Ames Housing Prices

Manuel Molina

Overview

Ames Housing Dataset

Introduced in 2011 by Professor Dean De Cock

Contains 2919 (2051) observations from 2006 -2010 in Ames, Iowa

80 features (23 nominal, 23 ordinal, 14 discrete, 20 continuous)

neighborhoods | overall quality | fireplaces | sq. ft.

6 most **important** features in predicting housing price?

What’s important as a buyer?

Method: Initial Model

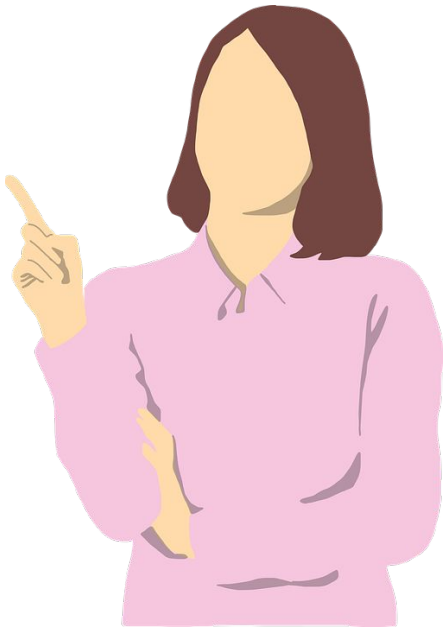
r2 (train)	76.72
r2 (test)	79.26
mae	25,912
rmse	36,317
cvs mean	75.55
resid mean	-1125

Handpicked-blind model

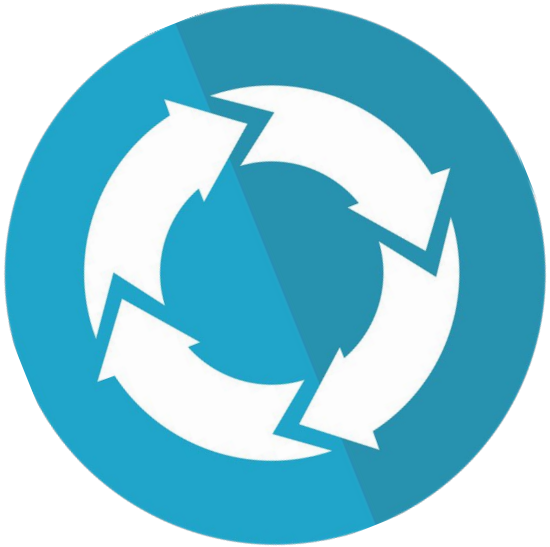
5 variables

Biased & underfit

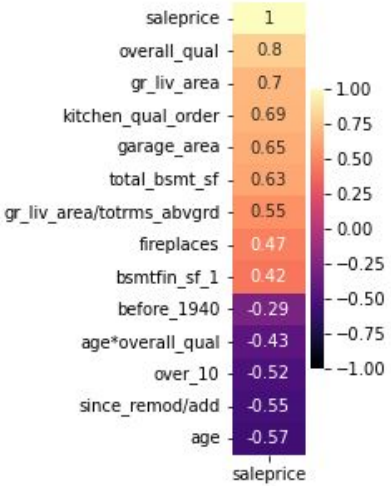
Add more variables



Iterations



- 8 official model iterations
- **All** except 1 were **underfit**
- Featured engineered in the hopes of ensuring **understandable interactions** and account for outliers
- ex: **age | over 10 | sqft per room | before 1940 | since remod**



Lasso Model (neighborhoods)

R2 (train) = 93.23%

R2 (test) = 92.53%

Mae = $15,433.07

Rmse = $21,799.93

Cvs = 89.98%

mean residual = -637.81

Polynomial Model (no neighborhoods)

R2 (train) = 91.22%

R2 (test) = 91.20%

Mae = $16,617.50

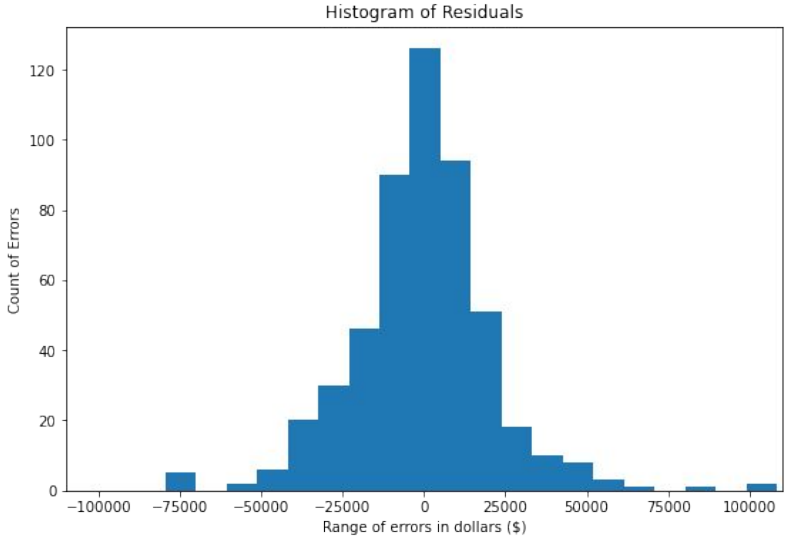
Rmse = $23,654.24

Cvs = 87.17%

mean residual = -1565.50

Without regularization the model **with neighborhood** has an R2 of 95.04% (train) and R2 = -7.45e19 (test)

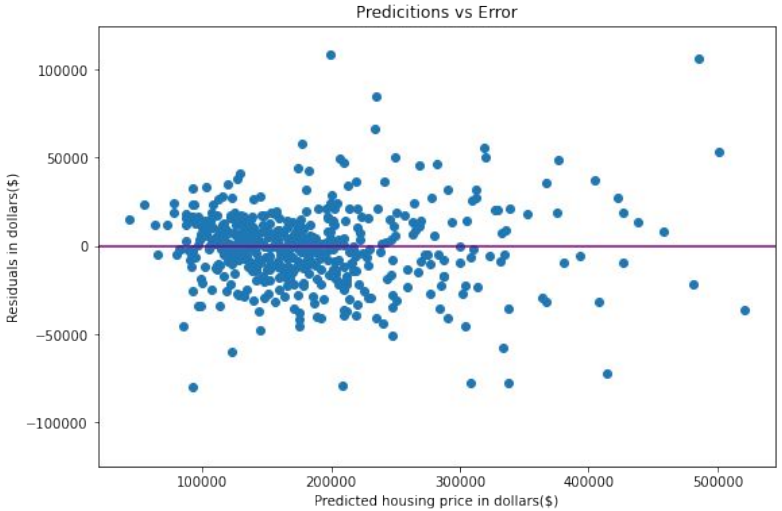
LINEM Worthy?



Linear? ✓

Normal? ✓

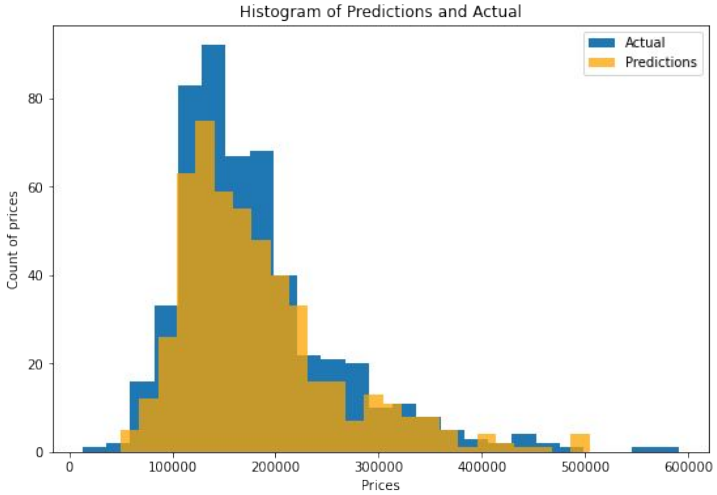
LINEM cont.

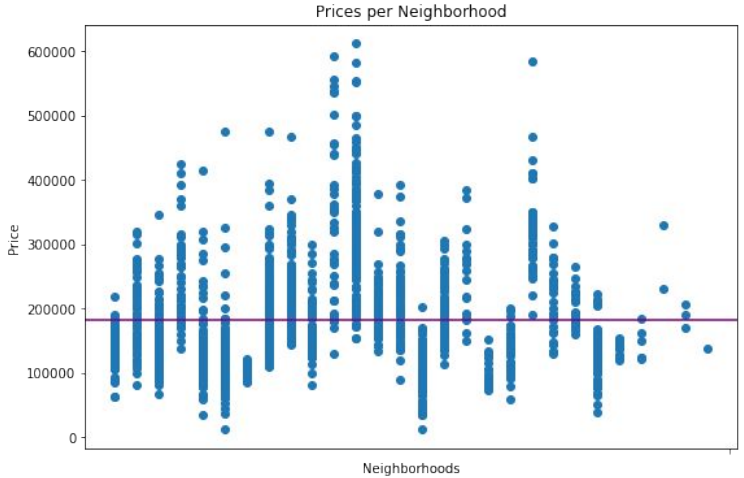


Equal Distribution?

Multicollinearity? Debatable

Interaction effects are magnified especially with interactions that are both highly correlated with each other and with price


Histogram of Predictions and Actual
Actual
Predictions
Count of prices
Prices
0
20
40
60
80
0
100000
200000
300000
400000
500000
600000


Prices per Neighborhood
Price
Neighborhoods
0
100000
200000
300000
400000
500000
600000

Top 6 Coefficients & Take-aways

coef	lr coefs
since_remod/add over_10	87462.029877
gr_liv_area overall_qual	45919.525831
overall_qual total_bsmt_sf	39091.189454
gr_liv_area/totrms_abvgrd^2	24652.577024
gr_liv_area kitchen_qual_order	24274.756863
gr_liv_area	23614.825330

Are neighborhoods important for you?

Preferences matter: Basements & living area in particular neighborhoods matter →

coef	lasso coefs
overall_qual total_bsmt_sf	20621.795222
gr_liv_area overall_qual	19117.145974
bsmtfin_sf_1	16269.328684

coef	lasso coefs
total_bsmt_sf^2	-18642.429168
gr_liv_area bsmtfin_sf_1	-10256.702616
since_remod/add neighborhood_NridgHt	-10112.126134

↖ **Remodeling** a house older than 10years matters most

Limitations & Recommendations

- Unused features
- Price inflation–natural overvaluation of features (need more research)
- External variables i.e. 2008 subprime mortgage crisis or soldiers returning from WWI&II
- No metric of competitors

- Further research into preferences per particular regions could yield more robust insights
- Test if generalizable to other regions
- More robust checks for multicollinearity
- Look for better ways to partition neighborhoods