Utilizing Demographics and Home Features to Predict Median House Value

By: Gayatri Chabra, Joel Beltran, Karthikeya Manchala, Rohit Varanasy, Richard Wang

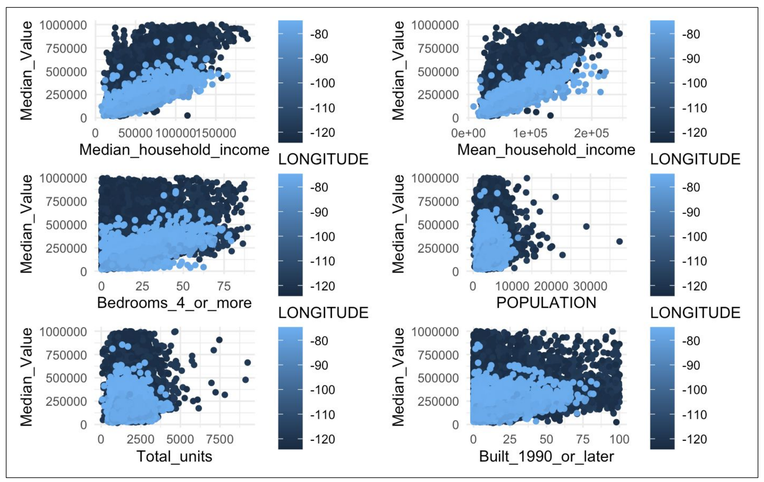
Background & Introduction

The goal of this project is to predict the median house value using various predictors such as population, location, housing characteristics, and socioeconomic factors. Analysis on the most important factors has the potential to provide valuable insights for a variety of stakeholders including both buyers and sellers, leading to more informed decision-making in the real estate sector. The benefits include enhancing decision-making processes, promoting sustainable urban development, and contributing to a more efficient and equitable housing market.

Exploratory Data Analysis

Exploratory Data Analysis included dealing with outliers, looking at correlations between the features, and seeing if any of the features obviously impacted the predictor variable.

Response vs Predictor Variables



It seems that `median & mean household income` has a positive correlation with `median house value`.

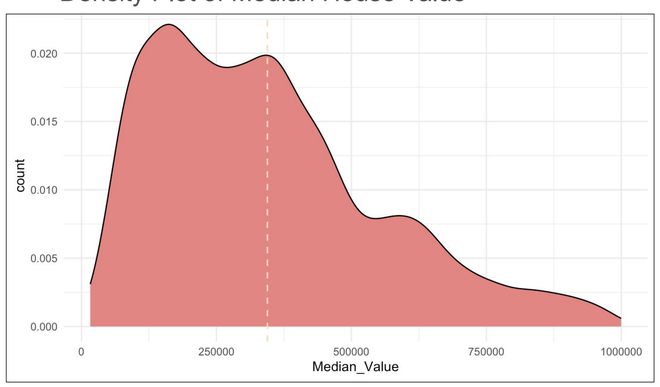
It also appears that -120 `longitude` has a higher `median house value` over all than -80 `longitude`.

We hypothesized that `median income` and `longitude` will have high feature importances in our models

According to our density plot, our response variable, the `median house value`, is right skewed.

It gives us a basic understanding of the distribution of `median house value`, and how the majority of houses have values around 100k, lower than the average of 300k

Density Plot of Median House Value



Methods

We split the data, which has 10,605 rows and 13 features, into a training and test set using a train-test split of 70/30. As this is a regression problem, we decided to used the following ML models:

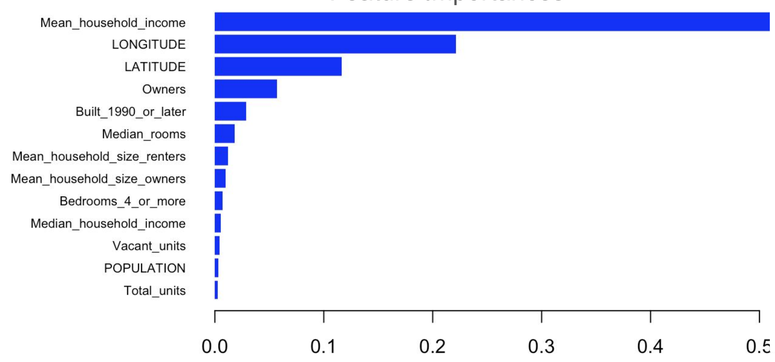
- **Linear Regression, Logistic Regression, Decision Trees, Random Forest, XGBoost, and K-Nearest-Neighbors**

Analysis

Model	MAE
Linear Regression	72,573
Decision Trees	133,863
Random Forest	52,671
XGBoost	**49,743**
K-Nearest Neighbors	182,787

It appears that XGBoost was the best performing mode, since it produced the lowest Mean Absolute Error, closely followed by Random Forest. In general, it makes sense that tree-based ensemble methods perform better given the bimodal distribution of some of the variables. We decided to use **mean absolute error** as our metric because our response variables have large values.

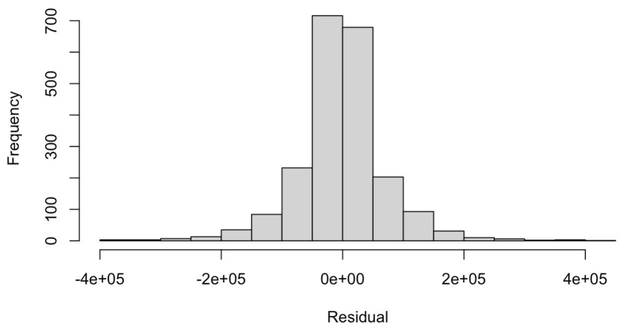
Feature Importances



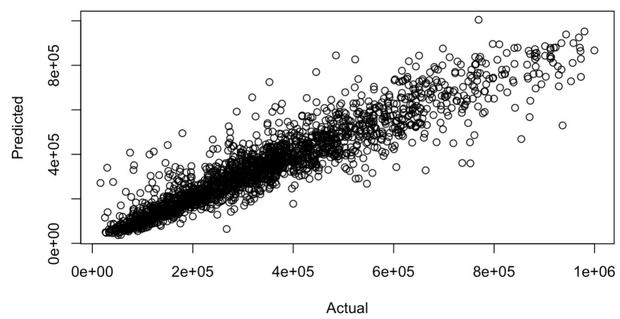
The bar plot shows the importance of each variable XGBoost was trained on. `Mean Household Income` was the most important variable, which makes sense considering income has a strong correlation with the cost of the house.

The diagnostic plot shows that out XGBoost model predicted values well, having a slope of about 1. However, the average errors ranges in about 50k.

Residual Plot



Diagnostic Plot



Conclusion

Mean household income is clearly important in predicting the median household value. This connection is pivotal for financial planning and real estate investment strategies. Recognizing the influence of mean household income on median household value allows individuals, businesses, and investors to make informed decisions about property acquisitions, developments, and market trends. Moreover, it provides a comprehensive understanding of the economic dynamics within a region, enabling stakeholders to anticipate market fluctuations, identify lucrative opportunities, and make strategic investments that align with the financial well-being of both homeowners and potential buyers.