Ames Housing Pricing Prediction

By Kennedy Bagnol

Scenario

I am a private real estate mogul looking to expand into Ames Iowa

Using my data science skills, I want to create a model to predict housing prices so I can identify up and coming areas to invest in and beat the market

Problem Statement

Using the Ames Housing Dataset, how can we create a model that accurately predicts the sale price of houses?

EDA

First things first, we have to see what we're working with.

The initial data had a decent amount of missing values and in this scenario we cannot just drop them

1: Limited dataset

2: Cannot drop any values from the test set

EDA

Solution:

Go through every column and deal with ‘Nan’s on a case by case basis

Continuous: Replace with 0

Discrete: Replace with 0

Nominal: Replace with ‘NA’

Ordinal: Replace with ‘NA’

Electrical: Replace with ‘?’

Initial data processing

Identify the features with the strongest correlation with price

Created a function that returns the features with a correlation higher than a specified value

First model (linear regression) with just the features that had a coef > .4 (abs val) had an avg cross_val_score of .82.

I used this as my baseline

Coef Calc

Next I used the features that had a corr > .2, this brought my cross_val_score to .84

After this I made a new model that looked at all of the columns in the dummified dataset and calculated if they changed the price by 20,000 or more. If they did, I kept them in the model

I then combined these two lists of features and my cross_val_score went to .86

Feature Engineering

After this I started to do some feature engineering

It initially did not go well

I would create manually create columns that had a higher coefficient than the rest of the features and it wouldn't affect my model performance or even make it worse

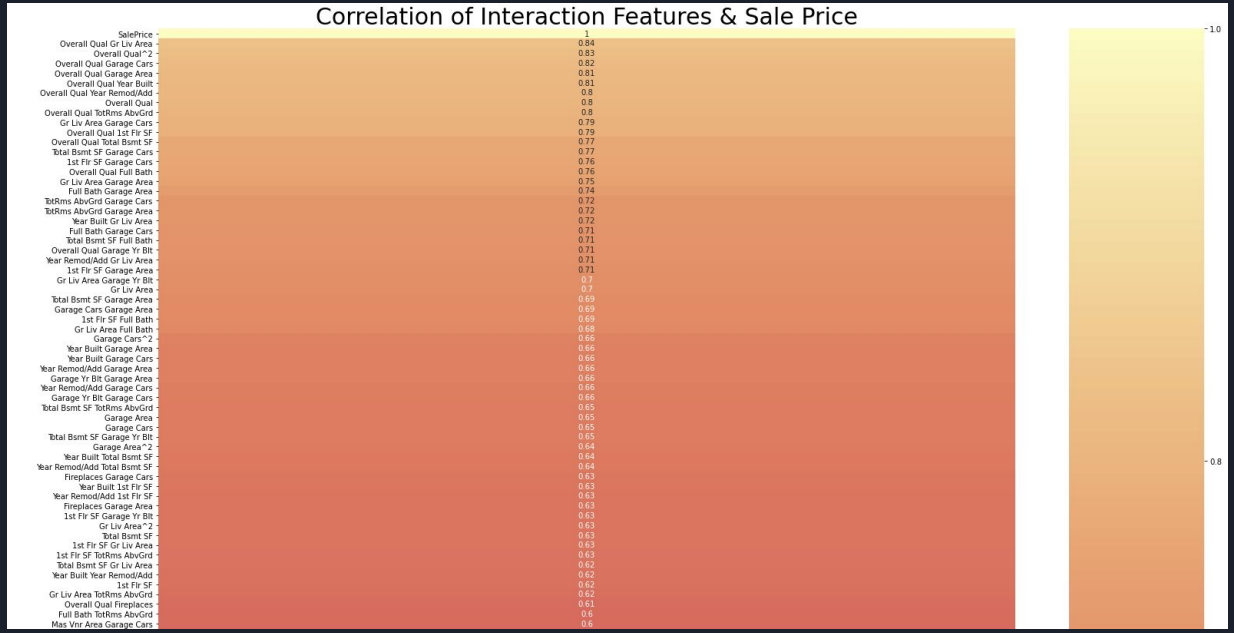
Decided to use polynomial features and go from there

Polynomial Features

Used polynomial features on all of the numerical features that had a corr > .3

Fit these into a ridge regression model

Chose ridge because of research that said it is good to use ridge when you have few variables that have small effects



Ridge Regression

Created 3 different Ridge models

1: Strictly polynomial features

Performed decently well, test score ~.88

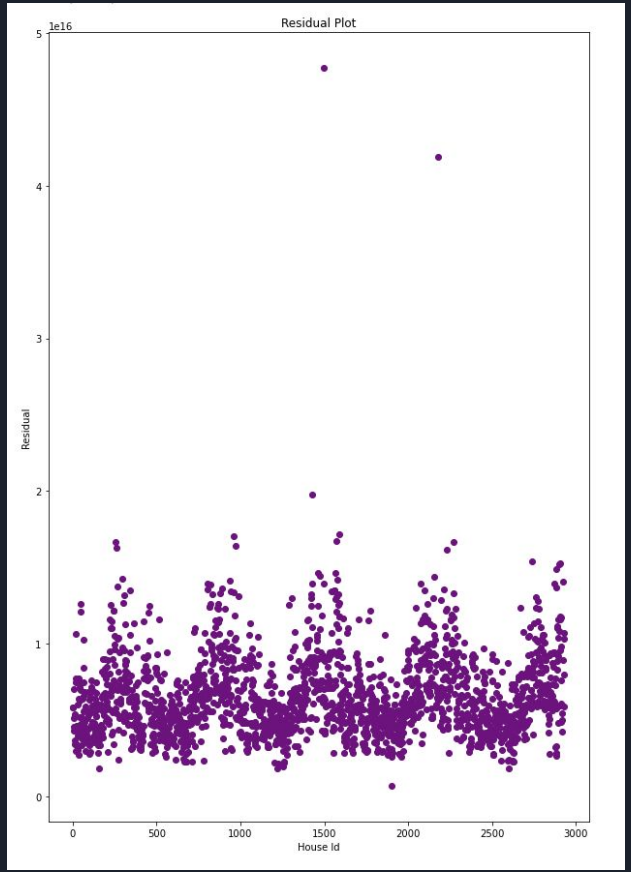
2: Polynomial features with corr > .3

Performed similarly to the first

3: Polynomial features with corr > .3 plus the categorical

Performed the best, test score ~.9

As you can see the final model had a decent residual distribution with only a few big outliers, will touch on that later in the presentation



Final Result

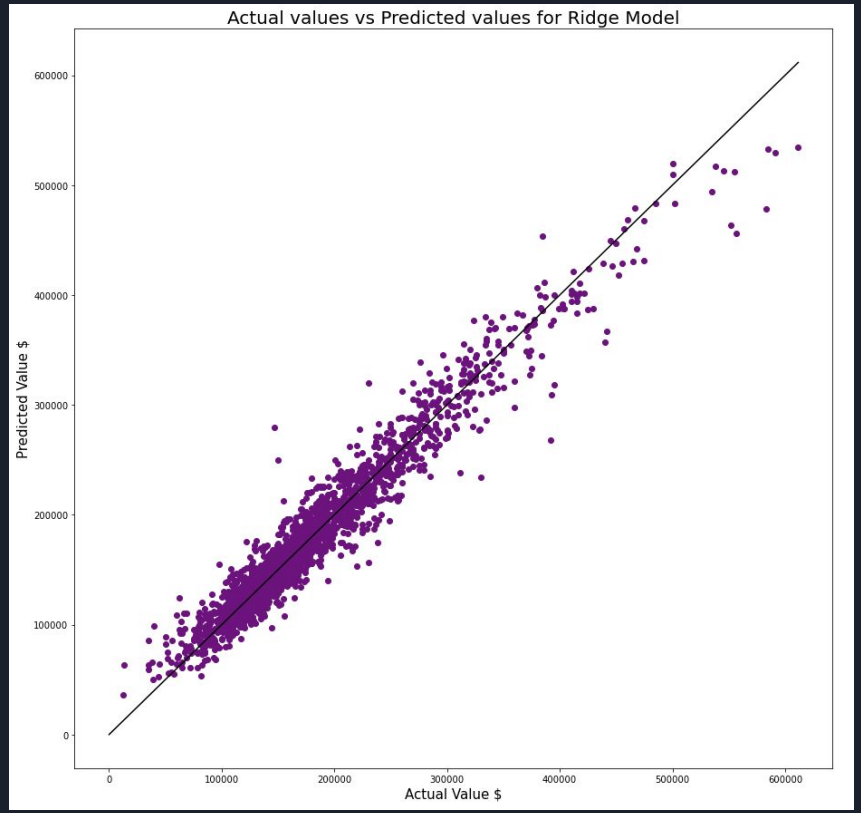
Best score was a RMSE of 23783

R2 Training Score | 0.9353597216837345

R2 Testing Score | 0.9089059397113372

Train RMSE | 20065.370868124563 |

Test RMSE | 24199.286495528653 |



Areas to Improve

More direct feature engineering, calculate the covariance of features and use that to feature engineer

Use gridsearch to optimize hyperparameters

Use Lasso and Ridge and compare the two

Identify and remove outliers, as you can see in the previous residual plot there were definitely big outliers within the data