Regression Modeling with Ames, IA Housing Data

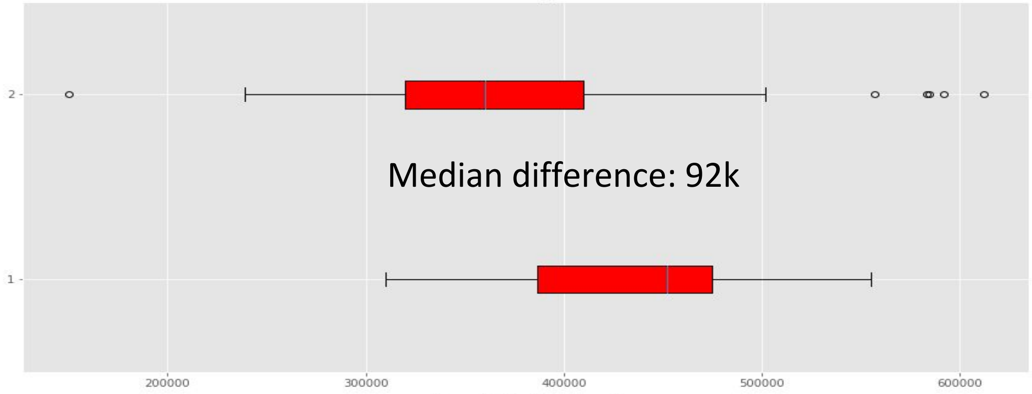
Eli Curme

Presentation Structure

- Takeaways from:
    - Cleaning
    - Feature engineering
- Modeling: best practices

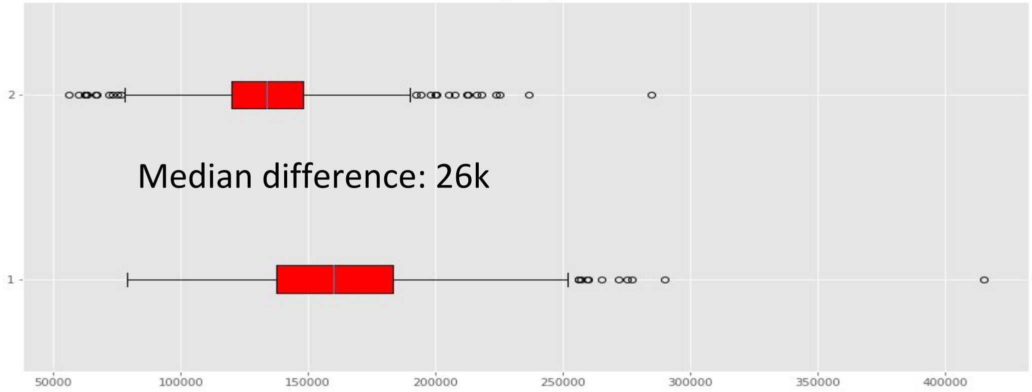
Feature Mapping

Overall Quality 10 vs. 9



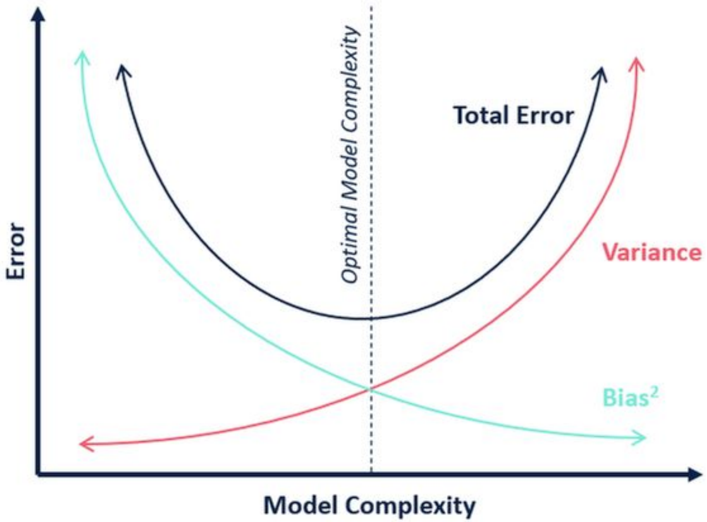
Median difference: 92k

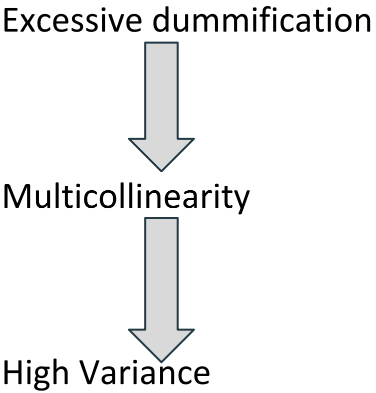
Overall Quality 6 vs. 5



Median difference: 26k

Feature Engineering



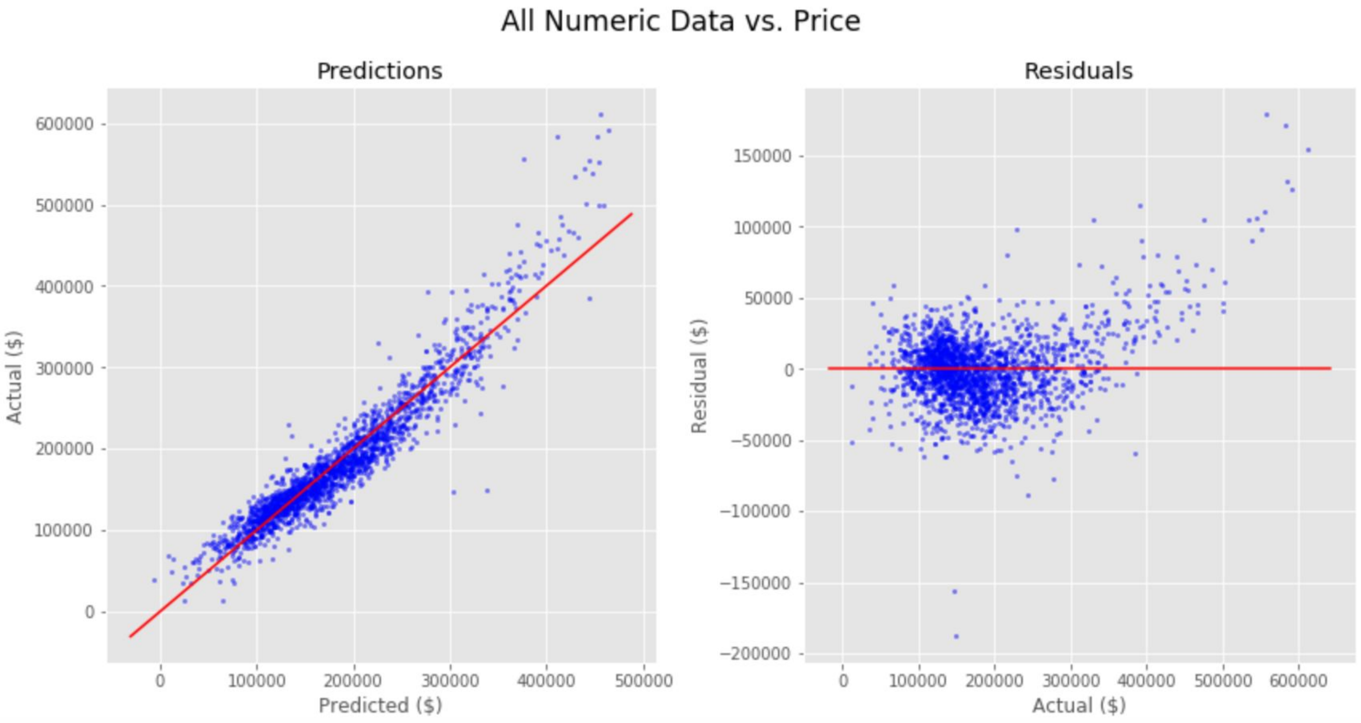


Excessive dummification

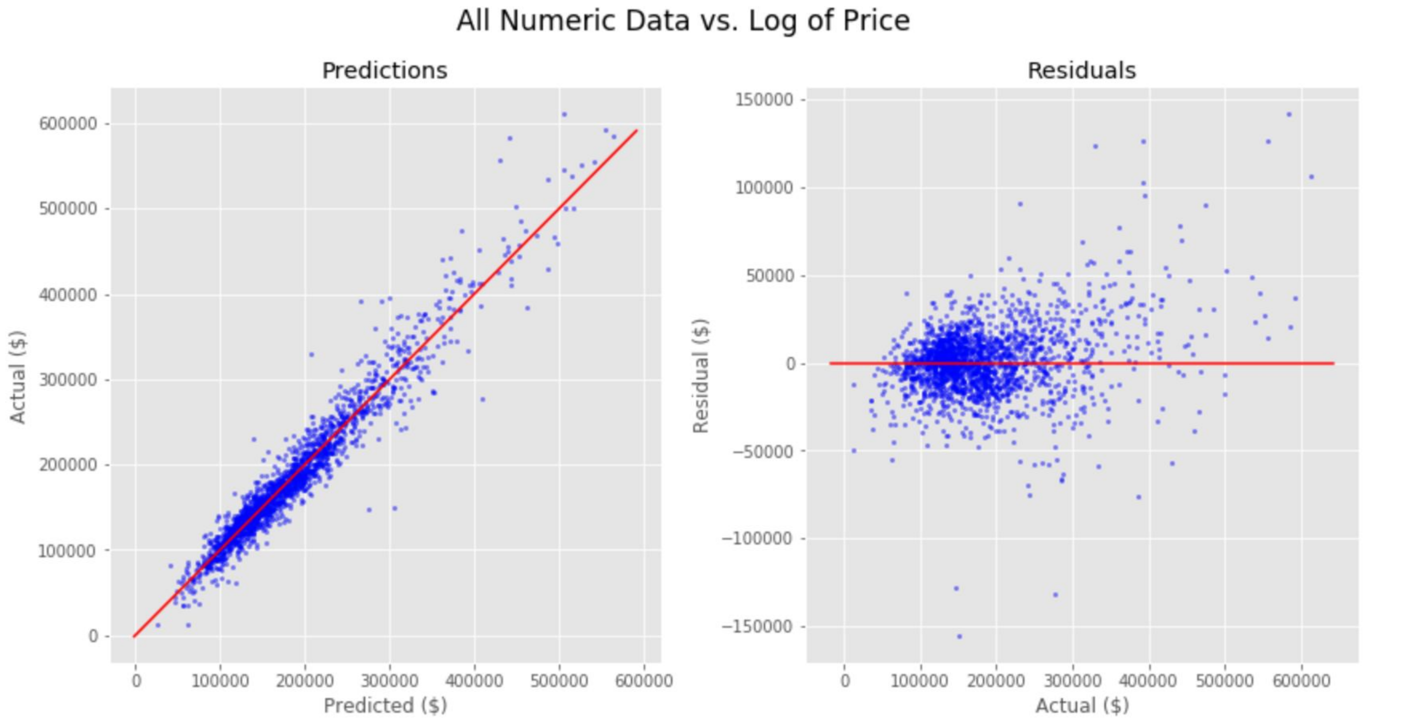
Multicollinearity

High Variance

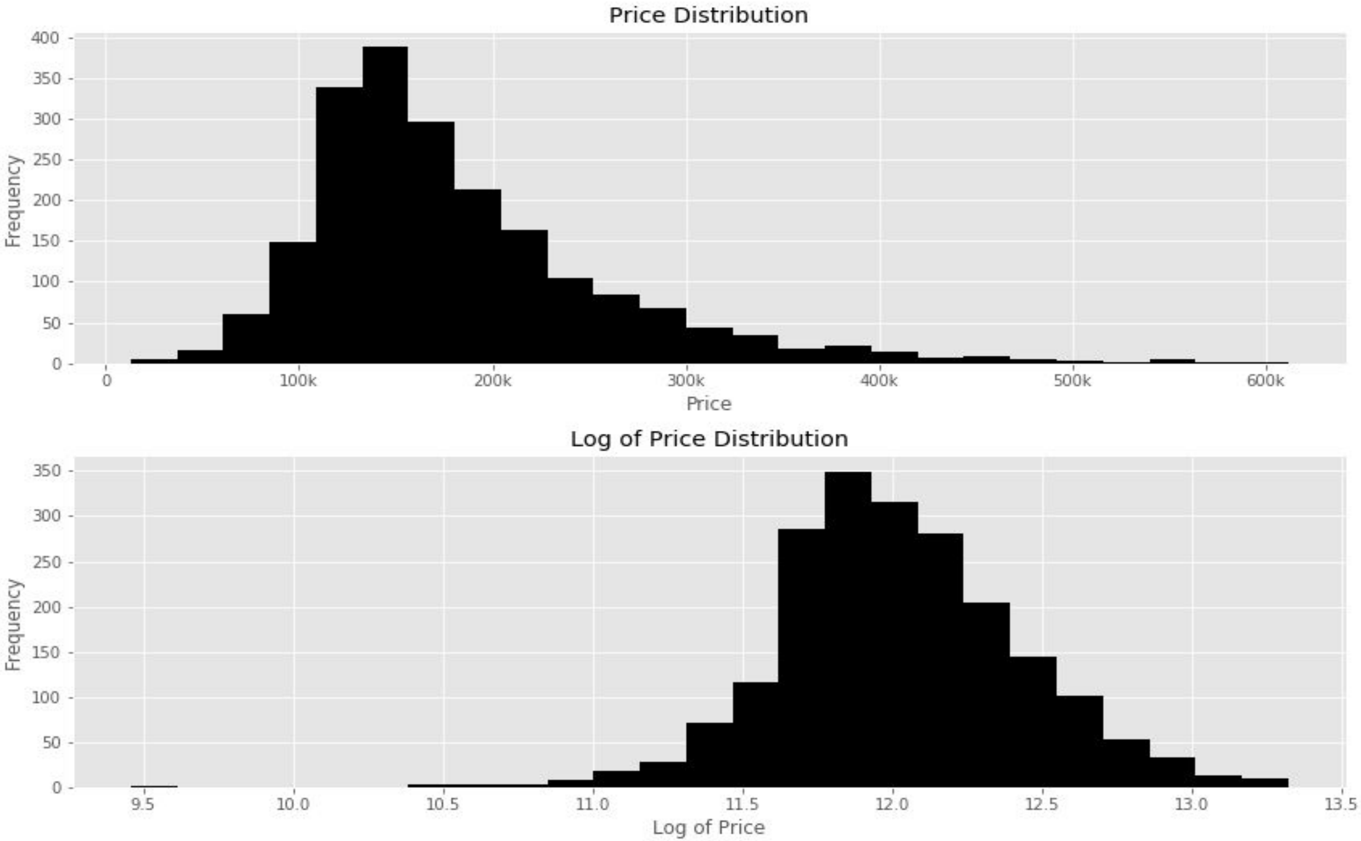
Modeling


All Numeric Data vs. Price
Predictions
Actual ($)
600000
500000
400000
300000
200000
100000
0
Predicted ($)
0
100000
200000
300000
400000
500000
Residuals
Residual ($)
150000
100000
50000
0
−50000
−100000
−150000
−200000
Actual ($)
0
100000
200000
300000
400000
500000
600000

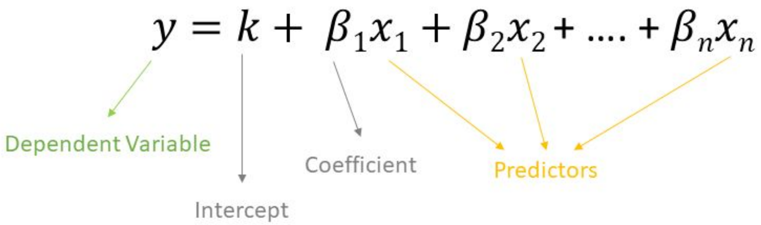
Modeling +


All Numeric Data vs. Log of Price
Predictions
600000
500000
400000
300000
200000
100000
0
Actual ($)
0
100000
200000
300000
400000
500000
600000
Predicted ($)
Residuals
150000
100000
50000
0
−50000
−100000
−150000
Residual ($)
0
100000
200000
300000
400000
500000
600000
Actual ($)

Price Distribution



Scaling



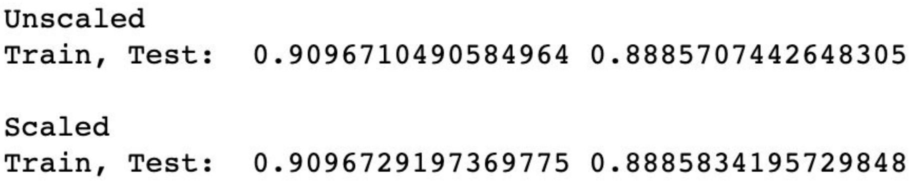
$$y = k + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_n x_n$$

Dependent Variable

Intercept

Coefficient

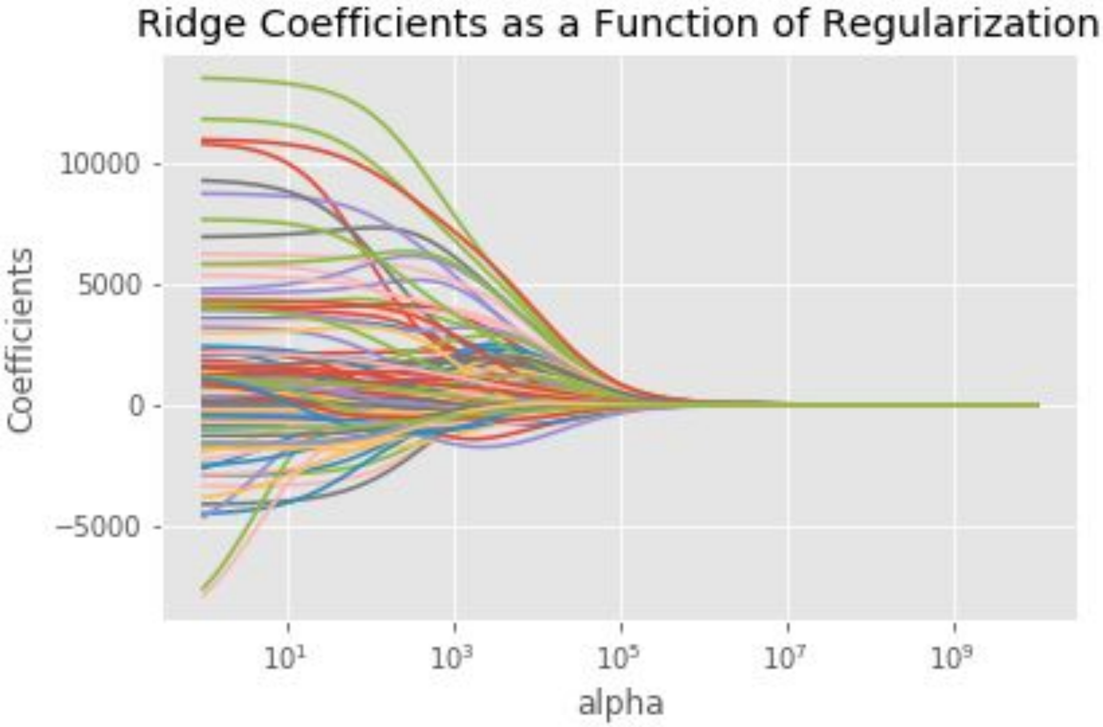
Predictors



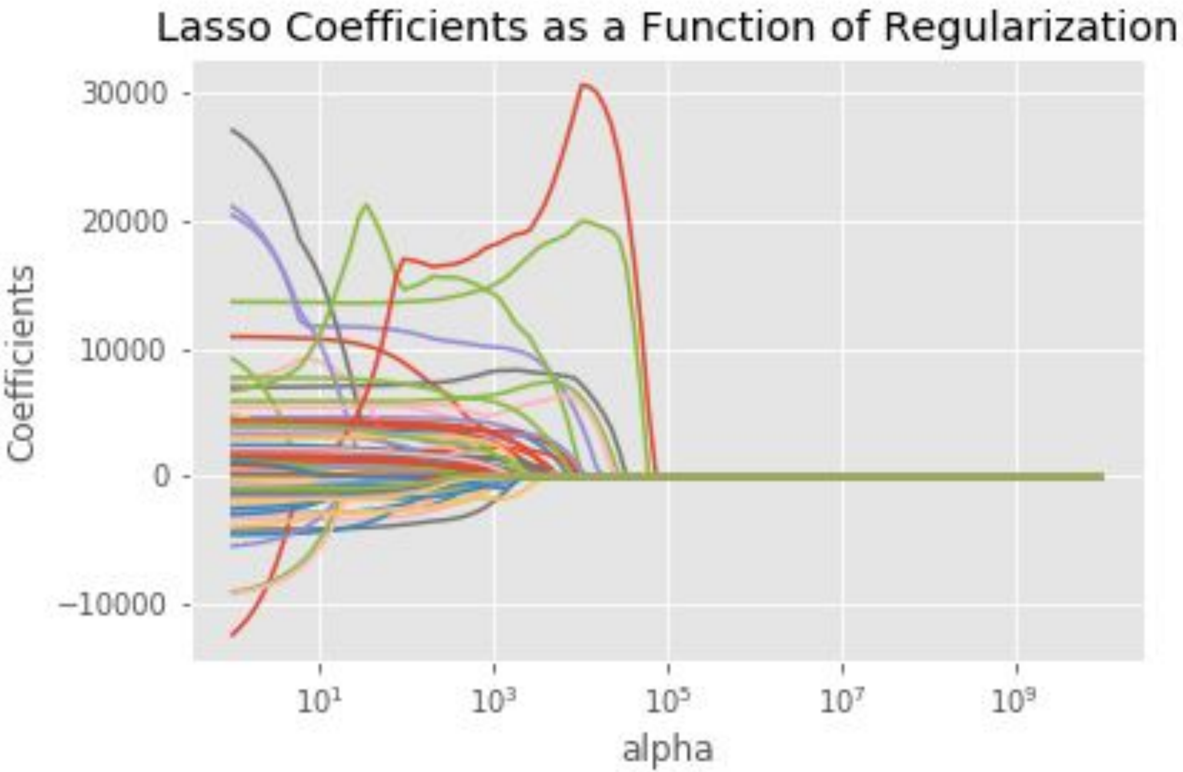
```
Unscaled
Train, Test:  0.9096710490584964 0.8885707442648305

Scaled
Train, Test:  0.9096729197369775 0.8885834195729848
```

Ridge Regularization



Lasso


Lasso Coefficients as a Function of Regularization
30000
20000
10000
0
−10000
Coefficients
10^1
10^3
10^5
10^7
10^9
alpha

References

https://stats.stackexchange.com/questions/29781/when-conducting-multiple-regression-when-should-you-center-your-predictor-varia

https://community.alteryx.com/t5/Data-Science-Blog/Bias-Versus-Variance/ba-p/351862

https://www.superheuristics.com/linear-regression-is-inaccurate-and-misleading/