Housing Price Prediction Dataset Documentation

Dataset Information

Title:

Hedonic prices and the demand for clean air

Authors:

Harrison, D. and Rubinfeld, D.L.

Source:

Journal of Environmental Economics and Management, 1978

Acknowledgement:

Belsley D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Objective:

The goal of this project is to predict housing prices based on various factors such as house area, number of bedrooms, furnished status, proximity to the main road, etc.

Data Overview:

Features:

House Area: The total area of the house in square feet.

Bedrooms: The number of bedrooms in the house.

Furnished: Binary variable indicating whether the house is furnished (1) or not (0).

Nearness to Main Road: Distance of the house from the main road.

Target Variable:

Sales Price: The price of the house.

Data Cleanup:

1. Loading the Data:

Loaded the dataset into a pandas DataFrame.

2. Exploring the Data:

Checked for missing values.

Examined basic statistics and distributions.

Explored data types.

3. Data Cleaning:

Handled missing values and outliers appropriately.

Encoded categorical variables (if any).

Investigated and addressed multicollinearity.

Exploratory Data Analysis (EDA):

1. Correlation Analysis:

Explored the correlation matrix to identify relationships between features.

Addressed multicollinearity issues by removing or combining highly correlated features.

2. Data Visualization:

Utilized scatter plots, histograms, and other visualizations to understand relationships between key features and the target variable.

Model Building:

Single Feature Model:

Feature Selection:

Selected a single feature based on correlation analysis and EDA.

Build Simple Linear Regression Model:

Split the data into training and testing sets.

Trained a simple linear regression model using the selected feature.

Evaluated the model using metrics like R-squared and RMSE.

Multiple Feature Model:

Feature Selection:

Considered features that contribute meaningfully to the target variable.

Build Multiple Regression Model:

Split the data into training and testing sets.

Trained a multiple regression model.

Evaluated the model using metrics like R-squared and RMSE.

Model Evaluation:

Metrics:

Evaluated both models using metrics like R-squared and RMSE.

Compared the performance of the single-feature and multiple-feature models.

Conclusion and Recommendations:

Findings:

Summarized the performance of both models.

Highlighted key factors influencing housing prices.

Recommendations:

Provided insights into potential improvements or additional features for better predictions.