Project 2

Predictive Modeling of the Ames Housing Data

Matt Devay, ATX-DSI-11

Problem:

To develop a predictive model of the Ames Housing Data:

Background

Two datasets were provided. The training set had all features including sale price. The test set had all features except for sale price. The goal was to develop a model that predicted the sale price in the test set.

Data Workflow

1. The data were imported via Pandas
2. Data cleaning was performed. Minimal cleaning was required with this dataset.
3. A correlation heatmap and basic model were generated
4. Feature generation was performed via polynomial expansion of the continuous features.
5. Categorical/Discrete features were One-Hot encoded and added individually.

Polynomial Expansion

1. The correlation heatmap indicated approximately 10 continuous variables with > 0.3 correlations to sale price.
2. A polynomial expansion was performed on those features
3. A LASSO CV regularization was performed on the polynomial features.
4. Features with > 0 betas were retained
5. Gross square footage was the most influential of these.

Categorical Variables

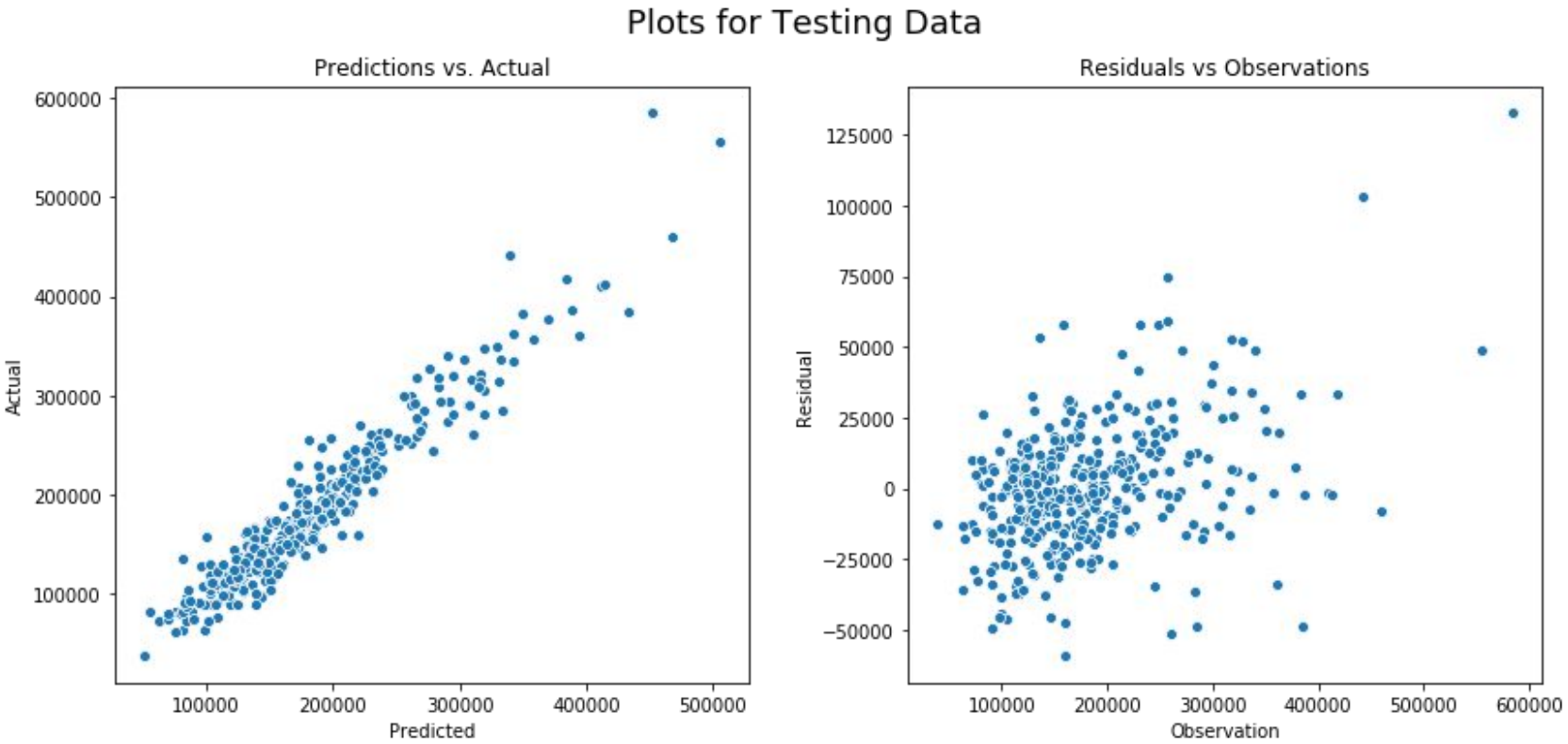
1. Categorical/ordinal variables were selected based on correlation.
2. Each were added via an automated function and analyzed individually
3. Overall Quality was the most influential of these.
4. Almost all categorical/ordinal variables improved the model, but interacted in unpredictable ways, leading to minimal gains if included.

Final Metrics

1. Training R^2 was approximately 0.92
2. Testing R^2 was approximately 0.91
3. 10 fold cross validation of the training data was approximately 0.89

Residual Analysis

Residual Analysis Cont



Conclusions:

1. Square footage and perceived quality are the dominant determining features

2. Outliers are common. One-Hot encoding of more esoteric features such as Pool QC helped rein in some of them, at the expense of higher variance in the model
3. Inclusion of Sale Condition would likely improve the model.