Home Price Prediction Model



By Garrett J.A. Hass



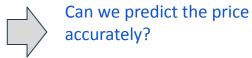
Problem Statement

- 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?





















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







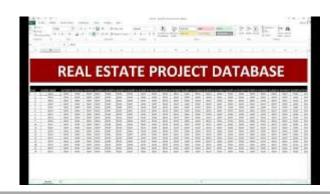
Solution to our Problem

- 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?





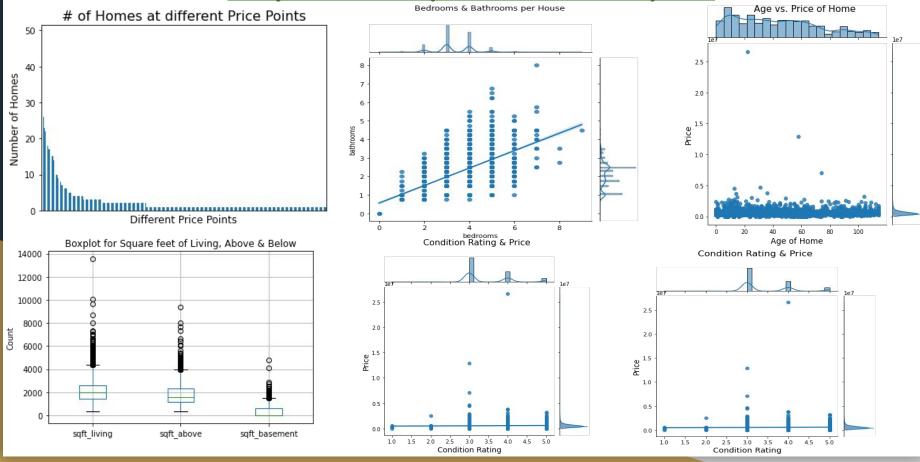


Data Wrangling

- 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

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 -	- % eq/pq	- age	yr_built -	sqft_above -	sqft_basement -	sqft_living -	sqft_lot -	floors -	view -	waterfront -	condition -	yr_renovated -

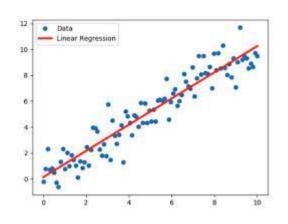


EDA Key Findings

- 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



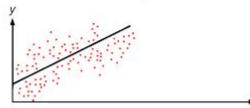
- Linear Regression
- Ridge Regression
- Random Forest Regression Model
- **❖** XGB Regression

Best Model was:

Linear Regression

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

Linear Regression



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

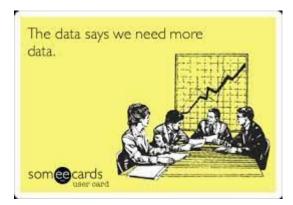
2) Information to Clients

- 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





