LinReg, Ridge, Lasso Score Comparison with Ames Housing Dataset

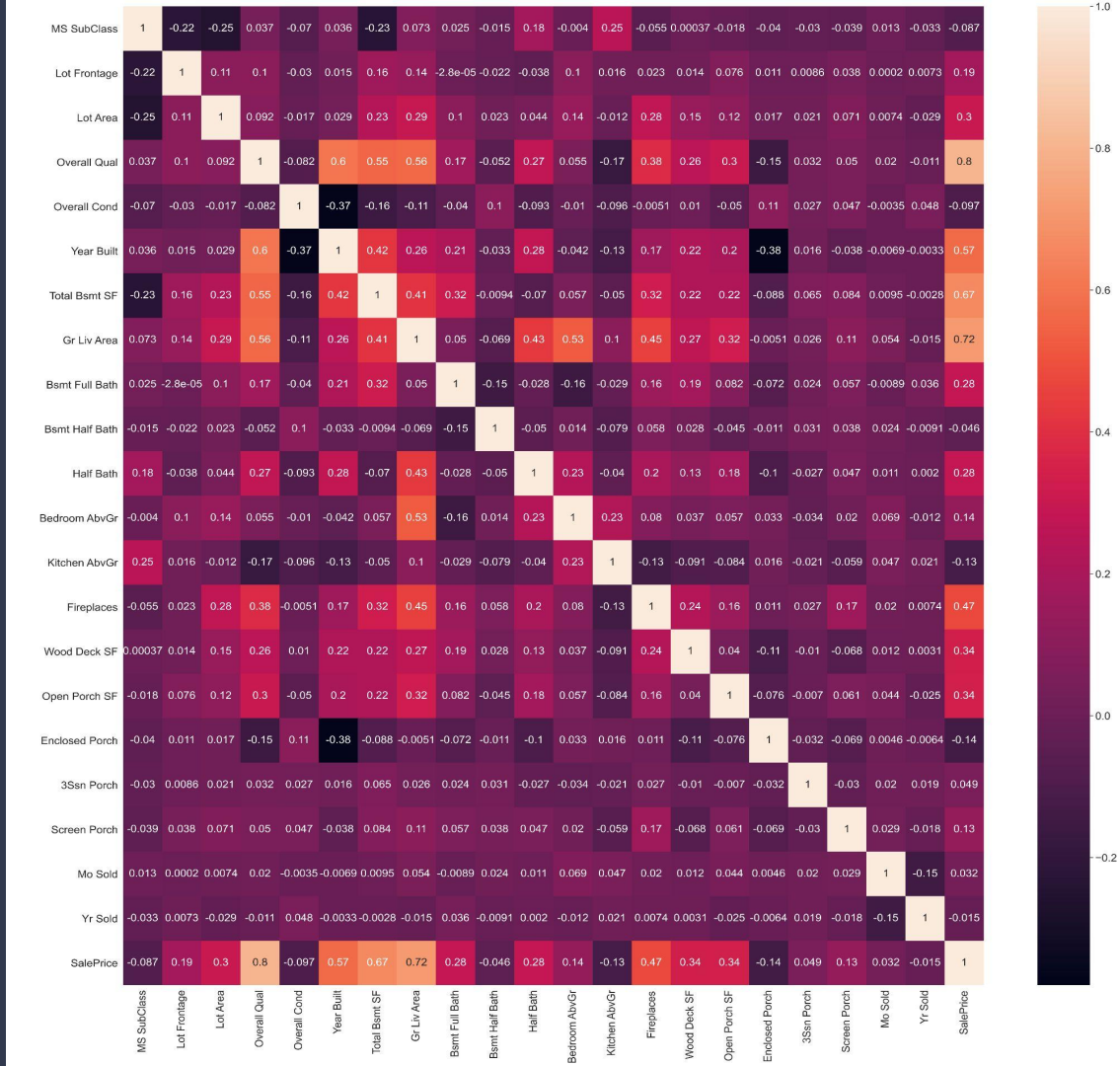
Julia Tsaltas
July 2nd, 2021

Outline

- Ames Housing Data
- Engineered Features
- Sale Price Prediction using Linear Regression, Ridge and Lasso

Ames Housing Data

- 74 features total
- Numeric features:
  - square footage (living areas, basement, garage, porch)
  - number of key features (bathrooms, fireplaces)
- Categorical features:
  - Quality rankings
  - House and building features (materials, heating type, neighborhood)



Engineered Features

Grouped category types that had small counts.

- Lot shape ➜ Combine any irregular shape
- Condition 1,2 ➜ Combined proximity to railroads
- Year Remod/Add ➜ Categorized to “Built/Remod < 10 yrs”, “Built/Remod 10-20 yrs”, “Built/Remod > 20 yrs”
- Roof Material ➜ Combined non composites
- Masonry Type ➜ Categorized to Has and None
- Electrical ➜ Combined fuse types
- Low Quality SF ➜ Categorized to “Finished”, “UnfinishedSF < 500”, “UnfinishedSF > 500”

Model Results - Using All Data

Linear Regression	RidgeCV	LassoCV
82.38% - Train	**83.70%** - Train	**89.85%** - Train
87.42% - Test	**89.95%** - Test	**90.41%** - Test
	alpha = 126.5	alpha = 424.3

LassoCV

- ***Slightly higher performance than baseline***
- ***Some overfitting - Low bias, high variance***

Model Results - Using Reduced Data

Linear Regression	RidgeCV	LassoCV
89.96% - Train	**90.21%** - Train	**90.10%** - Train
87.41% - Test	**91.75%** - Test	**91.95%** - Test
	alpha = 39.9	alpha = 1

After reducing features of high correlation the performance improves for all models, most notably for linear regression and ridge.

Model Results - Lin Reg With PCA

PCA All Features	Reduced Features	All Features
93.74% - Train	**89.96%** - Train	**82.38%** - Train
91.01% - Test	**87.41%** - Test	**87.42%** - Test

Best results with PCA!

100 components
82% cumulative explained variance

Model Improvements

Improve model accuracy:

- Continue feature engineering for numerical variables with PCA
- Try other regression models