Modeling Housing Prices in Ames, Iowa

Author: Michael Riedeman

Overview

- Problem Statement
- Feature Overview
- Baseline Model
- Modeling Process/Primary Findings
- Conclusions
- Recommendations
- Questions

Problem Statement

Create a linear regression model that maximizes the accuracy of predicting the price of a house at sale.

Feature Overview

81 features available to build the model

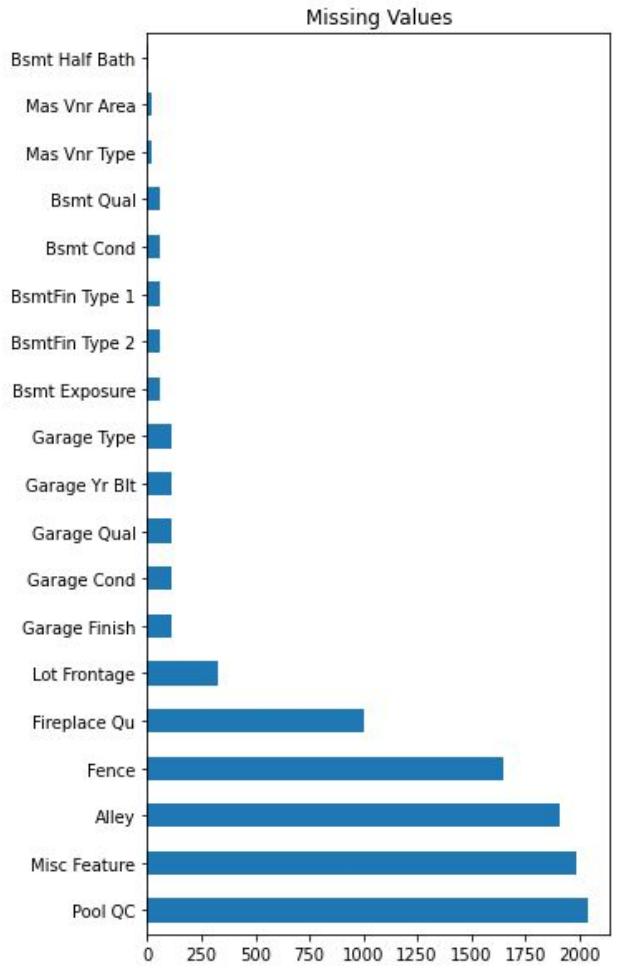
Ordinal data versus Nominal

Continuous and Discrete Data

Ordinal Features	Nominal Features
1. Utilities	1. PID
2. Land Slope	2. MS Subclass
3. Overall Qual	3. MS Zoning
4. Overall Cond	4. Street
5. Exter Qual	5. Land contour
6. Exter Cond	6. Lot Config
7. Bsmt Cond	7. Neighborhood
8. Bsmt Qual	8. Condition 1
9. Bsmt Exposure	9. Condition 2
10. BsmtFin Type 1	10. Bldg Type
11. BsmtFin Type 2	11. House Style
12. Heating QC	12. Roof Style
13. Electrical	13. Roof Matl
14. Kitchen Qual	14. Exterior 1st
15. FireplaceQu	15. Exterior 2nd
16. Garage Finish	16. Mas Vnr Type
17. Garage Qual	17. Mas Vnr Area
18. Garage Cond	18. Foundation
19. Paved Drive	19. Heating
20. Pool QC	20. Central Air
21. Fence	21. Garage Type

Feature Overview Continued...

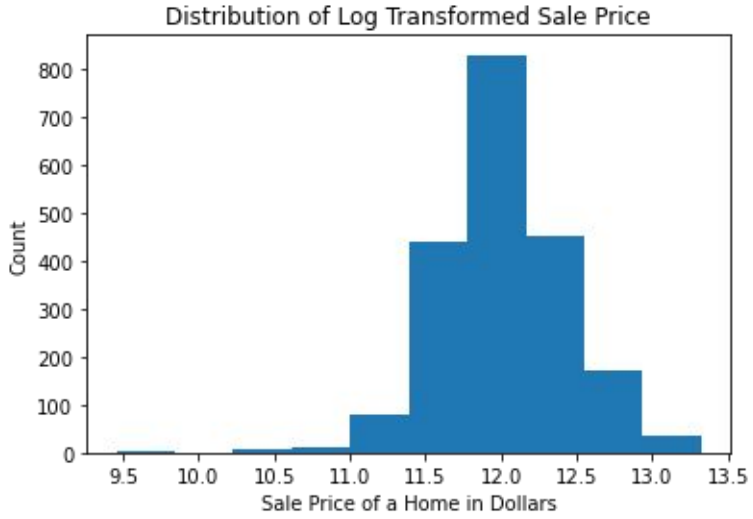
Missing Data/Null Values



Baseline Model

Average Price of Home: $180,904

Null MSE: 62,55,753,308



Modeling Process

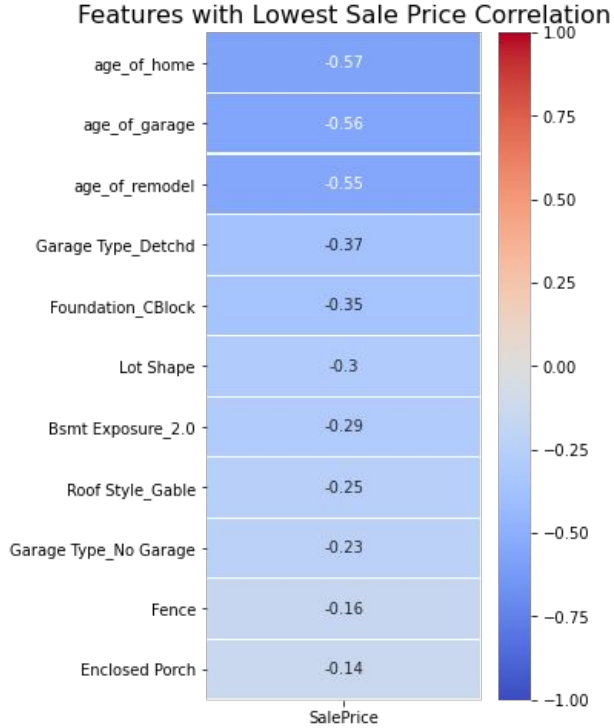
First Models:

Low bias, very high variance

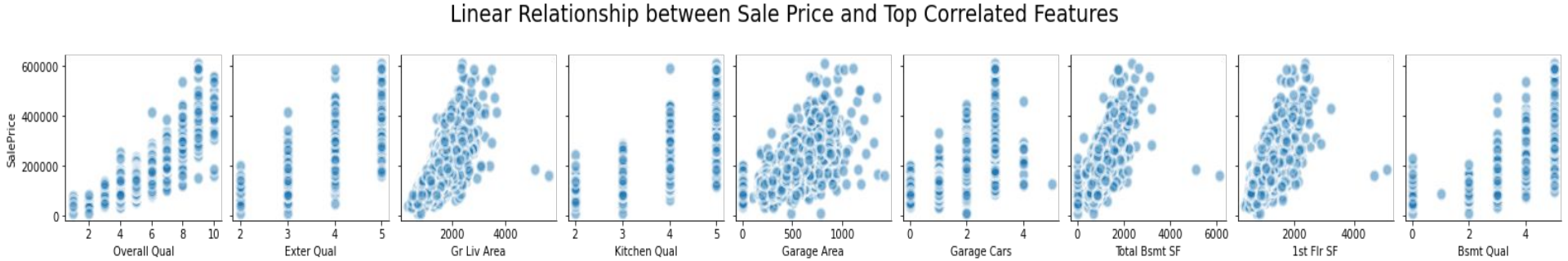
Used all possible features

Combinations of RidgeCV/LassoCV and StandardScaler

Modeling Process



Modeling Process

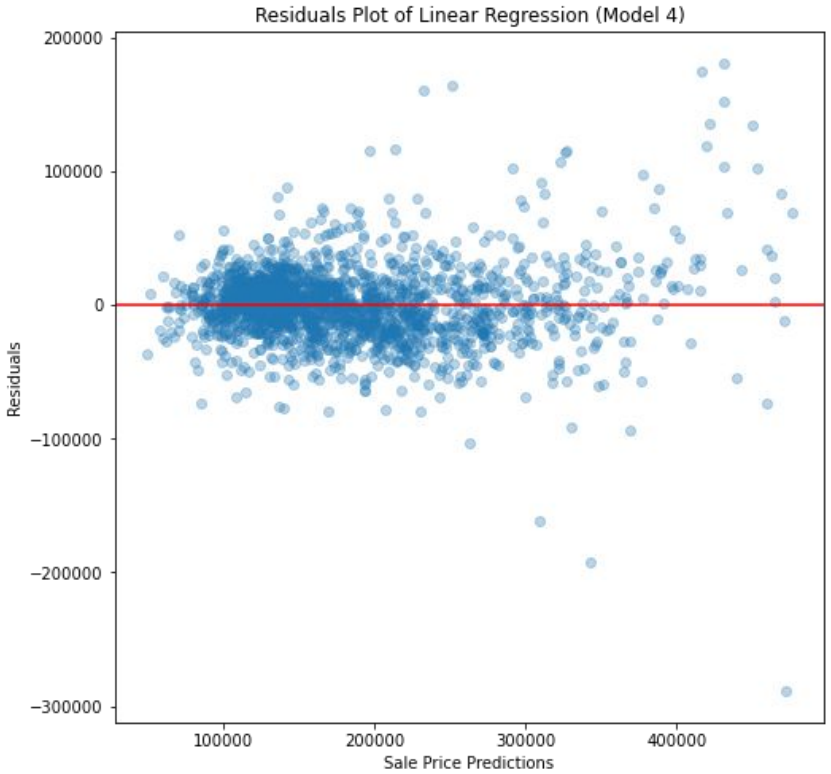


Primary Findings

Model Pipeline:

- Polynomial Features
- Standard Scaler
- Ridge Cross Validation

```
Train Mean Squared Error      : 689940168.7238624
Test Mean Squared Error       : 907058587.465089
Train Root Mean Squared Error : 26266.71217956032
Test Root Mean Squared Error  : 30117.413359468454
Train R-Squared Score         : 0.8856343001310519
Test R-Squared Score          : 0.8646291001336772
```



Conclusions

88% of the variability in sale price can be explained by the features in the model.

Feature interaction improved the models bias and eradicated the majority of the variance.

	Feature	Coefficients
30	Gr Liv Area Kitchen Qual	19909.745579
12	Overall Qual Gr Liv Area	18260.808656
57	Total Bsmt SF Bsmt Qual	17380.149910
16	Overall Qual Total Bsmt SF	14016.257713
21	Exter Qual Gr Liv Area	11803.422967
32	Gr Liv Area Garage Cars	11629.352244
17	Overall Qual 1st Flr SF	11237.237907
41	Kitchen Qual 1st Flr SF	10565.236284
52	Garage Cars 1st Flr SF	8696.365020
35	Gr Liv Area Bsmt Qual	8437.750249
38	Kitchen Qual Garage Area	8382.799746

Recommendations

The model would benefit from log transformations on the non-linear features.

Overall Quality of Facilities and Space contribute the most to sale price.

Model can be used on homes in Ames, Iowa priced less than $300,000 at this time.

Questions?