l'erreur quadratique moyenne (MSE) et le R^2 ajusté.
2. Sur le jeu de donnée test (disponible ici), effectuez maintenant la prédiction du modèle, avec intervalle de confiance. Qu'en pensez-vous ?

5 Conclusion

Rédigez un bref rapport qui résume vos résultats, vos choix de modélisation et les conclusions et interprétations que vous tirez de cette analyse.

6 Pour aller plus loin

Répétez l'analyse précédente sur un autre jeu de données. Vous pourrez pour ceci utiliser si vous voulez un jeu de données de votre choix, qui a un intérêt particulier d'un point de vue professionnel ou personnel. Choisissez dans ce cadre une variable quantitative à modéliser, et étudiez son comportement vis à vis des autres variables de votre jeu de données. Interprétez et tirez des conclusions en mêlant intuition et outils statistiques.

Vous pouvez sinon utiliser le jeu de donnée décrivant la note de dégustation donnée à des vins rouges et leurs caractéristiques chimiques et prix associés, disponible ici. Répétez l'analyse décrite par les questions précédentes.

Une fois cette analyse terminée, essayez de prédire les notes des vins blancs à partir de ces mêmes caractéristiques chimiques, jeu de donnée disponible ici. Qu'en pensez-vous ? Pensez vous à un test statistique permettant de savoir si le modèle fonctionne sur ce jeu de donnée ?

Si vous deviez maintenant faire un modèle global prédisant les notes de vins rouges et blancs, que feriez-vous ? Comment déterminer si les ressentis associés à chacune des caractéristiques chimiques des vins rouges et blancs diffèrent ?