Predicting Median House Values: Exploring Housing and Demographic Factors through Machine Learning

Ramya Ashish, Victoria Discua Santos, Annabel Frake, and Kathy Yu

36-600 – Overview of Statistical Learning and Modeling, Fall 2024, Carnegie Mellon University

Overview

Background: Housing market plays a critical role in shaping communities, making it essential to understand the factors that influence house values. Recognizing the importance of these factors, this project compared the predictive capabilities of six models to identify key drivers of house value.

Dataset Description and Methods: This project aimed to predict the median house value for specific locales using a dataset that includes diverse housing and demographic features. Features include variables such as longitude, latitude population size, household characteristics, and property features (e.g., number of bedrooms). This project evaluated the predictive performance of six models, using mean squared error (MSE) as the metric for comparison. The model with the lowest MSE was selected as the final, most accurate predictor.

Data

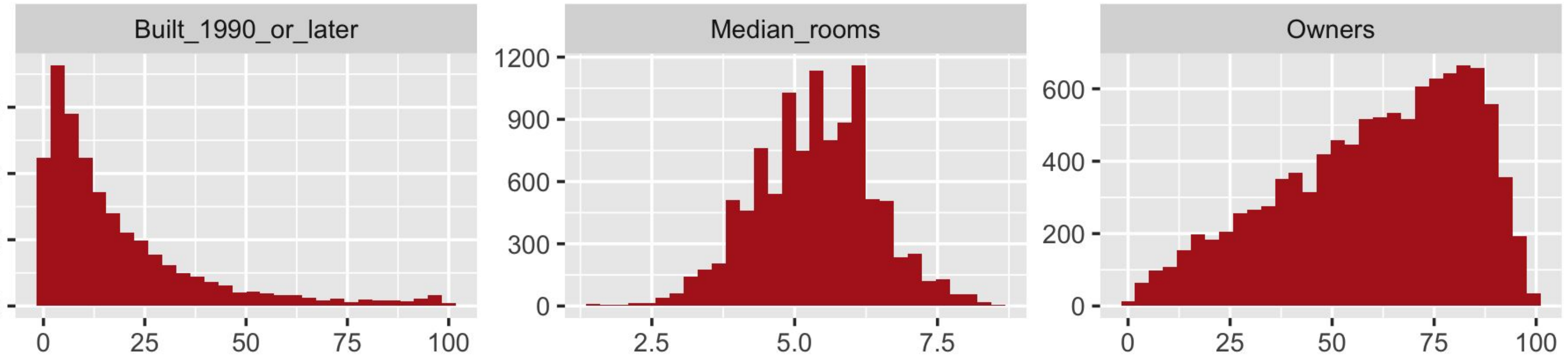

Figure 1. Example Distributions. From left to right are examples of features that roughly follow right skewed, normal, and left skewed distributions before processing.

- The dataset consisted of 13 predictor variables and 1 response variable (median house value) for a total of 14 variables. Of the 10,605 data entries, none contained missing values.
- A brief EDA analysis was conducted to visualize the distribution of each variable (representative examples shown in Figure 1):
  - ➢ **Normal Distribution:** `Median_rooms, POPULATION, Total_units`
  - ➢ **Skewed Right:** `Bedroom_4_or_more, Built_1990_or_later, Mean_household_size_owners, Mean_household_size_renters, Mean_household_income, Median_household_income, Vacant_units`
  - ➢ **Skewed Left:** `Owners`
  - ➢ **Other**: `LATITUDE, LONGITUDE`
- The following variables were transformed using a square root function: `Median_Value, Bedrooms_4_or_more, Built_1990_or_later, Total_units, Vacant_units, POPULATION.`
- The following variables were transformed using a log function: `Mean_household_income, Mean_household_size_owners, Mean_household_size_renters, Median_household_income.`
- After variable transformations, 119 outliers were removed by eye.

Analysis

Table 1. MSE Comparison

Model	Test-set MSE
XGBoost	3754.521
Random Forest	3920.685
KNN	6312.634
Linear Regression	6536.618
BSS	6536.618
Regression Tree	9180.338

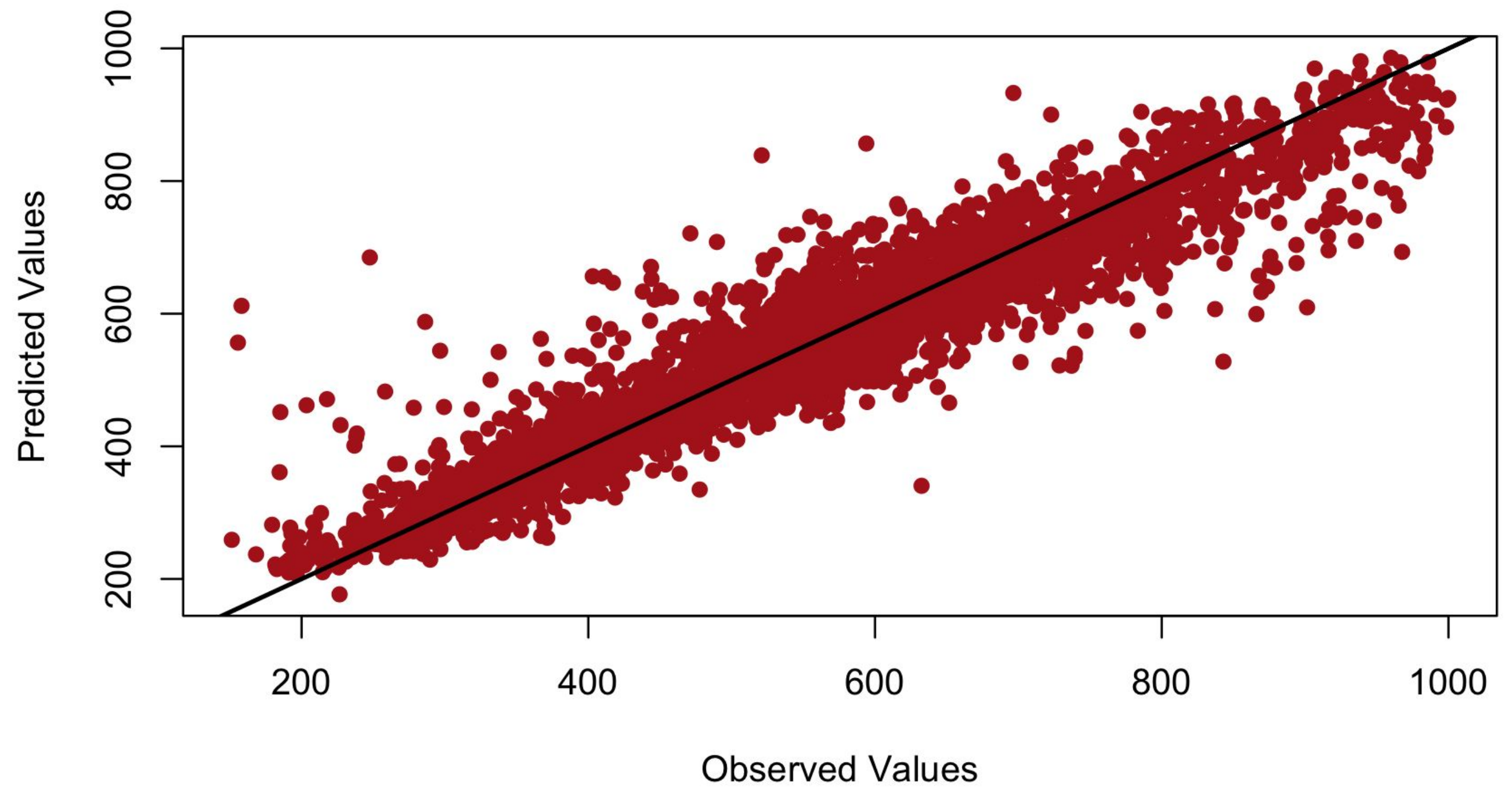

Figure 2. Final Model Observed vs Predicted response

- The data was then split into into train (70%) and test (30%) sets through random sampling.
- Linear regression, BSS, KNN, Regression Tree, RF, and XGBoost models were compared.
- Table 1 shows a comparison of the test-MSE values for each of the models. XGBoost had the lowest test-MSE and was therefore chosen as the final model.
- Figure 2 shows the observed square root median house value versus the predicted square root median house value for the XGBoost model. Since the data points generally follow the unit line, this indicates that the final model is a relatively good at predicting the response variable. However, there is still some variability that signals room for improvement.
- Feature importance was run using the RF model, and it was determined that longitude is the most important predictor for median house value, followed closely by mean household income. Given that housing prices can generally depend upon location and socioeconomic status, these results match expectations.

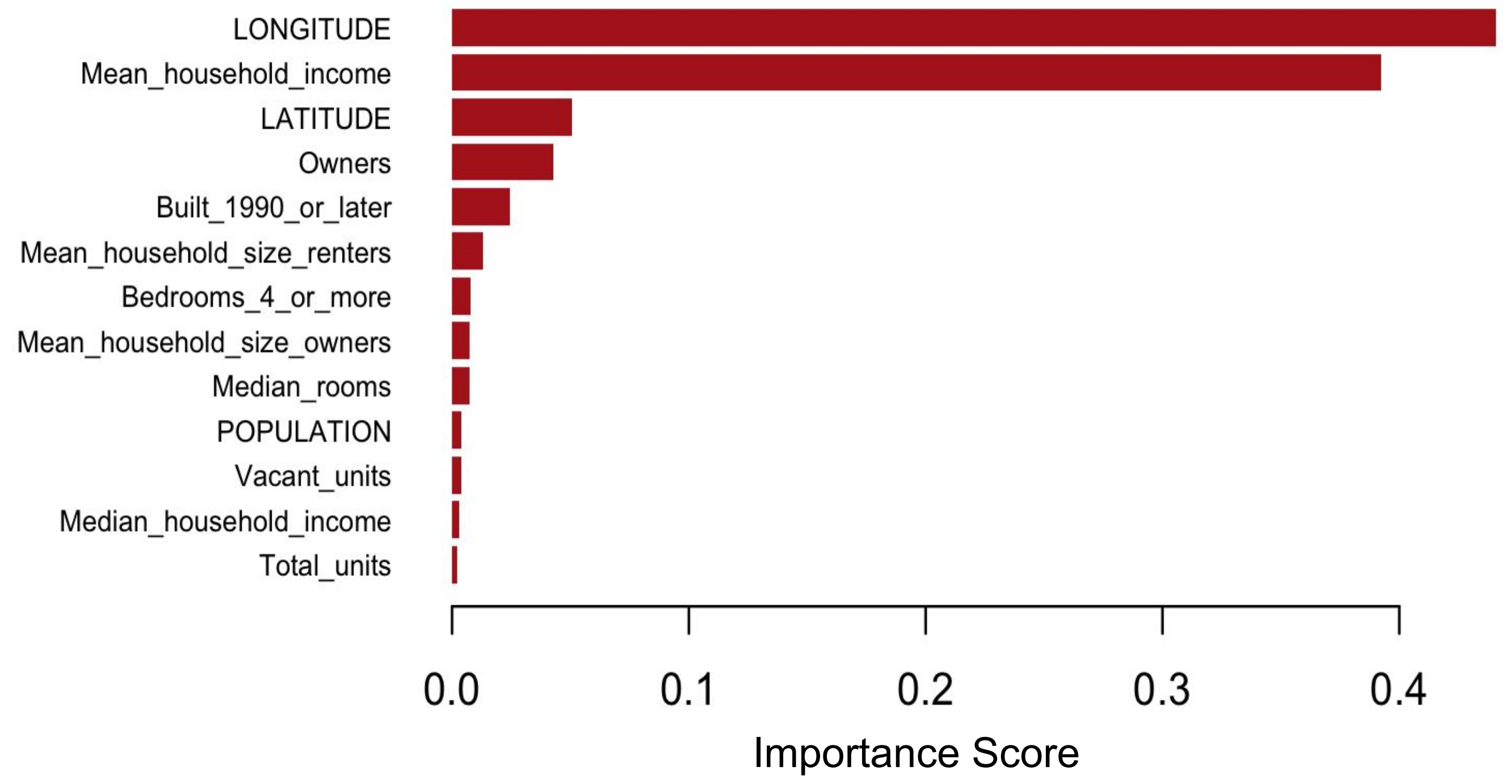

Figure 3. Variable Importance Plot

Conclusions

The XGBoost model was the best at predicting median house value out of the models tested in this project. XGBoost's iterative nature allows it to refine predictions at each step, making it effective at capturing complex, nonlinear relationships. While the predictive capabilities of the final model are good, further improvements could potentially be achieved using alternative models not tested here or different preprocessing (e.g., other transformations). In terms of feature importance, longitude and log mean household income were the most important predictors for median house value.