




Ames Iowa Pricing Predictability

Spencer Buckner

**My Background: Data
Scientist at XYZ, a
marketing analysis
company**

**Problem: How to accurately
predict home sale prices in a
model with a min. R^2 and
Cross Val Scores for
Marketing purposes?**



R2 Score - Statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model
In Practice - The features/target relationship accounts for X% of the variation

Cross Val Score - Assessing how the results of a statistical analysis will generalize to an independent data set.
Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations

Predicting accurate sale prices on homes is **not easy** - **Source: Zillow**

Factors that can affect Sale Prices:

1. Historical Sale Prices
2. Neighborhood
3. Market Forces
4. Size and Appeal
5. Age and Condition
6. Nearby Features

What our dataset can tell us!

Source: Inman.com

Quick Sense of the Numerical Data



Several features have high correlation with Sale Price

Good place to start our Linear Regressions!



Linear Regression # 1

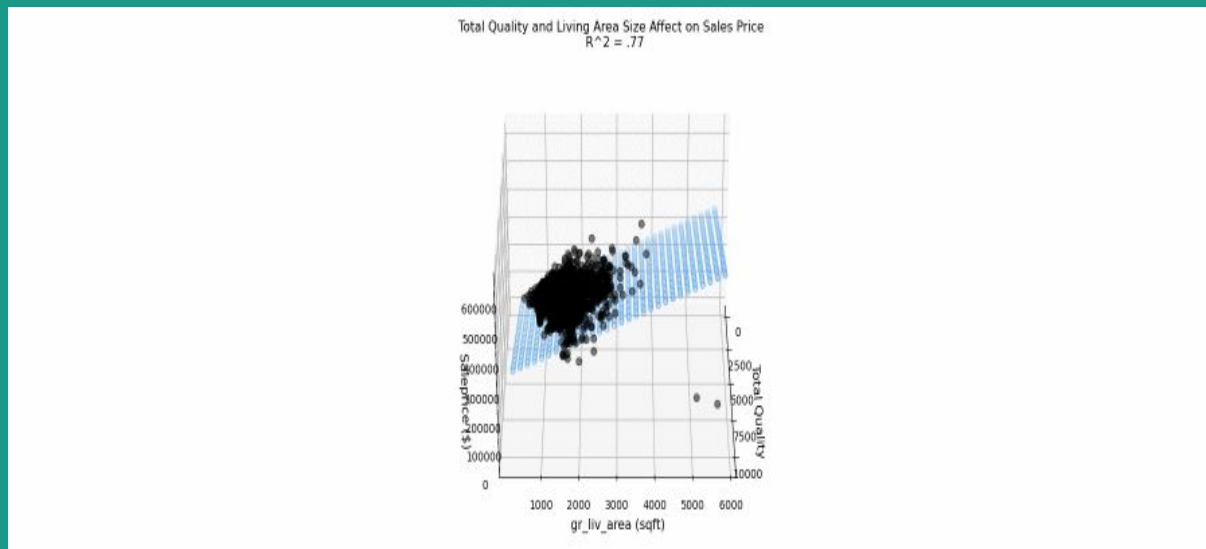
Features:

- Total Quality (Ftrd. Engr.)
 - Overall_qual
 - Exter_qual
 - Heating_qc
 - kitchen_qual
- Gr_liv_area

R2 Score = 0.77

Cross-Val Score = .76

Good Start!



**Now let's look at more advanced
multi linear regressions**

**First, let's look at a numerical linear
regression**

**After, we will look at a categorical
linear regression**

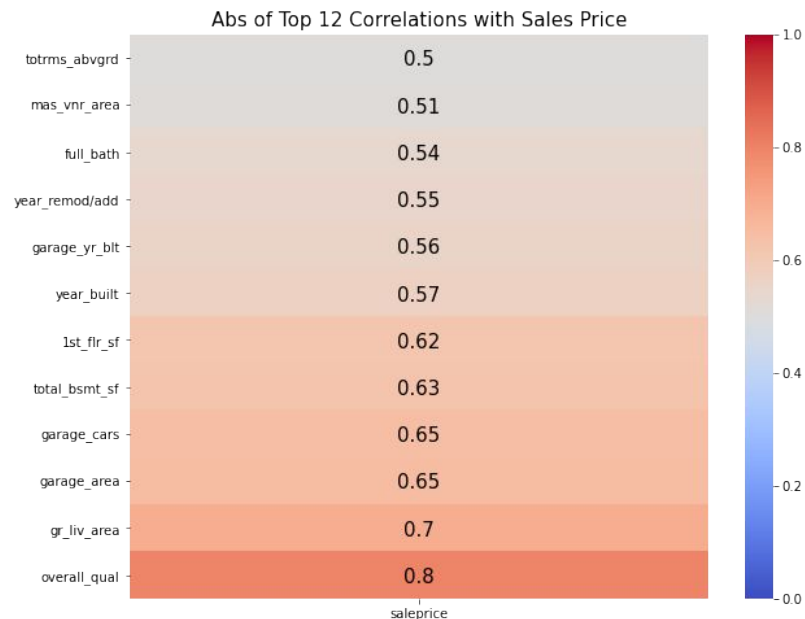
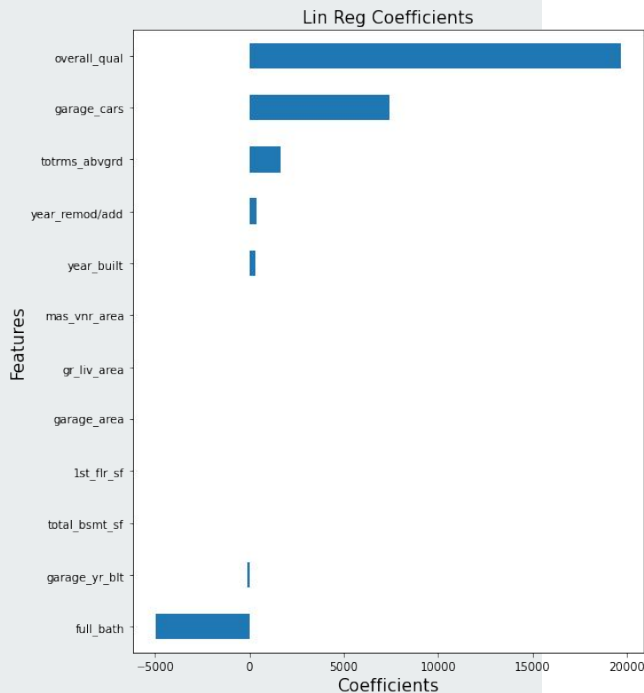
Linear Regression # 2 - Numerical



R2 score = .78
Cross-Val = .76

Several
coefficients not
zero, but close
in scale due to
large
coefficients

Score slightly
better than first



Linear Regression # 3 - Categorical



R2 score = .83

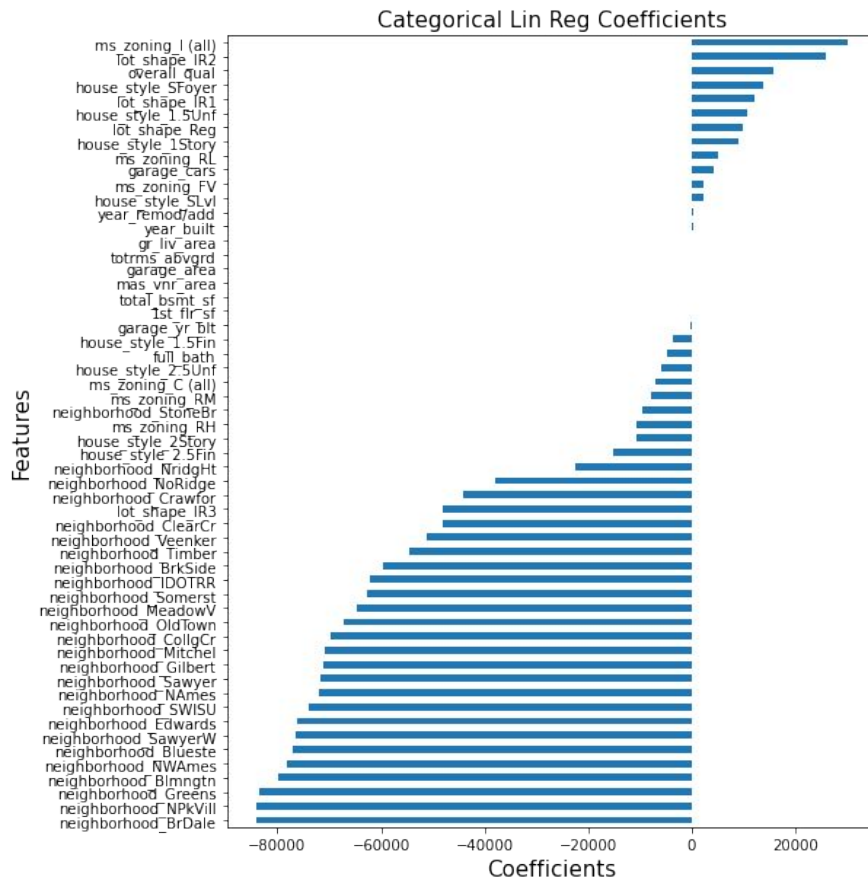
Cross-Val Score = .79

Features: House Style, MS Zoning, Lot Shape, Neighborhood

Several coefficients not zero, but close in scale due to large coefficients

Large negative coefficients for neighborhoods, most likely due to large swings in prices by specific neighborhood

Our best score yet!



Linear Regression # 4

43 features total - golden ratio

Polynomial Features

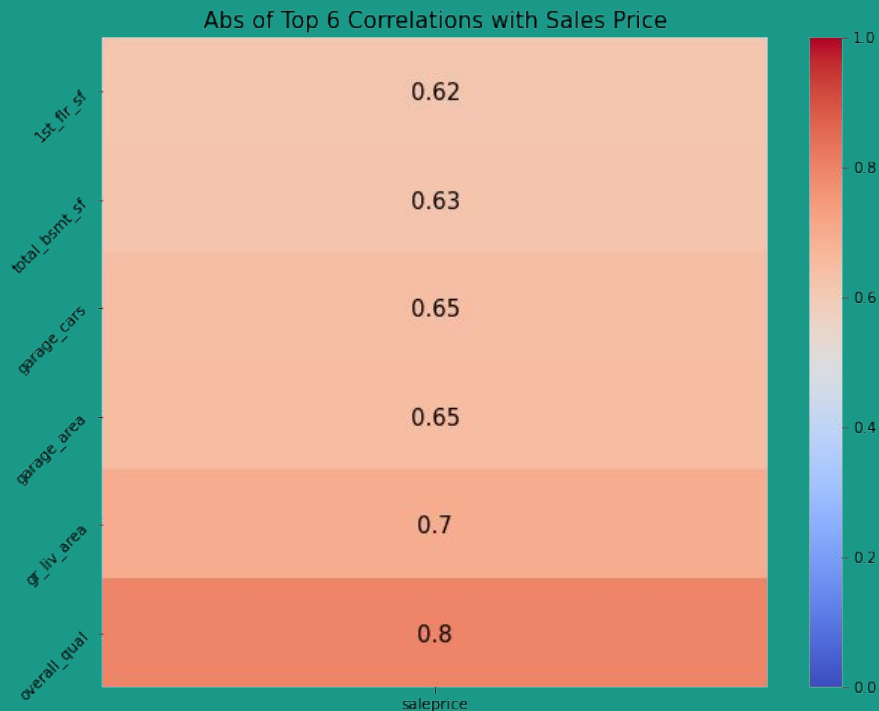
- 26 features

Dummy Variables

17 features total

- House Style
- Lot Shape
- MS Zoning

R2 Score = 0.87
Cross-Val = 0.85



Linear Regression # 4 - Lasso

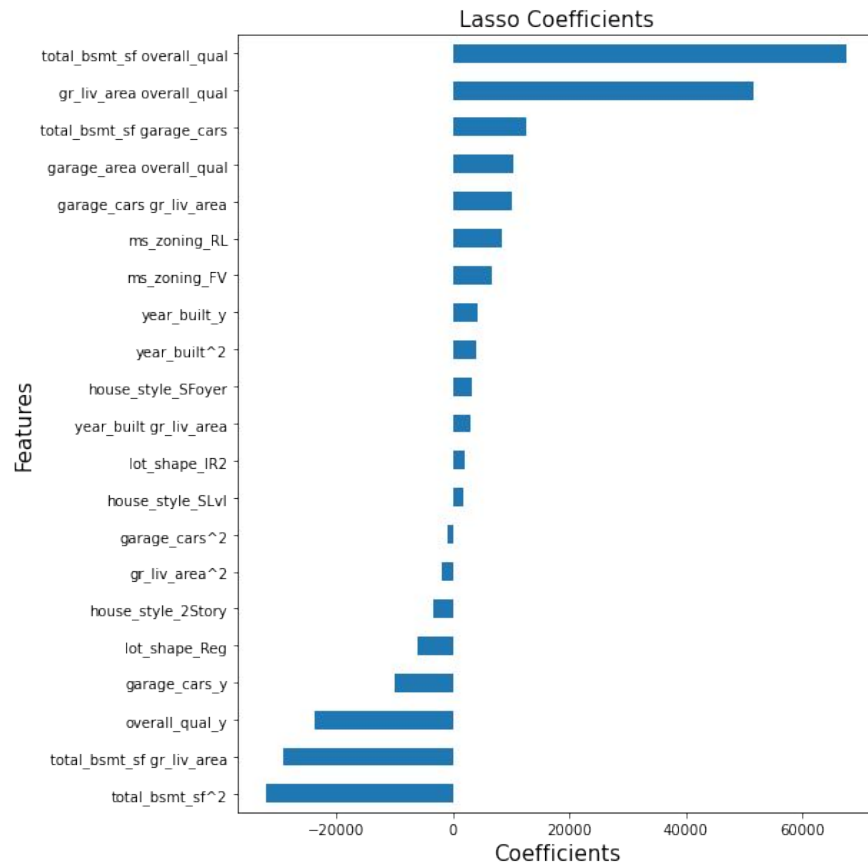


R2 score = .87

Cross=Val = .869

Lasso technique brings total number of features down from 43 to 21!

Same score as normal linear regression but allows us to use less features!





Conclusion

Winner - **Linear Regression # 4 - Lasso**

Through the use of several techniques and model iterations, I have been able to produce a Lasso model that has produced a R^2 score of 0.87, while using only 21 features for analysis.

This is a great start into this model, and with more time and analysis, the model will only be made better.