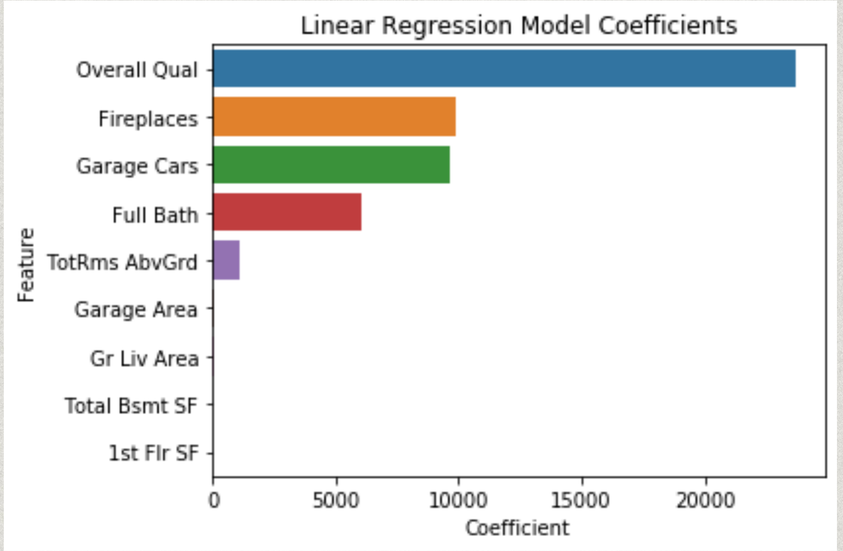
PREDICTING HOUSE SALES

IN AMES, IOWA

Multiple Linear Regression Model

- In this initial Linear Regression model, "*Overall Quality*" has the biggest influence over price. For every 1 change in overall quality, there is an increase in price by $23,678 with all other things being equal.



- Interpretation: The test score of 76.7% for train and 82.5% for test, suggest that our model can reasonably account for ~80% of the total variance of the target variable, price.

Then I built a separate data frame just to explore which dummy variables I should add to my existing list.

- POSITIVE correlations:
  - All baths: For every 1 change in the number of baths, there is an increase in price by $39,726 with all other things being equal.
  - Northridge Heights neighborhood: Increase in value by $31,368.
  - Basement Exposure: a house containing a good walkout or garden level basement increases in value by $21,732.
- NEGATIVE correlations:
  - Area per room: there is a slight positive correlation between the number of rooms in a house and sale price, but it seems like massively large rooms penalize sale price. This number could be refined.
  - House to lot ratio: the number of amenities afforded to a house provide many positive correlations, but it seems that too big of a lot to house ratio decreases house prices. This number could also be refined.

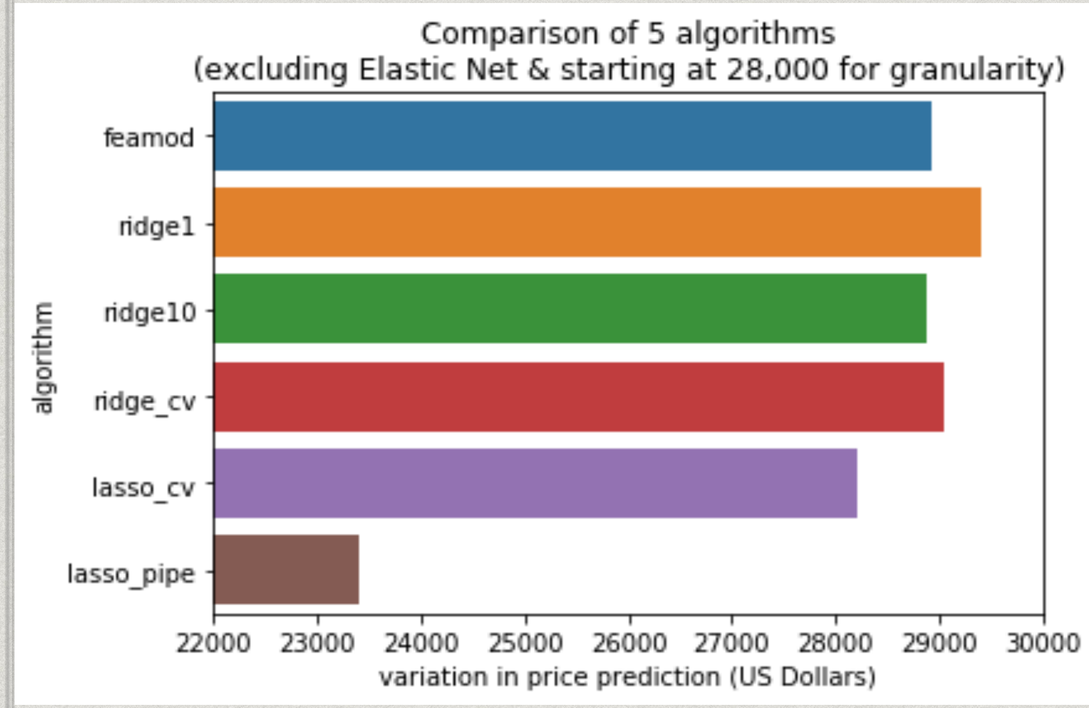
	col_name	coef
26	allba	3.972661e+04
20	Neighborhood_NridgHt	3.136874e+04
25	Bsmt Exposure_Gd	2.173249e+04
0	Overall Qual	1.667417e+04
7	TotRms AbvGrd	1.418225e+04
18	Foundation_PConc	9.408607e+03
3	Garage Cars	7.582658e+03
8	Fireplaces	4.618848e+03
23	Fireplace Qu_Gd	3.089238e+03
24	Fireplace Qu_Gd	3.089238e+03
27	beba	4.476283e+02
2	Garage Area	2.258713e+01
1	Gr Liv Area	1.466798e+01
4	Total Bsmt SF	9.328081e+00
5	1st Flr SF	2.871668e+00
29	fin_sqft	-2.381858e+01
16	Mas Vnr Type_None	-1.021363e+02
19	BsmtFin Type 1_GLQ	-2.376615e+02
6	Full Bath	-8.097526e+02
13	Fireplace Qu_NA	-8.135033e+02
12	Fireplace Qu_NA	-8.135033e+02
15	Garage Finish_Unf	-2.038226e+03
14	Bsmt Qual_TA	-2.874281e+03
17	Garage Type_Detchd	-4.339246e+03
11	Kitchen Qual_TA	-7.338118e+03
22	Exter Qual_Gd	-1.764940e+04
21	Exter Qual_Gd	-1.764940e+04
10	Exter Qual_TA	-2.053165e+04
9	Exter Qual_TA	-2.053165e+04
30	house_to_lot_ratio	-4.038322e+04
28	area_per_room	-1.068636e+07

MODELS

A COMPARISON OF FINER TUNED MODELS

Which model did best?

- Tried several models with differing regularization techniques, pipelines and gridsearches.
- The best performing model was the a polynomial Lasso CV pipe.

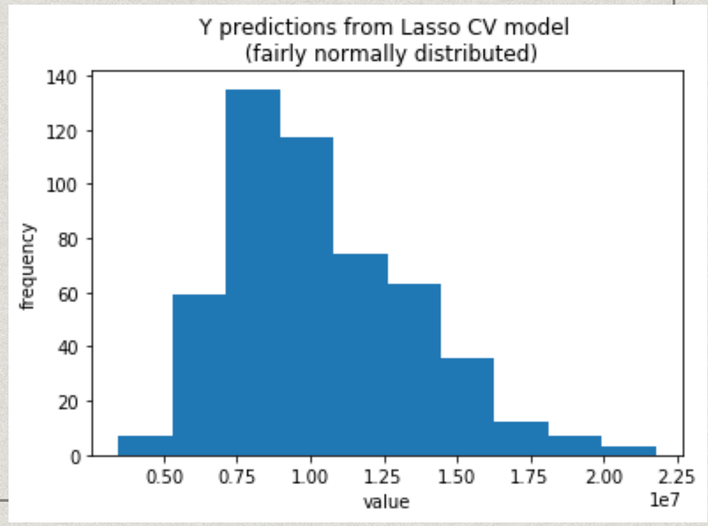


The Lasso CV Pipeline

- The pipeline that performed the best had:
    - Polynomial Features
    - Used the Lasso CV( )
- The test score without Polynomial Features was 0.872 but WITH Polynomial Features it had a test score of 0.912 and a low RMSE.
- But has problems:
    - There were many Convergence Warnings
    - The train score at 0.921 was higher than the test score
    - Meaning = The polynomial features has probably caused overfitting (too much variance)

The Lasso CV with no Polynomial

- The model that out-performed the Multiple Linear Regression and did not exhibit bias nor variance was the Lasso CV.
- 86.7% of the variation in y can be explained by the chosen X variables.
- The predictions were off, on average, by $28,756 which is low considering the range of prices ($12,789 and $555,000).
- The predicted values also are fairly normally distributed for real-world data.
    - It was a stable model with no errors.
    - Changing the parameters did not improve the score.



Improvements?

The model can be improved by engineering more exact features representation

- Finished vs unfinished square feet.
- Better representation of baths vs rooms in a house.
- Deeper GIS evaluation of the neighborhoods.
- More refined engineered columns to reduce clutter.