Home Price Prediction
Final Report

By
Garrett JA Hass

1. Introduction

a. Problem Statement

Real Estate companies need help in accurately predicting the price of a home based upon the appropriate variables such as square feet, bedrooms, etc.. This can be useful to their clients when listing or buying a home. Both the realtor and client need to know if a house is over or undervalued. It can help a company assist their clients better and give them an advantage over other real estate brokers by providing accurate price predictions.

b. Criteria for Success

Finding which elements have the highest correlation with the price of a home. Most importantly we need to accurately predict the price of a property. This will help a real estate company to have an advantage over other firms and provide value to their clients.

2. Data

The data is a CSV from a Seattle metro area real estate database on kaggle, (https://www.kaggle.com/datasets/shree1992/housedata?select=data.csv). The data timeframe is from May 1st, 2014 to July 9th, 2014. These are the columns from the dataset:

1. Date (of purchase)
2. Price
3. Bedrooms
4. Bathrooms
5. Sqft_Living
6. Sqft_Lot
7. Waterfront
8. View
9. Condition
10. Sqft_Above
11. Sqft_Basement

12. Yr_Bulit
13. Yr_Renovated
14. Street
15. City
16. State Zip
17. Country

Rows I added were:

1. Bedroom / Bathroom % (Bedrooms / Bathrooms)
2. Age (of Home) (Date - Yr_Built)
3. Renovation Age (Date - Yr_Renovated)

The rows I added were helpful variables for each transaction price. I also created encoding dummy variables for view and condition for Linear Regression. For my Random Forest Regression Model I used numeric values for all 44 different cities.

3. Data Wrangling and Cleaning

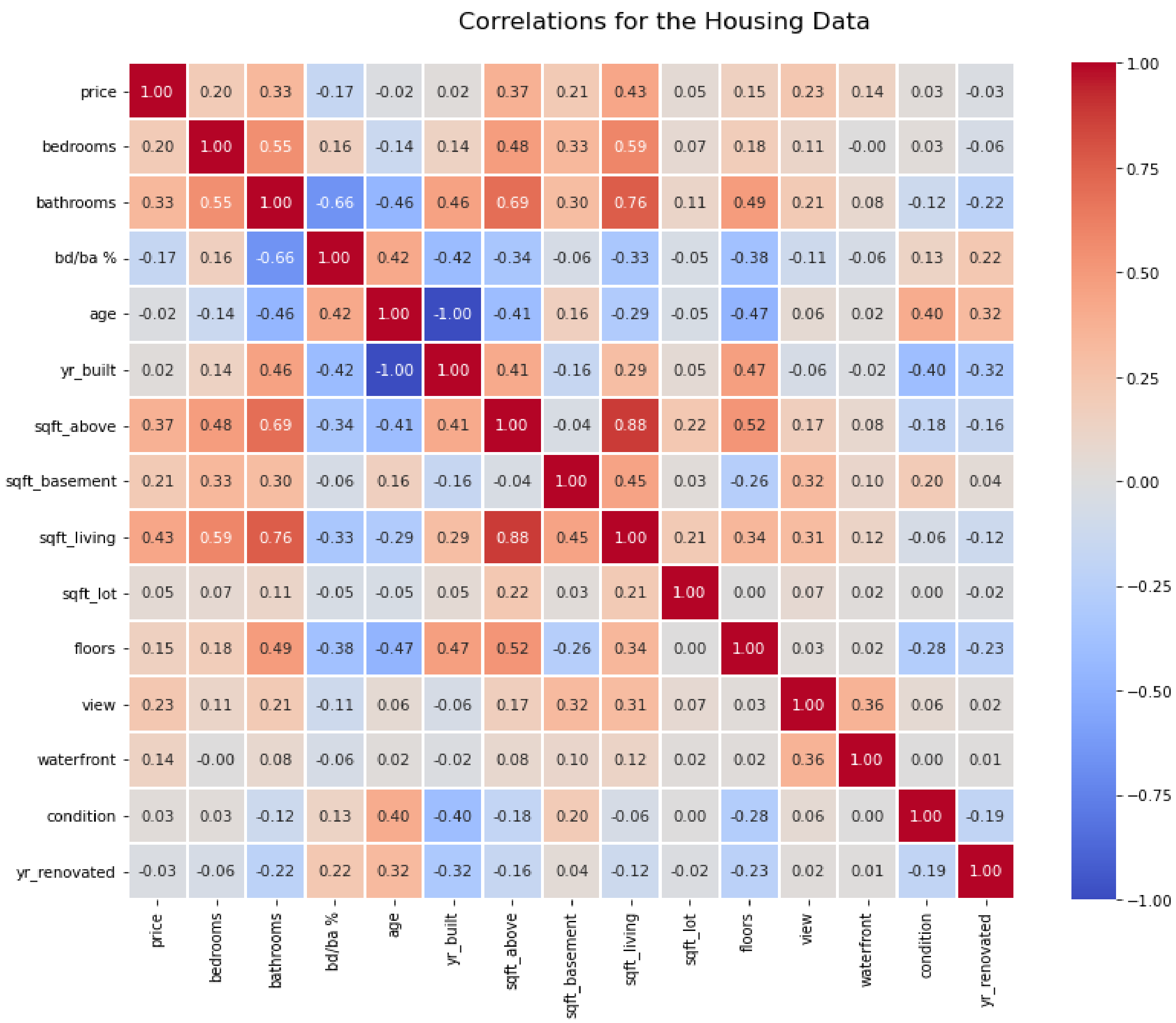
I initially had 4,600 rows and 18 columns from the kaggle dataset. During this stage I uploaded the data to a jupyter notebook. Cleaned the data such as separating the column ‘StateZip’ into a separate ‘State’ and ‘Zip Code’ columns. Added the rows ‘bd/bd %’, (Bedrooms / Bathrooms) and 'Age’ (Year **-** Yr_built). I rearranged the columns to put relevant variables next to each other and explored the mean, maxim, minimum, standard deviation of each numeric column. This helped to understand my data and see if there were any missing values or outliers.

4. Exploratory Data Analysis

a. Correlations

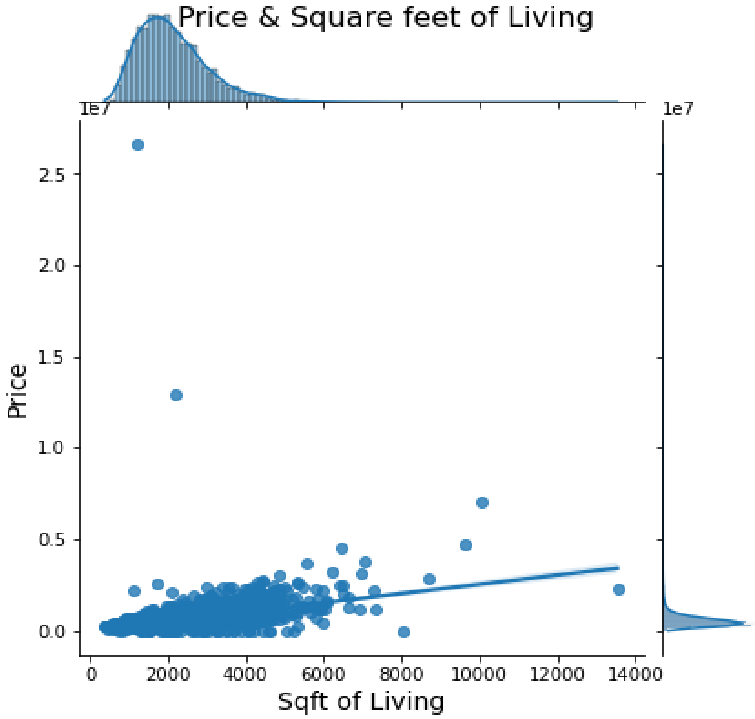
The correlation heat map was very useful to see which columns were relevant. More importantly I found out which columns had a positive or negative correlation with our Y variable of Price. There were a few variables that had a high positive correlation:

1. Square feet of living and square feet above are related at .88.
2. Square living space and bathrooms had a high correlation at .76.
3. Square living space above and bathrooms also have a high correlation at .69.
4. The biggest negative correlation was bed to bath ratio at -.66.


Correlations for the Housing Data
price 1.00 0.20 0.33 -0.17 -0.02 0.02 0.37 0.21 0.43 0.05 0.15 0.23 0.14 0.03 -0.03
bedrooms 0.20 1.00 0.55 0.16 -0.14 0.14 0.48 0.33 0.59 0.07 0.18 0.11 -0.00 0.03 -0.06
bathrooms 0.33 0.55 1.00 -0.66 -0.46 0.46 0.69 0.30 0.76 0.11 0.49 0.21 0.08 -0.12 -0.22
bd/ba % -0.17 0.16 -0.66 1.00 0.42 -0.42 -0.34 -0.06 -0.33 -0.05 -0.38 -0.11 -0.06 0.13 0.22
age -0.02 -0.14 -0.46 0.42 1.00 -1.00 -0.41 0.16 -0.29 -0.05 -0.47 0.06 0.02 0.40 0.32
yr_built 0.02 0.14 0.46 -0.42 -1.00 1.00 0.41 -0.16 0.29 0.05 0.47 -0.06 -0.02 -0.40 -0.32
sqft_above 0.37 0.48 0.69 -0.34 -0.41 0.41 1.00 -0.04 0.88 0.22 0.52 0.17 0.08 -0.18 -0.16
sqft_basement 0.21 0.33 0.30 -0.06 0.16 -0.16 -0.04 1.00 0.45 0.03 -0.26 0.32 0.10 0.20 0.04
sqft_living 0.43 0.59 0.76 -0.33 -0.29 0.29 0.88 0.45 1.00 0.21 0.34 0.31 0.12 -0.06 -0.12
sqft_lot 0.05 0.07 0.11 -0.05 -0.05 0.05 0.22 0.03 0.21 1.00 0.00 0.07 0.02 0.00 -0.02
floors 0.15 0.18 0.49 -0.38 -0.47 0.47 0.52 -0.26 0.34 0.00 1.00 0.03 0.02 -0.28 -0.23
view 0.23 0.11 0.21 -0.11 0.06 -0.06 0.17 0.32 0.31 0.07 0.03 1.00 0.36 0.06 0.02
waterfront 0.14 -0.00 0.08 -0.06 0.02 -0.02 0.08 0.10 0.12 0.02 0.02 0.36 1.00 0.00 0.01
condition 0.03 0.03 -0.12 0.13 0.40 -0.40 -0.18 0.20 -0.06 0.00 -0.28 0.06 0.00 1.00 -0.19
yr_renovated -0.03 -0.06 -0.22 0.22 0.32 -0.32 -0.16 0.04 -0.12 -0.02 -0.23 0.02 0.01 -0.19 1.00
price
bedrooms
bathrooms
bd/ba %
age
yr_built
sqft_above
sqft_basement
sqft_living
sqft_lot
floors
view
waterfront
condition
yr_renovated
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

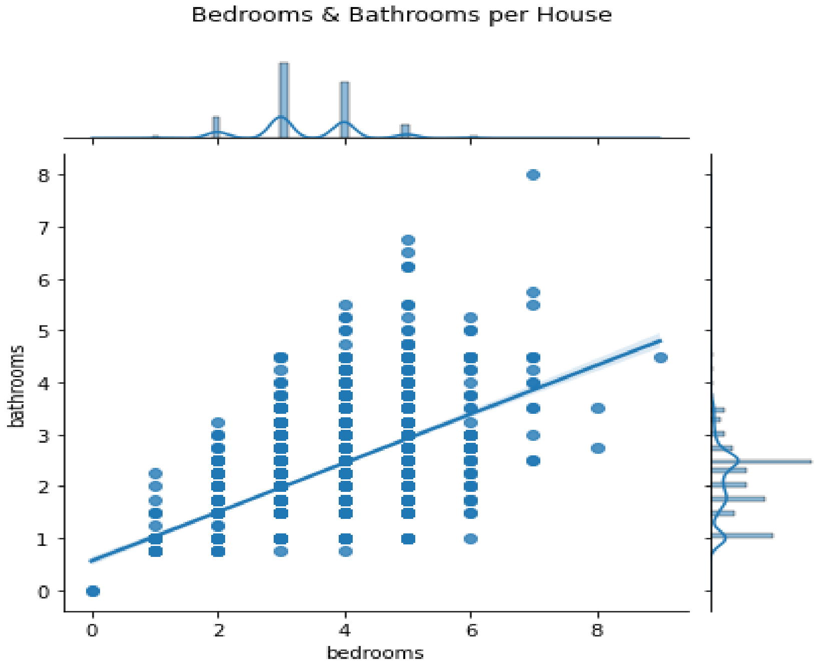
B. Price & Square feet of Living

There was a positive correlation between these two variables. There was one home where the square feet of living space was under 2,000 sq ft and was very expensive. Besides a few outliers home prices increased as the square feet of living space increased.



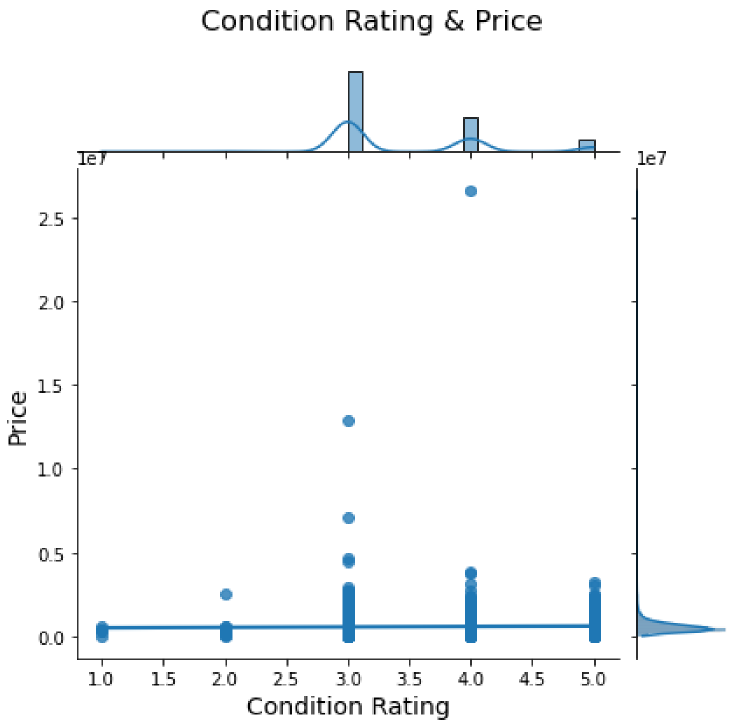
C. Bedrooms & Bathrooms

These two variables also had a positive correlation as well. As the number of bedrooms increased so did the amount of bathrooms in a particular house increase too. Most of the homes had between three and six bedrooms. When a home had six bedrooms they had a wide range of bathrooms from one to seven.



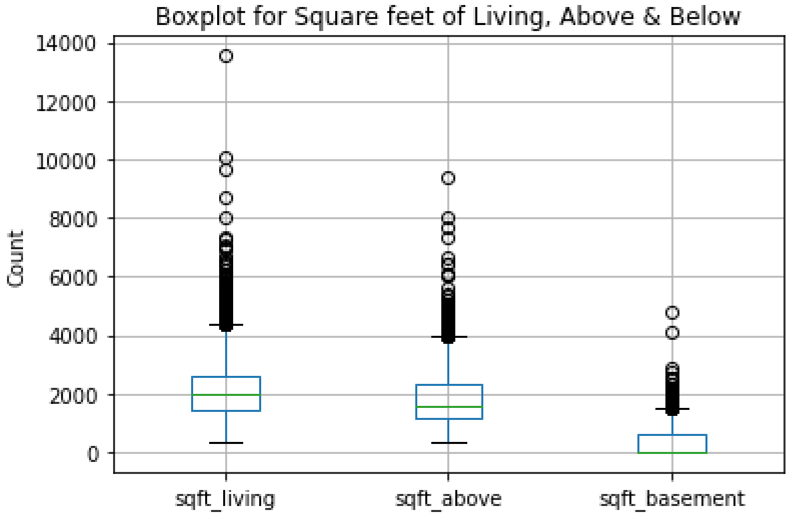
D. Price of Home & Condition Rating

The majority of homes that were sold had a condition rating of 3-5. Interestingly enough two homes had a very high price and a rating of only three. This is possible because the homes were in a nice neighborhood or a desirable location in the Seattle metro area. From this graph we can determine that a vast amount of homes sold were in a good or fair condition.



E. Square feet of Living, Above & Below

The boxplot of the three variables displays that there are many outliers with a high amount of square feet for living, above and below ground level. Houses come in all different shapes and sizes. These homes could potentially be very expensive. Most homes did not have very much basement space. Which would make sense since Seattle is in a coastal area.



5.Pre-Processing

In this stage of the project I prepared the data for modeling. I created a new column, ‘Renovation Age’, (Year - Yr_Renovated). I also removed the two NaNs from the ‘bd/ba %’ column and made them 0. They were NaNs since the two homes did not have a bed or bathroom. I also split the ‘Date’ column into ‘Year’, ‘Month’, ‘Day’ , and ‘Week’. I encoded the view and condition columns for three models, Linear Regression, Ridge Regression and XGB Regression. I did a 70% train and 30% test split on the data for my four models. I dropped a few columns for the X variables and used the following columns:

1. bedrooms
2. bathrooms
3. bd/ba %
4. age
5. sqft_above
6. sqft_basement
7. sqft_living
8. sqft_lot
9. floors
10. waterfront
11. year
12. month
13. day
14. week
15. renovation_age
16. view_0
17. view_1
18. view_2
19. view_3
20. view_4
21. condition_1
22. condition_2
23. condition_3
24. condition_4
25. condition_5

6. Modeling

I tested four different models with a standard scaler on each for my supervised regression price prediction:

1. **Linear Regression**
   This was the best performing model. The R2 score was .47 and a mean absolute error of 164,817.45. The MAE indicates that we could be off by that amount each time we predict the price of a home using Linear Regression. The RMSE, root mean square deviation of 248,464.50 indicates that we could be +/- off of the actual price when using this model. The MSE, mean squared error of 61,734,606,846.02 is the average squared difference between observed and predicted prices of the homes. The 10 cross validation mean score of .41 was slightly lower than the initial R2 score. The grid search CV best score was even lower at .33.

2. **Ridge Regression**
   This was the second best regression model. The MSE, MAE AND RMSE were barely higher than Linear Regression. These higher figures suggest that this model is less accurate at predicting the price of a home than Linear Regression.

3. **Random Forest Regression**
   This model was the third best at predicting the price with a R2 score of .4. The CV mean score of .36 and grid search best score of .28 indicates that this model probably has a lower R2 score than initially observed.

4. **XGB Regression**
   This particular model was the least accurate at price prediction. The R2 score was lower than the random forest model by .1. The MSE, MAE, RMSE were by far the highest for any model. This model also had a lower cross validation and grid search score.

	R2 Score	MSE	MAE	RMSE	CV Mean Score	Grid Search CV Best Score
Linear Regression	0.47	61734606846.02	164817.45	248464.50	0.41	0.33
Ridge Regression	0.47	61974982112.42	165587.71	248947.75	0.41	0.42

Random Forest Regression Model	0.40	70352537281.89	140215.58	248947.75	0.36	0.28
XGB Regression Model	0.30	81126092073.29	145701.32	248947.75	0.29	0.24

7. Recommendations

After running all four models on the data I recommend:

1. More Data

That we use more sales data to predict the price of a home. More information on homes that sold could lead to a model that has a higher R2 score and is more accurate at predicting. If there are more variables like garage size and school district could lead to an improved model. If we had the data on sales for every home sold in the Seattle area for an entire year we could potentially create a better model predictor.

2. Sharing Information to Clients

I would share with potential clients of the real estate firm the correlations we found from our exploratory data analysis. If they are interested in purchasing a home, we found certain traits that go together. Such as the more square feet a home has, the price typically increases. Thus the buyers would need more money. If a buyer with a large family values space in a home they will likely need more cash for the property. A good thing to mention to clients is that a majority of homes that sell have a condition rating of three or higher. This would help with buyer morale since they know the home could potentially be in a good shape and does not need many repairs. It would also help sellers to realize that they will need to fix issues with the house if they want to go under contract.

3. Unique Houses for Clients

If a client was looking to sell a home that had traits that were an outlier in the Seattle area we could compare it to other properties that had similar traits. We could use this data for comps to show a potential seller what they could possibly sell their house for based upon similar homes. The data could also help a buyer know if a house they really like is over or under priced. It could potentially be a better predictor of comps than what is on zillow.

8. Next Steps

The company needs to obtain or collect more data to create a better regression model to predict the price of a home for their buyers and sellers. The data department could collect the information from their local Multiple Listing Service or from county documents.The more information we can obtain the better we can help serve our potential clients. This would give us a competitive advantage over other real estate firms.

I also suggest running models on all the variables to help clients. This would help them understand how much square feet of living space you can expect with other independent variables. Or how many bedrooms they could expect with certain X variables.