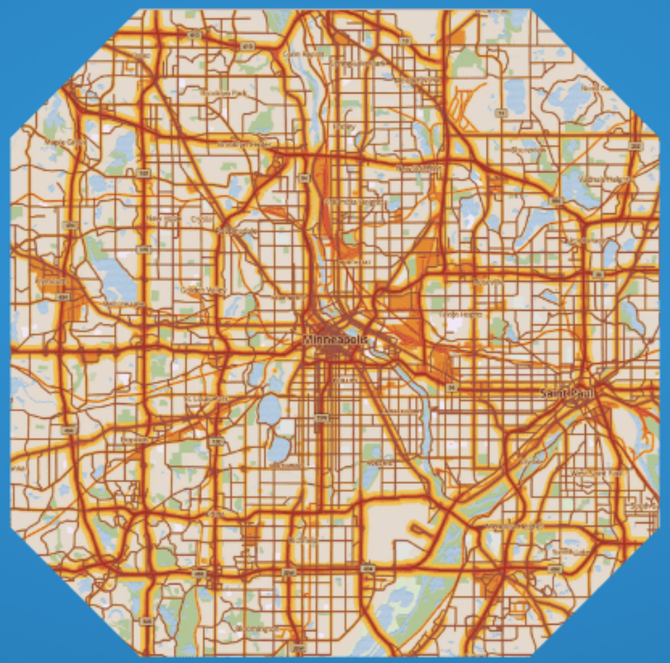
Data Science of Real Estate Data



Twin Cities Area

by Georgios Pligoropoulos <george@pligor.com>

Typical steps, like EDA, were skipped

Listing Price Distribution

Log Transformed Price Distribution

Question 1

"You are home developer looking to partner with the top real estate companies to acquire and then sell a large volume of properties in the Twin Cities area (the more the better).
You do not have the resources to manage too many real estate partners and a strict timeline to negotiate the deals.
What realty companies would you pick as your partners? Why would you make that choice? Demonstrate this through analysis, visual display of your results, and description of your methodology of selection."

Question Summary:

- Home developer wants high revenue by cooperating with only a few Realty companies

Answer:

- Build a linear regression model to get the value of each house
- Subtract the value of the actual price to get the potential of the house
- Sum the potentials per realty company
- Rank from most potential to least

Feature Engineering

Drop Attributes with no information

- State - same for all
- Last Sale Data - almost all values missing

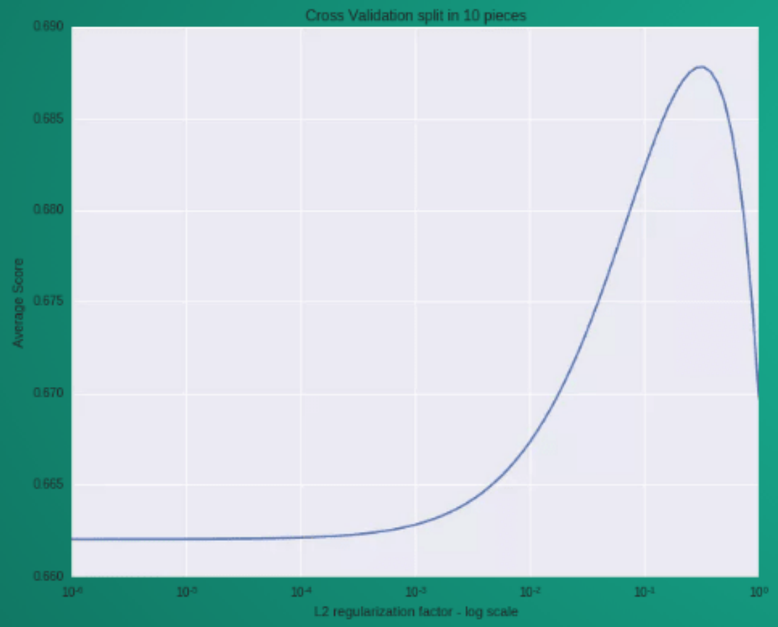
Drop Redundant Attributes

- DOM - not expected to affect
- Realty - affiliations not analyzed
- Sold Previously - second hand house or not
- Address - too much preprocessing
- Location - already used City and ZIP (one-hot encoded)

KNN imputation

- BATHS: 1 null
- LotSize: 64 nulls
- LATITUDE-LONGITUDE: 6 nulls

Ridge Linear Regression



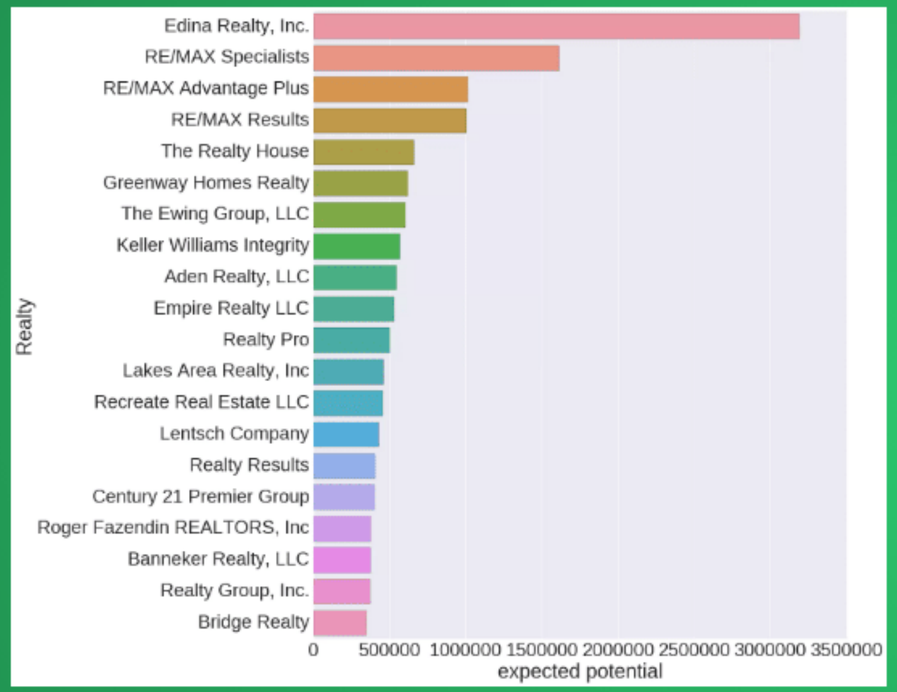
Optimal Regularization Factor: 0.327455

0.762776

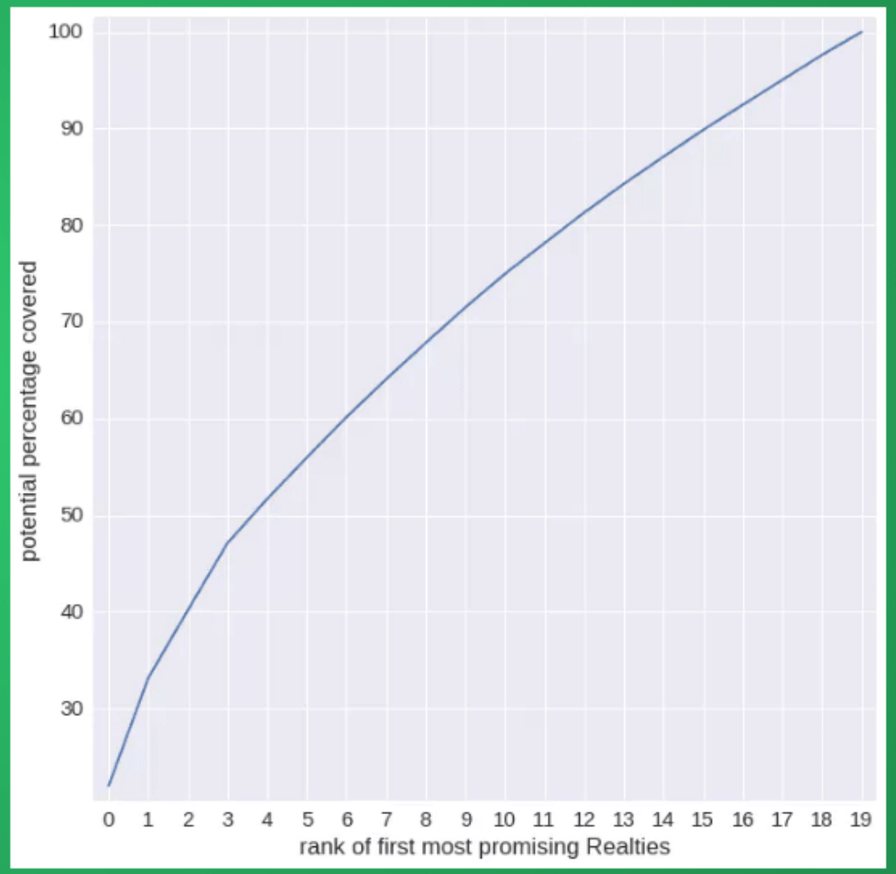
Correlation
Coefficient Score

testing dataset

Which are the best Realties to partner with?



potential from most promising Realties



The least amount of Realties you'll need for your investment

Question 2

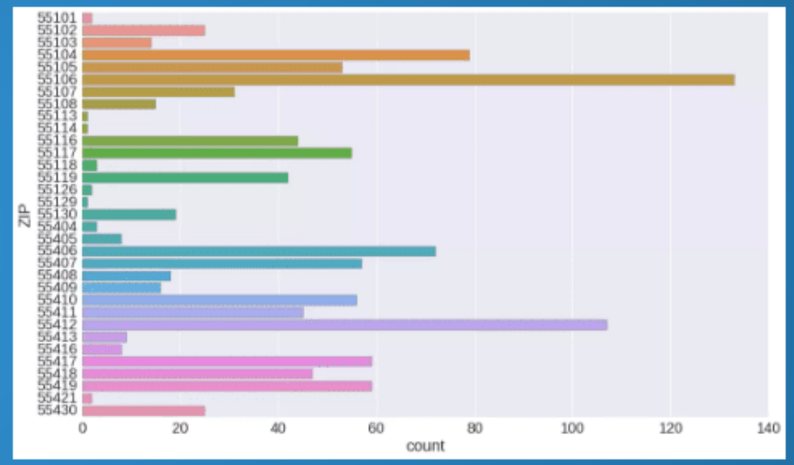
"All things being equal what would you predict as the listing price for a 2111 square foot house if that was the only information you had on a house in this area? How did you arrive at that estimate? Please explain."

Question Summary:

- Predict price of listing house for 2111 sqft for a certain area

Answer:

- Neighbors approach: Take all the houses near the value 2111 sqft for the particular area and average listing price (not implemented)
- One-variate regression: Where the only input is sqft against listing price (not implemented)
- Multivariate Regression: Set area and average rest: age, long-lat, bedrooms, bathrooms



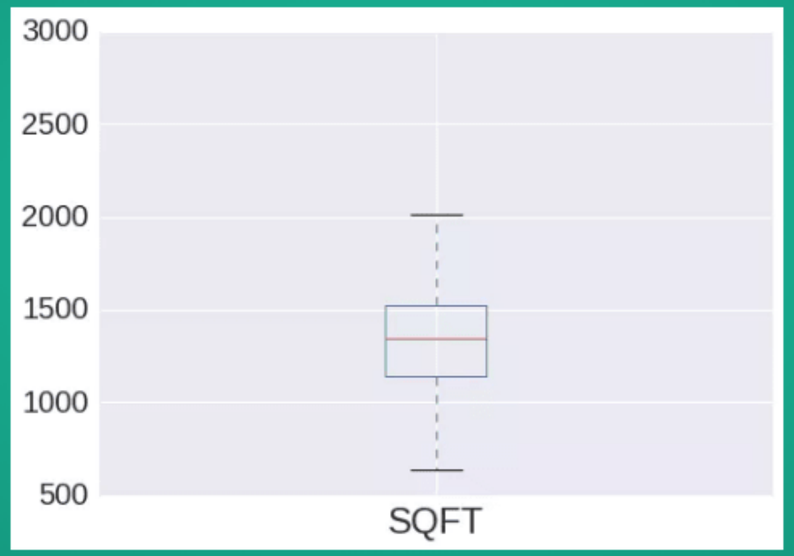
Average SQFT: 1460

SQFT 2111 for ZIP code 55107

Averages

- Age: 90
- Lat - Long: 44.925746, -93.085913
- Bedrooms: 3
- Bathrooms: 1.5
- Lot size: 5681
- Parking Spots: 2
- Has Garage: Yes
- Is Short Sale: No

- SQFT: 1460



2111 sqft is considered an outlier

319000

$

Most predicted price for a listing with 2111 SQFT in ZIP code 55107

Question 3

"Imagine you are an enterprising real estate agent who has the chance to buy up a
bundle of houses for sale – but you can only pick one zip code (either 55104 or 55108).
Also, you can only buy a bundle of properties priced within the middle 50% of the values
– you won't be able to buy the most expensive houses or the cheapest houses.
Assume the sample of houses in this dataset is representative for those zip codes. If
you had the choice to buy 1000 homes in either 55104 or 55108 which zip code would
you invest in and why? Provide your analysis and reasoning."

Question Summary:

- You have only a few houses in the dataset but if you needed to buy 1000 listings in zip codes 55104 and 55108, which would you choose?

Answer:

- Assumption 1: Bundle of houses does NOT mean adjacent houses
- "..prices within middle 50% of the values.." -> instances from Q1 (25%) - Q3 (75%)
- Assumption 2: The limit theorem holds and the potentials have gaussian distribution
- Alternative: Use hacker statistics - not implemented

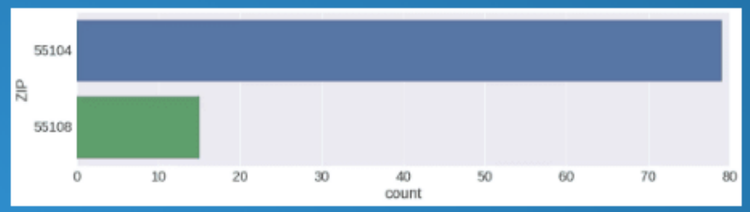
Potential Confidence Intervals

ZIP 55104

- **Average Potential: -48070$**
- **Standard Deviation: 65152$**

ZIP 55108

- **Average Potential: -23309$**
- **Standard Deviation: 68104$**



Number of available listings in each ZIP code

-64806$

to

-31333$

90% confidence interval for ZIP 55104

-65650$

to

19030$

90% confidence interval for ZIP 55108

Question 4

"We're looking to understand what features of the home are most important to potentially predicting the list price of a house. What has the strongest relationship to listing price: square foot, lot size, or number of bedrooms? How do they compare? Please explain."

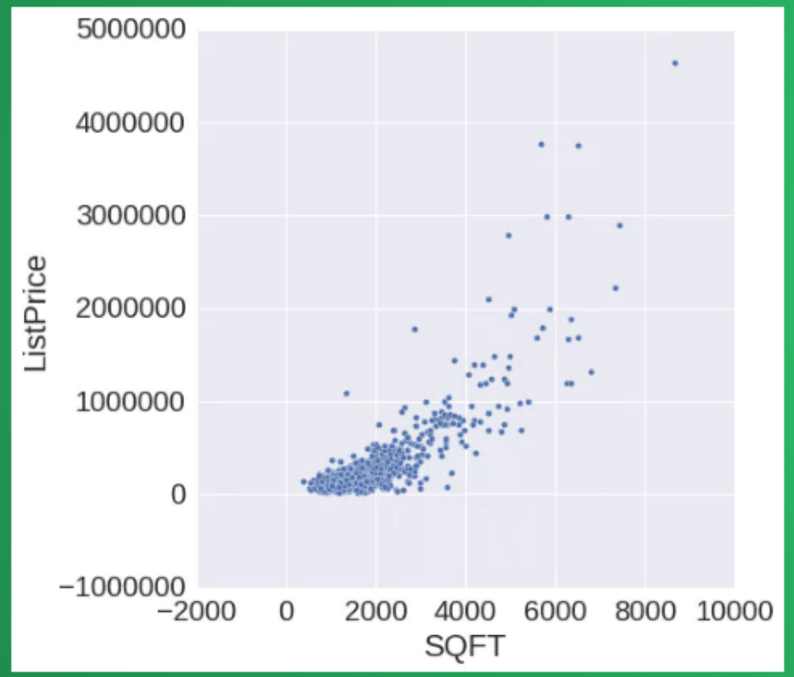
Question Summary:

- Sqft, Lot Size or number of bedrooms have the strongest relationship with listing price?

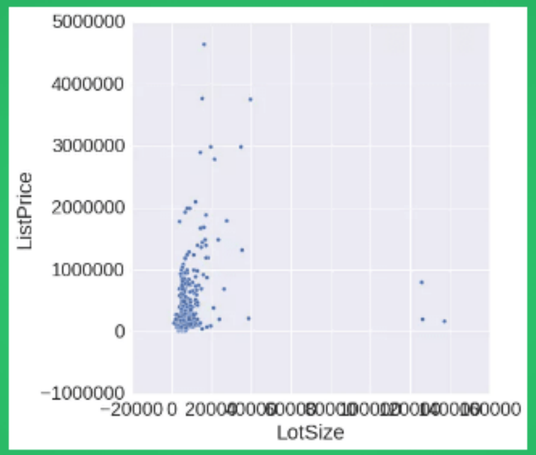
Answer:

- Build Machine Learning Model for each feature. Measure model performance. - not implemented
- Pearson Correlation: Listing Price versus square foot, lot size and number of bedrooms + p-value for no correlation

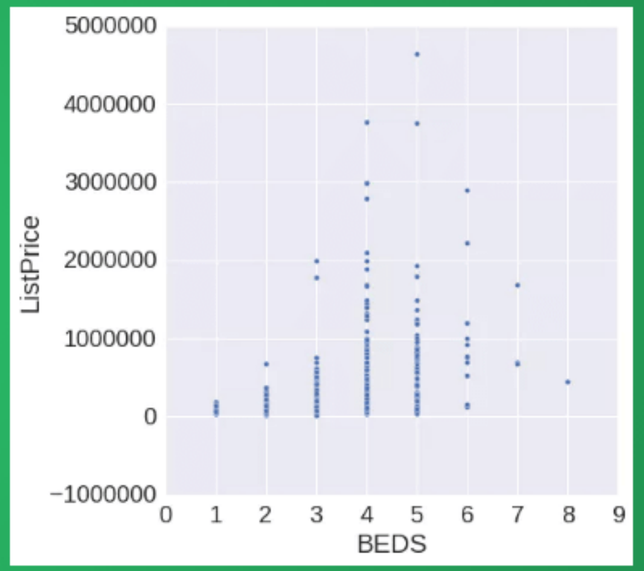
Correlations: LotSize < Bedrooms < SQFT



Pearson Correlation: 0.828
p-value: 1.29e-280



Pearson Correlation: 0.266
p-value: 1.73e-19



Pearson Correlation: 0.406
p-value: 1.48e-45

Question 5

"Short sales are sometimes good opportunities to get value for a house but there is also the risk the property will need a lot of work. How do short sales compare on the average square foot of the house, average price per square foot, and the average lot size?

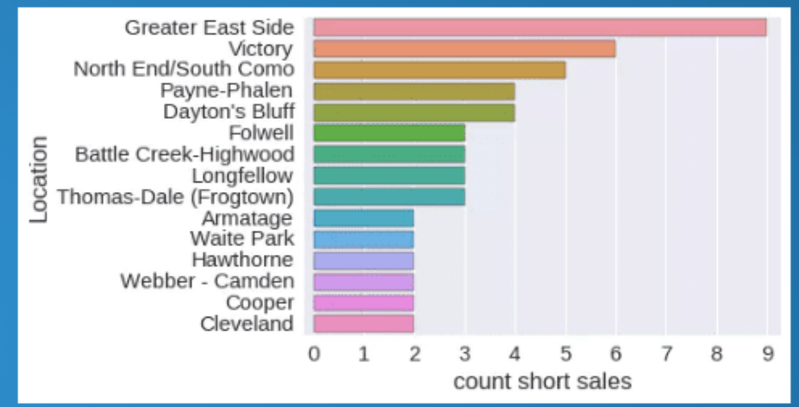
Now pivot this data based on the location field. Are short sales always a better deal than regular listings? Can we say with certainty this is true for every location? Why or why not?"

Question Summary:

- Compare averages between short sales and rest
- Are short sales always a better deal?

Answer:

- Understand very well what a short sale is => Assumption: Short sale houses represent the objective value (targets)
- Assumption: Government has a dummy model for the objective value = f(sqft, lot size) per Location
- Good deal:
  Objective Value — Listing Price



limited data of short sales per location

Short Sale Comparison

Location	is median SQFT smaller on short sale	is median Lot Size smaller on short sale	is median Listing Price smaller on short sale
West Side	No	No	No
Hawthorne	No	No	No
Wenonah	Yes	No	Yes
Victory	Yes	No	No
Jordan	Yes	Yes	Yes
ALL LOCATIONS	Yes	Yes	Yes

Sample of 5 Locations

Good Deal for Location: Greater East Side

Some deals could be better than Short Sales

Question 6

"We'd like to understand how listings compare with the population in their area. Take the zip code data in the 2nd sheet and match against the house listing data. What zip codes have the highest amount of listings per the population size? Show the top 10.

Separately, what zip code has the highest listing price per person? Google that zip code and provide some hypothesis and examples as to why this might be true."

Question Summary:

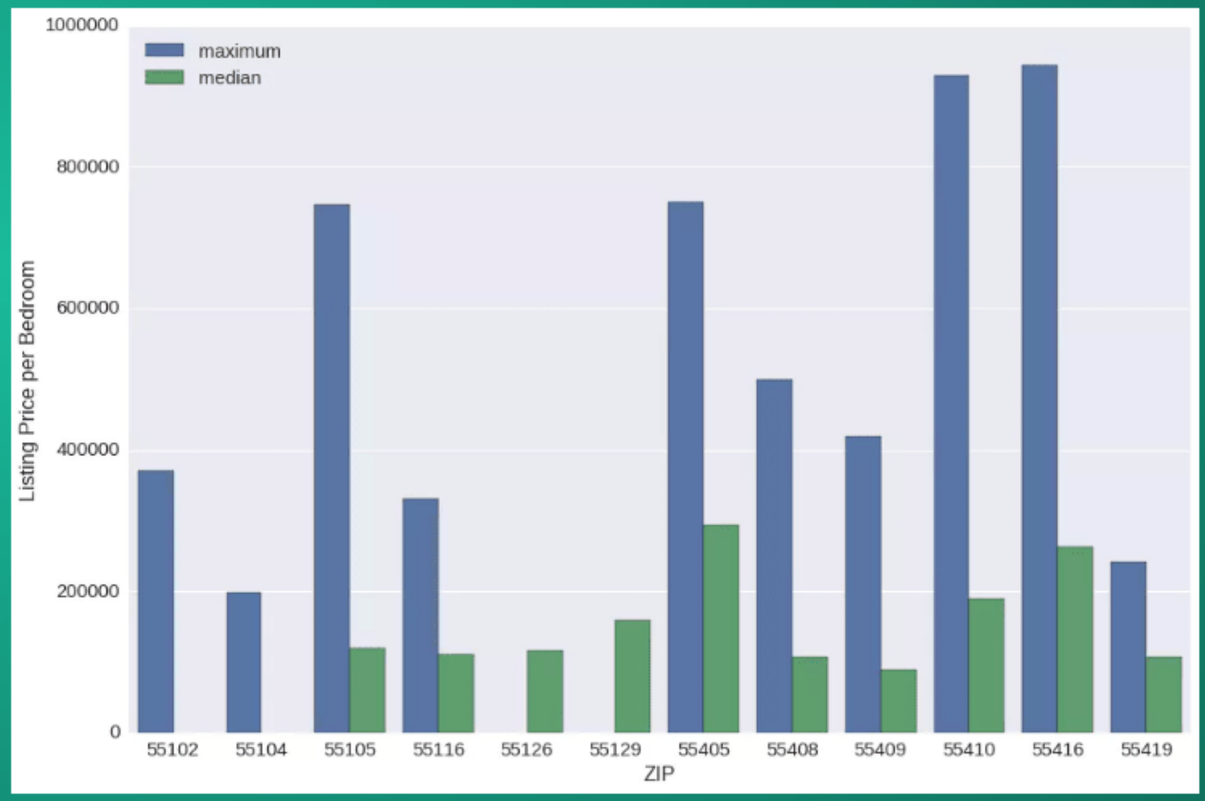
- Listings per population - show top 10
- Listing Price per person
- Verify with real examples

Answer:

- Divide number of listings with population per zip code
- Simplification: One bedroom corresponds to one person
- Divide maximum price with corresponding number of bedrooms per zip code
- Alternatively: Divide the median price with the median number of bedrooms per zip code

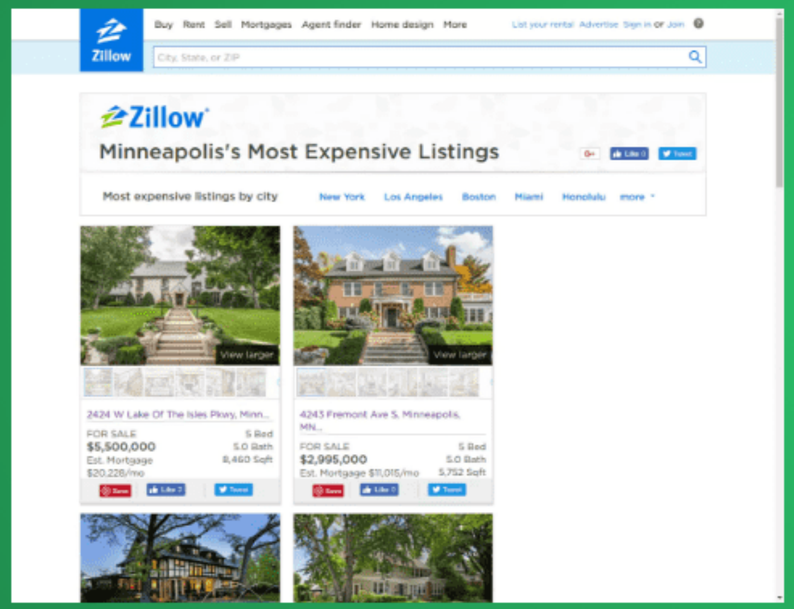
Top 10

	ZIP	Percentage of Number of Listings per Population
1	55412	48.31%
2	55410	28.96%
3	55106	25.22%
4	55417	23.72%
5	55406	22.42%
6	55419	22.34%
7	55107	20.98%
8	55105	18.63%
9	55116	18.45%
10	55104	18.27%



8 out of 10 ZIP codes are most expensive in both metrics

Cross check with online information



source: https://www.zillow.com/minneapolis-mn/expensive-homes/

ZIP codes Frequency:

- **55401**: 2
- **55403**: 8
- **55404**: 4
- **55405**: 8
- **55406**: 1
- **55408**: 3
- **55409**: 3
- **55410**: 15
- **55413**: 1
- **55416**: 2
- **55419**: 3

ZIP Codes we classified as most expensive:

1. **55105**
2. **55116**
3. **55405**
4. **55408**
5. **55409**
6. **55410**
7. **55416**
8. **55419**

6 zip codes match

Question 7

"You've just been hired as a data scientist at the premier real estate firm in this area.

They want to forecast the actual sales price for any of these listings (and future listings).

What variables from this example data do you think would be the most predictive of the actual sales price? What other kinds of data would you want have to provide the most accurate prediction of the actual sales price?

Assume you can get any data you want. Describe this data clearly and why you think it would help you build an accurate predictive model."

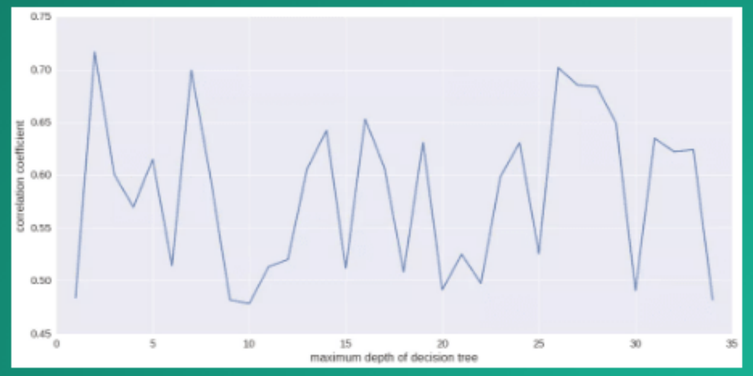
Question Summary:

- Which features are more important for predicting the price?
- Name useful features

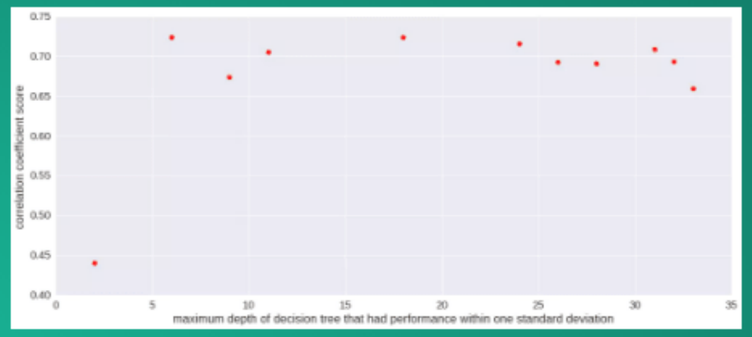
Answer:

- Decision Tree Model
- Get the Gini importance (purity separation) of each feature
- Use our domain knowledge to request for new data attributes

Decision Tree Model



Performance of DT model has very high variance => choose (almost) simplest model within one std

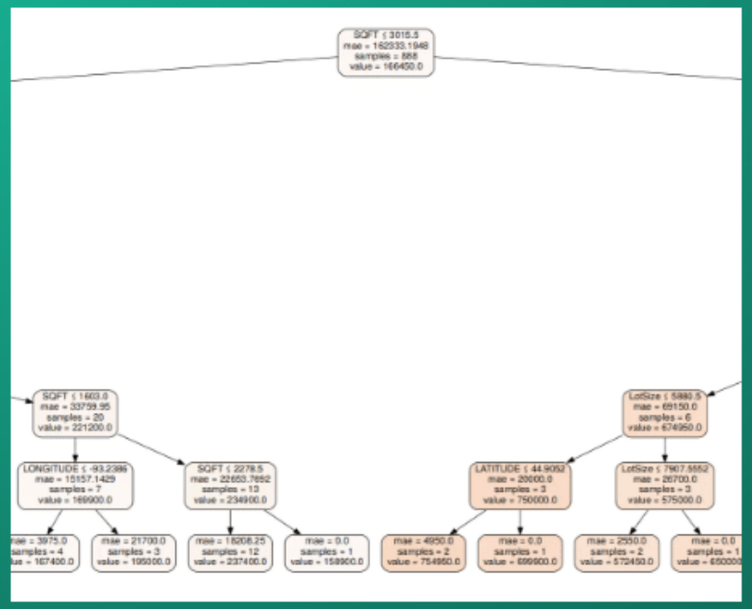


verifying that choosing the extreme case of the simplest model is not optimal

0.728530

Correlation Coefficient Score

testing dataset



Best DT Depth: 7

Features

10 most important Features (Gini):

- SQFT: 0.563189
- Latitude: 0.133847
- Longitude: 0.130789
- Age: 0.058908
- Lot Size: 0.041424
- Bedrooms: 0.018911
- Bathrooms: 0.011640
- Parking Spots: 0.011396
- Zip Code is 55410 or not: 0.006582
- Zip Code is 55405 or not: 0.006325

Extra features for more accurate predictions:

- how many floors?
- has swimming pool?
- current condition?
- year of last renovation?
- maximum earthquake resistance?
- Balcony?
- Beautiful view?
- Energy independence (solar panels)?
- Crime of the area?
- Energy efficiency certificate?

if price is your target let customer interactions reveal the important features

« The context changes the content and the content changes the design »

Thank you