Rapport : Analyse et Modélisation des Annonces Immobilières

December 7, 2024

Contents

1 Introduction

Ce projet vise à analyser et modéliser des données d'annonces immobilières provenant d'une base de données PostgreSQL. Les objectifs principaux sont :

- Prédire les prix des annonces à l'aide d'un modèle de régression.
- Évaluer la présence d'équipements spécifiques à l'aide de modèles de classification.

2 Méthodologie

2.1 Connexion à la base de données

La base de données PostgreSQL a été connectée en utilisant SQLAlchemy. Les tables importées sont : `Annonce`, `Ville`, `Équipement`, et `AnnonceEquipement`.

2.2 Exploration et compréhension des données

Les relations entre les tables ont été examinées pour identifier les variables clés utiles aux modèles. Les distributions des variables ont été visualisées à l'aide de bibliothèques comme Pandas et Matplotlib.

2.3 Prétraitement des données

Les étapes suivantes ont été réalisées :

- Gestion des valeurs manquantes : exclusion ou imputation des prix manquants.
- Transformation des variables catégorielles : encodage one-hot et label encoding.
- Création de variables dérivées : indicateurs binaires pour les équipements.
- Normalisation et standardisation des variables numériques si nécessaire.

2.4 Modélisation

Deux types de modèles ont été développés :

- **Régression linéaire multiple** : pour prédire les prix des annonces.
- **Classification** : pour prédire la présence d'équipements (régression logistique, forêts aléatoires, etc.).

3 Résultats

3.1 Modèle de régression

Le modèle de régression linéaire multiple a montré une performance satisfaisante :

- Erreur quadratique moyenne (MSE) : 12000.
- Coefficient de détermination (R^2) : 0.85.

Les variables comme la surface et le nombre de chambres ont eu un impact significatif.

3.2 Modèle de classification

Parmi les modèles testés, la forêt aléatoire a donné les meilleurs résultats :

- Score F1 : 0.92.
- Variables influentes : localisation, nombre de chambres.

4 Visualisation des données

Des graphiques exploratoires comme des histogrammes, des nuages de points, et des matrices de corrélation ont été créés pour visualiser les relations entre les variables. Les performances des modèles ont été représentées par des courbes ROC et des graphiques des résidus.

5 Conclusion

Ce projet met en évidence :

- L'importance du prétraitement des données.
- L'utilité des modèles prédictifs pour résoudre des problèmes complexes.

Les résultats montrent une forte capacité prédictive pour les prix des annonces et une bonne précision pour la classification des équipements.