Project 2 - Ames Housing Data Challenge

Introduction and Problem Statement

- When it comes to real estate pricing, there is a famous old saying:

  "There are three things that matter in property: location, location, location."

- This project aims to produce a model can be used as a tool by real estate agents to aid their price evaluation of residential properties and seek to identify any attributes that influence property prices.

Data Available

- Ames, Iowa housing price data set is used to train our model
- Recoded in 2006 to 2010
- 81 features

Data cleaning

Simple imputation:

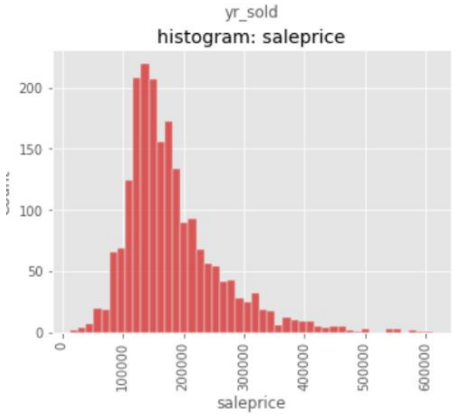
- 460 observations for `lot_frontage` imputed using median values.
- Observations Id 2237 and Id 1357 for `garage_yr_blt` imputed using median values
- Observation Id 1578 for `electrical` imputed using mode
- Observation Id 1357 `garage_finish` imputed using mode
- Observation Id 2237 `garage_finish`, `garage_cars`, `garage_area`, `garage_qual`, `garage_cond` imputed with median values or mode

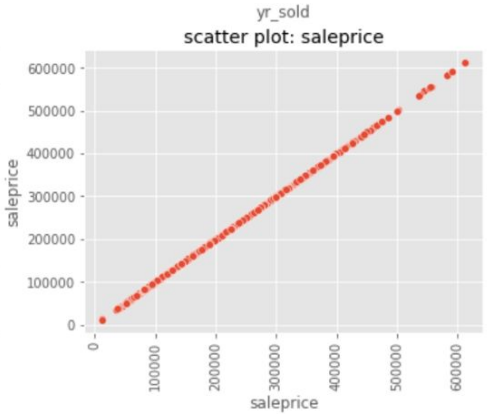
Logical imputation:

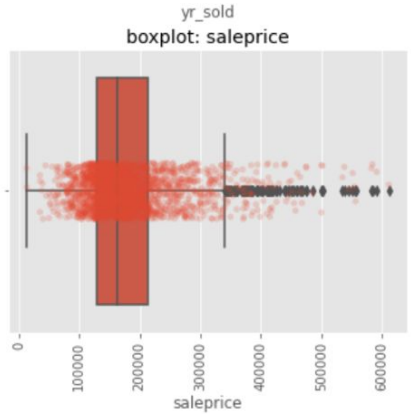
- For the remaining missing values, it was assumed that each of the value was meant to be "None" or "0" but was incorrectly left empty instead.

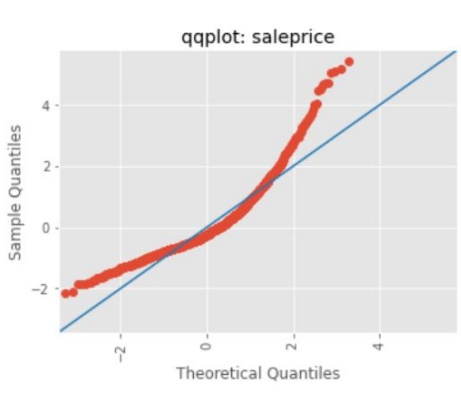
Exploratory Data Analysis

- Visualised the histogram, scatter plot, box plot, and qq plot
- Issues that could impact machine learning :
  - skewed data,
  - Outliers
  - Poor distribution









Modeling

- First iteration

Summary of results on unseen data set:

	Dummy regression	Linear regression	Ridge	Lasso	Elastic net
r2	-5.251×10^{-5}	-9.320×10^{19}	0.9322	0.9289	0.9309
rmse	0.3847	1.174×10^{12}	0.1002	0.1026	0.1011
rmse*	74990	N.A.	17400	17140	17210

Modeling

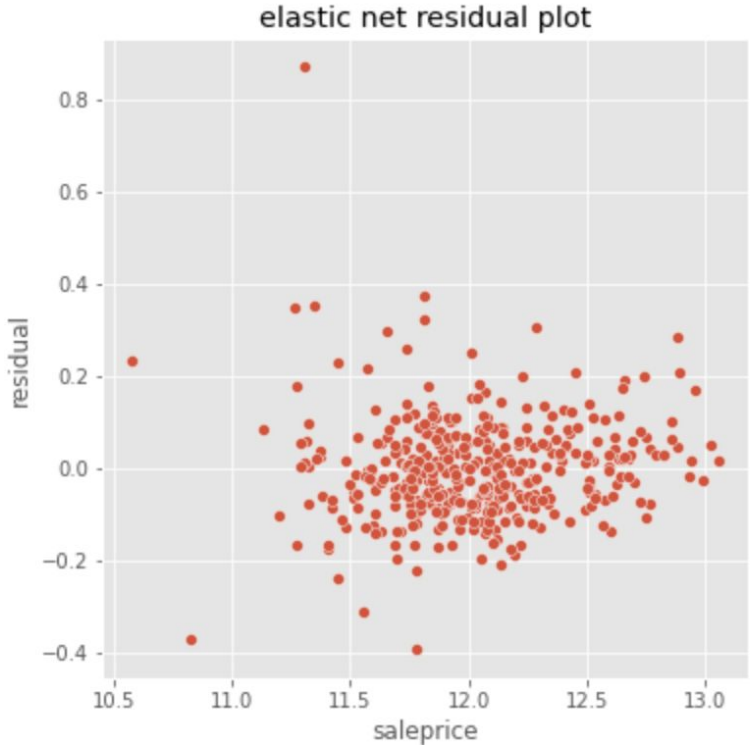
- Second iteration
- Top 30 features from lasso model in the first iteration
- The performance of all 3 models were very similar
- Elastic net is chosen instead. As final model. i.e. reduces the coefficients of predictors while not eliminating them

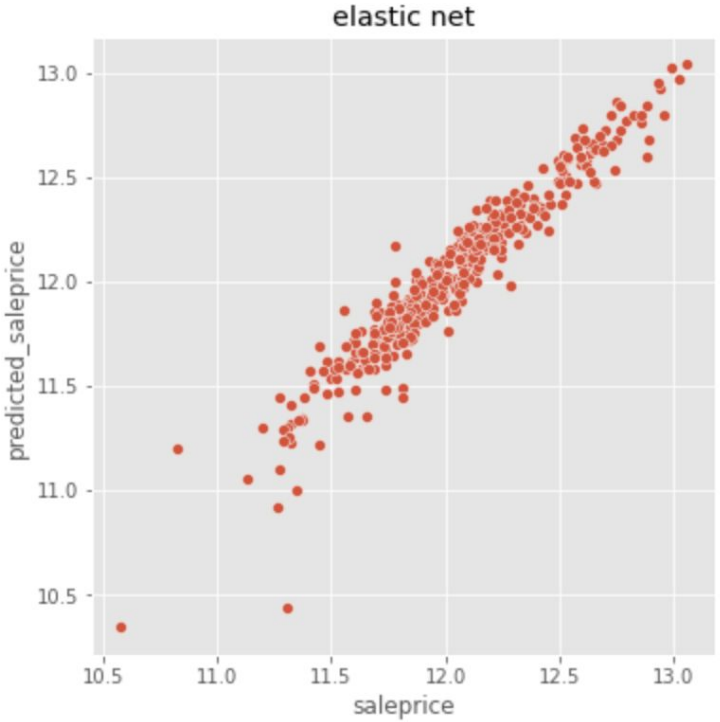
	Linear regression	Ridge	Lasso	Elastic net
r2	0.9164	0.9180	0.9171	0.9179
rmse	0.1112	0.1101	0.1108	0.1102
rmse*	18140	18047	18120	18070

Coefficients of the elastic model

Features	Coefficients	transformation
gr_liv_area	0.132990	log
overall_qual	0.084052	-
overall_cond	0.056268	log
year_built	0.069774	square
total_bsmt_sf	0.045902	log
lot_area	0.048071	log
ms_zoning_A (agr)	-0.036443	-
bsmtfin_sf_1	0.026404	log
ms_zoning_C (all)	-0.018633	-
bsmt_qual	0.015030	square
garage_cars	0.030172	-
neighborhood_NridgHt	0.015355	-
exterior_1st_BrkFace	0.015216	-
kitchen_qual	0.020225	-
bsmt_full_bath	0.018473	log
neighborhood_GrnHill	0.015582	-
functional_Typ	0.016411	-
heating_Grav	-0.014586	-
year_remod/add	0.012948	-
neighborhood_StoneBr	0.014512	-
screen_porch	0.012728	log
kitchen_abvgr	-0.012605	log
bsmt_exposure	0.011451	log
heating_qc	0.017885	-
fireplaces	0.015078	log
sale_type_New	0.016299	-
neighborhood_Crawfor	0.014046	-
paved_drive	0.008841	square
neighborhood_Edwards	-0.008357	-
condition_1_Norm	0.014311	-

Residual plots and distribution plots





Conclusions

Key housing attributes that influence prices

- RMSE score 19,953 in kaggle
- top predictors that influenced the target price the most were gr_liv_area, overall_qual, overall_cond, year_built, total_bsmt_sf, lot_area.
- Log transformation reduces the interpretability of the model's coefficients i.e. practical relationship between individual feature against sale price.

Improvement:

Review the features used in final model as of the featreview the features used in final model as of the features are poorly distributed and likely not a good candidate as a predictor e.g. paved_driveures

Recommendations

Final model is fit for price housing price prediction in Ames, Iowa. Nevertheless, as the most of the predictors were skewed, transformation was applied and hence complicate the interpretation of the linear coefficients. As such, for this model, it is not easy to directly quantify the influence of sale price for every unit increase of each predictor.