Predicting House Sale Prices for Ames, Iowa

Rashidi

SG-DSI-27

Content

- Background
- Feature Selection
  - Data Cleaning
  - Exploratory Data Analysis
- Simple Model
- Model Tuning/Feature Engineering
- Model Benchmarks
- Production Model
- Insights
- Kaggle Submission Score

Background

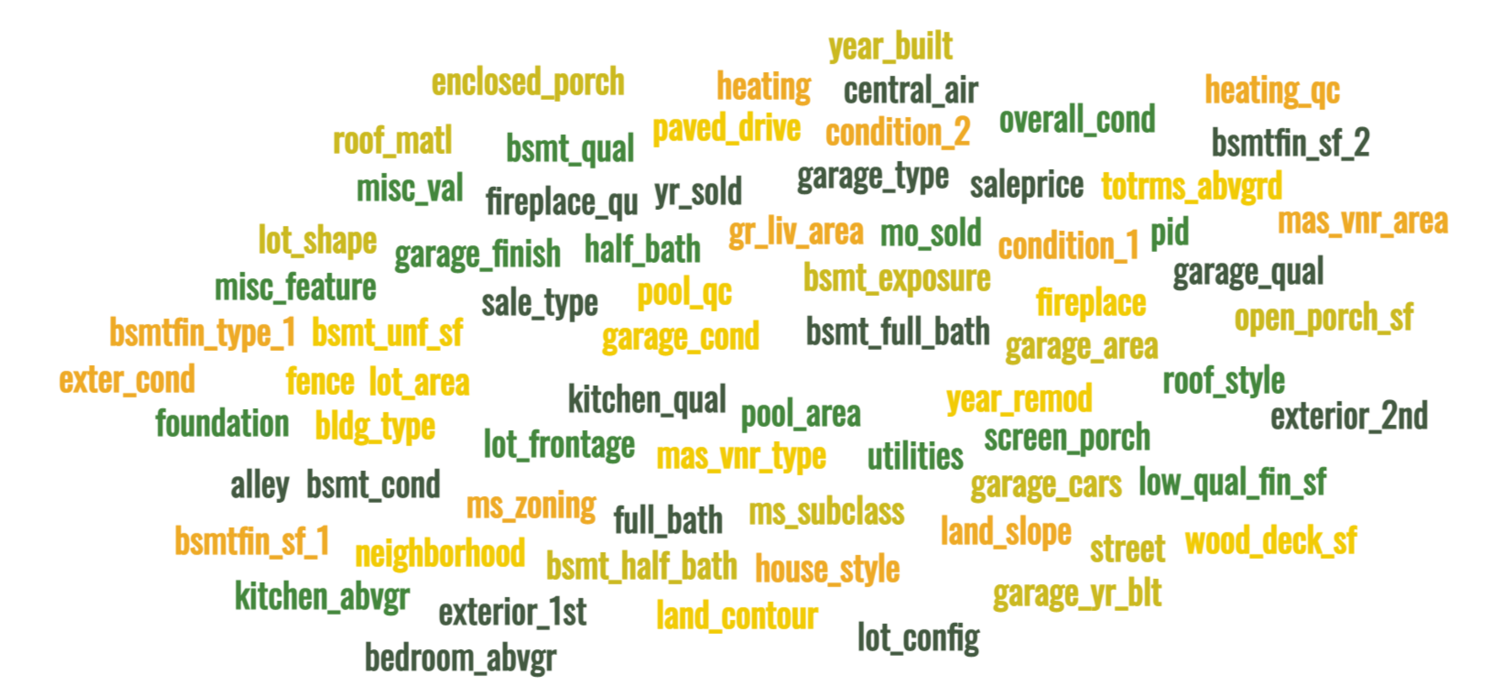
Problem Statement

As a new Data Scientist in ABC-XYZ Corp., a real estate agency, I was tasked to create a website that can estimate a property sale price for the whole of USA, starting with Ames, Iowa (where our HQ is based).

Feature Selection

Data Columns = 81

Which to choose? How to clean?



Initial Feature Filtering

- Filter out those that describes the same thing as another
  - ‘garage_cars’ < ‘garage_area’
- Filter out those that is a subset of another (*Except ordinal features)

  ‘bsmtfin_sf_2’
  ‘bsmtfin_sf_1’ $\subseteq$ ‘total_bsmt_sf’
  ‘bsmt_unf_sf’

- Filter out those that are identification features

Data Cleaning

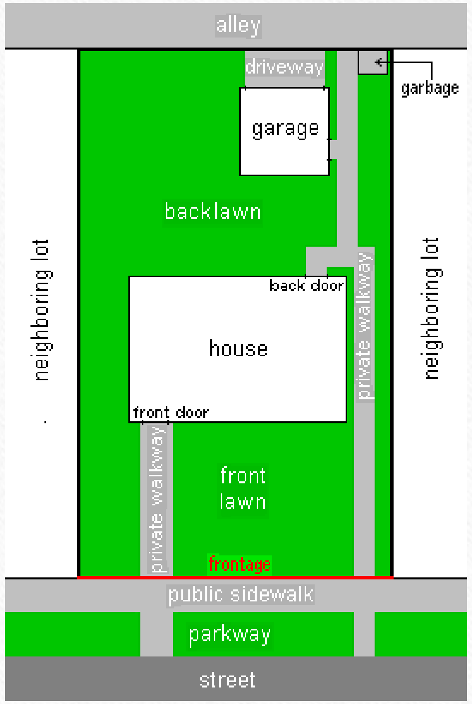
- Numerical features -> Check for null values
- Ordinal features -> Change to numerical scale
- Categorical features -> Manually Dummify
  - Done in order to determine easily which dummy column was dropped
    - Example: '150' was dropped from 'ms_subclass'.

Data Cleaning

‘lot_frontage’

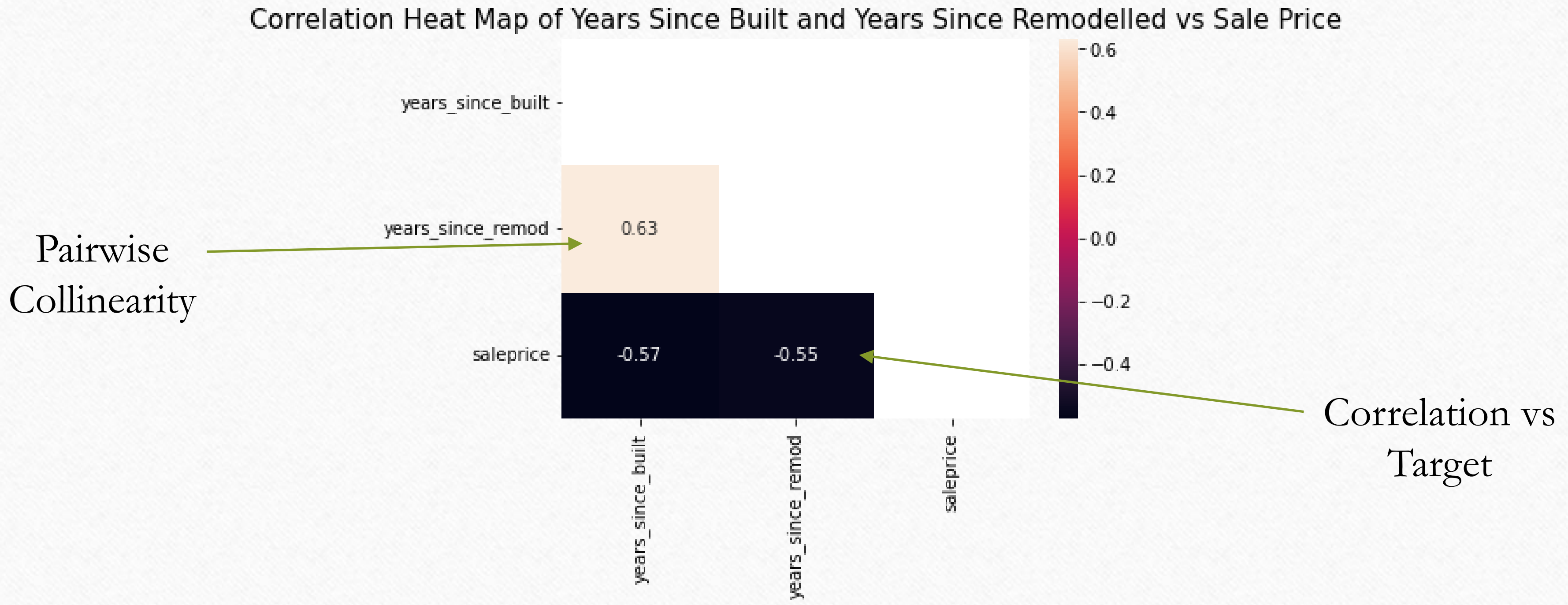
- Check any commonalities for 'ms_subclass’ or 'lot_config’ vs. ‘lot_frontage’
- Since there aren’t any, set NaN values as the mean value of each 'ms_subclass’

‘ms_subclass’	mean('lot_frontage')
20	77.03
30	61.04
40	51.75
45	54.82
50	63.00
60	78.27
70	64.32
75	70.47
80	79.87
85	73.33
90	69.40
120	44.82
150	44.82
160	27.59
180	26.60
190	71.60

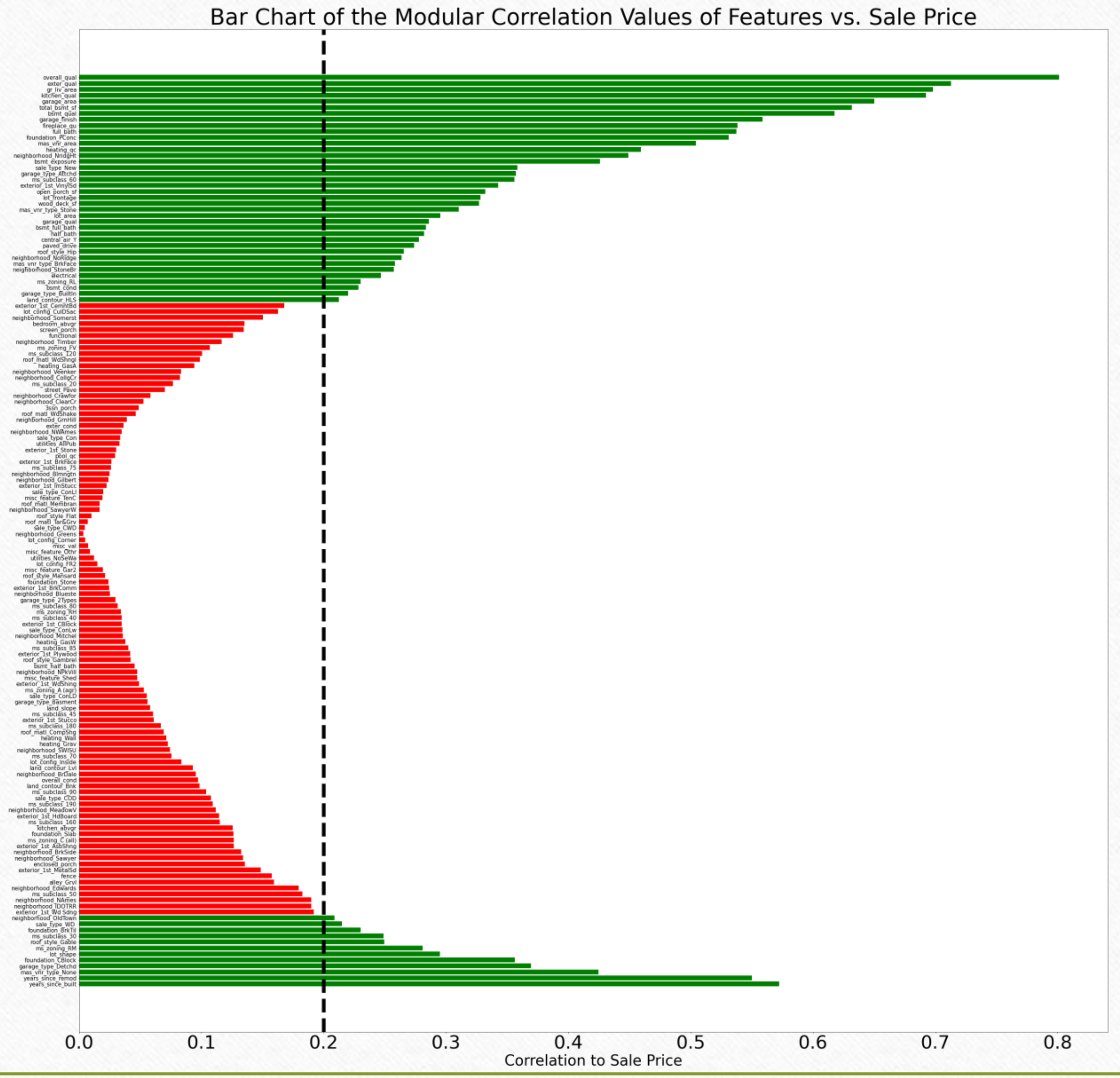


(from: https://upload.wikimedia.org/wikipedia/commons/b/bc/Lot_map.PNG)

Exploratory Data Analysis



Pairs	Pairwise Correlation	First Feature Correlation Vs. Sale Price	Second Feature Correlation Vs. Sale Price
ms_subclass_90 vs. bldg_type_Duplex	1.000000	-0.103817	-0.103817
ms_subclass_80 vs. house_style_SLvl	0.954549	-0.031484	-0.042176
garage_qual vs. garage_cond	0.950118	0.285858	0.265517
ms_subclass_50 vs. house_style_1.5Fin	0.942502	-0.182567	-0.196051
pool_area vs. pool_qc	0.904689	0.023115	0.029289
ms_zoning_FV vs. neighborhood_Somerst	0.874843	0.106749	0.150167
ms_subclass_45 vs. house_style_1.5Unf	0.869662	-0.060391	-0.066877
fireplaces vs. fireplace_qu	0.859621	0.470091	0.538252
gr_liv_area vs. totrms_abvgrd	0.812723	0.698046	0.502909



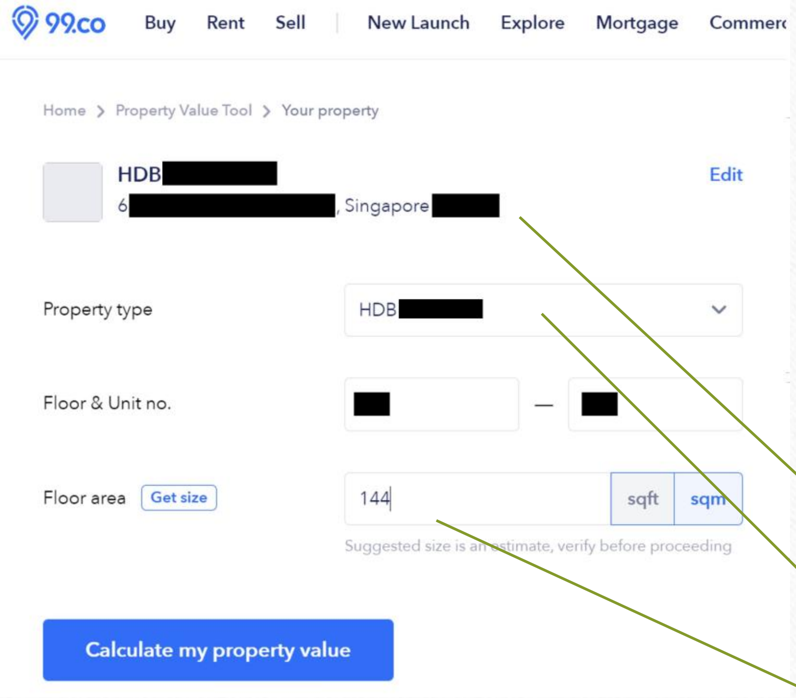
Where do we draw the line?

Simple Model

99.Co

Simple Model

99.co



(source: 99.co, https://www.99.co/)

Neighborhood

MS Subclass

Gr Liv Area

Linear Regression

Model Tuning/Feature Engineering

Steps Taken:

1) Lasso Regression (alpha=874.0802078515503)
2) Linear Regression after dropping Lasso Zero Coefficient Features
3) Ridge Regression after dropping Lasso Zero Coefficient Features (alpha=335.1602650938841)

Model Tuning/Feature Engineering

Lasso Regression

Pairwise Collinearity of lot_frontage vs. gr_liv_area: 0.360696

Feature	lasso_coef Value	Saleprice Correlation
lot_frontage	-0.000000	0.328149
gr_liv_area	24273.462134	0.69804

Number of Features dropped: 97

Model Benchmarks

Model	Train MSE	Test MSE	Cross Val Score
drop_0_coeff Ridge Regression	702796769	619778333	950438300
drop_0_coeff Linear Regression	651192793	666638781	1022462959
Lasso Regression	701584300	649567599	1015959823
99co Linear Regression	1464088294	1152320467	1576989381

~40% improvement

Production Model

Production Model Attributes

Train MSE	673341543.1
Cross Val MSE	833273475.8
Ridge Regression Alpha	335.16
Total Features Used	73
Kaggle Public Score	33203.03021

Insights

Insights

Production Model

- Overfit (Cross Validation MSE >> Train MSE)

Future Works:

1) Eliminating outliers from deep-diving into model predicted residuals.

2) Explore pairwise interactions.

3) Explore different cutoffs to see which will effectively eliminate poorly correlated features (vs. Target) and produce the best model.

Insights

Project

- Features used in model != features easily known by layman
  - It will be better to get data on which features are easily known/accessible by our platform users.
- Strive between simplicity (like 99.co) versus accuracy.
  - No one would want to sit down and complete a form with 70+ blanks to fill up.