Predicting Housing Prices: A Model Comparison

Katy Chow - DSI 6

Table of Contents

➔ Executive Summary

➔ EDA & Variable Selection

➔ Final Model & Implications

➔ Visually Comparing Model Performance

➔ Conclusions

Executive Summary

The purpose of this analysis is to be able to predict housing prices given a certain set of attributes. Although the Ames housing dataset had over 80 fields, it looks as though to achieve a model that can explain more than 80% of the variability we only need a small subset of the attributes and using simple regression methods with a penalty term. Additional model improvements (R^2 in the 90%) can be achieved with simple cleaning methods such as backfilling NULLs.

EDA & Variable Selections

	Model 1	Model 2
Rows Removed/Replaced	Gr Liv Area >= 4000 Lot Area >= 25000	Gr Liv Area >= 4000 Lot Area >= 25000 Replaced all Nulls with 0 values in numeric columns
Columns Removed/Replaced	Only kept 20 columns (15 numeric, 5 categorical)	Removed only non numeric columns

Variable Creation and Transformations Model 1

New Numeric Attributes Created

- Total_bath_abv_grd - total bathrooms above ground (half baths = 0.5*baths)
- Total_bath_bsmt - Total bathrooms in basement
- Outdoor Liv Area - Sum of Wood Deck, Open Porch, Screen Porch
- Overall Cond Bi - Binary 0,1 of overall condition sliced by 5
- Bldg Type Bi - Binary 0,1 of Building types
- Sale Type Bi - Binary 0,1 of type of Sale

Used standard scaler to fit Lasso Regression for Model 1

Top 10 Attributes for Model 1

Attribute	Coefficient
Gr Liv Area	0.230867629
Year Built	0.122832252
Overall Cond Bi	0.063372003
Year Remod/Add	0.059098986
Tot_bath_bsmt	0.049010267
Lot Area	0.039695006
Sale Type Bi	0.02616346
Outdoor Liv Area	0.023583795
Bldg Type Bi	0.012205136
TotRms AbvGrd	0.01057815

Model Attribute Importance

Model 1

Variable Creation and Transformations Model 2

New Numeric Attributes Created

- Used 2nd order polynomials to create both interaction terms and higher order polynomials for all numeric columns

Used standard scaler to fit lasso regression for Model 2

Top 10 Attributes for Model 2

Attribute	Coefficient
Overall Qual Gr Liv Area	22445.9982
Overall Qual Total Bsmt SF	17049.1677
Year Built Year Remod/Add	10376.2196
Overall Qual BsmtFin SF 1	8397.50178
BsmtFin SF 1^2	7750.41746
Overall Qual Garage Area	7437.46859
Overall Qual 1st Flr SF	5195.32073
Lot Area Overall Cond	5109.17072
Year Built^2	3861.52536
Total Bsmt SF Half Bath	3834.35777

Model Attribute Importance

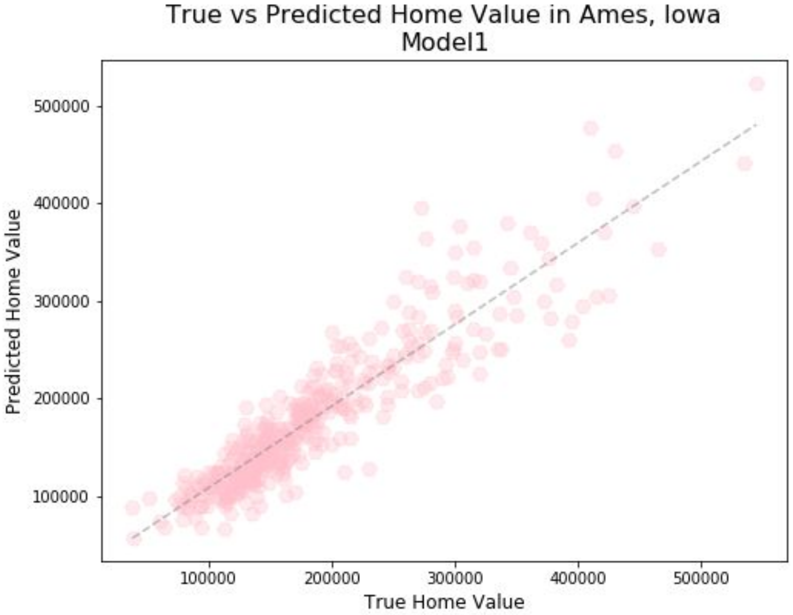
Model 2

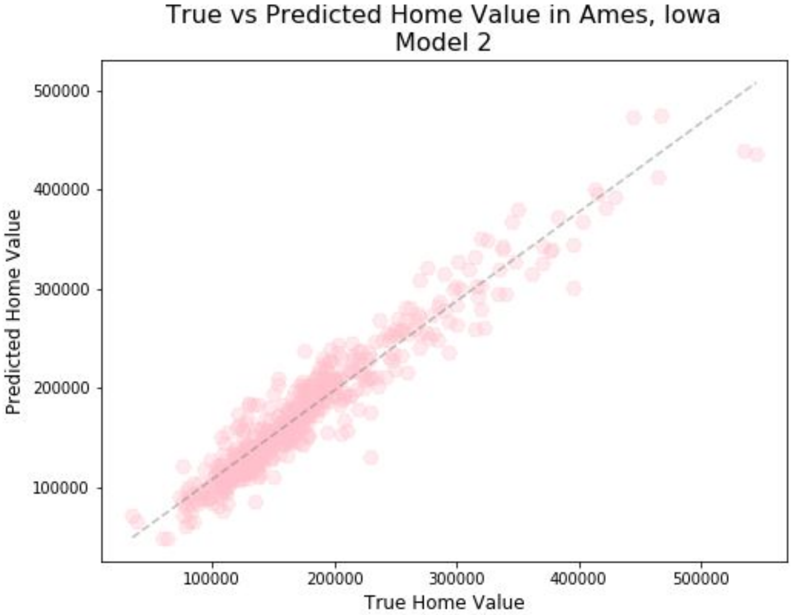
Screen
Abvgr
Pool
Fireplaces
Totrms
Bath
Bsmtfin
Yr
Garage
Abvgr
Built
Remod
Open
Mas
Ms
Qual
Misc
Add
Cars
Nd
Full
Add
Year
Year
Nd
Half
Sf
Total
Pool
Half
Cond
Sold
Fin
Flr
Gr
Wood
Blt
Unf
Area
Enclosed
Garage
Val
Overall
Ssn
Bedroom
Sold
Kitchen
Remod
Enclosed
Mo
Bsmtfin
Frontage
Vnr
Area
Fin
Porch
Kitchen
Low
Bath
Val
Liv
Yr
Flr
St
Low
Cars
Fireplaces
Full
Wood
Frontage
Sf
Mas
Bsmt
Gr
Qual
Mo
Screen
Porch
Bedroom
Open
Totrms
Liv
Deck
Lot
Abvgrd
Subclass
Lot
Vnr
Unf
Ssn
Bsmt
Ms
Cond
Deck
Misc
Total
Abvgrd
Built
Subclass
Overall

Final Models & Implications

	Model 1	Model 2
Score on Test	0.8376	0.9173
Score on Train	0.8194	0.9282
Number of Attributes	15	98
Best Fit Model	Lasso	Lasso

Visually Comparing the Models





Conclusions

Questions??

It is very easy to see that the more explainable models may not always be the best fit, but with some cleaning methods and adding more attributes it you can easily bump model performance and leaving you still slightly more confused.