Predicting Home Prices in Ames, Iowa

Matt Williams
July 10th, 2020

Problem Statement

We will be using home sale data from 2006-2010 to build and evaluate models to estimate home sale prices in Ames, IA. For modeling, will use ordinary least squares, Ridge and Lasso methods.

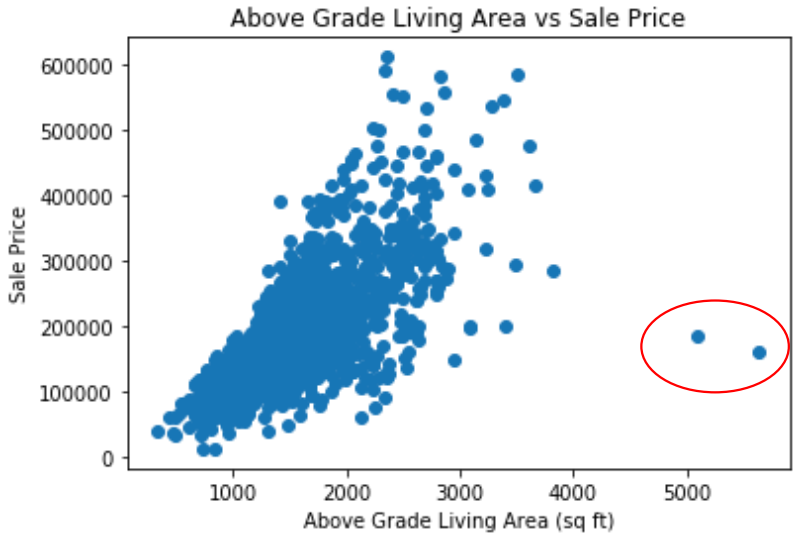
Image Source

Data Sources and Cleaning

Data Source: Ames Housing Data

- Data from 2006 - 2010

Shape of the Data	
Observations	2051
Variables	81

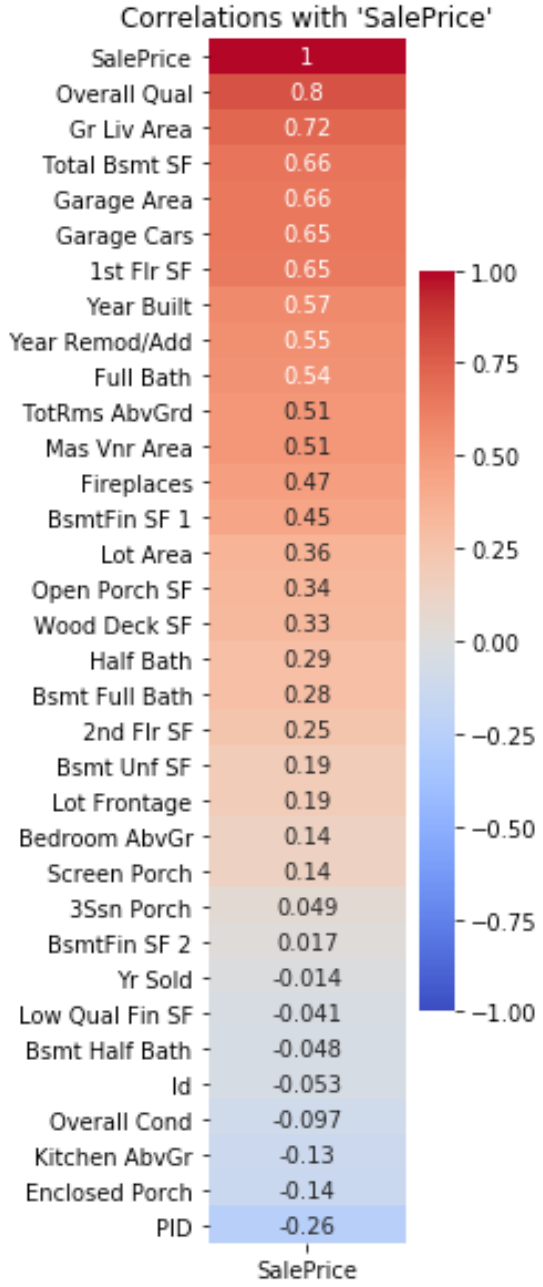


Data Cleaning Actions Taken:

- Reclassify 2 columns
- Drop 5 columns
- Replace null values with 'NA' or zero
- Drop 5 outliers

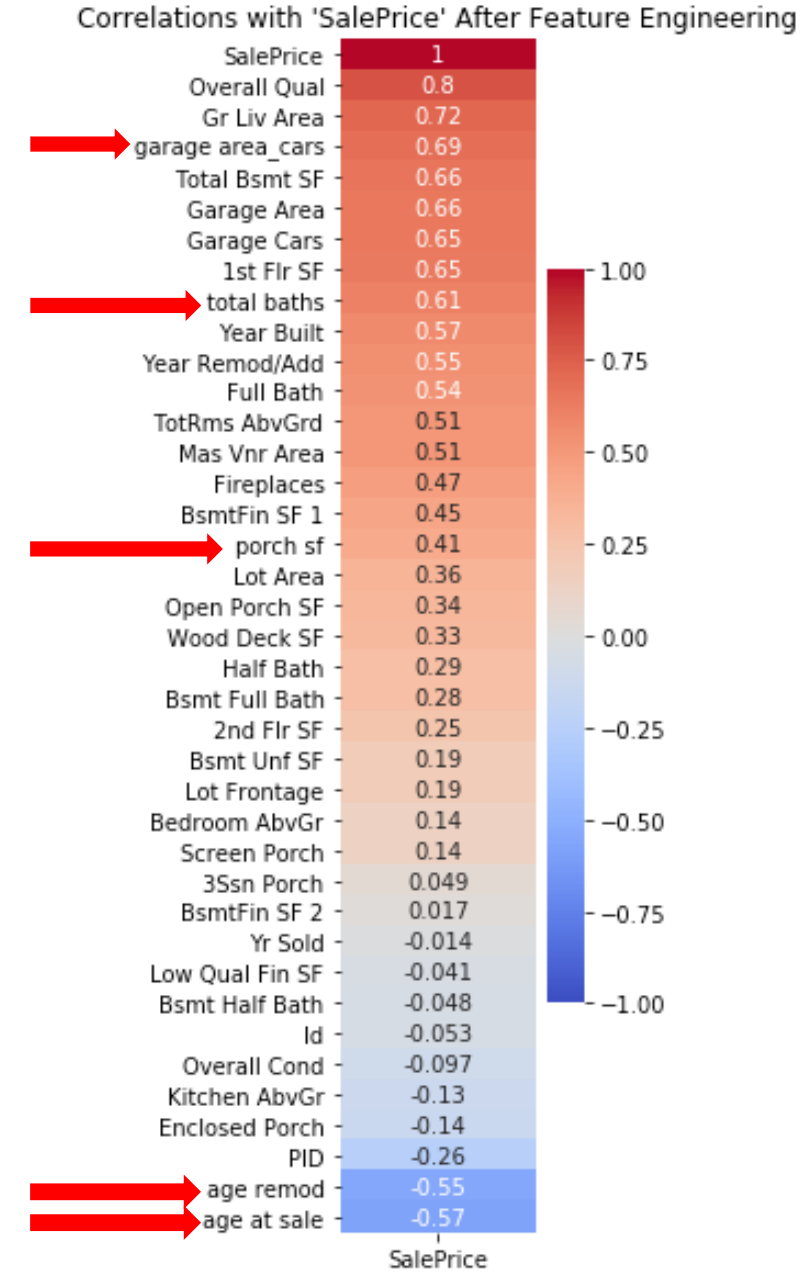
Exploratory Data Analysis

- Examined relationship between ‘SalePrice’ and numerical variables
- Examined distributions of numerical variables
- Examined relationship between ‘SalePrice’ and categorical variables



Exploratory Data Analysis

- Engineered Variables:
  - 'Garage Area' * 'Garage Cars'
  - 'total baths'
  - 'porch sf'
  - 'age at sale'
  - 'age remod'
- Dummy Variables:
  - Dummied all categorical variables



Exploratory Data Analysis

- Ran SLR models for each variable

Numerical Variables:

Quantitative:	R2 Score:		
	Train	Test	Crossval
Overall Qual	0.6977	0.7087	0.6782
GR Liv Area	0.4783	0.5083	0.4687
Garage Cars	0.4235	0.4098	0.4132
Garage Area	0.4162	0.4421	0.4001
1st Flr SF	0.3601	0.4483	0.353
Year Built	0.3148	0.3637	0.3126
Year Remod./Add	0.2985	0.3159	0.2956
Full Bath	0.2858	0.3001	0.2887
Garage Yr Blt	0.2996	0.2417	0.2864
TotRmsAbvGrd	0.2735	0.1917	0.2671
Mas Vnr Area	0.2726	0.2296	0.264
Fireplaces	0.2288	0.1996	0.2238

Categorical Variables

Qualitative	R2 Score:		
	Train	Test	Crossval
Neighborhood	0.5886	0.5397	0.5687
Exter Qual	0.5300	0.5035	0.5237
Kitchen Qual	0.5058	0.5015	0.4950
Bsmt Qual	0.3467	0.3495	0.3413
Foundation	0.2817	0.3073	0.2715
MS Subclass	0.2483	0.2789	0.2394
Garage Type	0.2431	0.2901	0.2392
Fireplace Qu	0.2380	0.1778	0.2339
Heating QC	0.2275	0.1878	0.2258
Bsmt Fin Type 1	0.2195	0.2289	0.2073

Predictive Models

- Three models:
  - OLS
    - Manually constructed
    - Key difference from interpretive model is inclusion of 'Neighborhoods'
  - Ridge
    - Included all numeric and dummy variables
  - LASSO
    - Included all numeric and dummy variables
- Observations
  - Ridge: interactions with 'Neighborhoods' performed best
  - LASSO suggests 'Lot Frontage' & 'Lot Area' are important features, but neither added much to manually constructed models.

Model	R2 (test set)
OLS	.9105
Ridge	.9110
LASSO	.8246

Interpretive Model

- 43 Features:
  - A mix of quantitative variables and their associated qualitative variables (i.e. basement sq. ft. and basement condition)
  - *MS SubClass

```
'Overall Qual', 'Gr Liv Area', 'garage area_cars', 'Total Bsmt SF', 'total baths',
'TotRms AbvGrd', 'Mas Vnr Area', 'porch sf', 'Lot Area', 'age remod', 'age at sale',
'Bsmt Qual_Fa', 'Bsmt Qual_Gd', 'Bsmt Qual_NA', 'Bsmt Qual_Po', 'Bsmt Qual_TA','Kitchen Qual_Fa',
'Kitchen Qual_Gd', 'Kitchen Qual_TA','Kitchen Qual_Po', 'MS SubClass_150', 'MS SubClass_160',
'MS SubClass_180','MS SubClass_190','MS SubClass_20','MS SubClass_30','MS SubClass_40','MS SubClass_45',
'MS SubClass_50', 'MS SubClass_60', 'MS SubClass_70', 'MS SubClass_75', 'MS SubClass_80', 'MS SubClass_85',
'MS SubClass_90','Garage Cond_Fa','Garage Cond_Gd','Garage Cond_NA','Garage Cond_Po','Garage Cond_TA',
'Exter Qual_Fa', 'Exter Qual_Gd', 'Exter Qual_TA'
```

- Model Performance:

Metric	Score
R2	0.9006
MSE	$648,338,397.66
RMSE	$25,462.29

Variable	Coefficient
MS SubClass_75	2.41E+04
MS SubClass_45	1.98E+04
MS SubClass_85	1.86E+04
MS SubClass_70	1.61E+04
MS SubClass_180	1.47E+04
MS SubClass_30	1.44E+04
MS SubClass_40	1.31E+04
MS SubClass_80	1.12E+04
Overall Qual	1.12E+04
MS SubClass_50	1.08E+04

Conclusions

- The factors that have the greatest impact on driving home prices in Ames are overall home quality, neighborhood, and type of home.
- The factors that have the greatest impact on decreasing home prices in Ames are age of the home and how long it has been since the home was remodeled.
- Actions for model improvement:
  - Aim to increase model performance while decreasing model complexity
  - Use results of the Ridge and LASSO models to incorporate more high-performing features
  - Use more current data

Questions?