InClass Prediction Competition

DSIR-1116-Ames Regression Challenge

Project 2 Competition site for GA's DSIR 1116

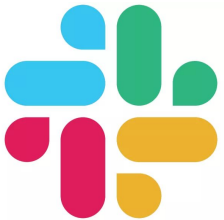
BONUS ROUND

hosted by Marta Lew

What do you think?

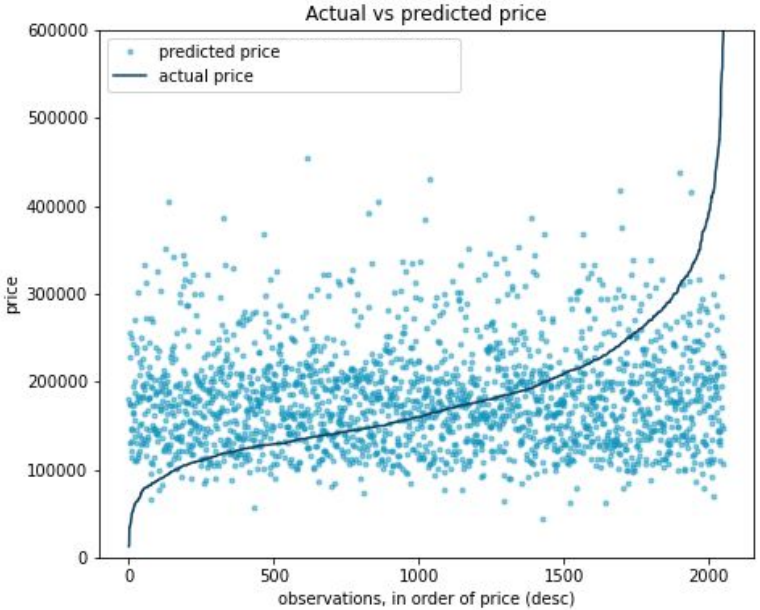
Which of these six techniques did/would you trust most to improve your linear regression model's score?

1. Including more dependent variables vs base model
2. Including even more dependent variables
3. Excluding dependent variables with p-values greater than 0.05
4. Removing a handful of outlier observations
5. Regularizing model coefficients (Ridge or Lasso)



One X variable

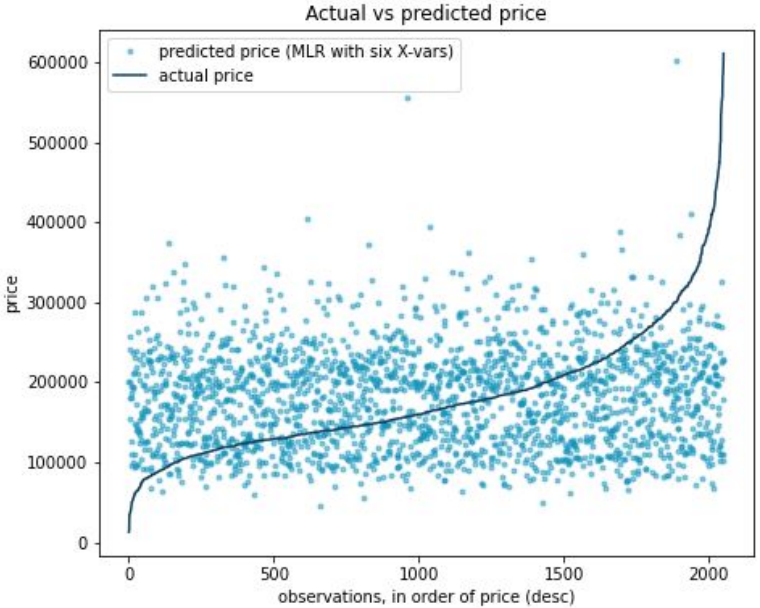
Baseline model: Price vs Square Footage

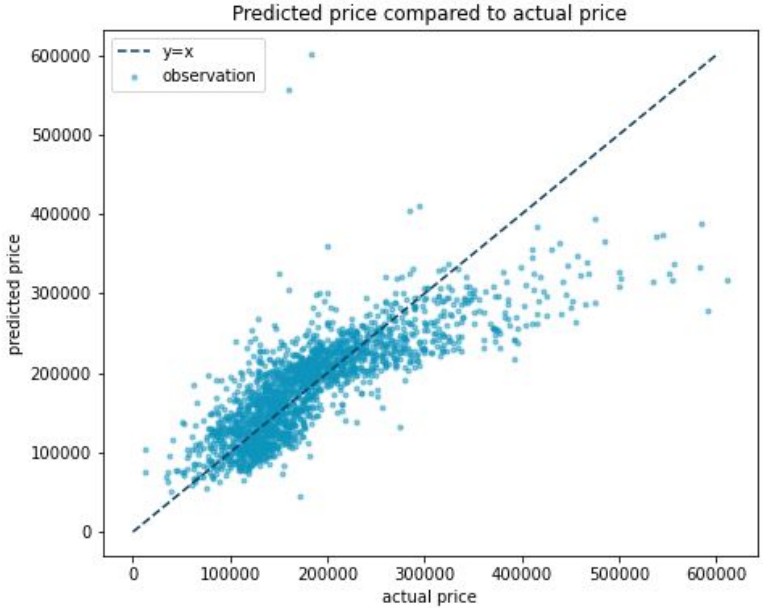


- R2 on training data = 0.5315
- R2 on test data = 0.3300
- Average **R2 in 5-fold cross-validation = 0.4843**

Six X variables

Iteration 2: More dependent variables

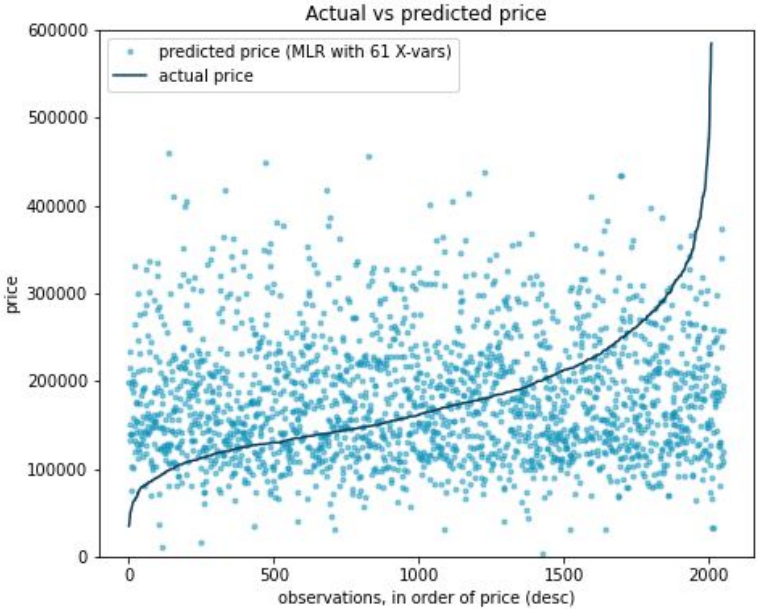




- R2 on training data = 0.6191
- R2 on test data = 0.6884
- Average R2 in **5-fold cross-validation = 0.6321** (std of 0.0352)

Sixty+ X variables

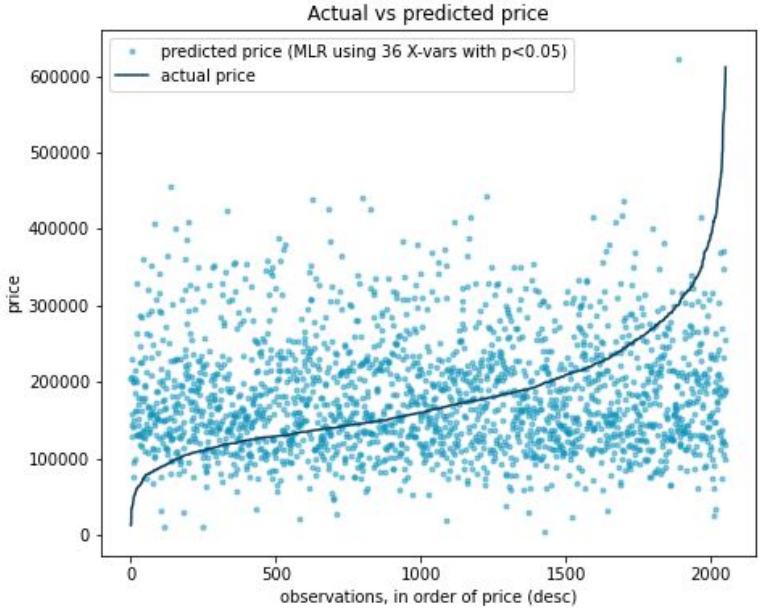
Iteration 3: Many more dependent variables

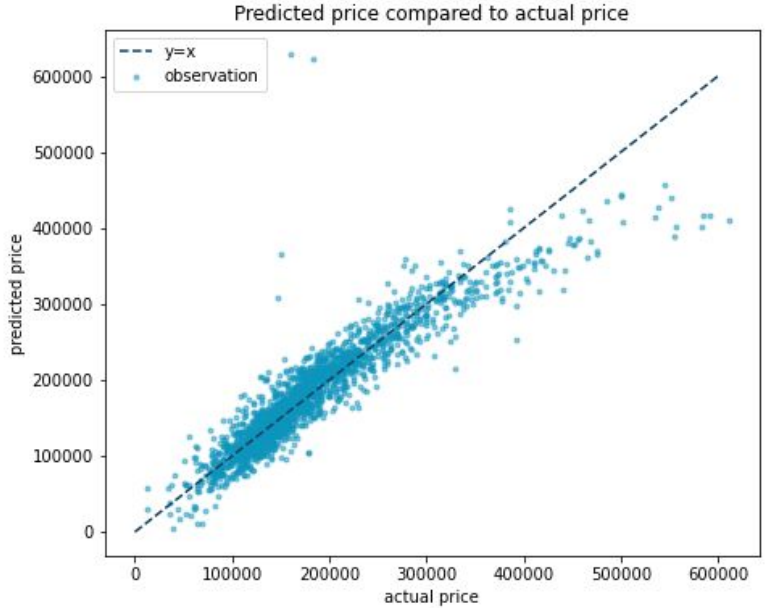


- R2 on training data = 0.8955
- R2 on test data = 0.6446
- Average **R2 in 5-fold cross-validation = 0.8364** (std of 0.0335)

36 X variables

Iteration 4: Only dependent variables with p<0.05

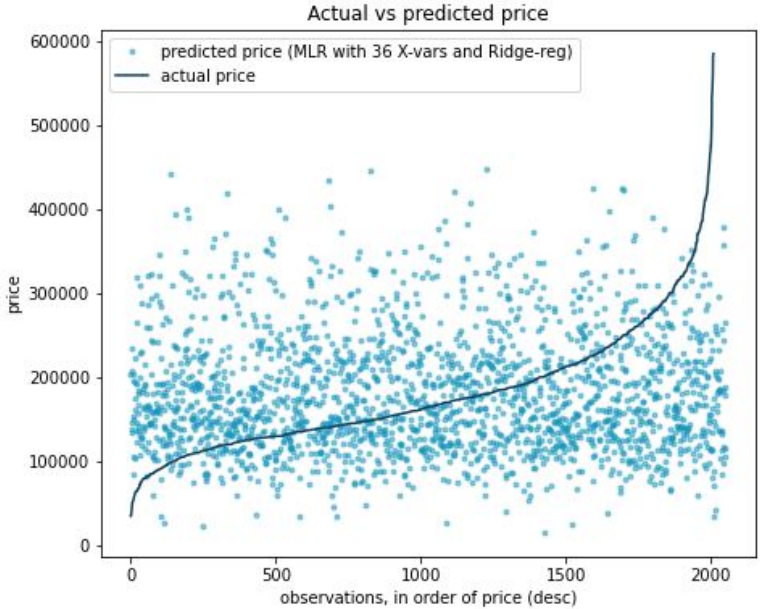


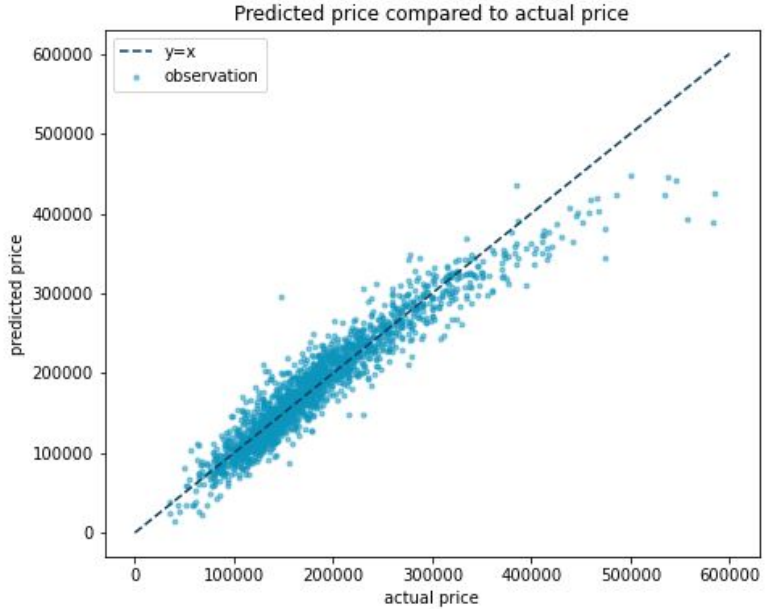


- R2 on training data = 0.8637
- R2 on test data = 0.8028
- Average **R2 in 5-fold cross-validation = 0.8396** (std of 0.0345)

36 X variables

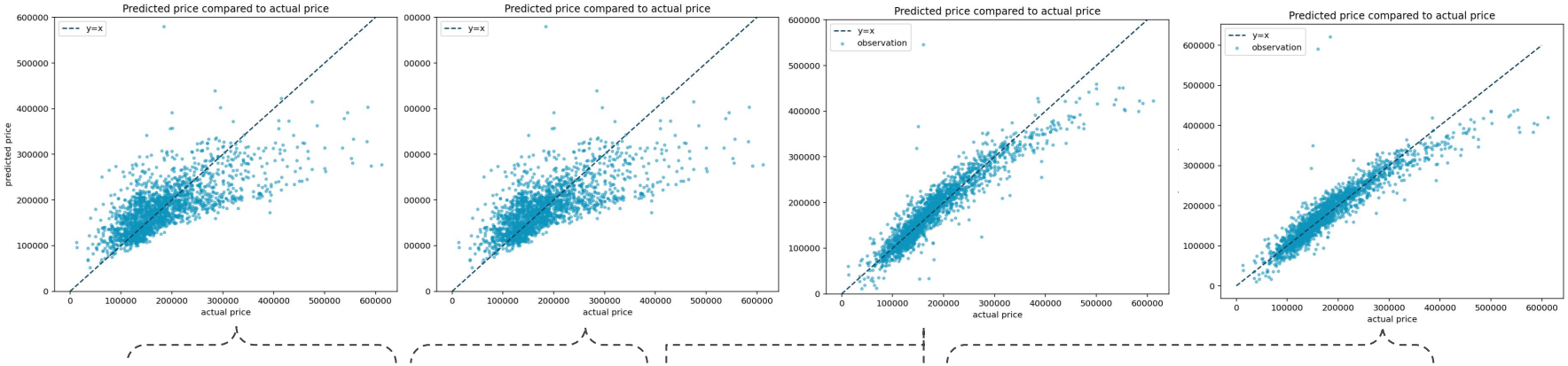
Iteration 5: Regularization to the rescue?

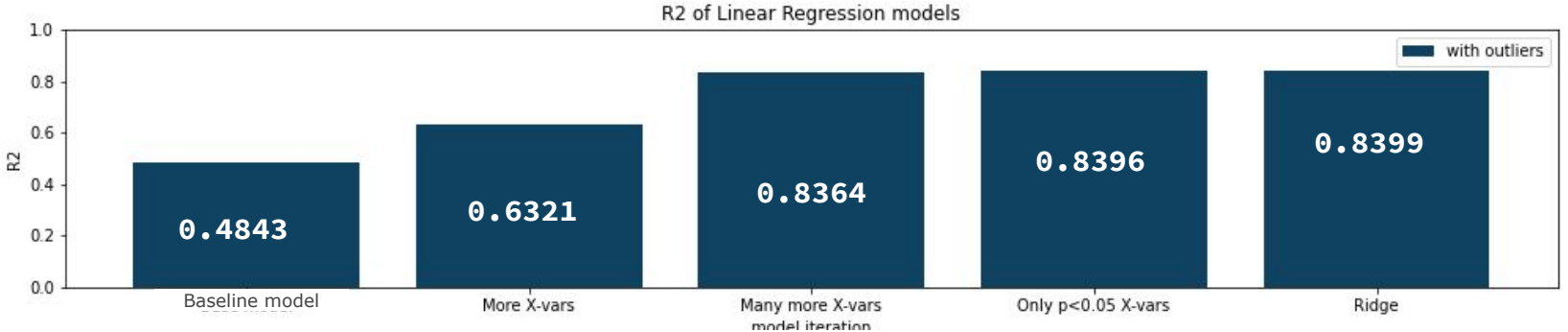




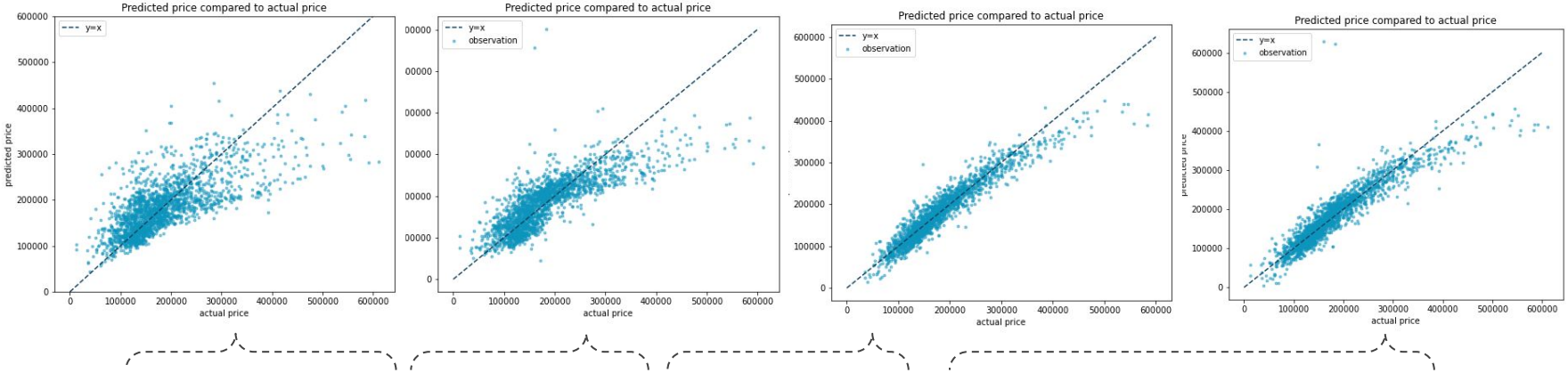
- R2 on training data = 0.8635
- R2 on test data = 0.803
- Average **R2 in 5-fold cross-validation = 0.8399** (std of 0.0342)

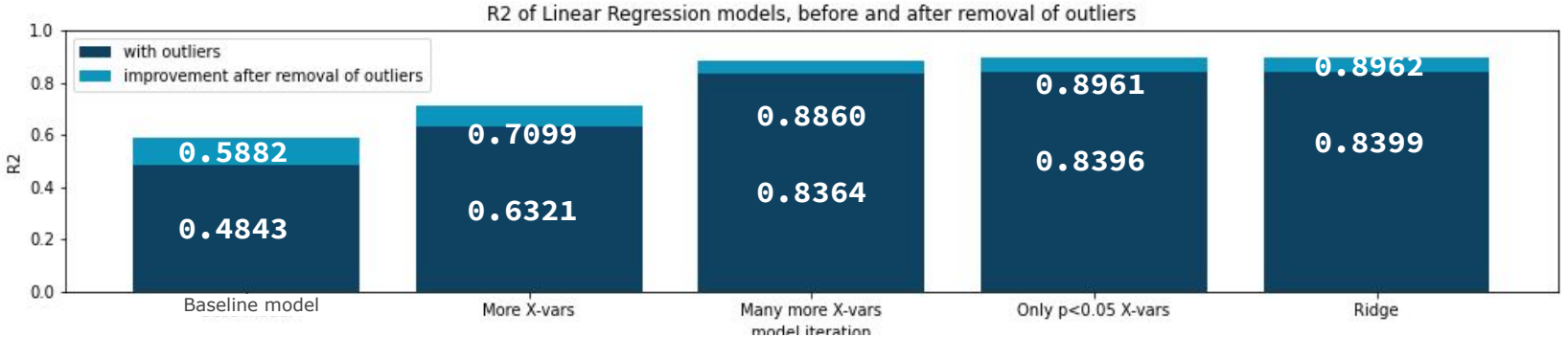
Iterations, iterations...





… never good enough with the outliers!

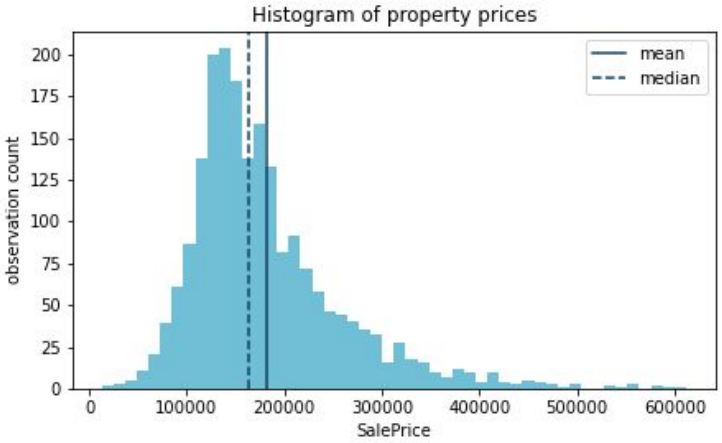


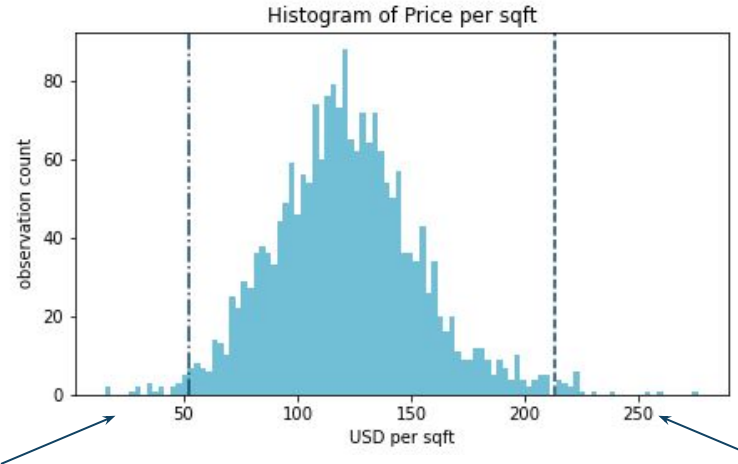


The Takeaway:

Don’t forget to clean your data, fellow Data Scientists!!!

Distribution of Price and Price per sqft





21 observations below
1st percentile

21 observations above
99th percentile