Project 2:Ames Housing Dataset

Saba Suhail

Objectives

To predict the sale prices of houses for various stakeholders

- Analyze the dataset
- Understand relationships between various features by EDA
- Clean data-imputation,deletion
- Graphs for visual understanding
- Develop regression models

What was done?

- Train.csv(has sale prices)
- test.csv(lacks sale prices)

Certain columns altogether dropped such as Alley

Comparison with data dictionary algorithmically

Checking for negatives algorithmically

Exhaustive cleaning of train.csv and test.csv and results stored back

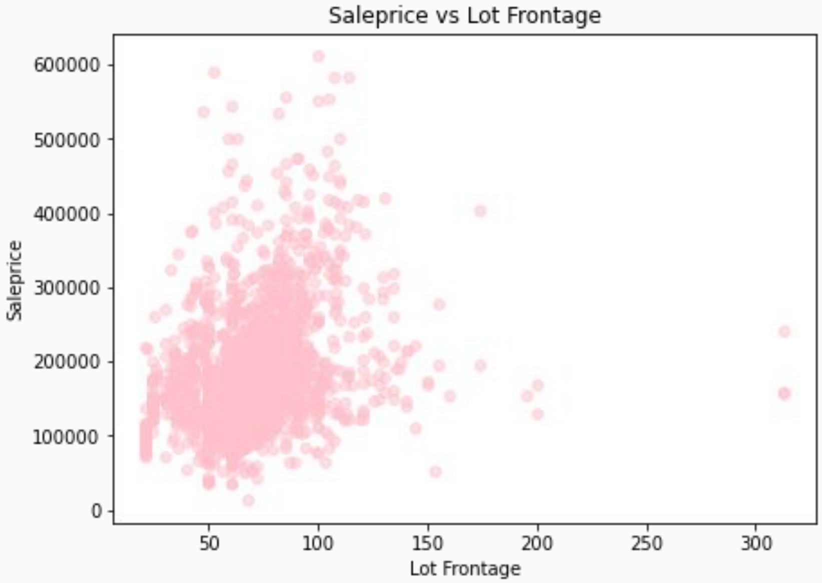
Can build any number of models in the future

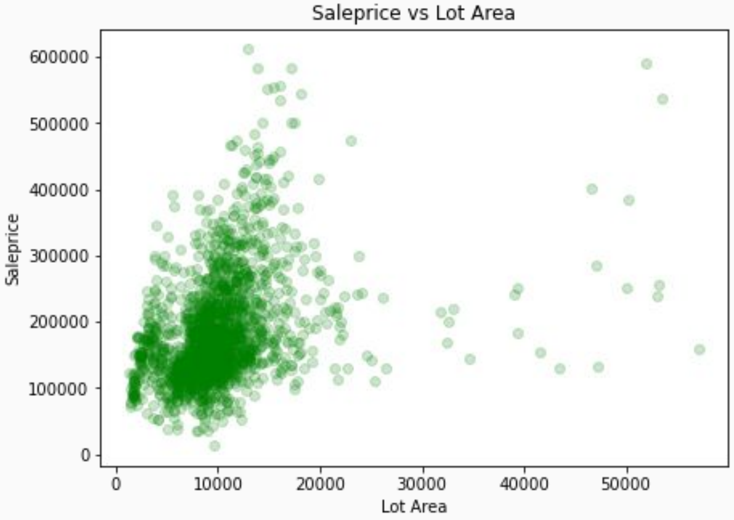
Cleaning of test.csv as per requirement

Scatter plot for all features

Preprocessing-Imputing,StandardScaler()

Illustrations Examples





Feature Engineering

Features Used:

'home_age','ms_zoning_RM','foundation_CBlock', 'overall_qual','total_sf','exter_qual','gr_liv_area',

'kitchen_qual','garage_area','total_bsmt_sf','baths','totrms_abvgrd',

'mas_vnr_area','fireplaces','year_remod/add','heating_qc','neighborhood_NridgHt',

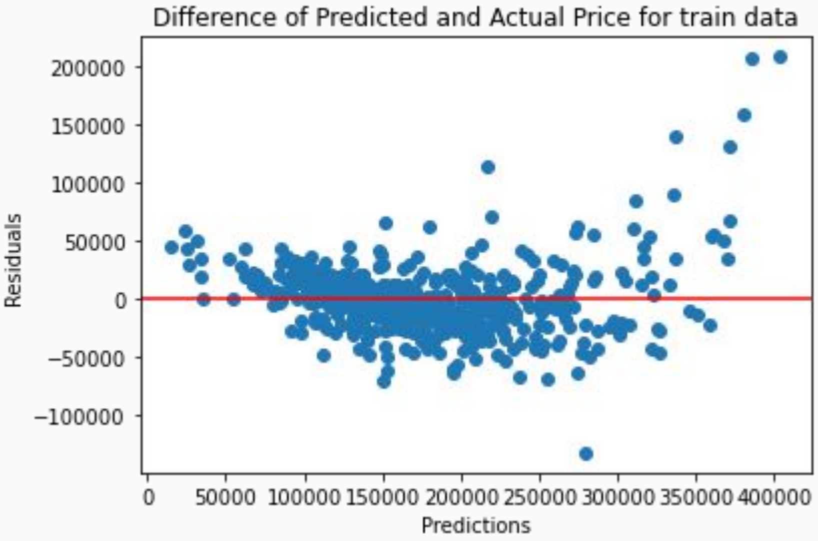
Features Generated:

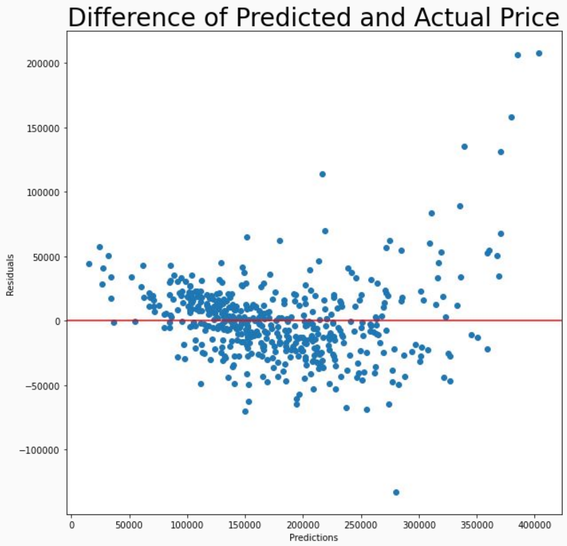
df['baths'] = df['bsmt_full_bath'] + df['full_bath'] + (df['bsmt_half_bath']/2) + (df['half_bath']/2)

df['home_age'] = df['yr_sold'] - df['year_built']

df['total_sf']=df['1st_flr_sf'] + df['flr2nd_sf']

Plots for residuals for different models





Models Employed

- Linear Regression
- RidgeCV
- LassoCV
- LassoCV with optimal alpha
- RidgeCV with optimal alpha

Comparable results across models

Low variance

Low bias

Model is not scoring as well on a few high price homes so there is still room to improve.

Limitations of models

Relying on discrete values for ordinal variables is very preliminary and really don't reflect the actual conditions.

The scope of this dataset is limited, both in terms of time frame and geographic range.

Any mathematical model can not really capture behavioral aspects.