Ames Housing Data

An analysis into the factors that influence housing price

Overview

- Problem Statement
- Background and Data
- Data Cleaning
- Exploratory Data Analysis
- Feature Selection
- Feature Engineering
- Modeling Process
- Results
- Takeaways
- Recommendations

Problem Statement

- As data analysts at a property agency in Ames, Iowa, we are tasked with conducting statistical analysis of housing transactions in Ames to identify prominent house features which affect house prices.
- Once we have identified these features, we are then able to provide advice and recommendation to potential home sellers to market their houses appropriately

Background and Data

- Features – 80 columns
- Training Data – 2051 rows
- Testing Data – 879 rows
- Time period – 2006 to 2010

CONTINUOUS	DISCRETE	NOMINAL	ORDINAL
20 variables	14 variables	23 variables	23 variables
Measurable	Countable	Categorical	Rankable
Examples: • Lot frontage • Lot area • Ground living area	Examples: • No. of bedrooms • No. of full baths • Year built	Examples: • Neighborhoods • Roof styles • Heating	Examples: • Basement quality • Kitchen quality • Garage finish

How did we clean the data?

Feature	Variable Type	Missing Value	Notes	Action
lot_frontage	Continuous	330	Missing value	impute NaN values with median value
alley	1911	Nominal	NaN because no alley access	drop data because too many NaN
mas_vnr_type	Nominal	22	Missing value	set null values to None, check 'None' rows should have mas_vnr_area = 0
mas_vnr_area	Continuous	22	Missing value	set null values to 0, check '0' rows should have mas_vnr_type = 'None
bsmt_qual	Ordinal	55	NaN because no basement	set null values to None
bsmt_cond	Ordinal	55	NaN because no basement	set null values to None
bsmt_exposure	Ordinal	58	3 more rows than bsmt_qual	drop the extra rows, set null values to None
bsmtfin_type_1	Ordinal	55	Missing value	set null values to None
bsmtfin_sf_1	Continuous	1	Missing value	set null values to 0
bsmtfin_type_2	Ordinal	56	55 no basement, 1 missing value	delete row with missing value, set null values to None
bsmtfin_sf_2	Continuous	1	NaN because no basement	set null values to 0

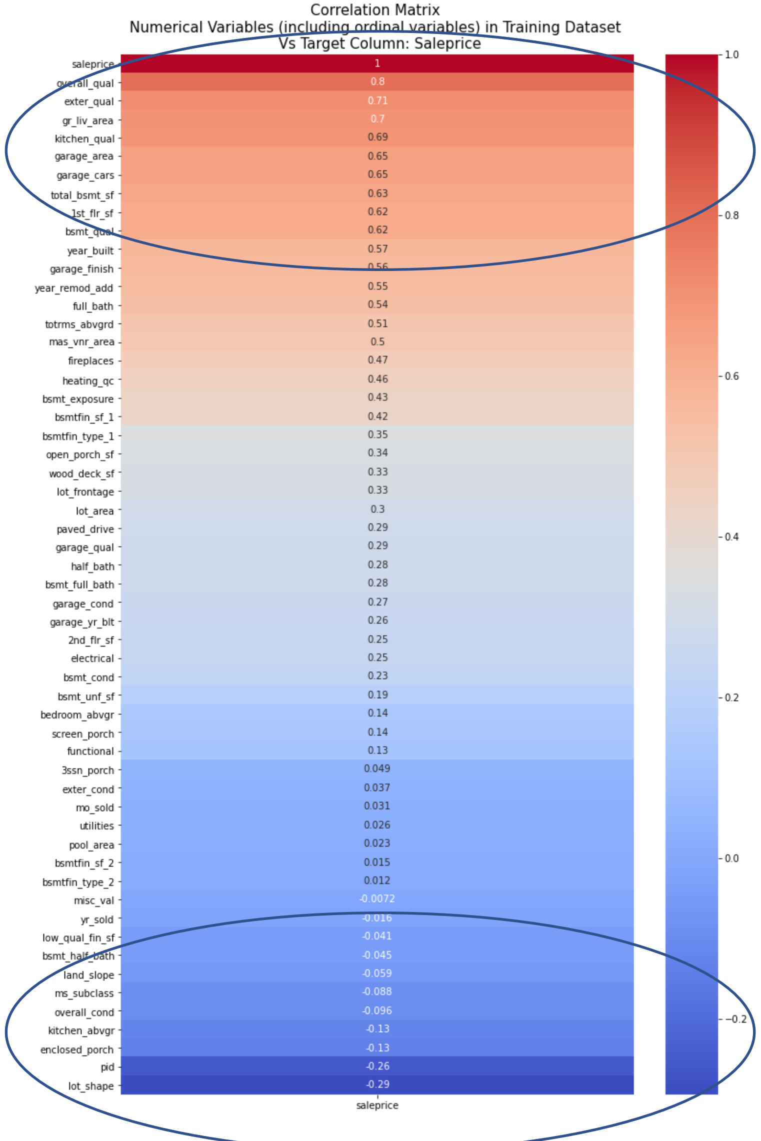
```
Bsmt Qual (Ordinal): Evaluates the height of the basement

       Ex	Excellent (100+ inches)
       Gd	Good (90-99 inches)
       TA	Typical (80-89 inches)
       Fa	Fair (70-79 inches)
       Po	Poor (<70 inches
       NA	No Basement
```

Handling null values

a) Need to understand the data dictionary provided
b) Most null values actually mean ‘NA’ or 0
c) Not possible to remove all rows with ‘NA’
d) Only dropping rows when null values cannot be intepreted

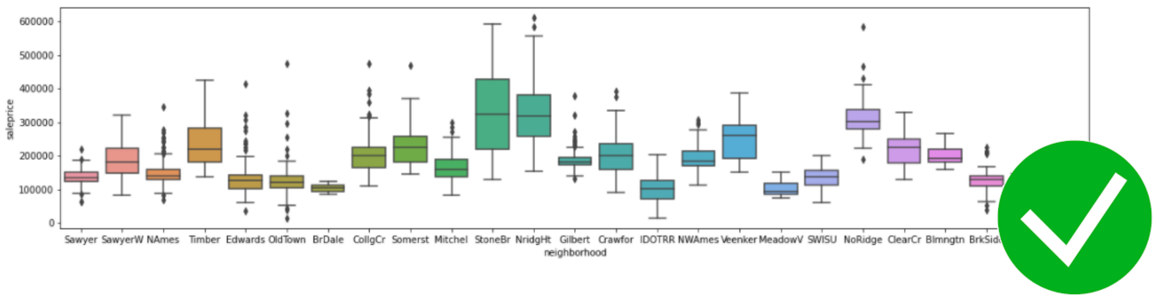
Digging into the training data shows which features correlates to house prices

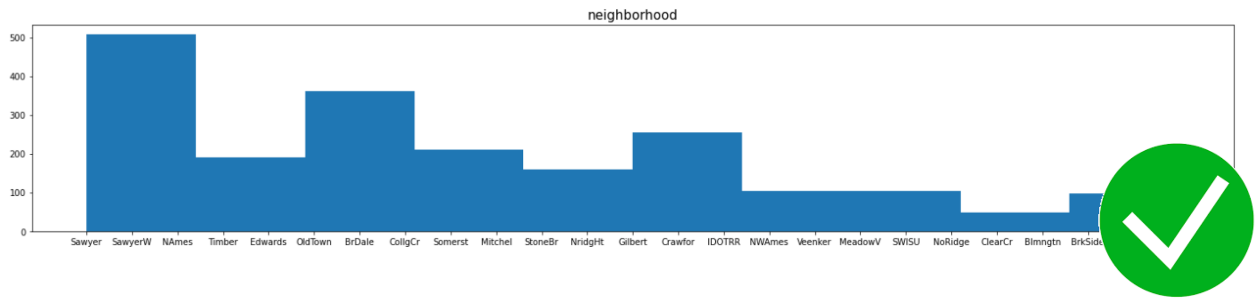


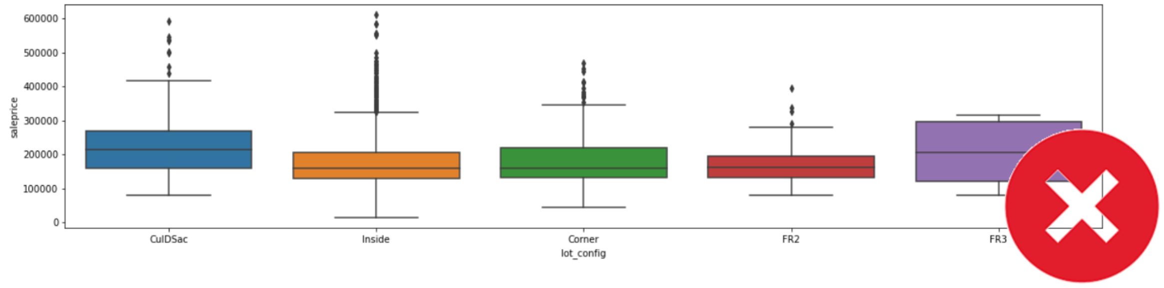
2 criteria in selecting nominal features as a good predictor of sale price

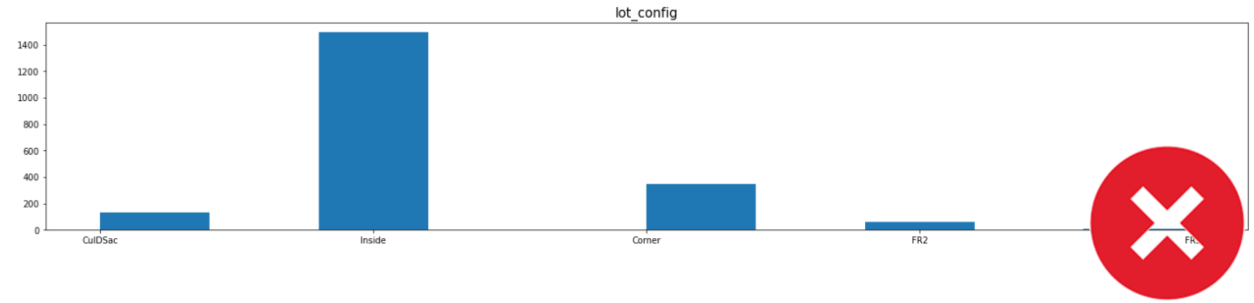
Selecting features showing **saleprice variability** within its categories

Selecting features showing **balanced distribution** of its category frequency









Feature engineering to fit our data into the models

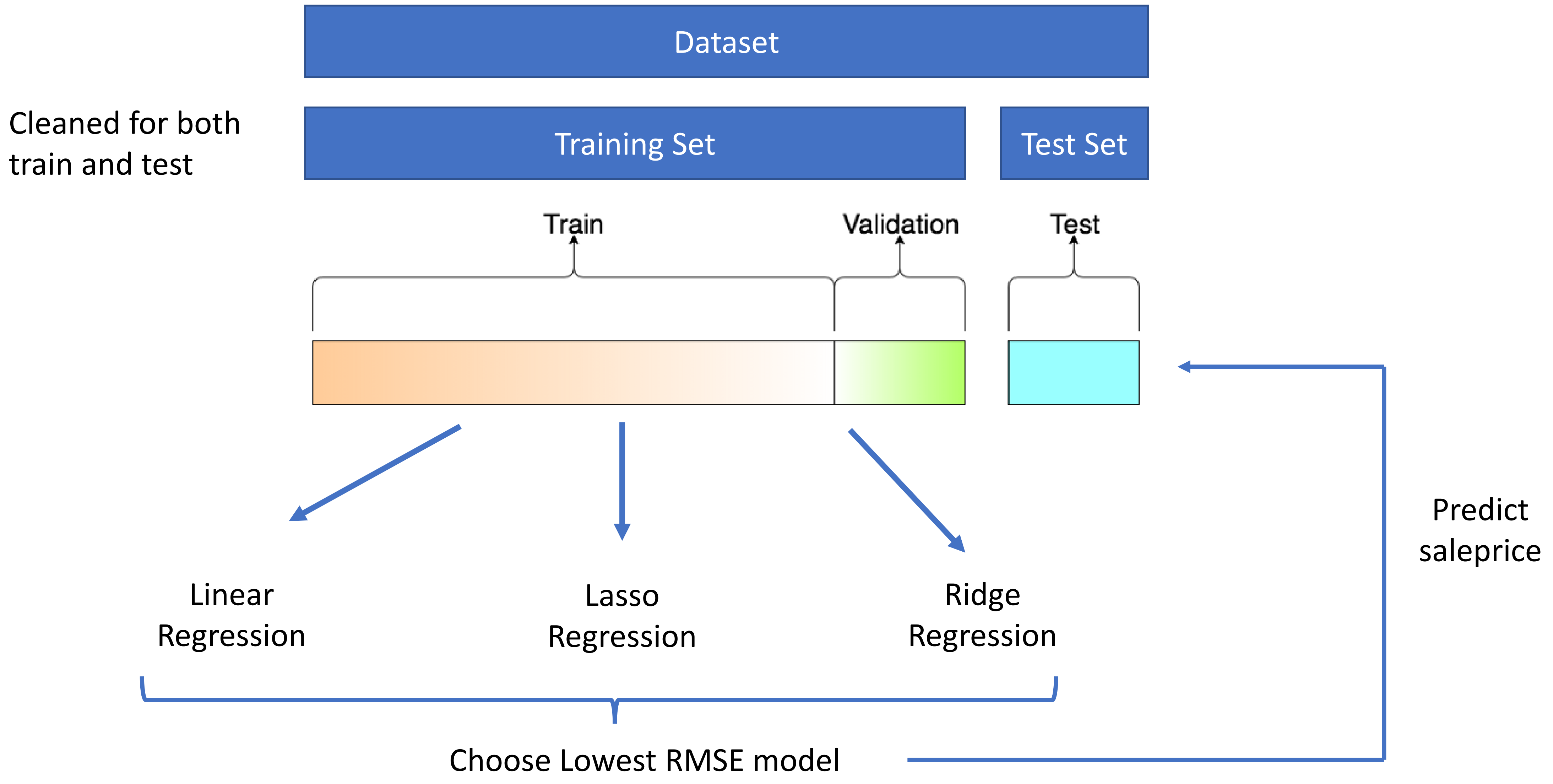
Continuous	Discrete	Nominal	Ordinal
Measurable	Countable	Categorical	Rankable
Course of action: Remain as is	**Course of action:** Remain as is	**Course of action:** Dummify via OneHotEncoder	**Course of action:** Dummify via value ranking

	mas_vnr_type_BrkFace	mas_vnr_type_None	mas_vnr_type_Stone
0	1.0	0.0	0.0
1	1.0	0.0	0.0
2	0.0	1.0	0.0
3	0.0	1.0	0.0
4	0.0	1.0	0.0
...	...	...	...
2035	0.0	1.0	0.0
2036	0.0	1.0	0.0
2037	0.0	1.0	0.0
2038	0.0	1.0	0.0

```
Bsmt Qual (Ordinal): Evaluates the height of the basement

5   Ex      Excellent (100+ inches)
4   Gd      Good (90-99 inches)
3   TA      Typical (80-89 inches)
2   Fa      Fair (70-79 inches)
1   Po      Poor (<70 inches
0   NA      No Basement
```

Modeling Process



Modeling Results on Training Set

Model	No. of Features Selected	RSME Score on Train Set	RSME Score on Validation Set
Linear Regression	31	30,252	32,260
Lasso Regression	31	30,818	31,916
Ridge Regression	31	30,975	32,142

* The lower the score, the better the model is at predicting saleprice

- Our goal is to get RMSE as close to zero as possible
- Linear Regression has the lowest RMSE on the train set
- Lasso Regression gives the lowest RMSE on the validation set
- We decide to base our choice from the RMSE of the validation set as this set assesses the model's ability to generalize to unseen data
- Therefore, we will be using Lasso model to predict the saleprice on our test dataset.

Takeaways: features as good predictors

	features	coeff
3	overall_qual	12719.028165
14	1st_flr_sf	9189.113527
16	gr_liv_area	8823.436159
2	neighborhood	8687.323258
17	kitchen_qual	8229.270898
6	mas_vnr_area	7264.888200
18	totrms_abvgrd	5280.468776
9	bsmt_exposure	5209.731898
23	garage_area	5117.516787
19	fireplaces	5038.462592

1. Space and area is one of the most important features
   a. Evidenced by highly ranked features like 1st floor square feet, above ground living area, masonry veneer area, garage area
2. Location, location, location (neighborhood)
3. Total number of rooms (excluding bathrooms)

Recommendations to our potential home sellers

2

Home buyers in Ames **prioritize space**

- consider to enhance their homes
- <u>increase livable space</u> especially on the first floor and garage area to increase its valuation

1

Real estate has a huge emphasis on **location**, therefore potential home sellers should be mindful of which neighborhoods their homes are located at

3

Increase **number of rooms** in home

- Consider reducing the size of your spacious living room or large kitchen and converting the space into another room e.g. study room or guest bedroom