Predicting Median House Values Based on Demographic and Housing Characteristics

Abdulhamid Mathkur - Carnegie Mellon University

Introduction

The objective of this analysis is to predict the median house value across different locales using a dataset that encapsulates various house-related attributes. Our model aims to identify key predictors such as population density, location (latitude and longitude), age of structures, the number of bedrooms, and more, ultimately seeking a sound estimate of median house values.

Data and EDA

Dataset Overview

• Number of Records: 10,605 • Number of Predictors: 13

Predictors:

name	Description
POPULATION	population of the tract in question
LATITUDE	latitude of the tract
LONGITUDE	longitude of the tract
Total_units	the total number of housing units
Vacant_units	the total number of vacant units
Median_rooms	median number of rooms per unit
Mean_household_size_owners	average number of people in owned homes
Mean_household_size_renters	average number of people in rented homes
Owners	the percentage of units that are owned
Median_household_income	self-explanatory
Mean_household_income	also self-explanatory
Built_1990_or_later	the percentage of units built after 1989
Bedrooms_4_or_more	the percentage of units with more than three bedrooms

The response variable: Median_value

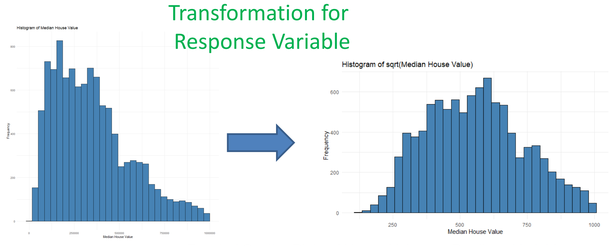
Exploratory Data Analysis (EDA)

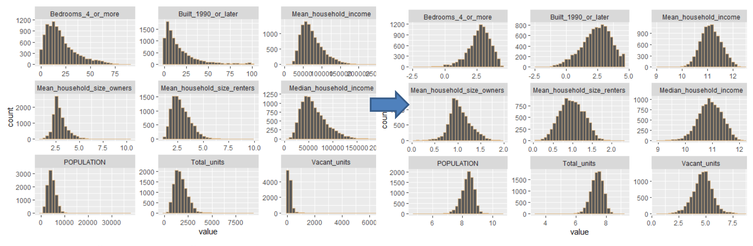
The dataset contains no missing values.

We apply a log-transformation to those predictor variables whose distributions exhibit a pronounced positive skew. Additionally, we apply a square-root transformation to the response variable.

We overview relationships between predictors and response variable, and then drop the bullet point about correlations

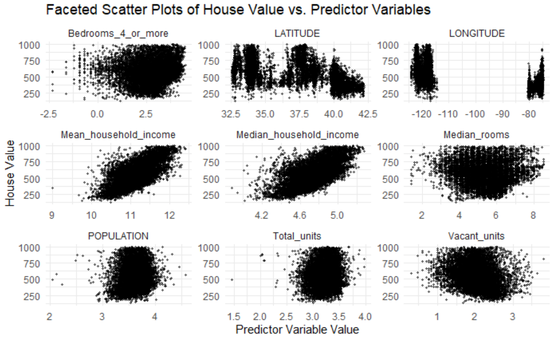
Transformation for Response Variable





Before transformation

After transformation



Analysis

Data Splitting:
The dataset was split into training (80%) and test (20%) sets.

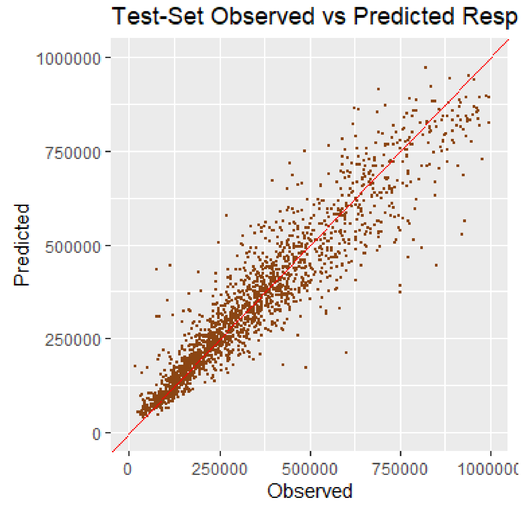
Method:
We learn the following regression models: multiple linear regression, MLR with subset selection, Random Forest, Gradient Boosting.

Best Subset Selection(BSS): We used BIC model for BSS. For the BiC model, 10 variables are retained.
(LATITUDE, LONGITUDE, vacant_units, Median_rooms, Mean_household_size_renters, Owners, Median_household_income, Mean_household_income, Built_1990_or_later, Bedrooms_4_or_more)

We adopt the model with the lowest test-set mean-squared error (MSE).

Model Performance:

• R-squared value to assess goodness of fit:
The R- squared was 0.797 that suggests that the linear model is effective and is useful to look into potential modifications to improve the model further.



Model	MSE
Linear Regression	6608.5
BSS	6627.7
Random Forest	3771.8
Gradient Boosting	3531.7

Conclusion

The analysis confirms that median house values can be effectively predicted using demographic and housing characteristics.

The model performances were compared in terms of Mean Squared Error (MSE), demonstrating its capacity to capture complex relationships within the data and choose best model. Gradient Boosting is the best model among the three based on the provided MSE values. Hence, Gradient Boosting is the best model among the three based on the provided MSE values. Additionally, the study highlights key predictors influencing house prices, providing stakeholders with valuable insights into the housing market dynamics. Future work could explore further refinements in model accuracy and the incorporation of additional data sources to enhance predictive capabilities.

References

[1] Zhang, J. Q., Du, J., “House Price Prediction Model Based on XGBoost and Multiple Machine Learning Methods,” Modern Information Technology, vol. 4, no. 10, pp. 15-18, May 2020.

[2] House Price Prediction using Machine Learning in Python. Available: https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/

[3] Sobana, P., Balakumaran, M., Bharathkumar, S., Boopathi, P., “House Price Prediction using Machine Learning,” Challenges in Information, Communication and Computing Technology, pp. 704-708, November 2024.