Use Linear Regression to Predict House Price Based on Auto Selected Parameters from Known Dataset

- Andy Xu

GENERAL ASSEMBLY

Project Statement

Problem:

- When predicting house price, many parameters could influence the price. What parameters should we select?

Project Statement:

- The project utilized linear regression method to predict the house sale price by analyzing the existing information that provided house price and selected highly correlated factors with the price.

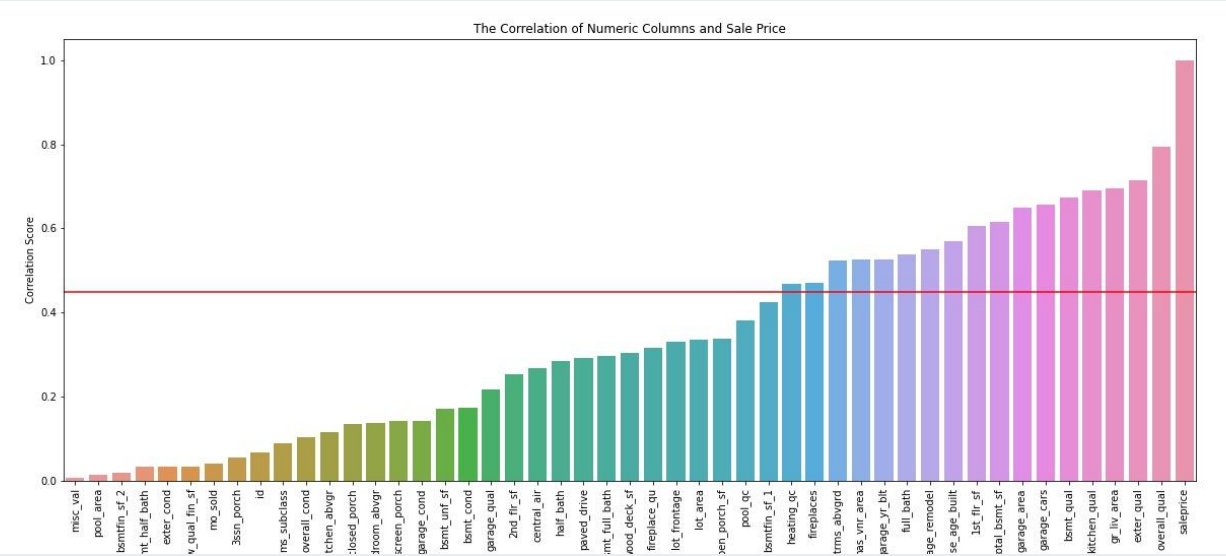
Basic Approach

Data Source: Ames, Iowa Assessor's Office

- Two Dataset: Train and Test
  - Train dataset: Create linear model.
  - Test dataset: Predict house price with the linear model.
- Train dataset was splitted into 2 parts
  - Train (80%): Create linear model
  - Validation (20%): Evaluate the model accuracy
- Data Cleaning, Feature Engineering, Preprocessing
- Linear Regression
- Model Evaluation

Data Cleaning: Numeric Columns

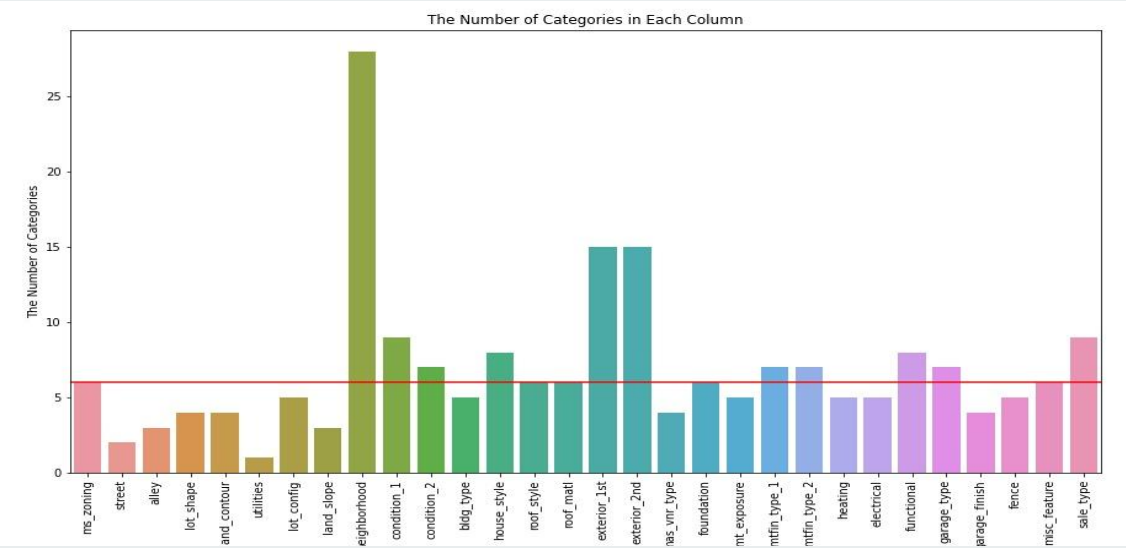
- Columns were splitted into numeric and object segments.
- Object columns only contained ratings were converted into numeric.
- The correlation score of each column versus sale price was calculated.



Filtration Rule:

1. Columns had correlation score over 0.45.
2. Columns had less than 5% of the missing values.

Data Cleaning: Object Columns



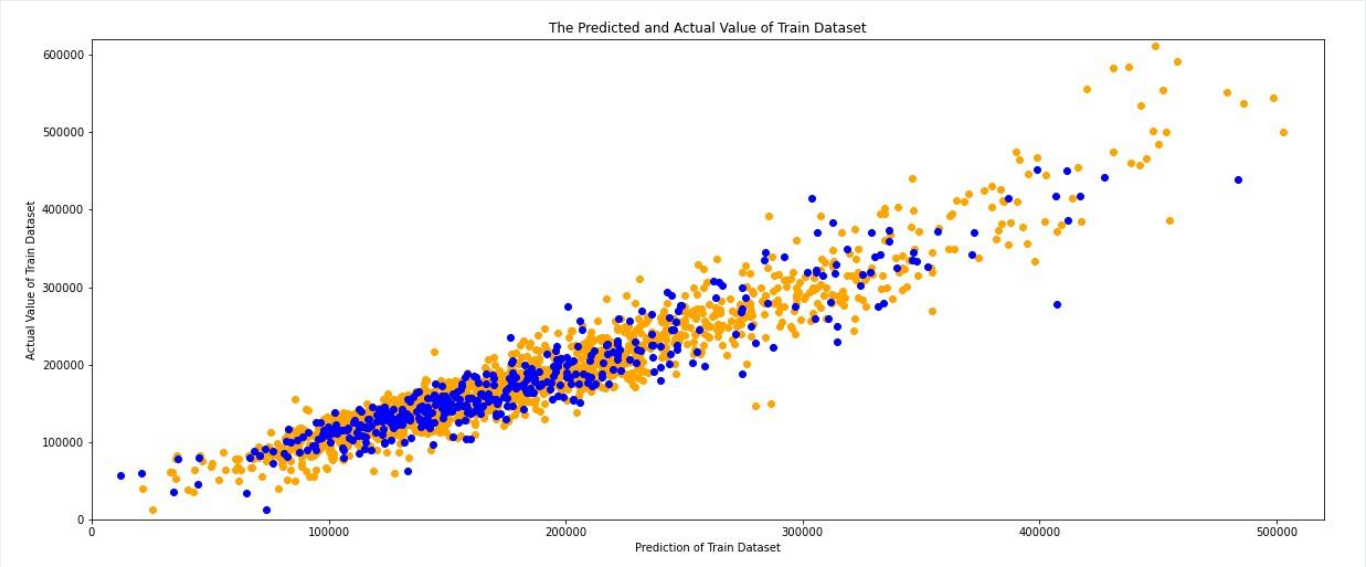
Filtration Rule:

1. Columns had no more than 6 categories.
2. Manually added the column of neighborhood

Next Step:

- Combine numeric and object columns.
- Apply feature engineering and preprocessing, include Simple Imputer, OneHotEncoder and Standard Scaler

Linear Regression



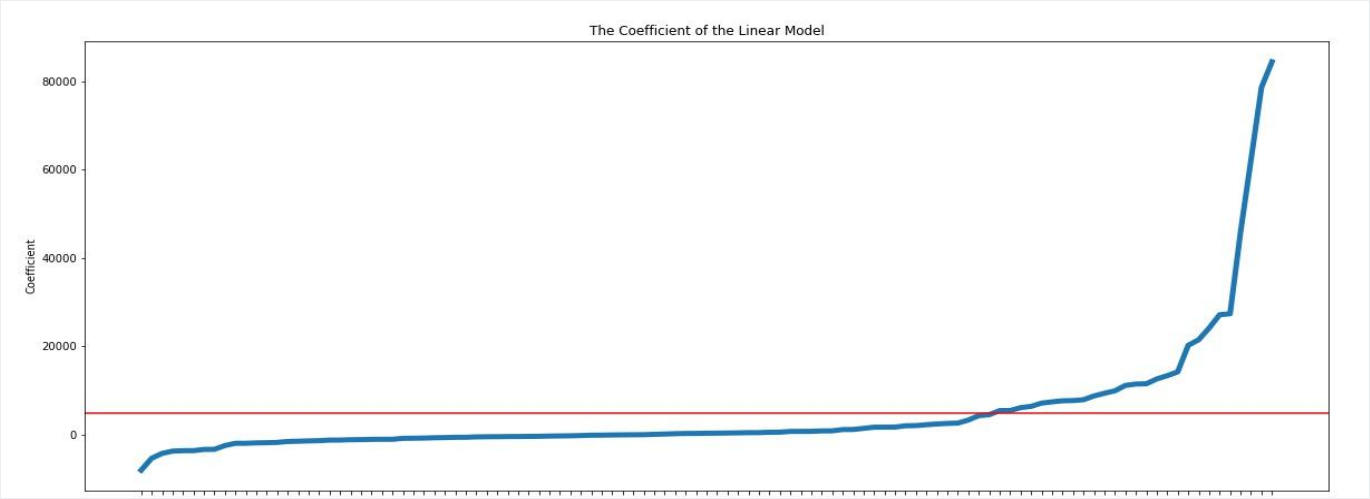
	Train (orange)	Validation (blue)
R^2 score:	0.908	0.888
Root Mean Squared Error:	24,142	25,842

Low bias, Low Variance

Linear Regression

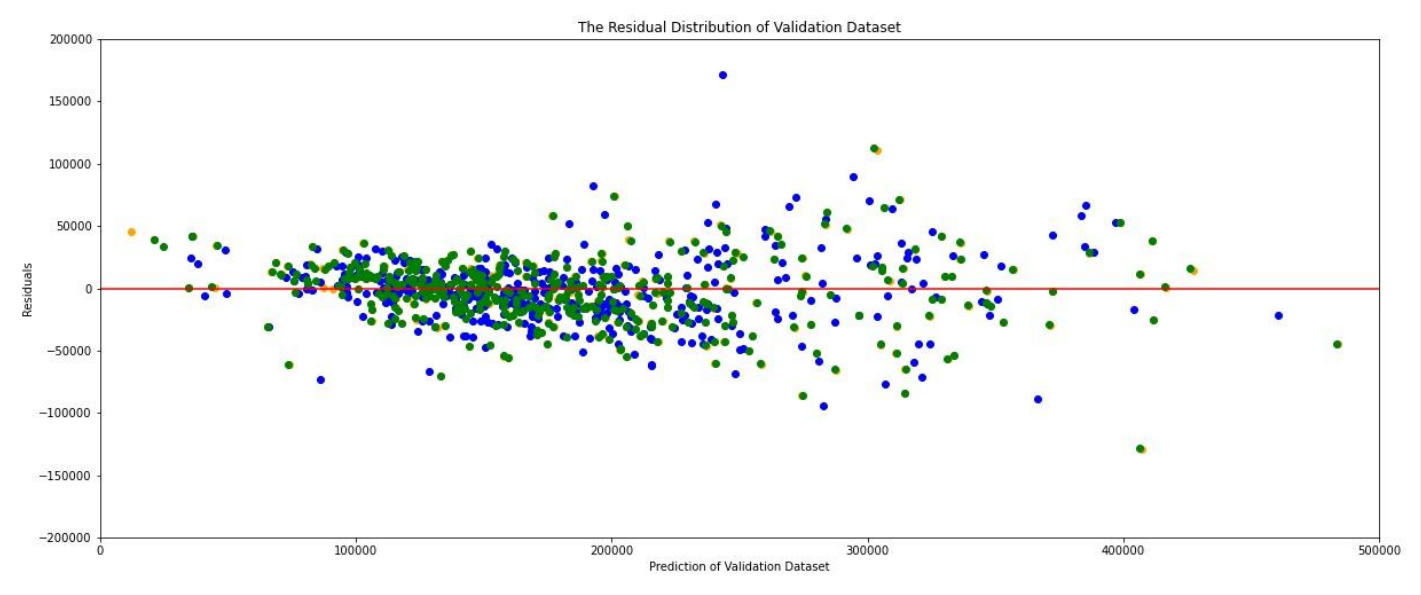
- Low residuals between the range of $100K to $250K .
- The distribution of residuals was close to normal distribution.

Linear Regression



The coefficients showed the scales of impact on house sale prices.
Top Positive Features: Miscellaneous features not covered, Roof materials, Above ground living area
Top Negative Features: Building type - townhouse, House age, Garage - rough finish

Ridge and Lasso



Original LR: orange, Ridge: blue, Lasso: green

Conclusion and Recommendation

- Linear model can predict!
- Prediction accuracy depends on the selected feature parameters.

Suggestions:

- **For sale**: roof material, masonry veneer area, kitchen, basement, heating system
- **For investment**: northridge heights, stone brook, green hills, northridge, crawford

Thank You for Listening

Questions?