Housing Price Prediction Using XGBoost

1. Introduction

The goal of this project is to predict housing prices using the housing.csv dataset. The dataset includes multiple features such as average area income, house age, number of rooms, and population that may influence housing prices. The model used for this task is XGBoost, a powerful gradient boosting algorithm that is highly effective in handling structured/tabular data.

2. Dataset Overview

The dataset includes the following features:

- **Avg. Area Income**: Average income of the area.
- **Avg. Area House Age**: Average age of the houses in the area.
- **Avg. Area Number of Rooms**: Average number of rooms per house.
- **Avg. Area Number of Bedrooms**: Average number of bedrooms per house.
- **Area Population**: Population of the area.
- **Price**: Target variable representing the house price.
- **Address**: Textual column (excluded from modeling).

3. Data Preprocessing

3.1. Dropping Irrelevant Columns

- The Address column was dropped since it does not contribute to numerical modeling and is non-numeric.

3.2. Handling Missing Data

- After initial inspection, no missing values were found in the dataset, so no imputation was necessary.

4. Exploratory Data Analysis

4.1. Correlation Analysis

The correlation matrix was used to understand the relationships between the features and the target variable Price. Notable findings include:

- **Avg. Area Income** had the highest correlation with Price (0.64), making it a strong predictor.
- **Avg. Area House Age** and **Area Population** also had moderate positive correlations with Price, at 0.45 and 0.41 respectively.
- **Avg. Area Number of Bedrooms** had the weakest correlation with Price (0.17).

4.2. Outlier Detection

- Outliers were detected using the IQR method and visualized through box plots. Significant outliers were found in Avg. Area Income, Area Population, and Price.

5. Model Development

5.1. Model with Outliers

The model was initially trained without removing outliers to assess its baseline performance.

Hyperparameter Tuning

- Hyperparameters were optimized using grid search over the following parameters:
    - max_depth
    - learning_rate
    - n_estimators
    - colsample_bytree
    - alpha

Results Without Removing Outliers:

- **Best RMSE**: 104,090
- **Best R-squared**: 0.912
- **Best Hyperparameters**:
    - max_depth: 2
    - learning_rate: 0.15
    - n_estimators: 350
    - colsample_bytree: 0.3
    - alpha: 25

5.2. Model with Outliers Removed

The model was retrained after removing the outliers identified using the IQR method to observe the effect of outlier removal on performance.

Results With Outliers Removed:

- **Best RMSE**: 105,817
- **Best R-squared**: 0.905
- **Best Hyperparameters**:
    - max_depth: 2
    - learning_rate: 0.1
    - n_estimators: 350
    - colsample_bytree: 0.3
    - alpha: 25

6. Conclusion

6.1 Model Performance Without Removing Outliers

The model performed better without removing the outliers:

- **RMSE**: 104,090
- **R-squared**: 0.912 The model was able to capture the variability in the target variable (house prices) better when the outliers were included.

6.2 Model Performance With Outliers Removed

After removing outliers, the performance slightly declined:

- **RMSE**: 105,817
- **R-squared**: 0.905 This shows that removing outliers in this case led to a **1.7% increase in RMSE** and a **0.007 reduction in R-squared**, indicating that the outliers might represent important high-value homes, and removing them led to a loss of critical information.

6.3 Explanation for Results

- **Outliers carry valuable information**: In this dataset, the outliers may represent legitimate extreme cases such as luxury homes or very high incomes, which are important in predicting housing prices. Removing these data points caused the model to lose accuracy.
- **Noisy but useful data**: While outliers can sometimes represent noise, in real-world datasets, they often contain valuable signals, particularly in domains like housing, where high-value properties can significantly influence the market.

6.4 Recommendation

- Based on the results, **keeping the outliers** in the dataset yields better model performance. It is recommended to retain the outliers in similar datasets where extreme values might have significant meaning, such as luxury home markets or high-income areas.