CSE422 : Artificial Intelligence

LAB PROJECT REPORT

Index

7. **Model Selection and Comparison**
8. **Conclusion**

1.Introduction:

The aim of this project is to predict housing prices based on features like square footage, number of bedrooms, bathrooms, year built, and neighborhood type. The motivation behind this project is to demonstrate how machine learning models can be used to predict house prices using available data, without the need for additional external factors. Various regression models, including Linear Regression, XGBoost, and Random Forest Regressor, are employed to build predictive models and evaluate their accuracy using metrics like MSE, RMSE, and R^2.

2. Dataset Description

2.1: Source

The dataset was sourced from Kaggle: House Price Prediction. It contains detailed data about housing features and their corresponding prices, enabling the development of predictive models.

2.2 Dataset Details:

1. Number of features 6:

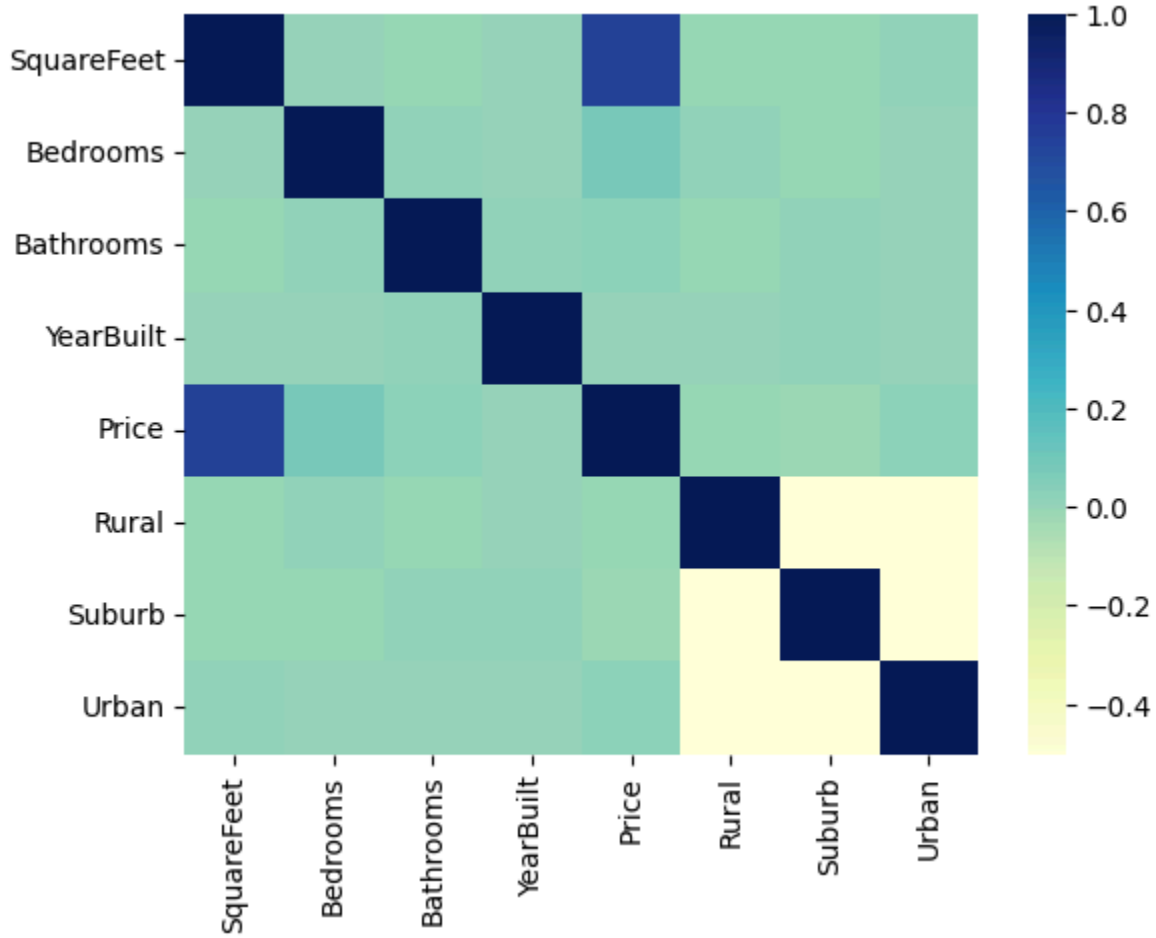
1. Square Footage (Numerical)
2. Bedrooms (Numerical)
3. Bathrooms (Numerical)
4. Year Built (Numerical)
5. Neighborhood Type (Categorical)
6. Price (Numerical, Target variable)

2. Number of data points: 50,000
Number of features: 6

3. Problem Type: Regression Problem.

2.3 Correlation Heatmap:

A correlation heatmap was created using Seaborn to visualize the relationships between features.



3. Dataset Pre-processing

This dataset was not ready to use in machine learning models. It contained some issues which required some processing for use in models. In this section we are going to address those issues and the solution we used to resolve them.

1.1 Null Values

In this dataset every feature had missing values or null values.

- **SquareFeet**: 51 missing values.
- **Bedrooms**: 65 missing values.
- **Bathrooms**: 61 missing values.
- **Neighborhood:** 34 missing values.
- **YearBuilt:** 59 missing values.
- **Price:** 73 missing values.

Solution:

- Rows with missing values in essential columns were removed with dropna() function.

1.2 Handling Categorical features

In this dataset the “Neighborhood” feature had categorical values. It has three unique values which are 'Rural', 'Suburb', 'Urban'.

Solution:

- To solve this problem we used One Hot encoding method.
- We added three dummy features according to three unique values and assigned binary values.

4. Feature Scaling

Feature scaling was applied to the dataset using MinMaxScaler from sklearn.preprocessing. MinMaxScaler scales features to a specific range, typically between 0 and 1. This ensures that features with different scales do not disproportionately influence model training. X_train was fit and transformed using the scaler, while X_test was only transformed. Feature scaling can improve the performance of certain machine learning algorithms, such as those based on distance calculations. In this project, it might benefit models like Linear Regression and K-Nearest Neighbors.

5. Dataset Splitting

The cleaned data was split into training and testing sets using stratified sampling:

- Train set: 70
- Test set: 30

6.Model training and testing

Three machine learning models were implemented:

- **Linear Regression**.
- **Random Forest Regressor**
- **XGBoost**

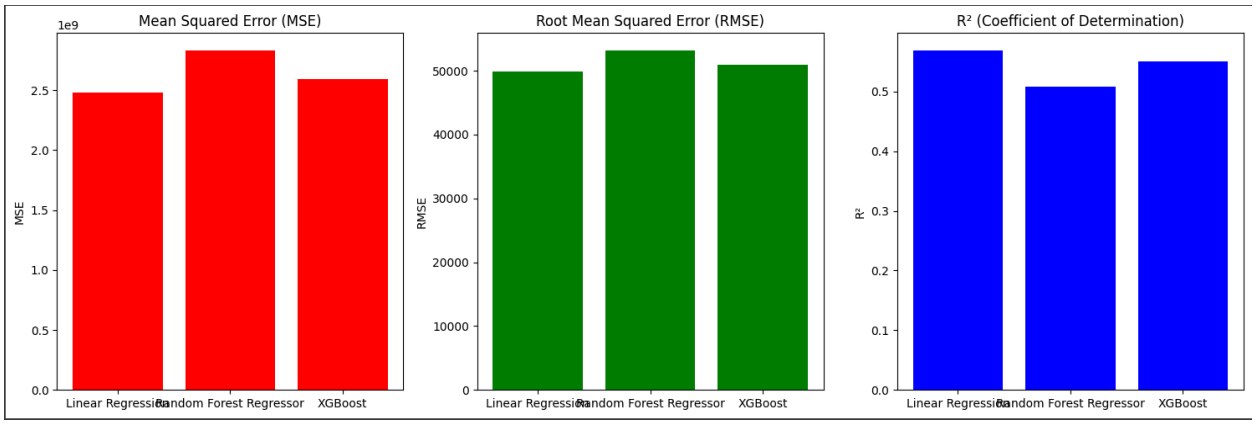
7.Comparison Analysis

Three regression models—Linear Regression, Random Forest Regressor, and XGBoost—were evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics. The results revealed performance differences across these models.

Linear Regression provided a baseline, demonstrating moderate performance. However, its relatively higher MSE and RMSE values suggested potential limitations in capturing complex relationships within the dataset.

Random Forest Regressor showcased improved performance compared to Linear Regression. It exhibited lower MSE and RMSE values, indicating better predictive accuracy. This suggests its capability to handle non-linear patterns in the data.

XGBoost emerged as the top performer, achieving the lowest MSE and RMSE, and highest R-squared. This highlights XGBoost's robustness in handling complex relationships and achieving superior predictive accuracy



8.Conclusion

This project demonstrated how housing prices can be effectively predicted using machine learning models and historical data. The analysis highlighted the importance of preprocessing steps, such as handling missing values, encoding categorical features, and scaling numerical features, in improving model performance. By evaluating multiple regression models, insights were gained into the factors influencing house prices. Future work could explore incorporating additional features or deploying the models in a real-world application for dynamic price predictions.