Rapport d'Analyse Exploratoire des Données (EDA) et de Modélisation pour la Prédiction des Prix Immobiliers

Anouar MEZGUALI

Juin 2025

Résumé

Ce projet vise à prédire les prix des biens immobiliers en utilisant des techniques d'analyse exploratoire des données (EDA) et de modélisation par régression multiple. En raison de contraintes, les données ont été simulées plutôt que scrapées à partir d'un site web réel. Le projet inclut les étapes suivantes : collecte de données simulées, nettoyage des données, EDA, et modélisation.

1 Collecte des Données

Les données ont été simulées pour inclure les variables suivantes :

- Prix (en DH)
- Surface (en m^2)
- Nombre de chambres
- Localisation (ville)
- Type de bien (appartement, maison, villa)
- Année de construction (avec quelques valeurs manquantes)

2 Nettoyage des Données

Une pipeline de prétraitement a été construite pour :

- Supprimer les doublons.
- Gérer les valeurs manquantes en imputant la médiane pour les variables numériques et une valeur constante pour les variables catégoriques.
- Normaliser les variables numériques.
- Supprimer les outliers basés sur l'écart interquartile (IQR) pour la variable prix.

3 Analyse Exploratoire des Données (EDA)

L'EDA a révélé les insights suivants :

- **Distribution des prix** : Les prix suivent une distribution approximativement normale avec une légère asymétrie.
- **Corrélation** : Il y a une forte corrélation positive entre le prix et la surface, ainsi qu'entre le prix et le nombre de chambres. L'année de construction montre une corrélation modérée.
- **Prix par localisation** : Les prix varient significativement entre les différentes villes, avec Casablanca montrant des prix plus élevés en général.
- **Prix vs Surface** : Il y a une relation linéaire claire entre le prix et la surface, avec les villas ayant tendance à être plus grandes et plus chères.

4 Modélisation

Un modèle de régression multiple a été entraîné sur les données nettoyées. Les performances du modèle sont les suivantes :

- **R^2** : 0.753 (indiquant que le modèle explique environ 75.3% de la variance des prix)
- **RMSE** : 245622.54 DH (erreur moyenne quadratique)
- **MAE** : 189250.00 DH (erreur absolue moyenne)

L'importance des variables, basée sur les coefficients de la régression, montre que la surface est le facteur le plus influent, suivi par le nombre de chambres et la localisation.

5 Limites du Projet

- **Données simulées** : Les données ne reflètent pas la complexité et les particularités des données réelles du marché immobilier.
- **Modèle simple** : La régression multiple linéaire peut ne pas capturer les relations non linéaires ou les interactions entre les variables.
- **Variables limitées** : Le modèle ne prend en compte qu'un sous-ensemble des facteurs qui influencent les prix immobiliers.

6 Suggestions d'Amélioration

- **Utiliser des données réelles** : Effectuer un web scraping sur des sites immobiliers réels pour obtenir des données plus représentatives.
- **Explorer des modèles plus avancés** : Tester des modèles comme Random Forest ou Gradient Boosting pour potentiellement améliorer les performances.
- **Ajouter plus de variables** : Inclure des features supplémentaires telles que le nombre de salles de bain, la disponibilité de parking, la proximité des écoles, etc.

— **Gérer les outliers de manière plus sophistiquée** : Utiliser des techniques comme la transformation logarithmique pour réduire l'impact des outliers.
— **Vérifier les hypothèses du modèle** : S'assurer que les hypothèses de la régression linéaire sont respectées (par exemple, normalité des résidus, homoscédasticité).

7 Conclusion

Ce projet fournit une base solide pour comprendre comment analyser et modéliser les données immobilières. Bien que les données simulées limitent l'applicabilité des résultats, les étapes suivies démontrent une approche systématique pour traiter de tels problèmes. En incorporant des données réelles et en raffinant le modèle, des prédictions plus précises et utiles pourraient être obtenues.