Project 2 - Ames Housing Data

By Afolabi Cardoso

Problem Statement

I am a Data Science Phd student at MIT and my Supervisor asked me to determine the best Linear Regression technique to use in determining the Sale Price of houses in his home town of Ames, Iowa.

Overview

- Problem Statement
- Dataset
- OLS vs Ridge Regression vs Lasso
- Data Cleaning and Feature Engineering
- EDA
- Model Evaluation
- Recommendations and Conclusion

Dataset

Loction	Number of features	Samples
State Office Ames, Iowa	**79**	**2050**

Data Cleaning and Feature Engineering

Features Dropped

'Lot Frontage'

'Alley'

Fireplace Qu'

'Pool QC'

'Fence','Misc Feature'

Top Features

'Overall Qual'

'Gr Liv Area'

'Garage Area',

'Total Bsmt SF'

'1st Flr SF'

'Year Built'

Dummied Features

42

Models

OLS

Minimizes the MSE

LASSO

Penalized for the sum of absolute values of the weights

$$\frac{1}{2m}\sum_{i=1}^{m}(y - Xw)^2 + alpha\sum_{j=1}^{p}\left|w_j\right|$$

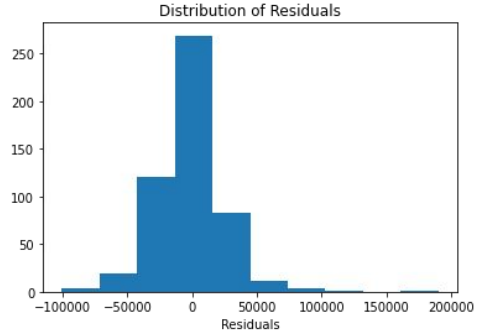
RIDGE

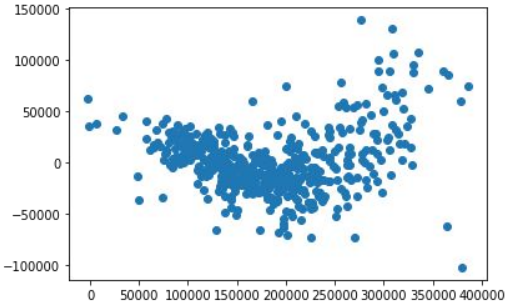
Penalizes the model for the sum of squared value of the weights

$$\sum_{i=1}^{n}(y - Xw)^2 + alpha\sum_{j=1}^{p}w_j{}^2$$

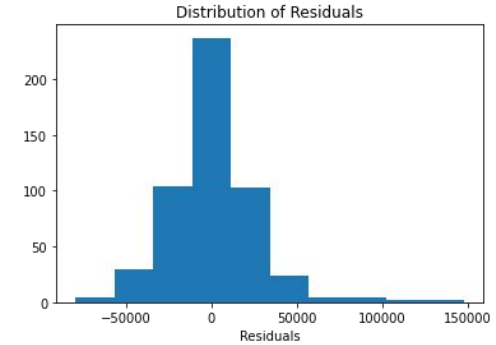
Distribution

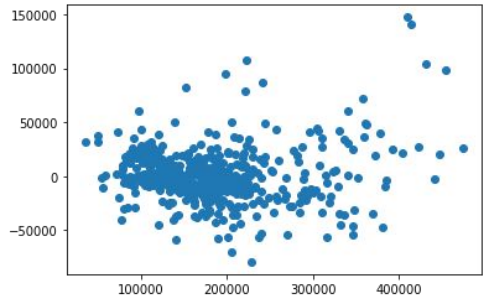
OLS



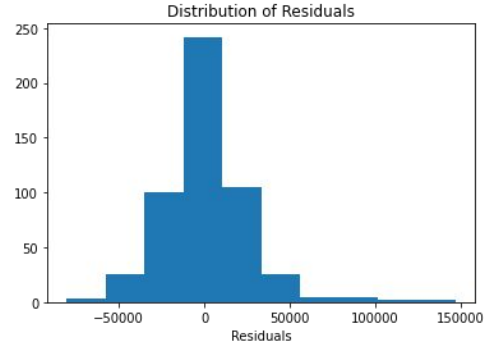


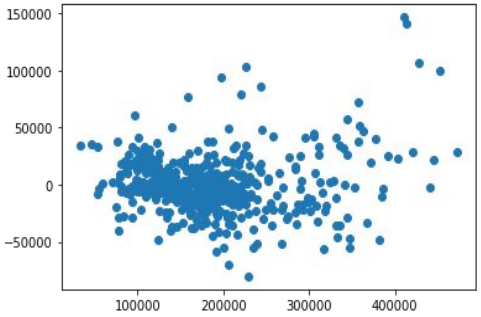
Lasso





Ridge





Evaluation

	OLS (numerical)	OLS (dummied)	Lasso	Ridge
Train set R2	80%	89%	89%	89%
Test set R2	85%	88%	90%	91%
Cross evaluation R2	74%	79%	80%	82%
RMSE	30138	26437	25212	25024

Conclusion

- The Ridge Model performed best overall
- The Lasso had similar results
- OLS performed better after dummying the categorical features

Recommendation

- I recommend to use either the Ridge or the Lasso model because they both gave a better overall result.

- As more data is collected the model will be reevaluated to see if the results still holds true.

“Data Science is an Art.”

—Riley