Home Price Prediction Model

By
Garrett J.A. Hass

Data Science Final Capstone 9/29/2022

Springboard

Problem Statement

- Need to predict the price of a home to assist clients
- Competitive advantage over other companies

What traits have
a high correlation with the price?

Can we predict the price accurately?

KELLERWILLIAMS ADVANTAGE REALTY LLC

Stakeholders

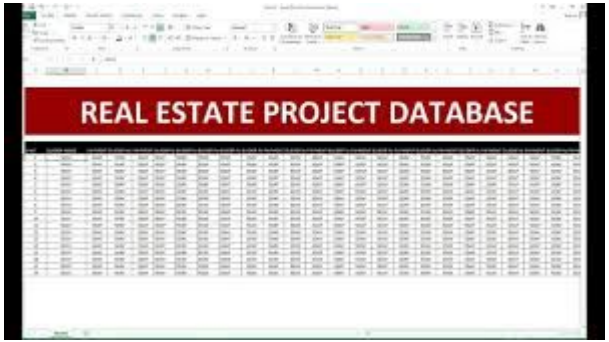
- Realtors
- Buyers
- Sellers
- Potential Clients of Real Estate Companies
- NAR
- Zillow
- Redfin
- Investors
- Housing Market Companies
- HOAs
- Mortgage Companies
- Banks
- Real Estate Companies
- Appraisers

Solution to our Problem

- Create a model to predict the price of a home
- Finding variables that are correlated with price
- Identifying unique outliers
- Using four types of Models:
  - Linear Regression
  - Ridge Regression
  - Random Forest
  - XGB Regression

Dataset

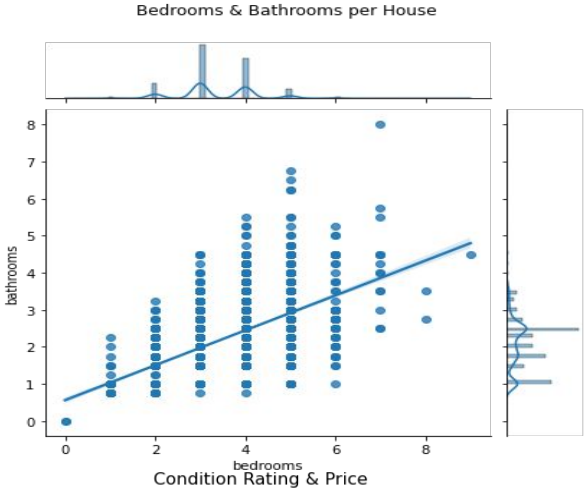
- Kaggle compilation of Real Estate transactions
- https://www.kaggle.com/datasets/shree1992/housedata?select=data.csv
- Seattle Metro Area
- Data is from May 1st, 2014 to July 9th
- Is there enough data to predict the price?

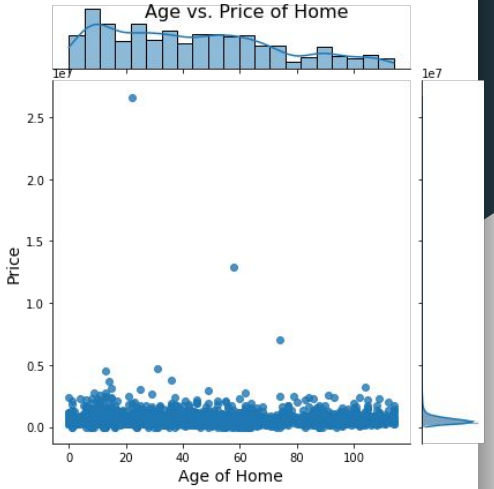


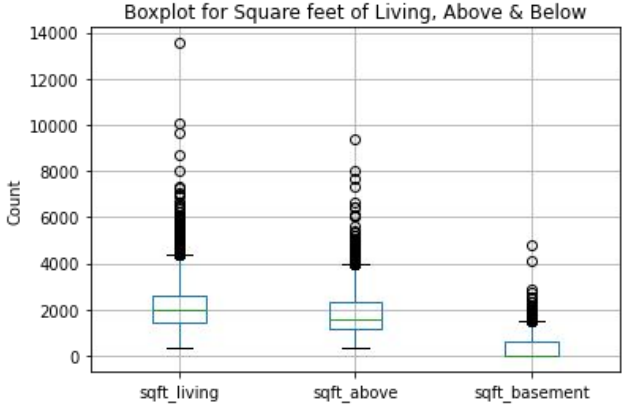
Data Wrangling

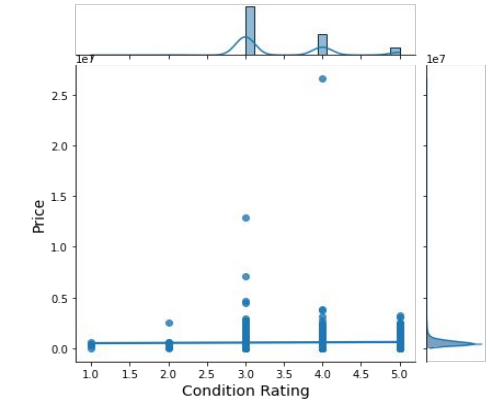
1. Started with 17 columns and 4,600 rows
2. Dropped certain rows
3. Added three columns
4. Only two NaN values
5. Dummy variables for Condition & View
6. Ended up with 25 columns and 4,600 rows
7. Target Variable is Price

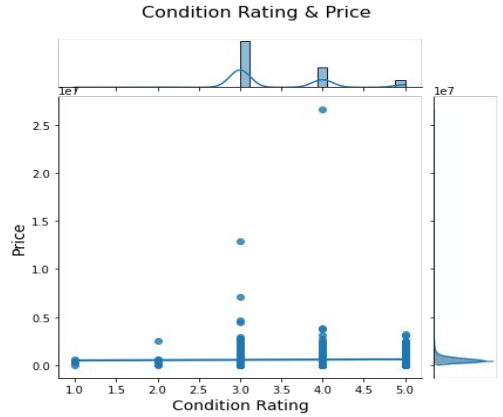
Exploratory Data Analysis



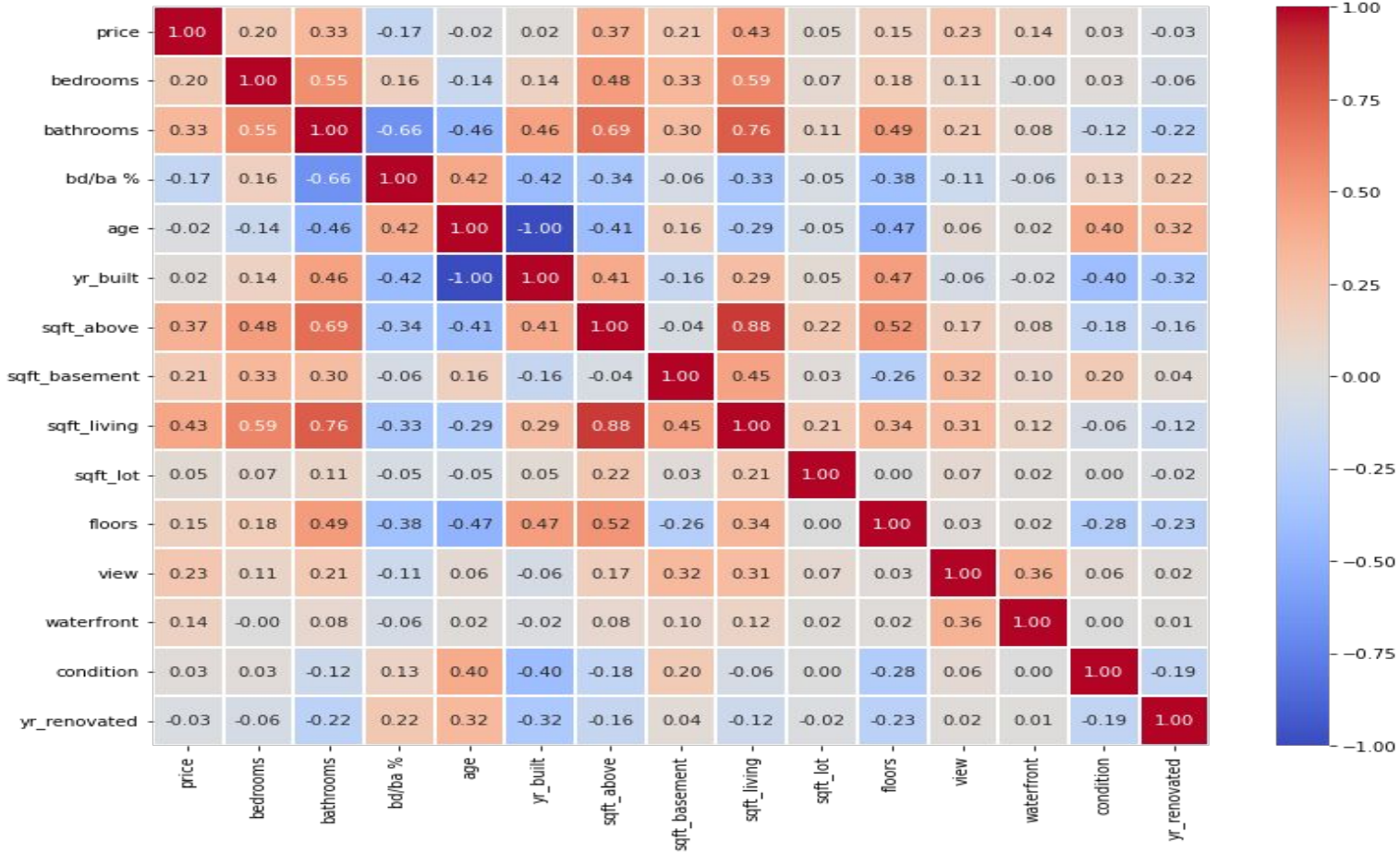








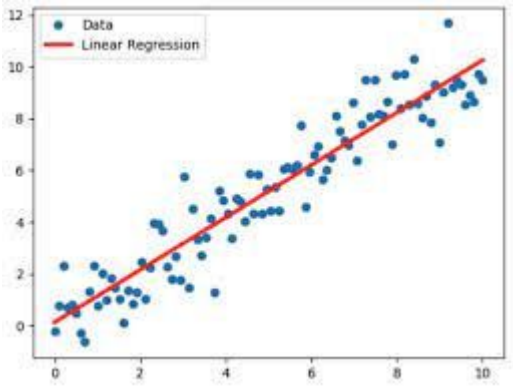
Correlations for the Housing Data



EDA Key Findings

- Many outliers for square feet of living, above and below.
- Most homes had a condition rating of 3-5.
- As # of bedrooms increased so did the # of bathrooms.
- Most of the homes had three to six bedrooms.
- Price and square feet of living space, positive correlation.
- One outlier home was under 2,000 sq ft and $10 million.

Regression Model Selection

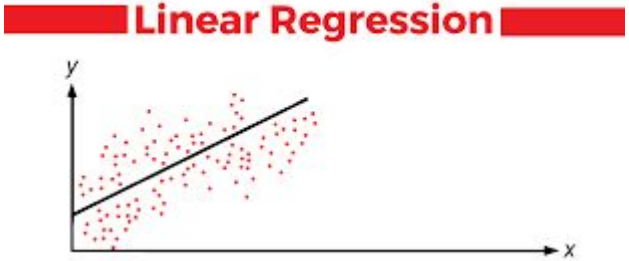


- Linear Regression
- Ridge Regression
- Random Forest Regression Model
- XGB Regression

Best Model was:

Linear Regression

- Typically best model for predicting price of real estate
- R2 score of .47
- MAE of 164,817.45



Model Results Analysis

a. Linear Regression

1) R2 = .4696
2) MSE = 61,734,606,846.02
3) RMSE = 248,464.50
4) MAE = 164,817.45

b. Ridge Regression

1) R2 = .467
2) MSE = 61,974,982,112.41
3) RMSE = 248,947.74
4) MAE = 165,587.70

c. Random Forest Regression

1) R2 = .395
2) MSE = 70,352,537,281.89
3) RMSE = 248,947.74
4) MAE = 140,215.58

d. **XGB Regression**

1) R2 = .303
2) MSE = 81,126,092,073.28
3) RMSE = 248,947.74
4) MAE = 145,701.31

Recommendations

1) **Need more data**
   a) Could improve R2, MAE, etc. scores
   b) More X variables

2) **Information to Clients**
   a) Share with them our EDA findings
   b) Improve buyer morale
   c) Help sellers fix issues with their house

3) **Unique Houses for Clients**
   a) Our model could be a better comp than zillow
   b) Helping buyers if the house they like is over or under priced
   c) Find similar houses with certain traits

Whats next?

- Collect more data
- More variables for our model
- Improve our R2 score
- Run models on different variables