
1. Data Cleaning
2. Feature Engineering
3. Model Selection
4. Pipeline
5. Predictions
6. Present/Interpret Results
Sean Cleary
THE PROCESS

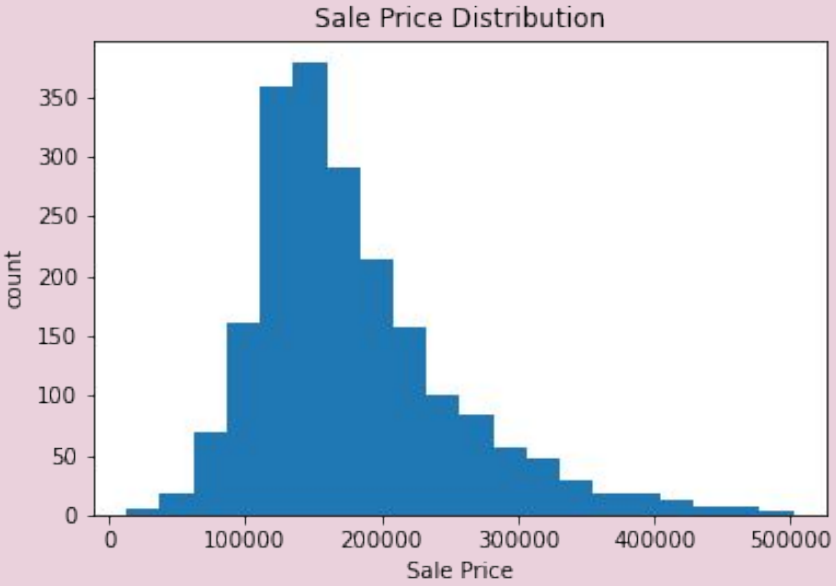
The Data

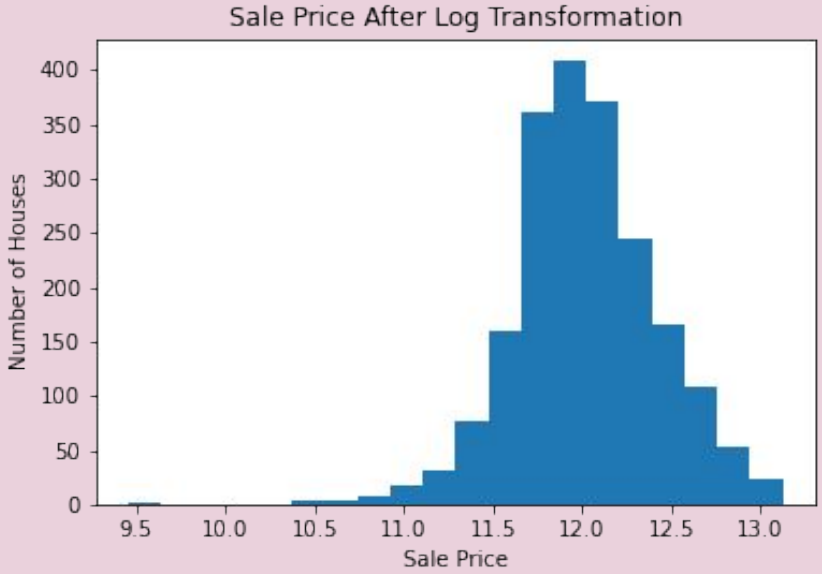
- Two datasets for Training and Test that need to be combined into 1.
- Train consisted of 2051 rows and 82 columns.
- Test consisted of 878 rows and 81 columns.
- The features consist of 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, in addition to 2 ID variables

Visualize Distributions

Outliers that need to be removed.

Feature Engineering

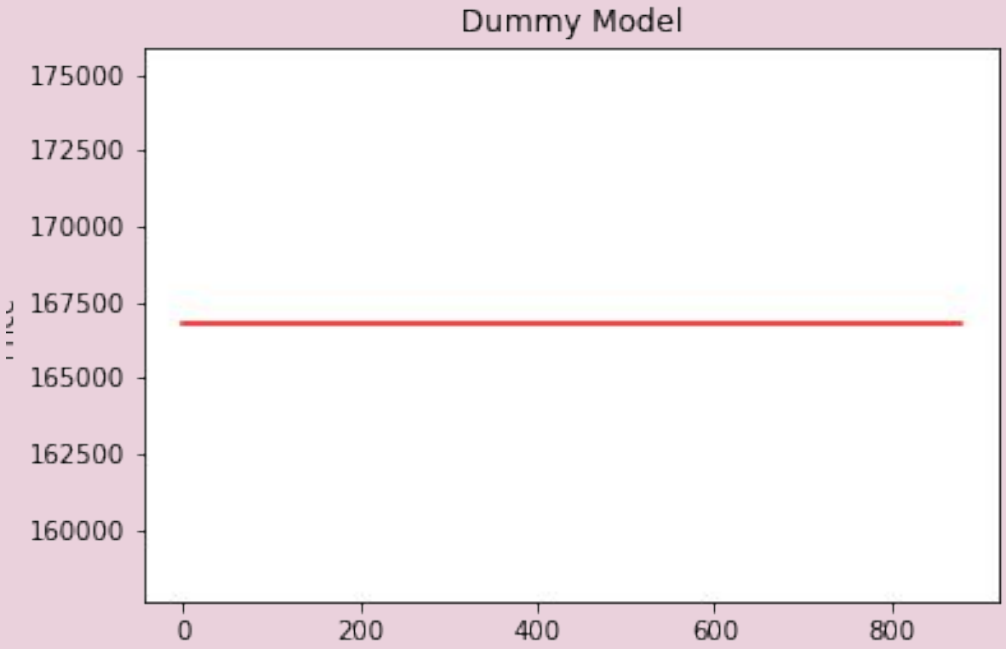




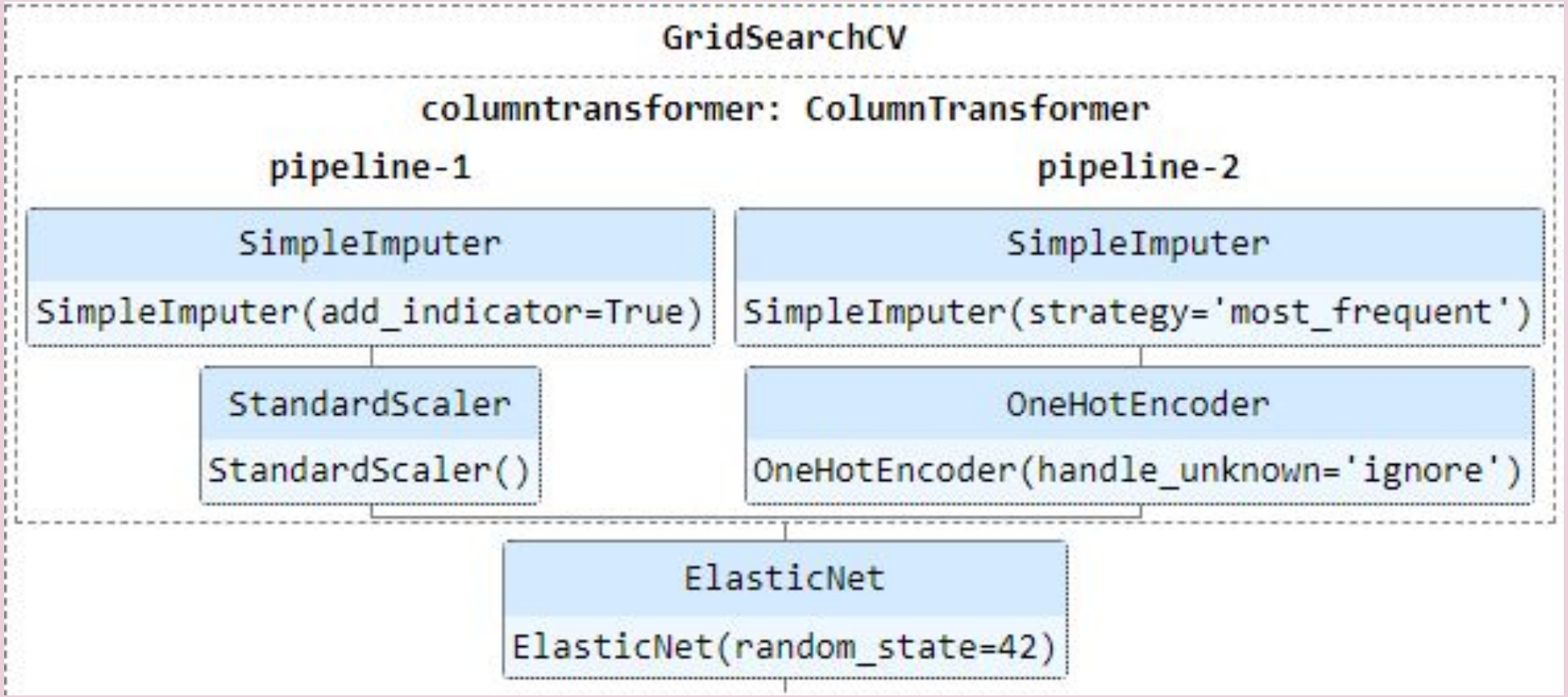
Baseline Model

Dummy RMSE:

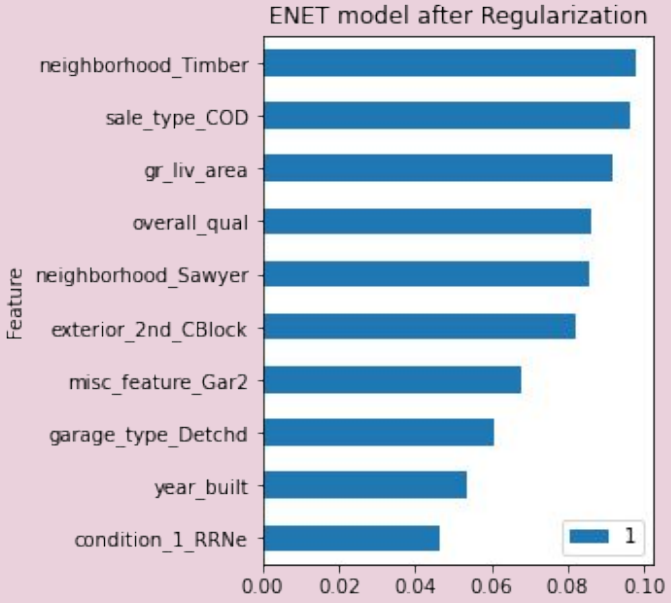
79050.31

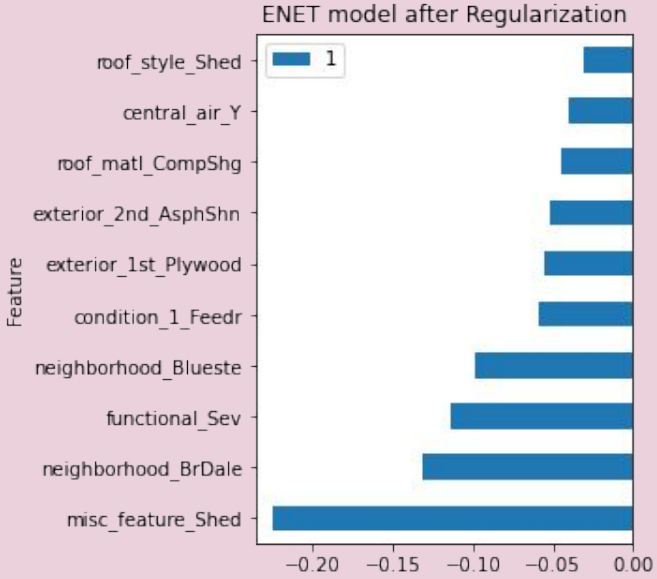


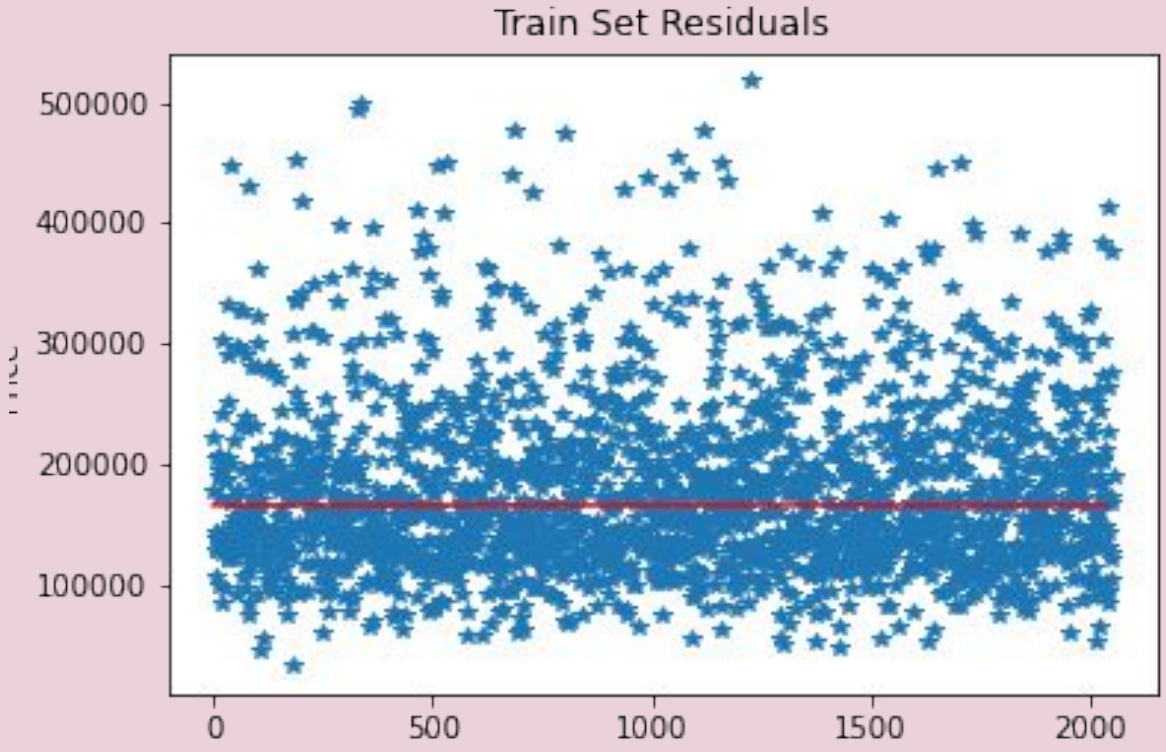
Pipeline Structure



Feature Importance






Train Set Residuals
500000
400000
300000
200000
100000
0
500
1000
1500
2000

Recommendations

-LOCATION, LOCATION, LOCATION

-Overall Quality and Gross Living Area

-Remove the shed from your property.

- Sawyer and Timberland are the best neighborhoods to live in.

- Avoid Briardale and Bluestem neighborhoods.

What's Next?

- Remove all foreclosures from dataset
- Impute constant values for most categorical features missing values
- Ensemble multiple models together to weight predictions.