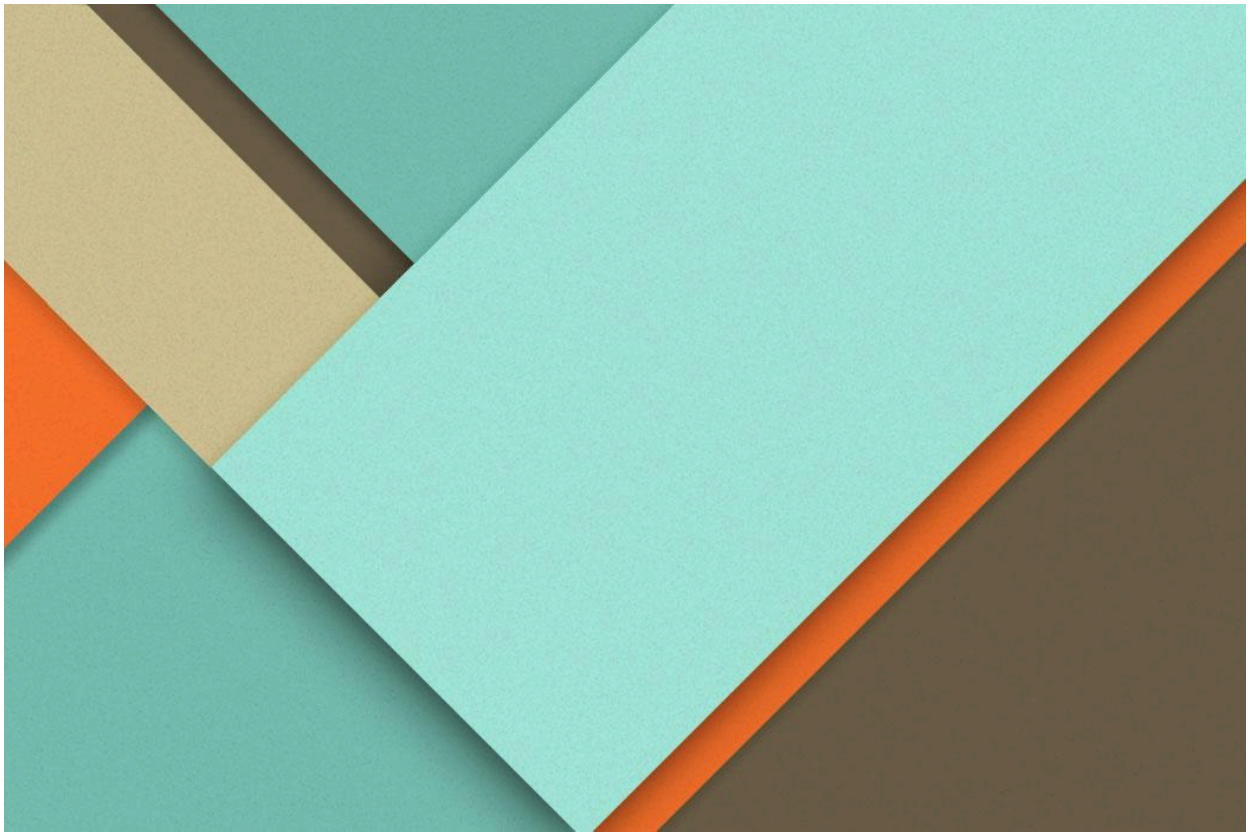


Capstone 3 - House Price Prediction

05.11.2024

Rohan Mani

Why?

Home price prediction is important for several reasons that impact various stakeholders in the real estate market.

For Home Buyers:

- Budgeting and Financial Planning: Accurate home price predictions help buyers plan their finances and budget effectively. Knowing the expected cost helps them determine the size and location of the home they can afford.
- Investment Decisions: Buyers looking to invest in real estate can use price predictions to gauge the potential appreciation of a property, helping them make informed investment decisions.

For Home Sellers:

- Setting the Right Price: Sellers can use price predictions to set competitive and realistic listing prices. This can lead to quicker sales and ensure they receive a fair market value for their property.
- Timing the Sale: Understanding market trends can help sellers decide the best time to sell their property, maximizing their return.

For Real Estate Agents and Brokers:

- Advising Clients: Agents can provide better advice to both buyers and sellers based on accurate price predictions. This enhances their reputation and effectiveness.
- Market Analysis: Accurate predictions allow agents to analyze market trends and identify opportunities or risks in the market.

Data source

The source data is a Kaggle dataset with house prices and other important features like the number of bedrooms, bathrooms, zip code, city and the house size.

Data

The USA Real Estate Dataset was derived from realtor.com data https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data.

- The file had realtor listings with the broker,city, state, and area codes for each home in the dataset.
- No duplicate data was contained in the file.

Data wrangling

House price data (~2M) homes was downloaded from the Kaggle website. The original data file had 2,226,382 house prices with 12 attributes. Rows with missing values were dropped from the dataset.

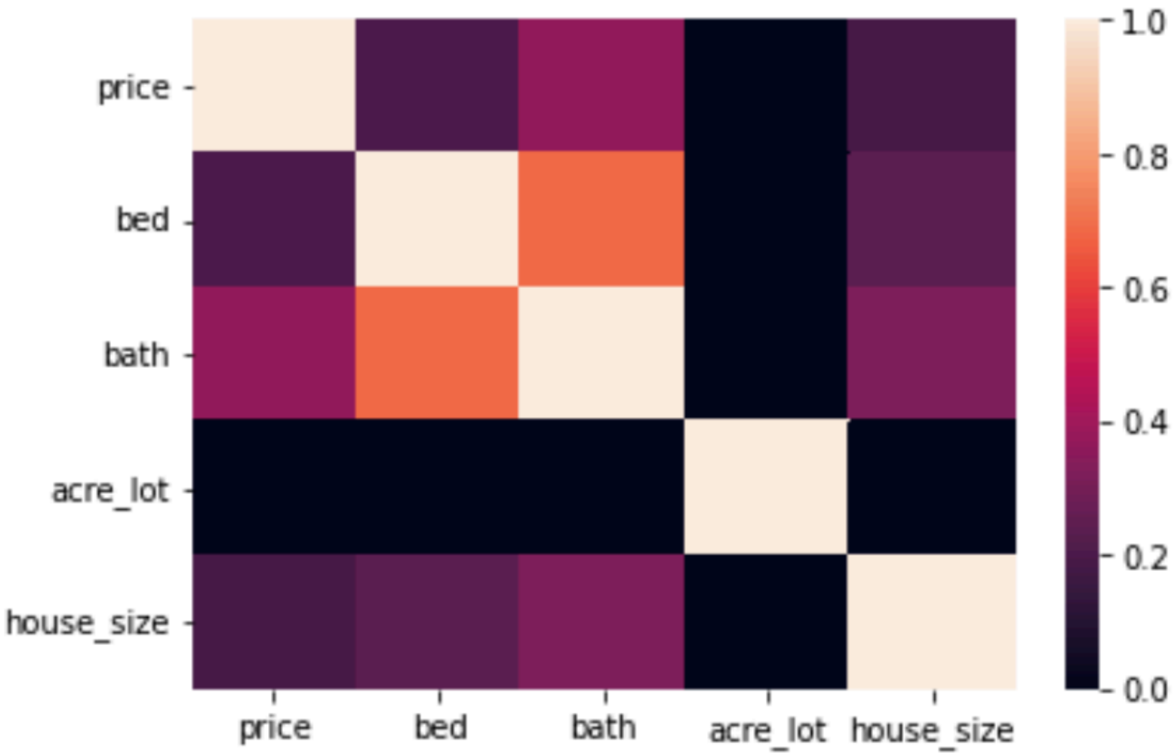
Columns with missing values

- House size had 25.5% missing values, bath had 22.9% missing values, and bed had 21.6% missing values.
- All rows with missing values were dropped. After rows with missing values were dropped, the resulting dataset had 1,084,909 rows.

Exploratory Data Analysis

Here are the key findings from EDA:

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

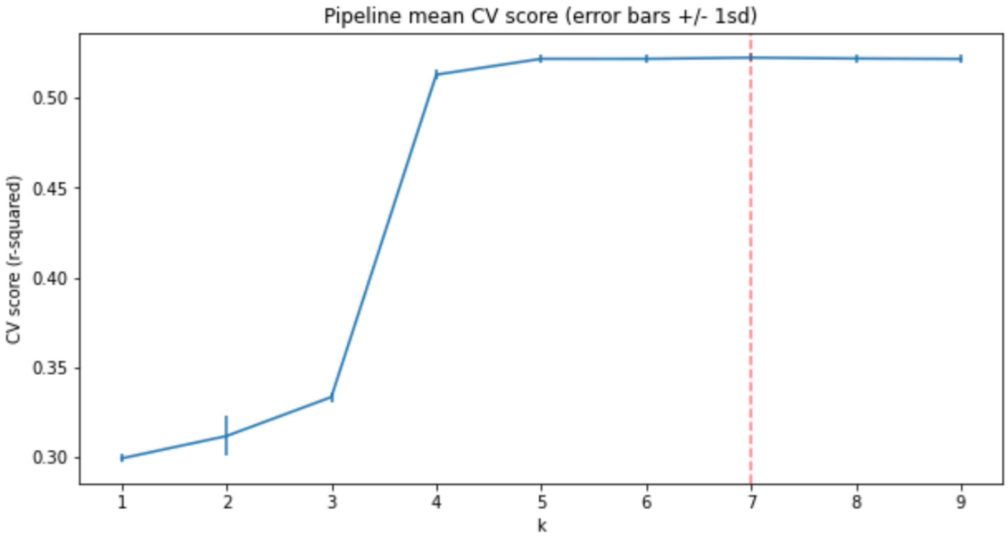


Preprocessing and training development

Here are the preprocessing steps that were performed:

- Removed outliers based on the upper threshold of price which excluded 1.2% of the houses.
- City and state columns were encoded using label encoder.

As shown in the chart below, performed 5 fold cross validation to select the ‘k’ best features out of the total 9 features available in the dataset.



As part of the 5 fold cross validation, the data is split into training data and test data. The test data is saved for model evaluation. The training data is split into 5 subsets and each subset is one fold. The model is trained on 4 folds of the data and tested on the one fold leftover. This is one iteration. Since there are 5 folds there are 5 iterations in total. The variance explained in the house prices from the leftover fold is calculated for each iteration and averaged over all 5 iterations. The graph below shows the average variance explained for each combination of features selected with the Random Forest model. Based on this graph, selecting 7 features has the same variance explained as selecting 8 or 9 features. So we can forgo 2 features.

Algorithms & Machine Learning/Modeling

The following models were implemented:

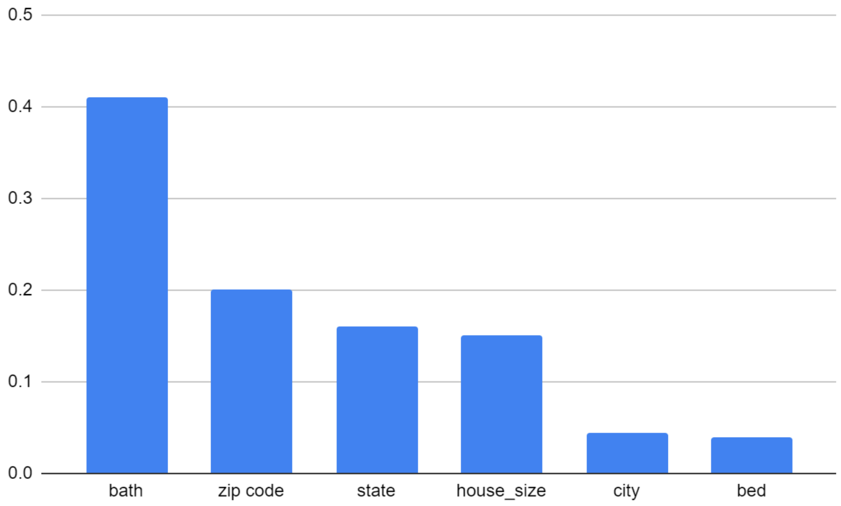
1. Linear Regression
2. Random Forest Regression
3. Gradient Boosting Regression
4. XGBoost Regression

The table below shows the model performance metrics.

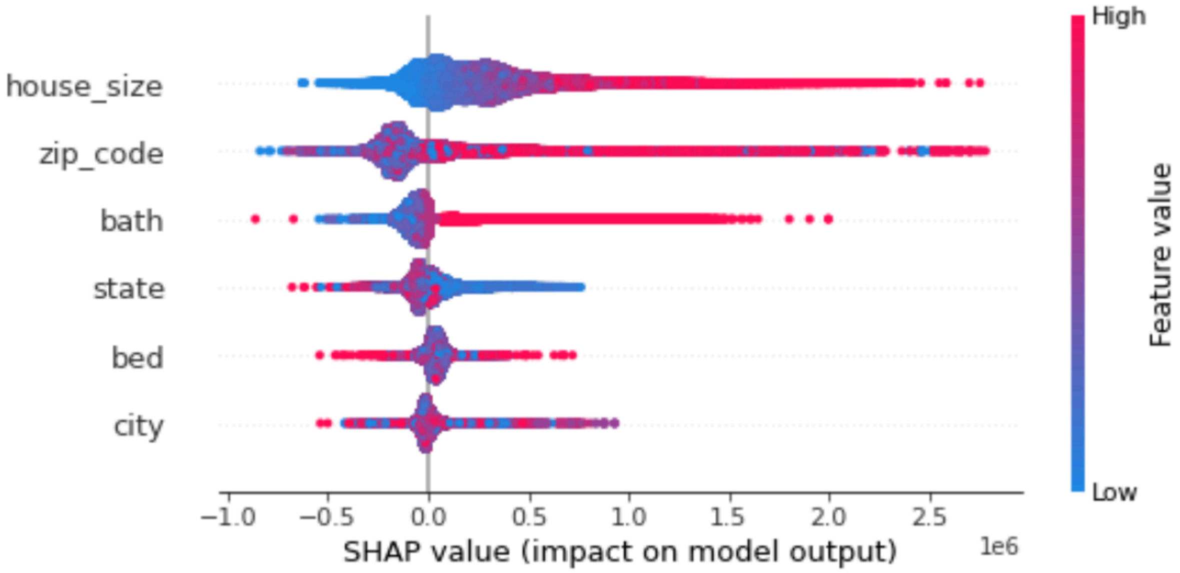
Model	R^2	RMSE	MAE
Linear Regression	0.30	428988	250076
Random Forest Regression	0.52	355520	200812
Gradient Boosting Regression	0.56	272683	143216
XGBoost Regression	0.77	212837	113642

Model Predictions

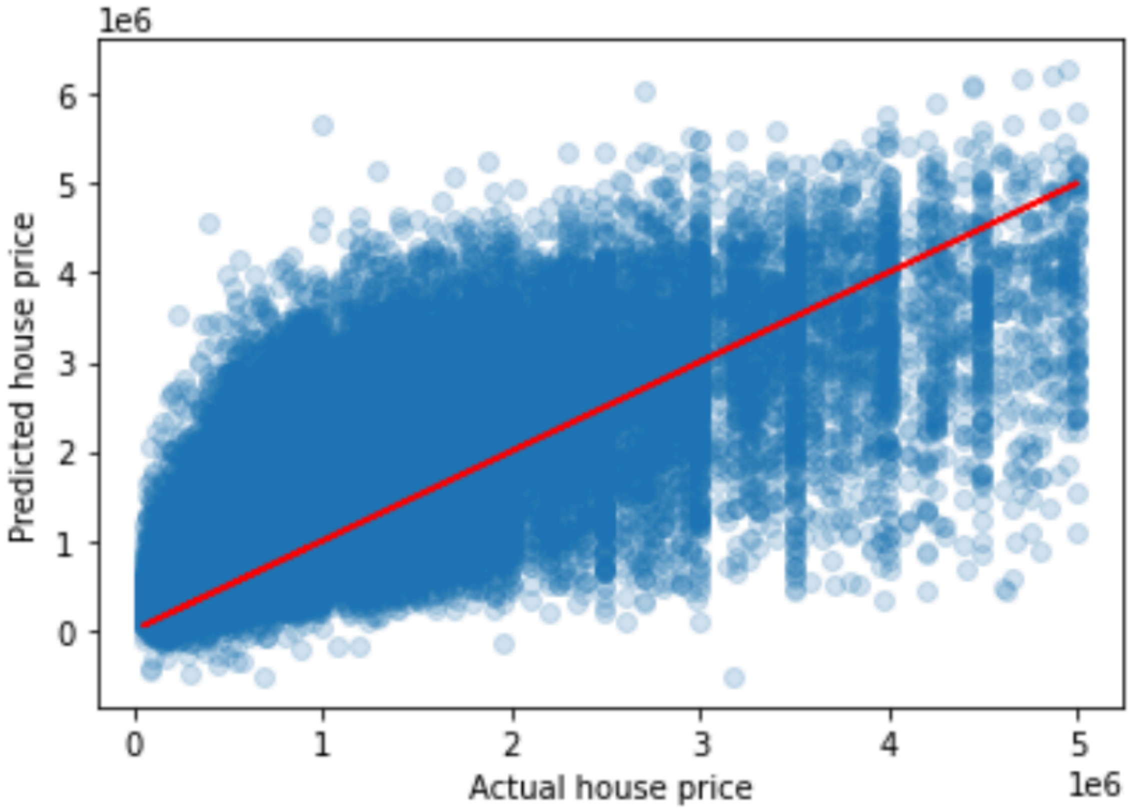
Some of the interesting and significant features and their importance (coefficient/weight) as predicted by the model is given below.



SHAP values provide a powerful tool for understanding model behavior and identifying important features for making predictions. The top 6 features in the SHAP plot were house size, zip code, number of bathrooms, state, number of beds, and the city. When the model uses the median value for each feature in the dataset, the SHAP value is 0. This means that the features have no impact on the baseline prediction. The positive SHAP values on the right indicate that the feature increased the predicted house price. The negative SHAP values on the left indicate that the feature reduced the predicted house price.



The below graph shows the actual versus predicted house prices. The red line represents points where there is no difference between the actual and predicted values.



How a stakeholder will use the model?

This model is useful for homebuyers, homesellers, and realtors to use as an additional guide to predicting house prices.. By providing the 6 attributes needed to predict the house price (# of beds, # of baths, etc.), running the model.predict() function will generate the predicted home price.