AMES, IA Housing Sales Price

Roxana Ruvalcaba

Agenda

- Problem Statement
- Data Cleaning and EDA
- Model
- Next Steps

Problem Statement

We have 81 features related to housing sold in Ames, IA. I have been tasked with using this data to create a model that will predict the sales price of Ames housing. I can use as many or as few features as I like keeping in mind that I don’t want to over or under fit the model.

Data

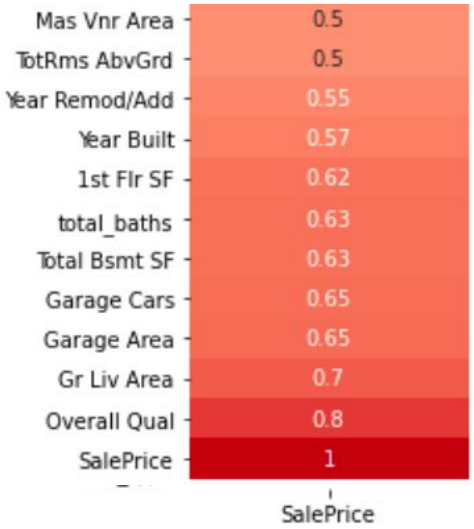
- Data from 2,051 sales
- 81 detailed features
- Selling Price from 2006 to 2010
- Predict 878 selling prices

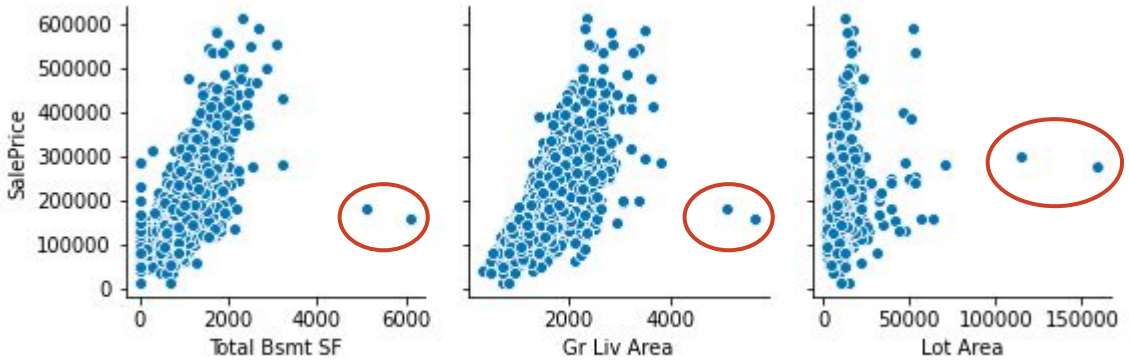
Data Cleaning and EDA

- Null values
  - 29 features with NaN values; deleted columns with > 1,000 NaN
- Simplifying columns
  - Multiple columns referencing sq ft of different sections of the basement
    - Total column already existed
  - Multiple columns referencing bathrooms
    - Created a total column
- Removed outliers for lot area, ground living area, and basement sq ft
- Removed classifiers that did not visually show much variation

Data Cleaning and EDA

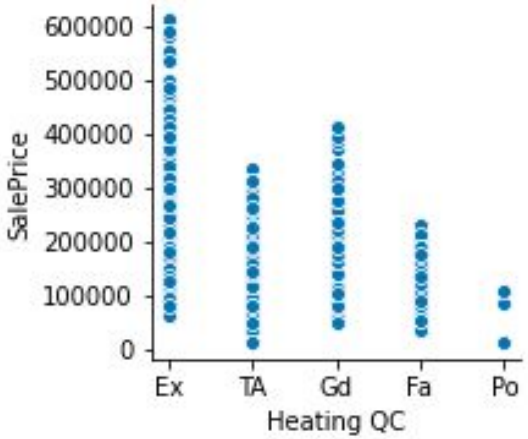
- Removed outliers for lot area, ground living area, and basement sq ft
  - Only two data points
  - External research suggests that the higher the total area, the higher the cost





Data Cleaning and EDA

- Kept categorical values with visual trends and dummified
  - Different categories show price differences
  - Data is not very unbalanced
- Did not create additional interactions
  - Current categories plus encoded variables were already providing high variance in model



Model Selection

I selected the Lasso model to reduce model complexity and help with overfitting given that I had so many features to test

LASSO RESULTS	
R2 train	95%
R2 test	86%

Model: Lasso

1. Train, Test, Split data
2. Log y
3. Scaled data
4. Entered into Lasso Model

Alpha = 0.001

Intercept = $168,187 after exponentiating

Model mean = $179,872

Compared to sample set mean of $181,469

count	878
mean	$179,872
std	$82,475
min	$44,233
25%	$127,493
50%	$159,198
75%	$211,653
max	$848,023

Housing Takeaways

Best Neighborhood: Crawford

What can owners do to increase value:
Improve material and finish quality

Ground living Area was the most important feature

Next Steps

Continue refining features included to reduce variance

- Simplify model
- Try different alphas
- Drop features where p-value is higher than 0.05

External Research

- Location - schools, employment, entertainment
- Comparable properties
- Interest Rates
- Economy