Predicting Housing Prices (Ames, USA)

Modeling and Analysis

Content

- Problem Statement
- Exploratory Data Analysis (EDA) & Data Cleaning
- Feature Engineering / Selection
- Prediction Model
- Findings
- Business Insights / Conclusions

Problem Statement

~

Using the available data to build an optimum model that can predict the sale price of homes in Ames, Iowa for real estates agents to use as a reference. To analyze and study the important features, allowing the agents to advice their client (homesellers) on ways to increase their homes' value.

~

Exploratory Data Analysis (EDA) & Data Cleaning

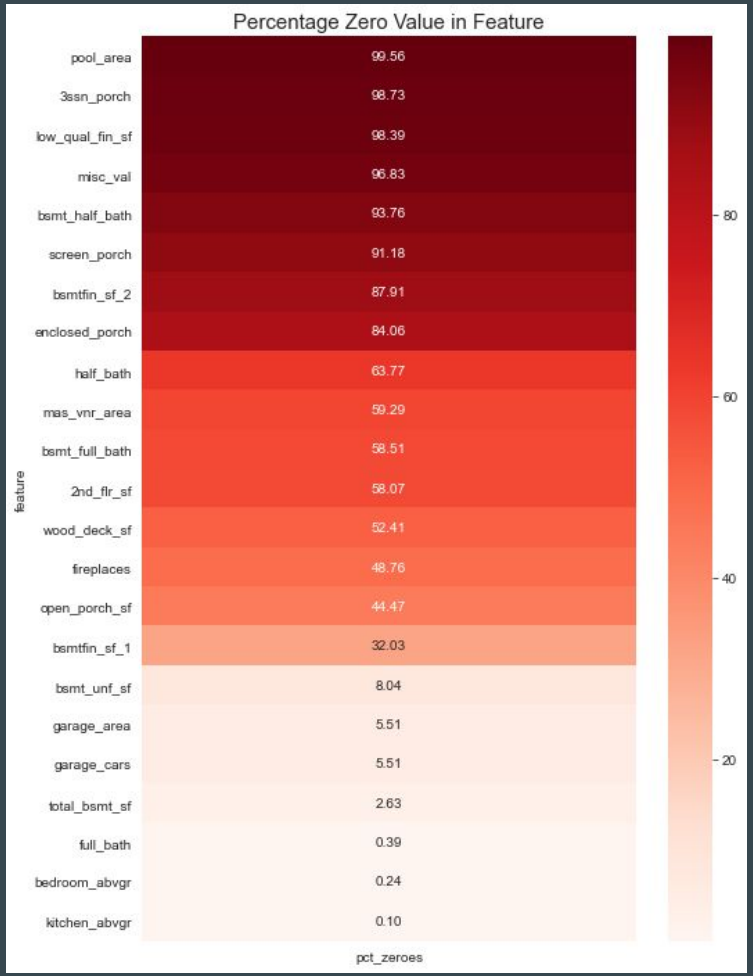
- Features with Missing Values
  - Null Values
  - Zero Values
- Features with High Percentage of Single Value
- Features with High Pairwise Collinearity
- Features with Low Correlation to Sale Price
- Other Preprocessing:
  - Removing Outliers
  - Handling Categorical Data

Dealing with missing values

- Dropped 4 features with over 80% null values.
  (Pool QC, Alley, Misc Feature, Fence)
- Systematically assign values to missing data via inference.
  (e.g. Impute Garage type with 'None', if all the other garage variables are null or 0)

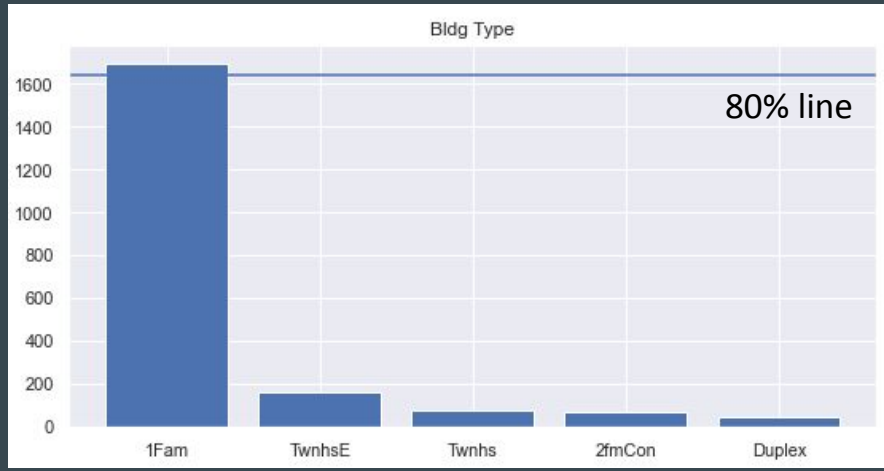
Dealing with Zero Values

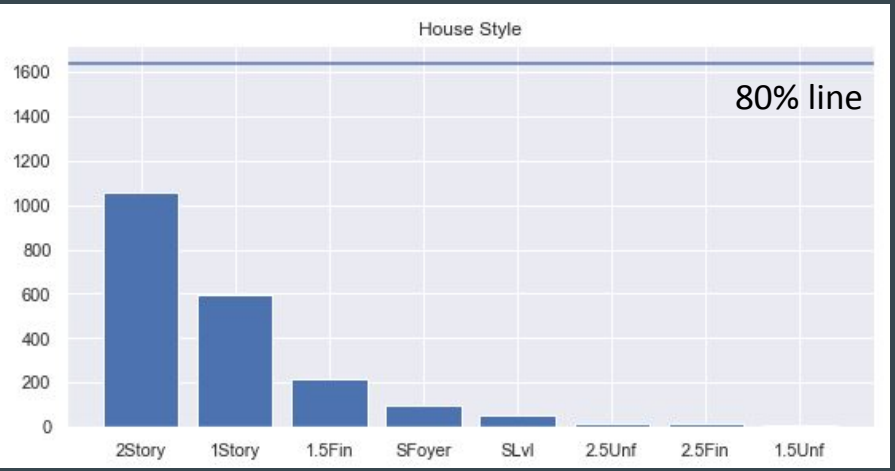
- Features such as Pool Area, 3 Season Porch, Screen Porch, Low Quality Finish Surface Area have very significant percentage of zero values.
- Dropped numerical features with more than 80% zero values.
- Systematically assigned values to missing data.



Features with High Percentage of Single Value

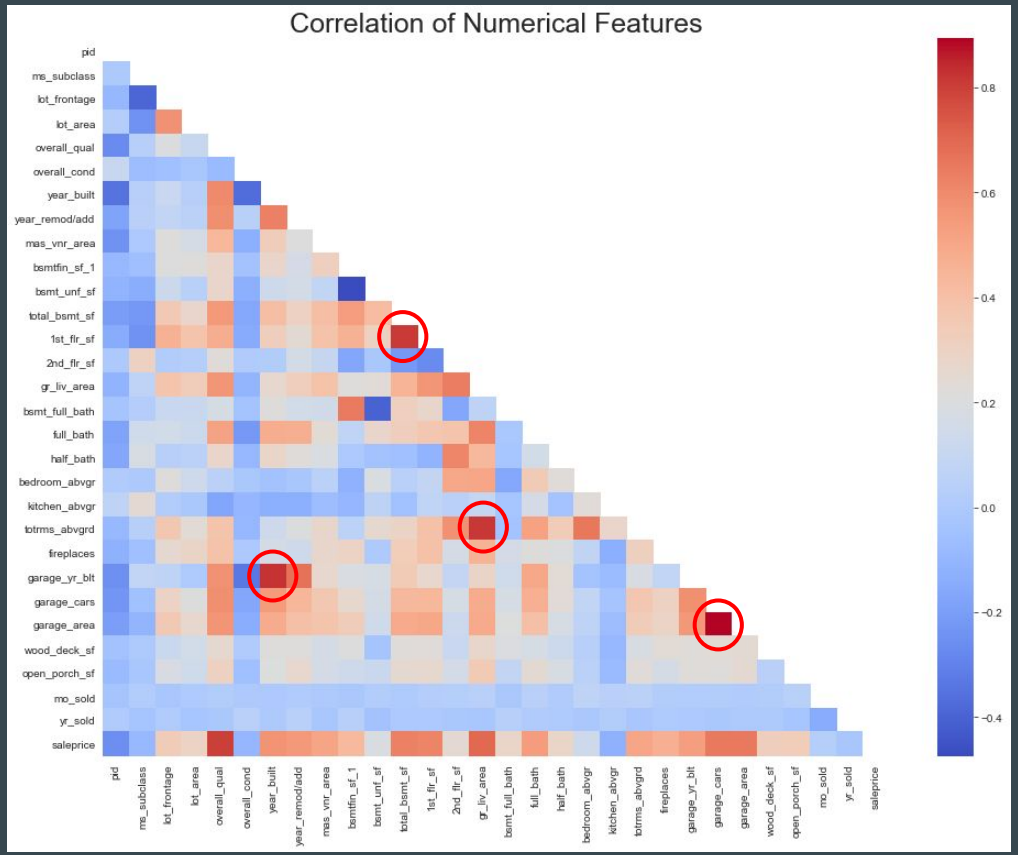
- Bar chart was plotted for categorical features.
- Any feature with a class exceeding the 80% was removed as they have insufficient variance and does not contribute to understanding of sale price.





Features with High Pairwise Collinearity

- 4 Pairs of features had high pairwise collinearity (correlation > 0.8).
- The pairs were explored, and only one feature was retained (with better correlation with sale price).



Features with Low Correlation to Sale Price

- Explored features with low correlation to sale price.
- Removed features whose distribution showed that it does not contribute to the understanding of sale price.

Feature Engineering

Interaction Terms

Code Snippet:

1.8 Using Interaction Terms to combine features

```
def interaction_terms(ames_train):

    ames_train['House Age']=ames_train['Yr Sold']-ames_train['Year Built']

    ames_train['Years since Remod/Add']=ames_train['Yr Sold']-ames_train['Year Remod/Add']

    ames_train['Garage Age']=ames_train['Yr Sold']-ames_train['Garage Yr Blt']

    ames_train.drop('Yr Sold', axis=1, inplace=True)

    ames_train.drop('Year Built', axis=1, inplace=True)

    ames_train.drop('Year Remod/Add', axis=1, inplace=True)

    ames_train.drop('Garage Yr Blt', axis=1, inplace=True)
```

executed in 4ms, finished 18:53:26 2020-11-22

```
interaction_terms(ames_train);
```

executed in 9ms, finished 18:53:26 2020-11-22

Dropping “Unnecessary” Features

Code Snippet:

```
train_num[['1stflrsf','2ndflrsf','lowqualfinsf','grlivarea']].sum()
```

```
1stflrsf         2378578
2ndflrsf          674504
lowqualfinsf       11307
grlivarea        3064389
dtype: int64
```

- Sum of area of 1st floor, 2nd floor and low quality finish were exactly equal to area of general living (above ground).

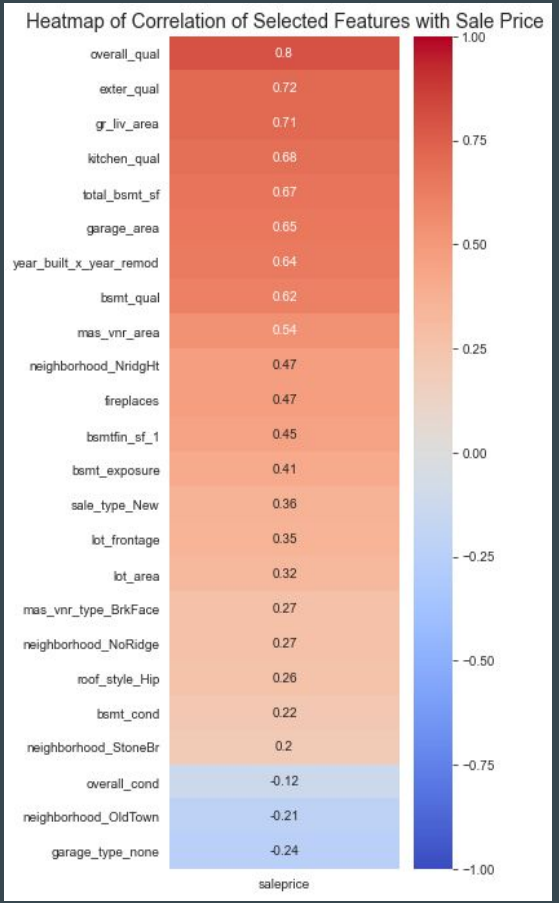
Feature Selection

Recursive Feature Elimination

1. Using the RFE function, selected 30 features
2. Modeling using the chosen 30 features using Lasso and Ridge model. It produced the results below:
    - Lasso: R2(Test) - 0.898
    - Lasso: RMSE(Test) - 26k
    - Ridge: R2(Test) - 0.899
    - Ridge: RMSE(Test) - 27k

LassoCV and RidgeCV Coefficient

1. Ran LassoCV and RidgeCV on all features
2. Generate two lists of 30 features with the highest absolute coefficient from LassoCV and RidgeCV
3. Among the two lists, there are 39 features in total (21 features overlapped)
4. 24 features with the highest correlation to sale price were selected for final model building



Results

Features Used	Model	R2 (Train)	R2 (Test)	RMSE (Test)	Improvement over Baseline (%)
All Features (191)	Linear Regression	0.928982	-1.25069e+26	9.02279e+17	-1.11822e+15
	Lasso	0.926415	0.905644	24782.9	69.29
	Ridge	0.927557	0.905264	24832.7	69.22
	ElasticNet	0.927158	0.905484	24803.7	69.26
Selected Features (24)	Linear Regression	0.904043	0.90003	25509.4	68.39
	Lasso	0.904043	0.90003	25509.4	68.39
	Ridge	0.903969	0.900032	25509.1	68.39
	ElasticNet	0.903928	0.900009	25512.1	68.38

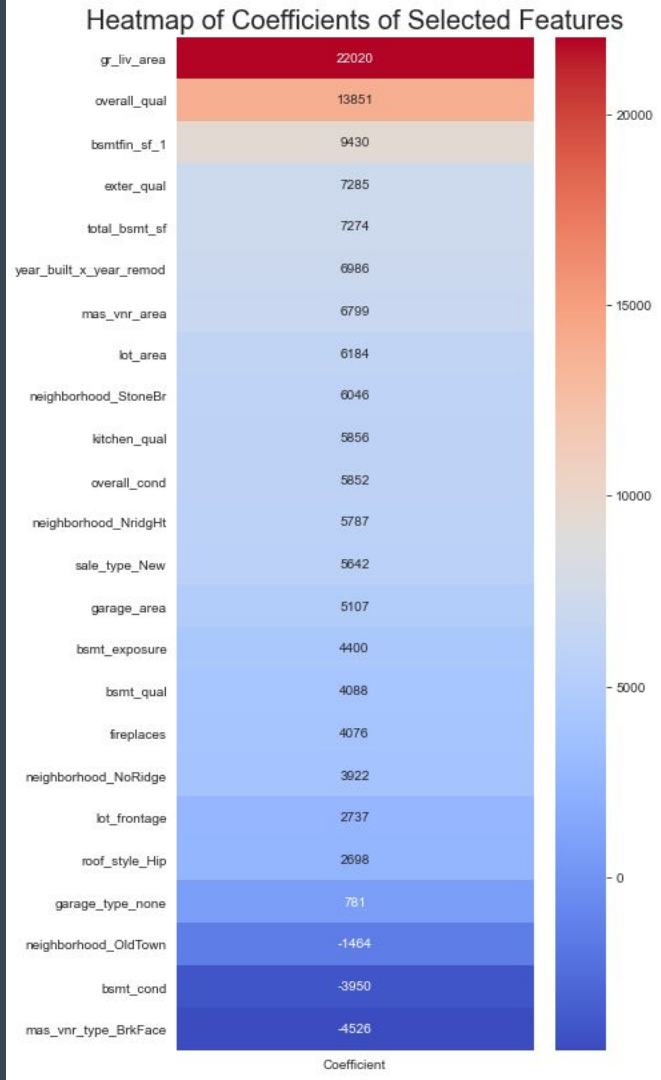
Baseline RMSE:
80688.95

Observations

- Linear Regression with all features performed very badly → Highly overfitted
- With all features, the Lasso model performed the best
- With selected feature, the Ridge model performed the best
- With selected feature, the Lasso model had the same result as Linear Regression → Features selected were all useful predictors

Business Insights / Conclusions

- Number of features was trimmed for better interpretability
  - Having more features would have led to better predictions.
- Size, quality (and condition), being located in particular neighborhoods and age of the house were the best predictors of sale price.
- Home owners wishing to sell their properties can increase the home value by:
  - Repair any defects in the house, Repaint the house → Improve the overall material and finish quality
  - Pursue expansion projects or re-purpose parts of the house to → Increase living area
  - Repairing / Renovate kitchen → Improve kitchen quality
  - Add a fireplace (if there is none)



Limitations

1. As mentioned previously, the sale price predicted by the model should only be used as a reference.

2. The data used is between 2006 to 2010 and may not fit well to present-day data.

3. Might not fit well to other locations.

Potential Improvements

1. Location specific features should be removed. (i.e. feature such as neighborhood)

2. Use data from a longer period of time. Allow the sale price to be better generalized.

3. New model for target location with different local culture.


Thank you!