Predicting Sale Price of Homes

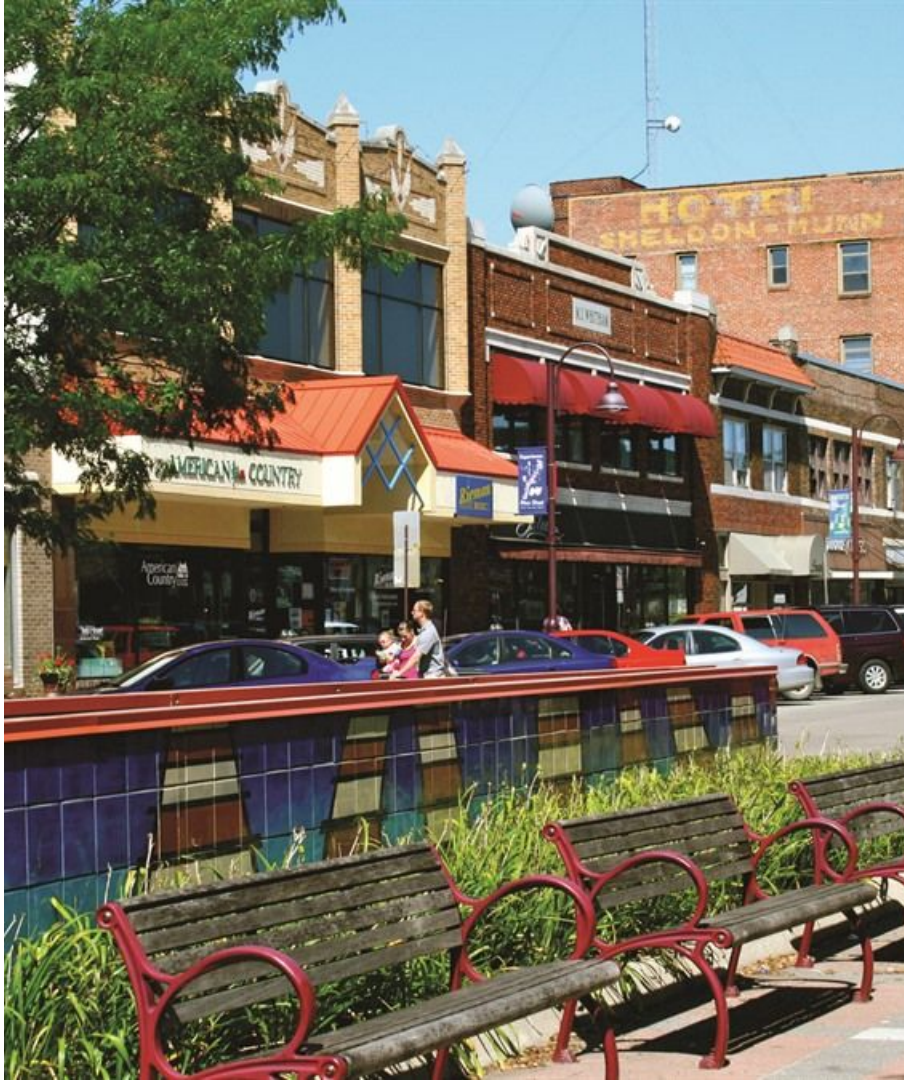
Using the Ames Iowa Housing dataset

Bryce Gillespie

The problem

Using the Ames Housing dataset, can we build a regression model that is predictive of home value?

Data Source



The methodology

- Exploratory data analysis
- Handling missing values and outliers
- Data transformations and scaling
- Model selection and cross validation
- Generating predictions

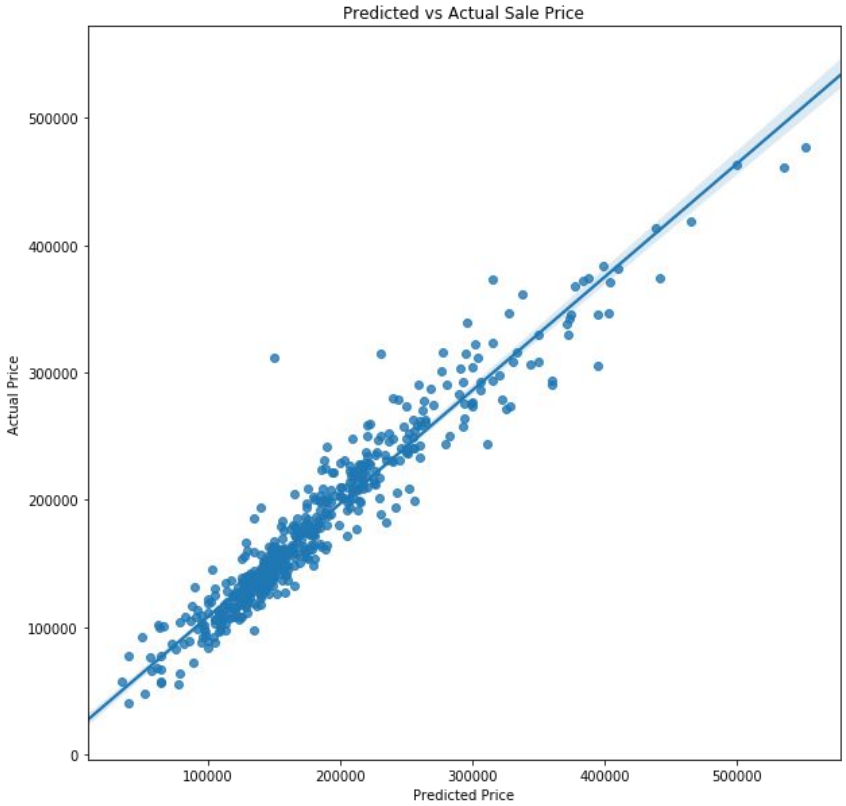
Key findings

- Intuition only gets us so far with this dataset
- Many features had small but significant correlation with Sale Price
- Lasso Regularization provided the best R2 score

Results

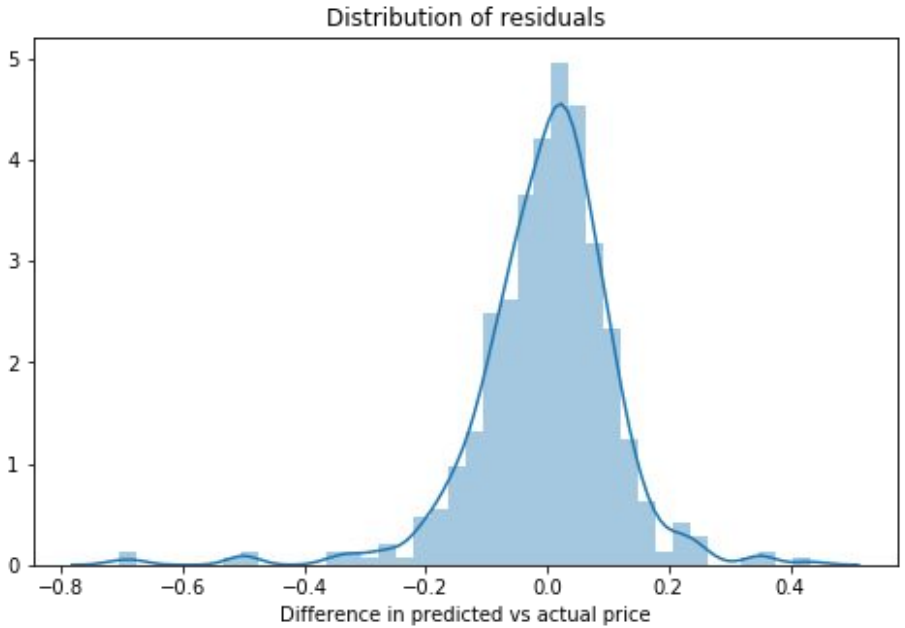
- 105 features used after the Lasso coefficient penalty

- R2 score of 0.93
  - “**93%** of the variability in our dataset can be explained by our model”



Residuals

- The distribution of residuals from the holdout set had a fairly normal distribution centered around zero with a few outliers.



Coefficients

- Coefficients of the top ten features, sorted descending.

- Intuitively, we see total square feet and overall quality at the top.

Feature	Description	Coefficient
gr_liv_area	Total Size (sq. feet)	0.119911
overall_qual	Overall Quality (1-10)	0.081587
total_bsmt_sf	Basement size (sq. feet)	0.047703
lot_area	Lot size (sq. feet)	0.040154
year_remod/add	Year remodeled	0.036557
ms_zoning_C (all)	Zoned as 'Commercial'	0.021261
bsmtfin_sf_1	Basement area (sq. feet)	0.021160
functional_Typ	Typical functionality	0.020425
condition_1_Norm	Normal condition	0.019918
year_built	Year Built	0.019695