Home price prediction

Rohan Mani
05.10.2024

Problem Statement Worksheet (Hypothesis Formation)

Predict house prices based on house attributes/price and market factors.

H

1 Context

Homebuyers and realtors need accurate information on the factors influencing home prices.

2 Criteria for success

Measure model's performance using MAE, R^2, MSE. Monitor model's performance and ability to adapt to market conditions.

3 Scope of solution space

Using house prices and attributes from 2012 to 2021.

4 Constraints within solution space

Limited to house price data and market factors available in the datasets.

5 Stakeholders to provide key insight

Homebuyers
Realtors

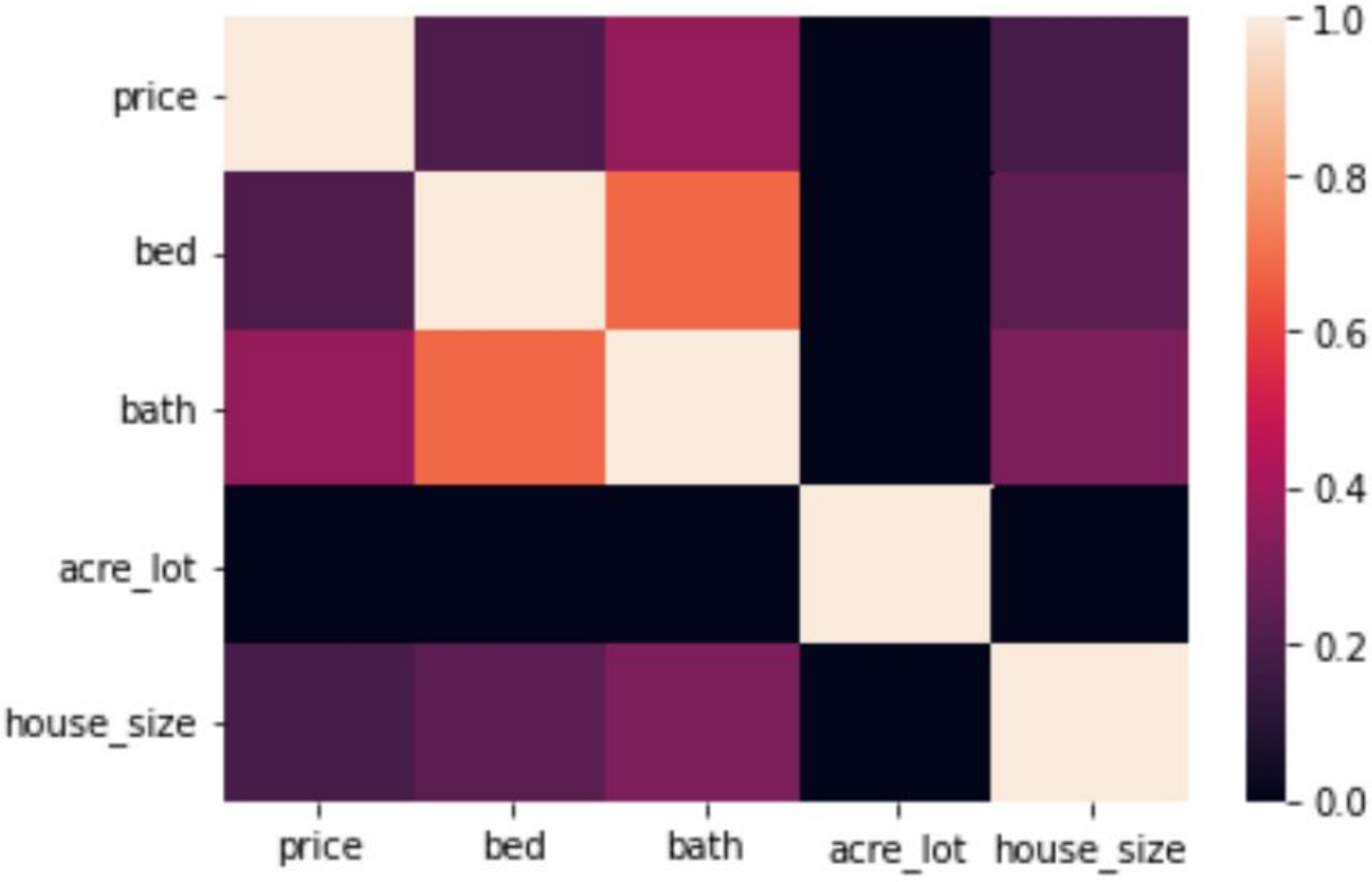
6 Key data sources

USA house price dataset derived from realtor.com-https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data

Data wrangling

- House price data from Kaggle
  - The original data file had 2,226,382 house prices with 12 attributes.
- House size had 25.5% missing values, previous sold date had 32.9% missing values, bath had 22.9% missing values, bed has 21.6% missing values.
- All rows with missing values were dropped resulting in 1,084,909 rows.

Exploratory Data Analysis (Correlation heatmap)



Exploratory Data Analysis Findings

- There is high correlation between variables bed and bath.
- There is medium correlation between price and bath.
- There is medium correlation between house size and bath.
- There is no negative correlation.

Feature engineering

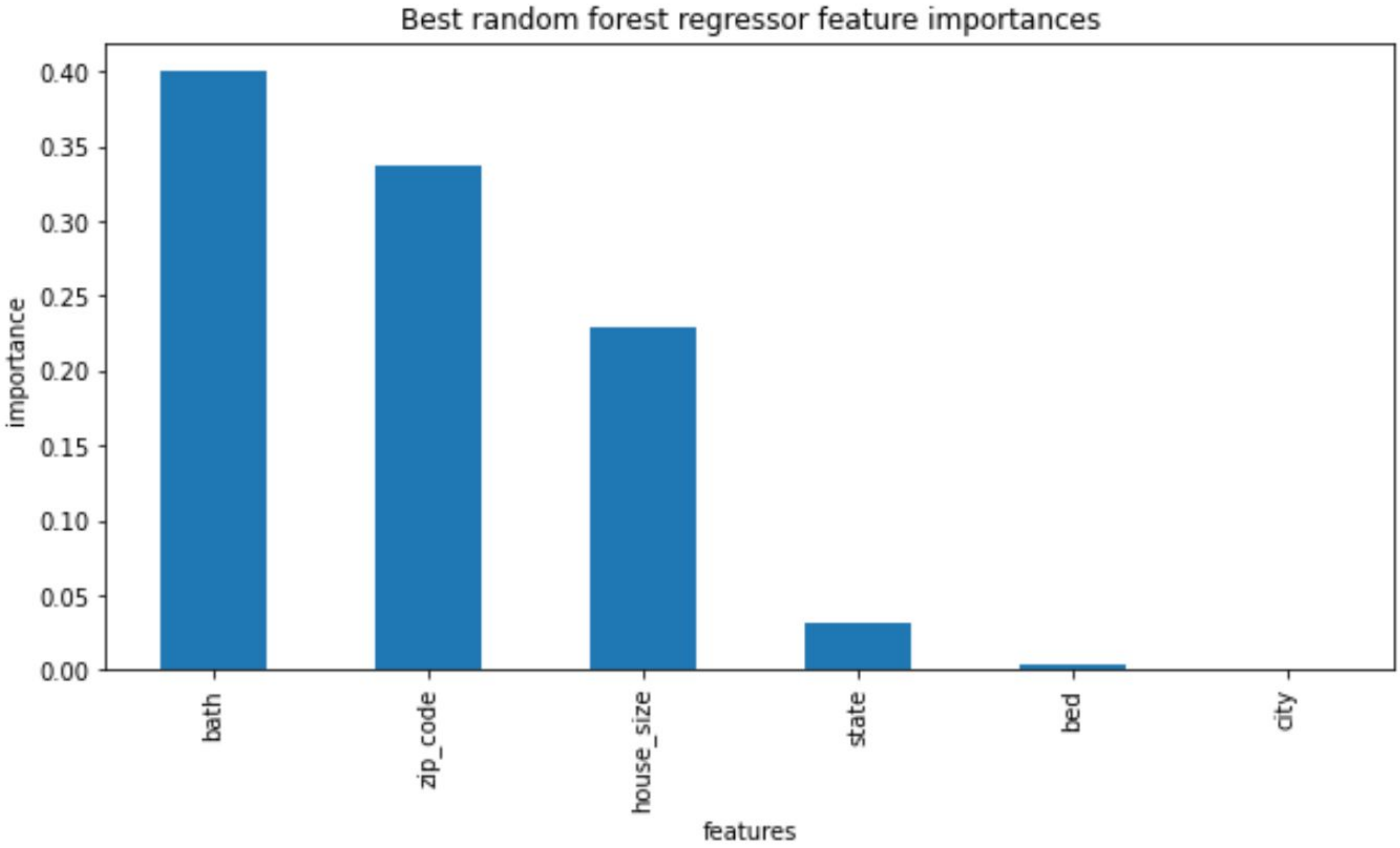
- Removed price outliers based on upper and lower thresholds of the house price. Excluded 5% of the houses based on a lower threshold of $50k and .02% of the houses based on a upper threshold of $5M.
- The dataset consisted of ‘sold’, ‘for sale’, and ‘ready to build’ houses. The status column identified each category.
- Dropped the status column. This column is no longer needed.
- Encoded categorical columns such as city, state.

Model Evaluation Metrics Comparison

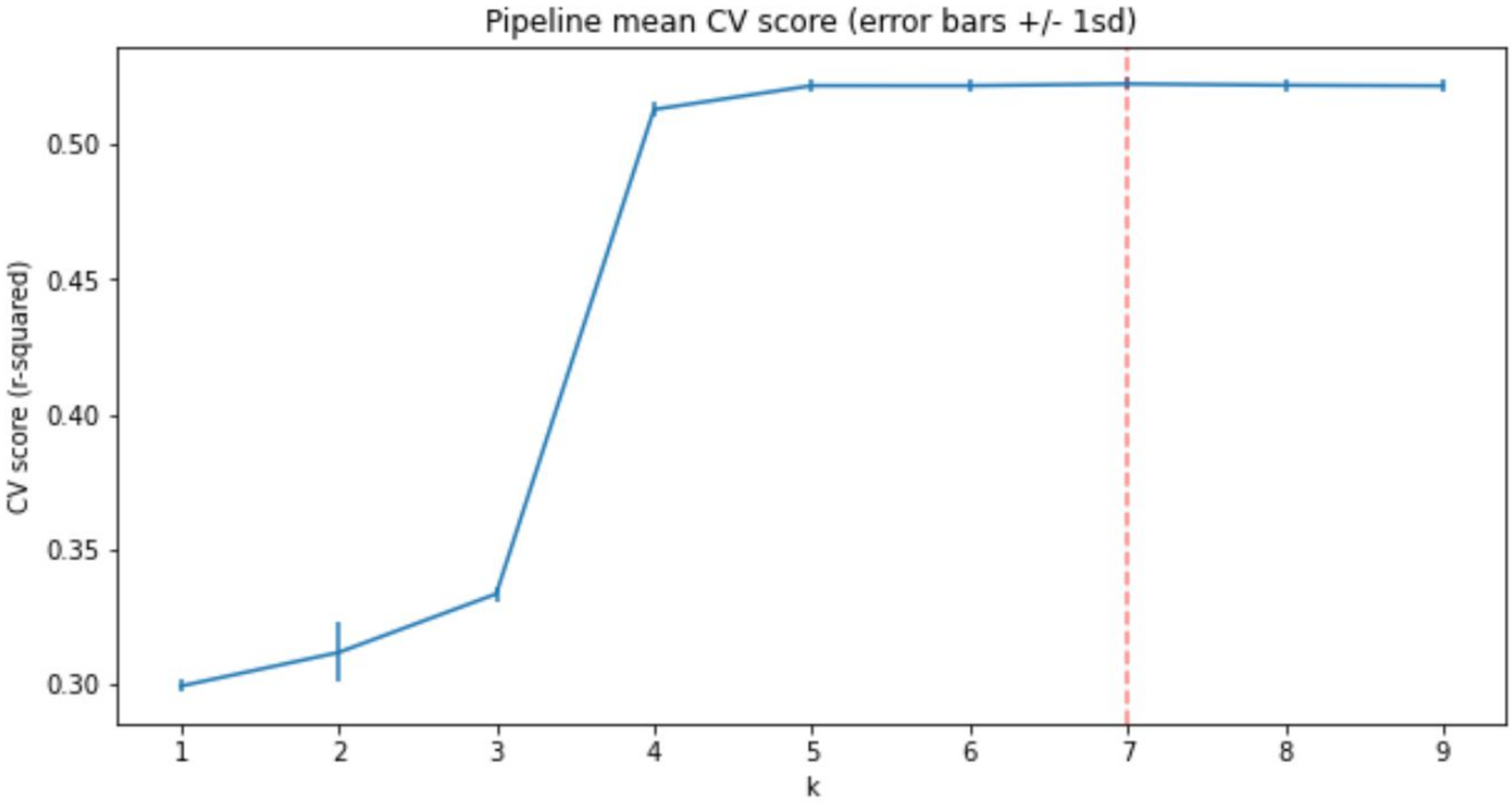
- The home price prediction problem is a regression problem so multiple regression models varying from simple to complex were used for training.
- R^2, MSE, and MAE were the metrics chosen for evaluating the regression models.

Model	R^2	RMSE	MAE
Linear Regression	0.30	428988	250076
Random Forest Regression	0.52	355520	200812
Gradient Boosting Regression	0.56	272683	143216
XGBoost Regression	0.77	212837	113642

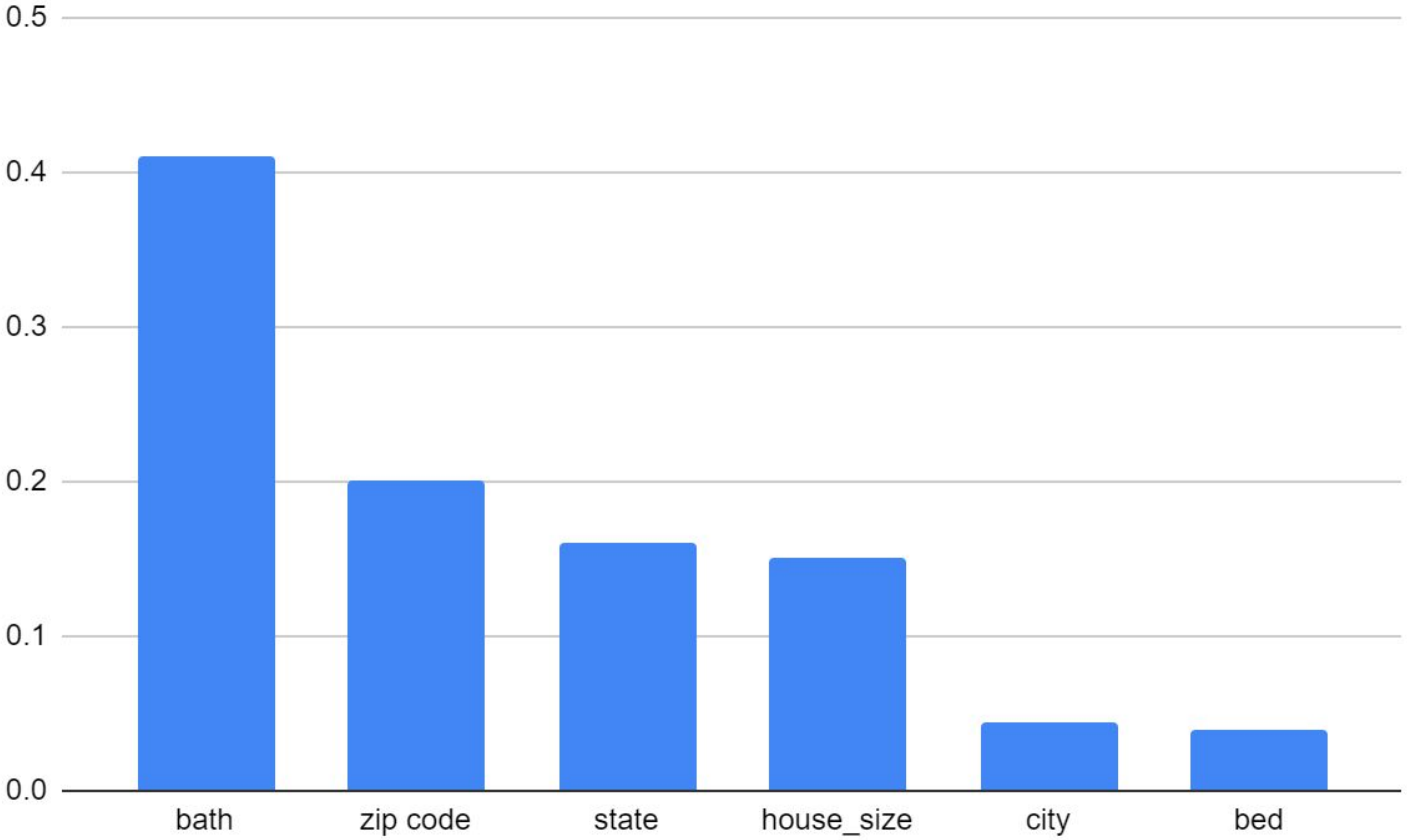
Random Forest Feature Importance



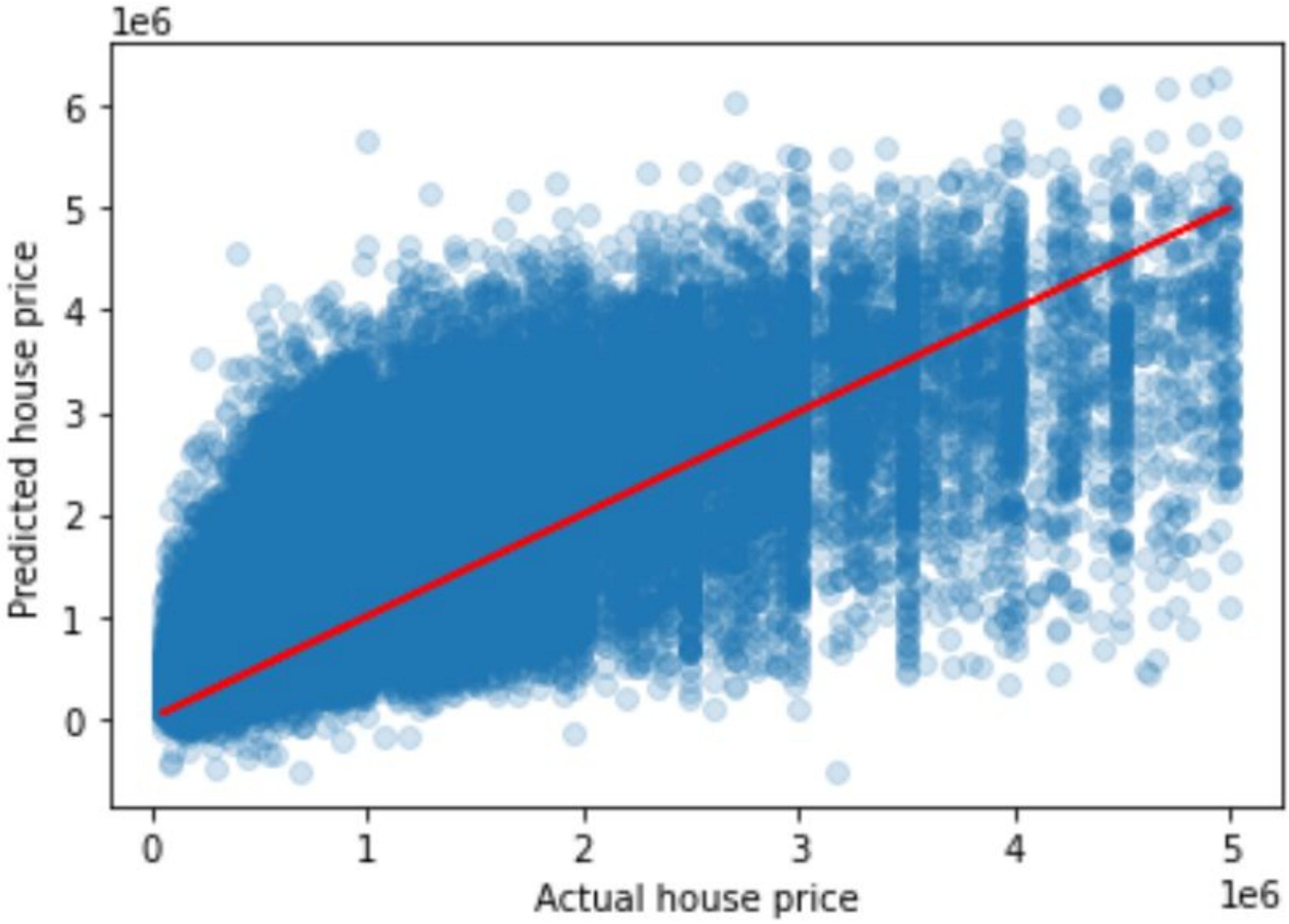
Explained Variance for cumulative features in Random Forest 5-fold CV



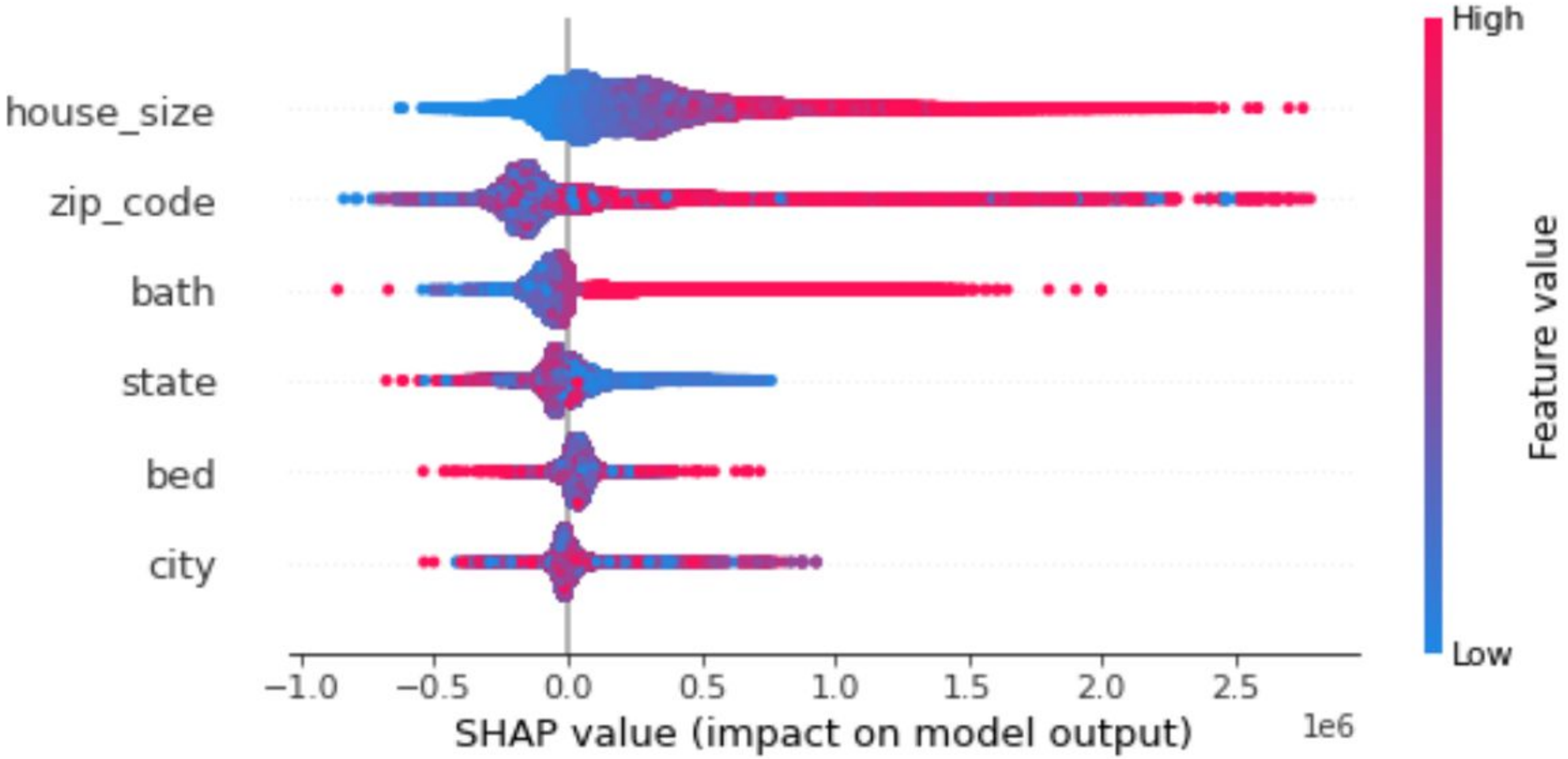
Feature importance for XGBoost model



XGBoost House Price Prediction on Test Dataset



SHAP summary plot to interpret model results



How stakeholders will use this model?

- Useful for homebuyers, homesellers, and realtors to use as an additional guide to predicting house prices.

- User should provide 6 attributes needed to predict house price.
    - House_size,bed,bath,city,zip_code, and state.

- Run the model.predict() function, passing the data into the function to output the price predicted by the model for the home provided.