Price Prediction of Ames Homes

Alex Walsh

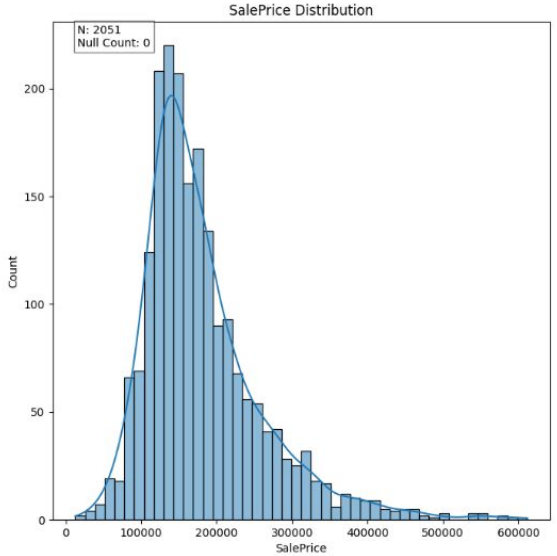
Problem Statement

- Many factors influence the price of a home, making it a difficult quantity to assess
  - Realtors can struggle to create an appropriate listing price
  - Buyers have no way to know if a home priced fairly

- Linear regression techniques can be used to estimate price of homes

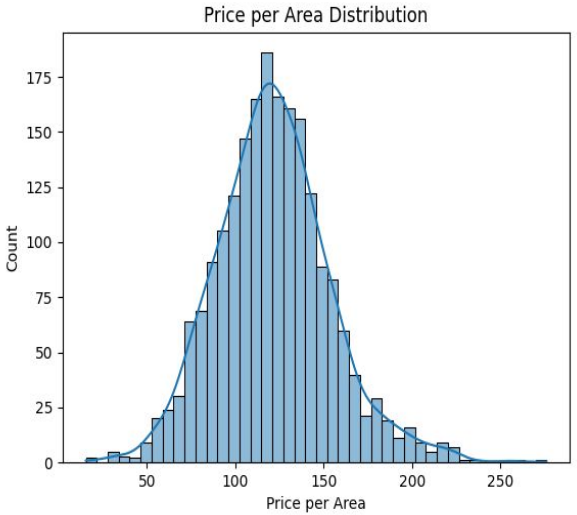
The Data

- The dataset is properties sold in Ames, Iowa
- The data is from 2006 to 2010
- Data contains both numerical and categorical features



Alternative Target: Price per Area

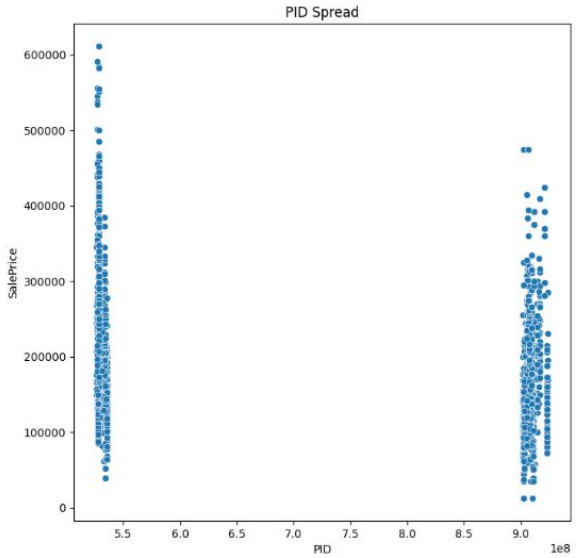
- Sale price may not be the ideal target for regression
- Property size can make other features harder to learn
- In addition, I tested models which tried to predict price per unit area



Data Splitting

- Certain features can be used to define two “types” of data points

- Each type of data point can be fitted with its own linear model



The Approach: Grid Search

- Model performance was used to determine what choices to make with the data
- 51,840 different preliminary models were tested
- Power transformations were applied to every feature
- Feature selection was done mostly automatically

$$y_i^{(\lambda)} = \begin{cases} ((y_i+1)^{\lambda}-1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y_i+1) & \text{if } \lambda = 0, y \geq 0 \\ -((-y_i+1)^{(2-\lambda)}-1)/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y_i+1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

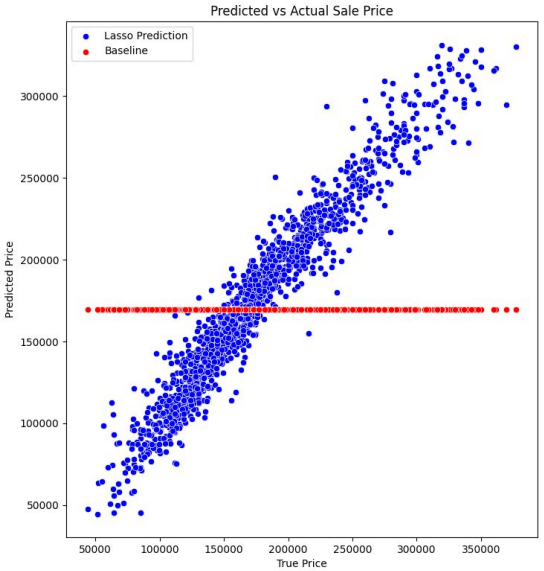
Results

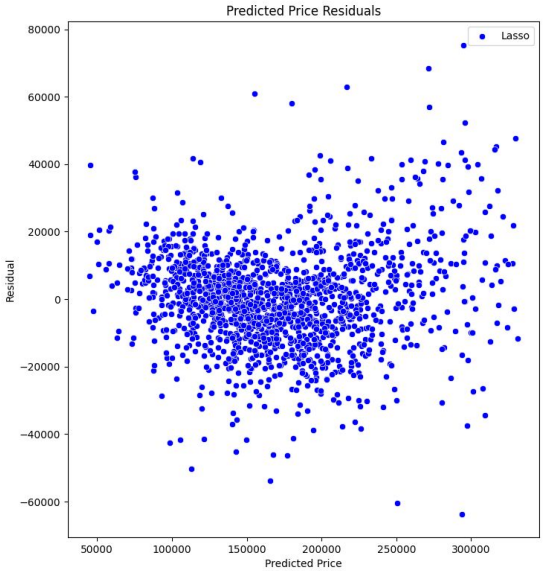
Lasso(λ=0.00497)

R^2 = 0.937

RMSE = $14, 401

Baseline RMSE = $57,178





Summary

- Many different hyperparameters were tested and the Lasso model was chosen as the best
- Data splitting led to better training fits, but did not generalize well
- Price per area was a better target on average, but the best models used price