Linear Regression + Cross Validation

Hayden Hawley

Overview

Goal: to predict house prices using the House Prices Kaggle dataset. My pipeline had to address several challenges:

- Every column has missing values, so I needed an imputer.
- Dataset contains roughly 80 features, and I needed to figure out if they were all important.
- Features are a mix of numerical and categorical values, so I needed to use one-hot encoding.

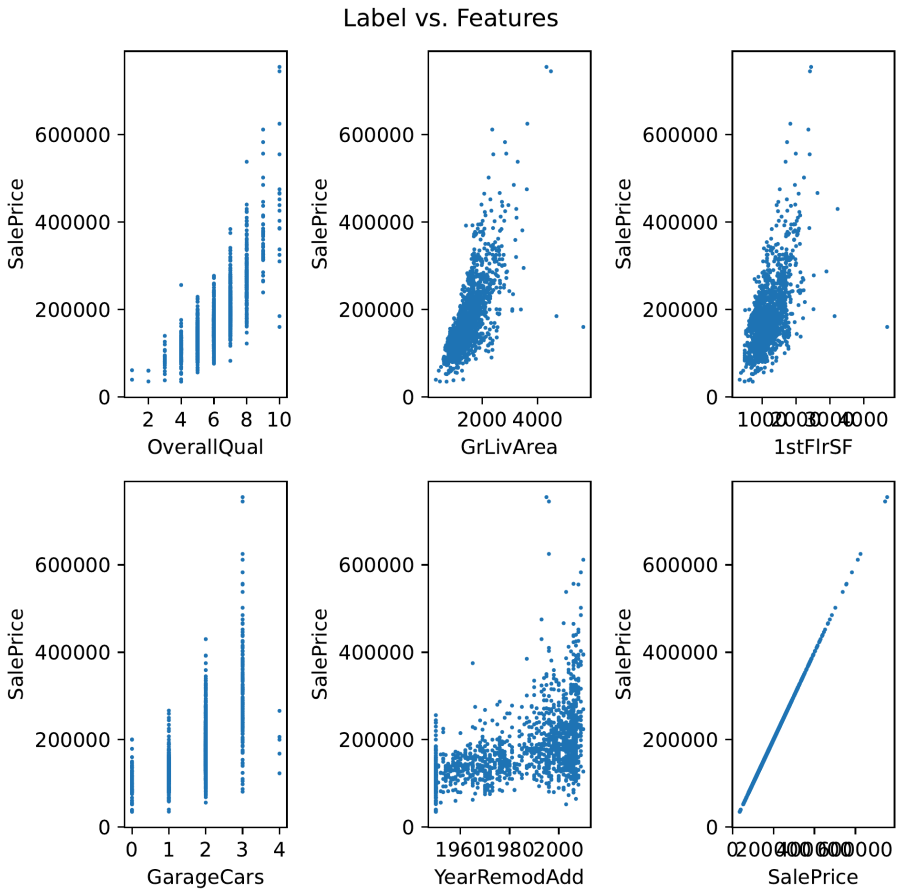
I also used cross validation and a method to output a prediction file to submit to the Kaggle competition.

I started with SGD as a baseline and then experimented with Ridge regression. I also applied a log-transformation to the target variable to try and mitigate outliers at the very high end of the price range.

Data Exploration

I chose a few seemingly important features to visualize:

- OverallQual – The overall material/finish rating (scores 1–10).
- GrLivArea – Above-grade (ground) living area in square feet.
- 1stFlrSF – First floor area.
- GarageCars – Size of the garage in car capacity.
- YearRemodAdd – The year the house was last remodeled.



Upon further investigation, I learned that a small number of extremely expensive homes seemed to be skewing the results. I filtered out homes with a SalePrice over $500,000 and then applied a log-transformation to the target, which improved performance by a decent amount.

Model Selection

These were my initial hyperparameters:

Missing value imputation: Median

Polynomial expansion: Degree 1

Scaling: True

SGD Regression with only a handful of features was quite poor at first:

R2: 0.71769

MSE: -1,769,916,693

MAE: -26,914

Adding all 80+ features improved results significantly:

R2: 0.82694

MSE: -1,104,605,257

MAE: -19,293

Switching to Ridge (alpha=1.0) didn't do as much as I'd hoped:

R2: 0.80893

MSE: -1,153,330,671

MAE: -18,248

But I noticed an outlier when I switched to 7 cross-validation folds:

R2: [0.88879703
0.89844166
0.81766405
0.86697165
0.88238054
0.83223409
0.47604398]

I decided to exclude all homes with a sale price of over half a million dollars, as well as predict the log of the sale price. These were my hyperparameters using the Ridge model with these adjustments:

Missing value imputation: Median

Polynomial expansion: Degree 2

Scaling: True

And these were my best results:

R2: 0.83911

MSE: -835,478,964

MAE: -16,680

Evaluation

The model still struggles with very high prices (the folds are grouped in ascending order of sale price):

Fold	R^2	MSE	MAE
1	0.8888	-587,971,247	-16,927.55
2	0.9054	-565,844,042	-16,745.44
3	0.8347	-1,012,236,900	-16,792.03
4	0.8587	-807,592,443	-18,231.91
5	0.8973	-386,801,293	-14,585.48
6	0.8701	-513,658,017	-16,174.52
7	0.6198	-1,974,248,810	-17,309.81
Mean	**0.8391**	**-835,478,964**	**-16,680.82**