Features Influencing Housing Value in Ames, Iowa

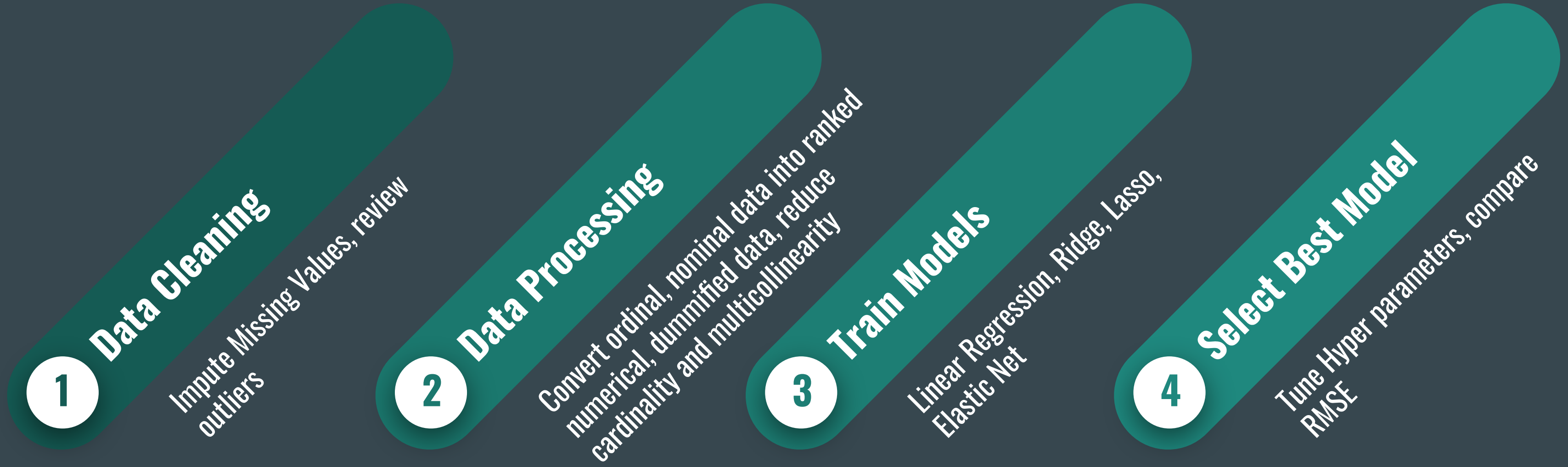
•••

Loh Chuan Hui

Introduction

In this project, we analyse home features which influence its selling price in Ames, Iowa

Analysis Methodology



1 **Data Cleaning**

Impute Missing Values, review outliers

2 **Data Processing**

Convert ordinal, nominal data into ranked numerical, dummified data, reduce cardinality and multicollinearity

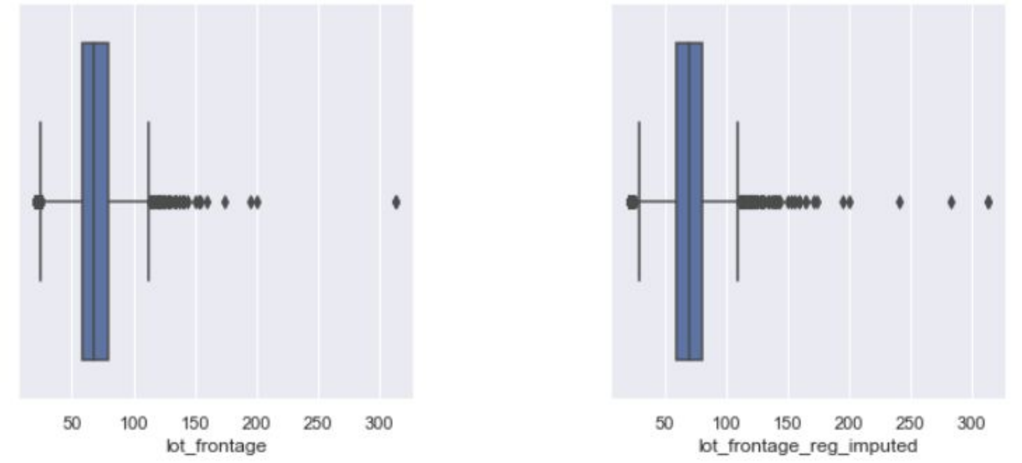
3 **Train Models**

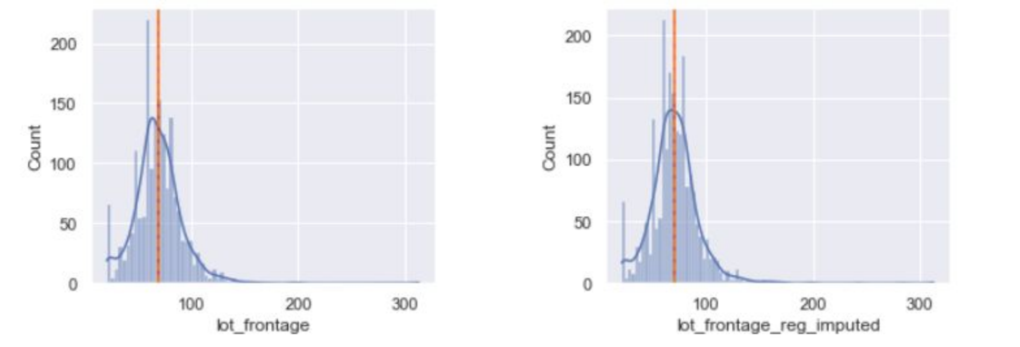
Linear Regression, Ridge, Lasso, Elastic Net

4 **Select Best Model**

Tune Hyper parameters, compare RMSE

Imputing Missing Values



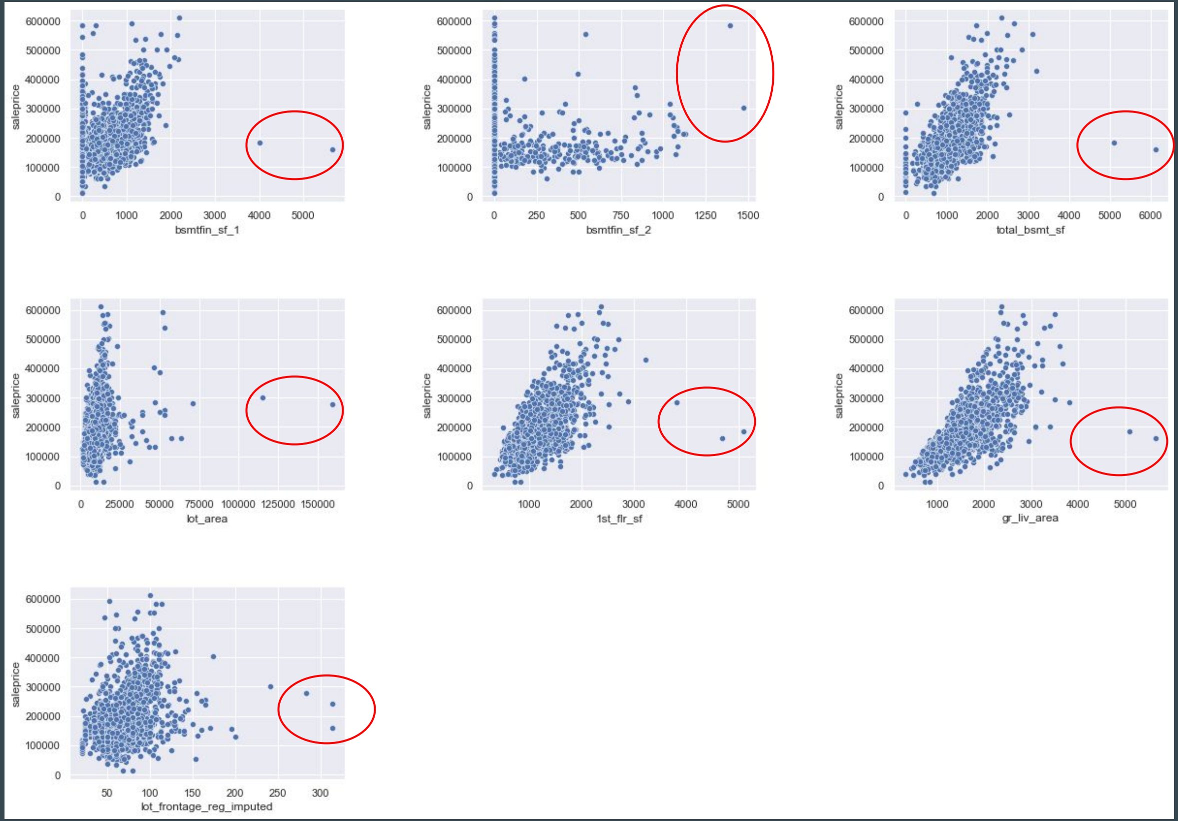


330 Missing Values in lot_frontage

- Impute through Linear Regression
- Distribution remain similar after imputation

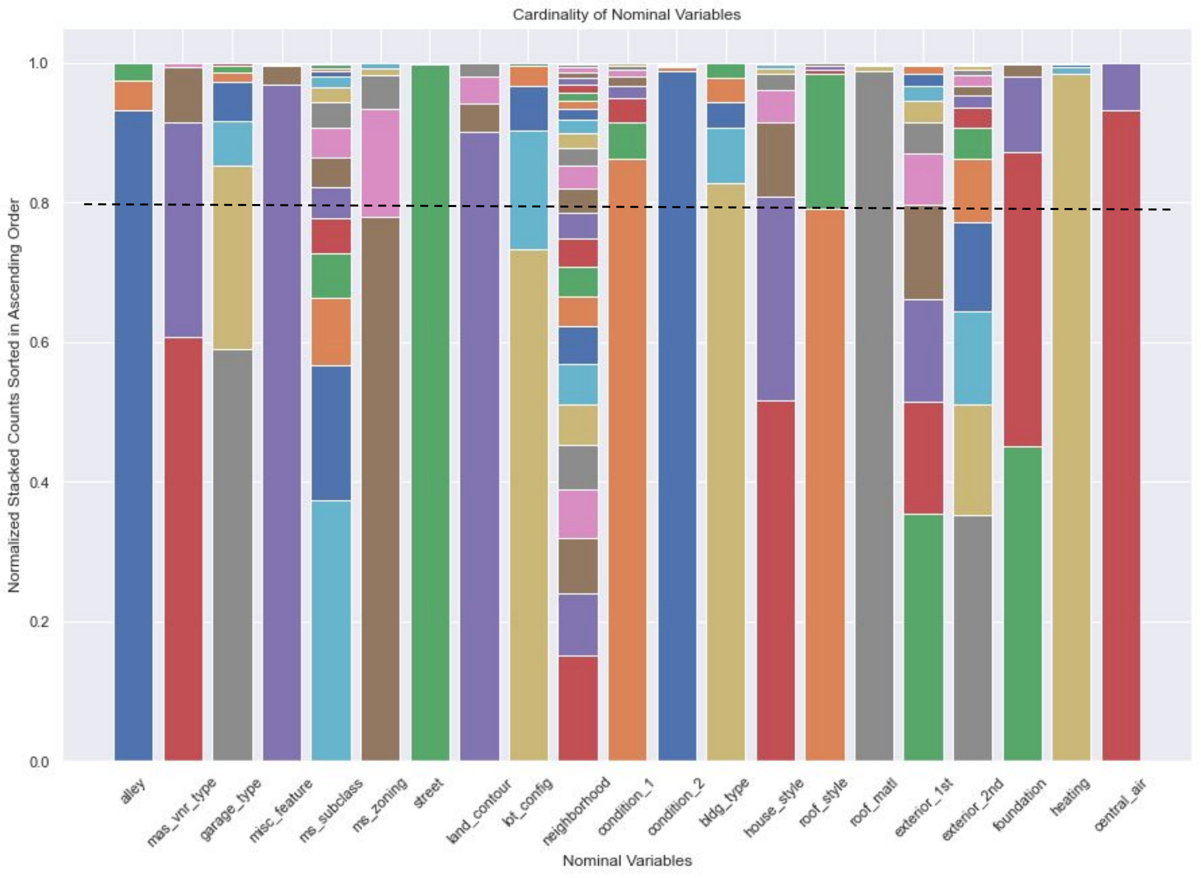
Outliers

- Scatterplot of continuous variables against saleprice
- Outliers were removed if < 5%



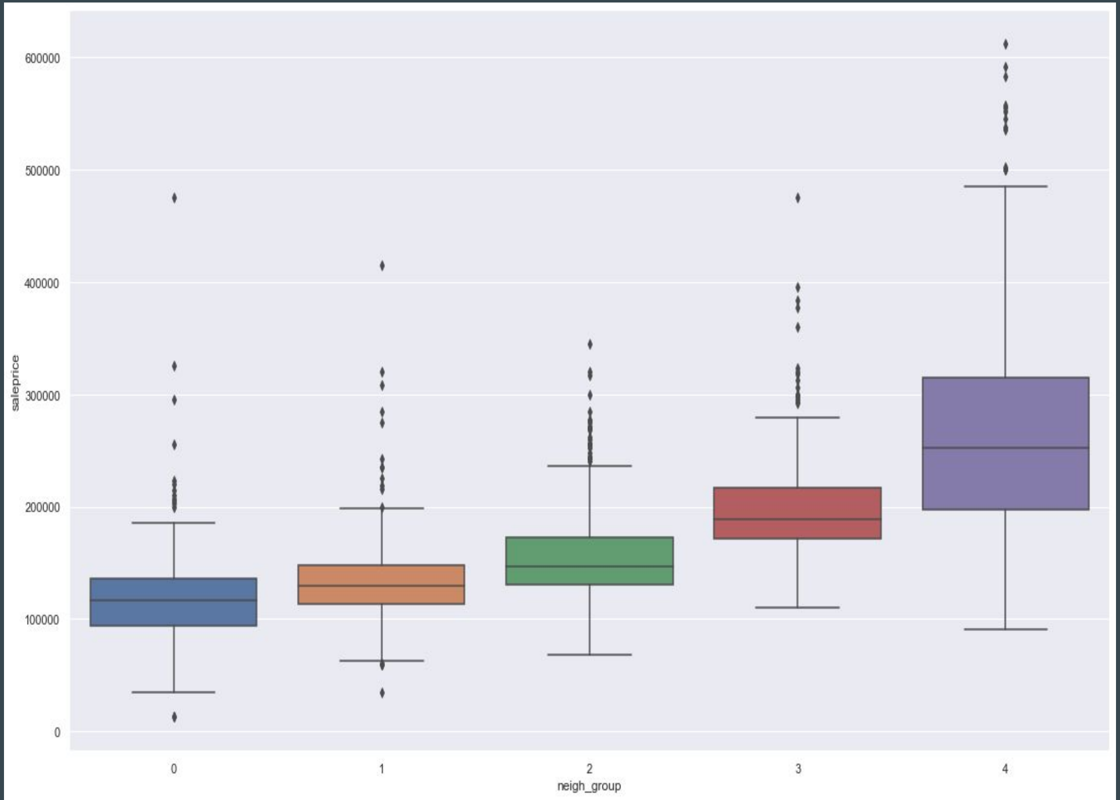
Reduce Cardinality

- Categories with < 20% representation were grouped together
- Reduce number of features after one-hot encoding



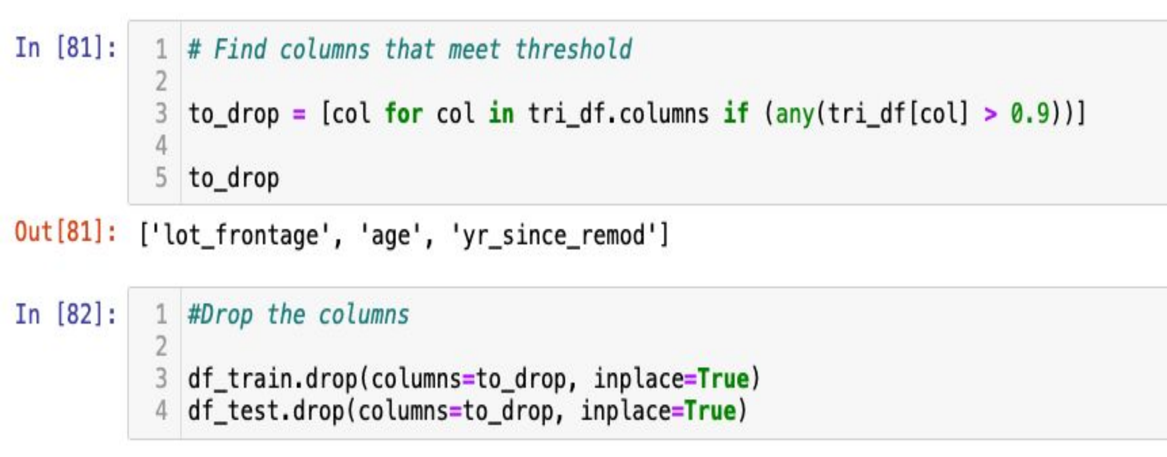
Reduce Cardinality

- Neighborhoods were split into 5 groups based on mean saleprice



Reduce Multicollinearity

- Drop one of pairs of variables if correlation exceeds 0.9

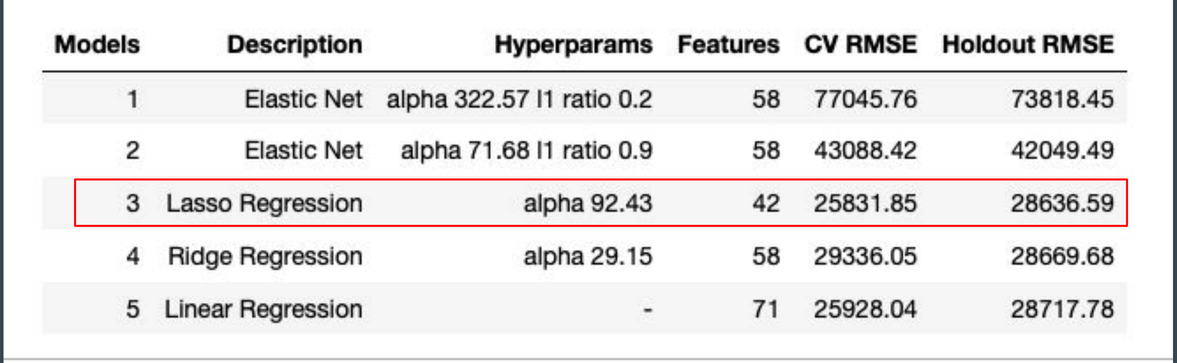


```
In [81]: # Find columns that meet threshold

         to_drop = [col for col in tri_df.columns if (any(tri_df[col] > 0.9))]

         to_drop
Out[81]: ['lot_frontage', 'age', 'yr_since_remod']
```

```
In [82]: #Drop the columns

         df_train.drop(columns=to_drop, inplace=True)
         df_test.drop(columns=to_drop, inplace=True)
```

Model Selected - Lasso Regression

- Best RMSE Score of 28636
- Reduce features from 75 to 42 through L1 regularization
- Achieved Kaggle score of 30403

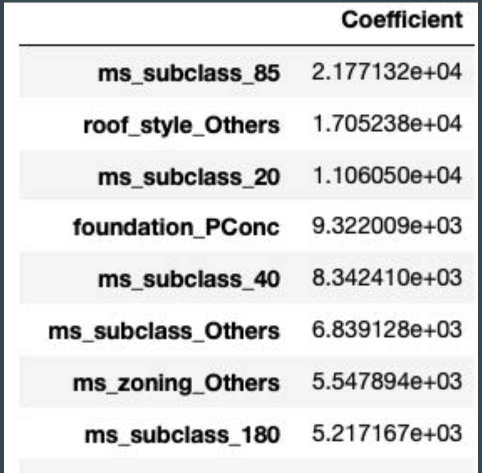


Models	Description	Hyperparams	Features	CV RMSE	Holdout RMSE
1	Elastic Net	alpha 322.57 l1 ratio 0.2	58	77045.76	73818.45
2	Elastic Net	alpha 71.68 l1 ratio 0.9	58	43088.42	42049.49
3	Lasso Regression	alpha 92.43	42	25831.85	28636.59
4	Ridge Regression	alpha 29.15	58	29336.05	28669.68
5	Linear Regression	-	71	25928.04	28717.78

Influence of Features

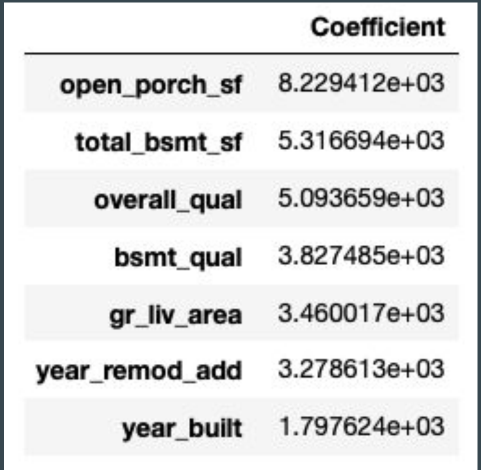
- Living Area Space and Quality are top continuous predictors
- House types are top categorical predictors
- Neighborhood groups was not selected by Lasso Regression

Top Influential Categorical Features



	Coefficient
ms_subclass_85	2.177132e+04
roof_style_Others	1.705238e+04
ms_subclass_20	1.106050e+04
foundation_PConc	9.322009e+03
ms_subclass_40	8.342410e+03
ms_subclass_Others	6.839128e+03
ms_zoning_Others	5.547894e+03
ms_subclass_180	5.217167e+03

Top Influential Continuous Features



	Coefficient
open_porch_sf	8.229412e+03
total_bsmt_sf	5.316694e+03
overall_qual	5.093659e+03
bsmt_qual	3.827485e+03
gr_liv_area	3.460017e+03
year_remod_add	3.278613e+03
year_built	1.797624e+03

Expand your porch!

Other recommendations and observation:

- Avoid houses with huge garage and precast coverings
- Neighbourhood are not good predictors for house price in Ames