Housing Sale Price Prediction

Introduction

The following is a data science challenge problem. In order for us to assess your coding as well as problem-solving skills, please solve it and prepare (1) a working version of your script (e.g. Jupyter Notebook), and (2) brief documentation that clearly address the following aspects:

(1) An explanatory data analysis (EDA) that elaborates the distribution of the features, and any interesting observations one can learn from the plots (optional);

(2) Details of data pre-processing steps and/or feature engineering that you think are necessary;

(3) Compare the model performance among (i) LASSO (or kernel ridge regression), (ii) random forests regression, (iii) XG-Boost, and (iv) stacking (v) Any other method you find effective.

approach. Which one has the best performance, and why?

(4) A screenshot showing the best submission score on public leader board that you achieve. You will be judged on the score and rank you achieve on the leader board primarily. You will also be judged on the simplicity of your solution. For a given score, we prefer the simplest possible solution.

We would stress that the key to success here is to demonstrate your creativity and proficient capability of (i) data analysis, (ii) feature engineering, and (iii) building efficient machine learning models, etc. Ultimately, we are interested in seeing the simplest possible solutions that are maximally effective.

We thank you in advance for your time and efforts, and we are looking forward to your creative solution!

Introduction(explanatory data analysis (EDA):

We first analyze the data and for finding trends in data. We perform dimensionality reduction on the dataset using PCA algorithm and feature selection module in sklearn package for python. We predict the final house prices using linear regression models like Ridge and Lasso. We also use advanced regression techniques like gradient boosting using XGBoost library in python.
Most of the variables were directly related to property sales. Description for these variables can be found here. Approximately 80 variables focus on the quality and quantity

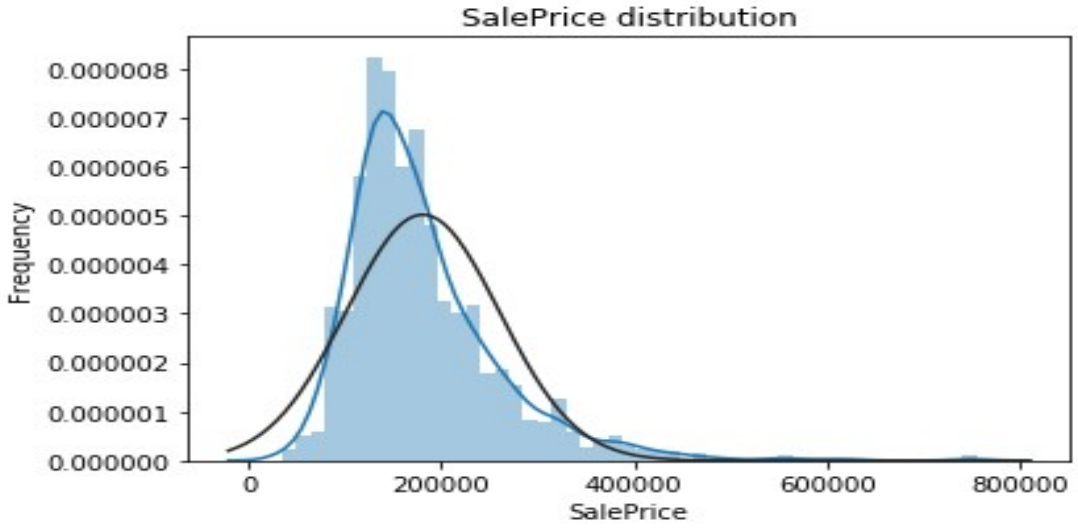
Explanatory data analysis (EDA)

2.1 Input

This is a small step which involves taking input train:csv and test:csv as dataframes. The Train data has 81 columns and 1460 rows. These columns include 79 explanatory variables to describe every aspect of house. Test data is fairly similar to Train data. Test Data has 80 columns and 1459 rows. It has no SalePrice column as it's value is to be predicted by our model.

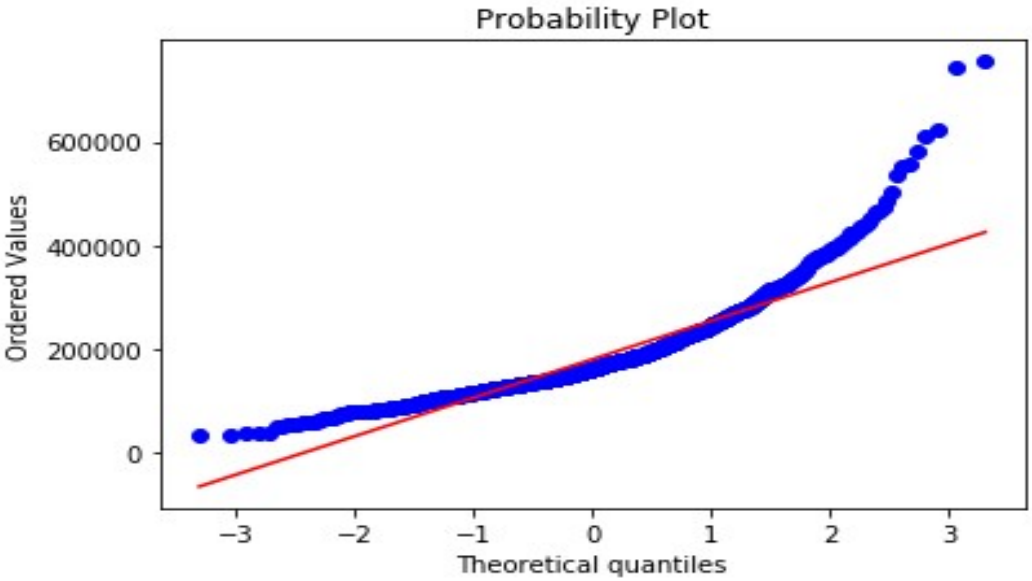
SALES DRITRBUTION

Early Analysis We study the Train data closely, through various graphs, to determine any trends in the data. The first graph is a distribution of Sale Price distribution. The below graph is close to a bell curve. We use the seaborn package to plot the graph. The below graph appears to be somewhat right skewed. This suggests that the mean for Sale Price is greater than its median.



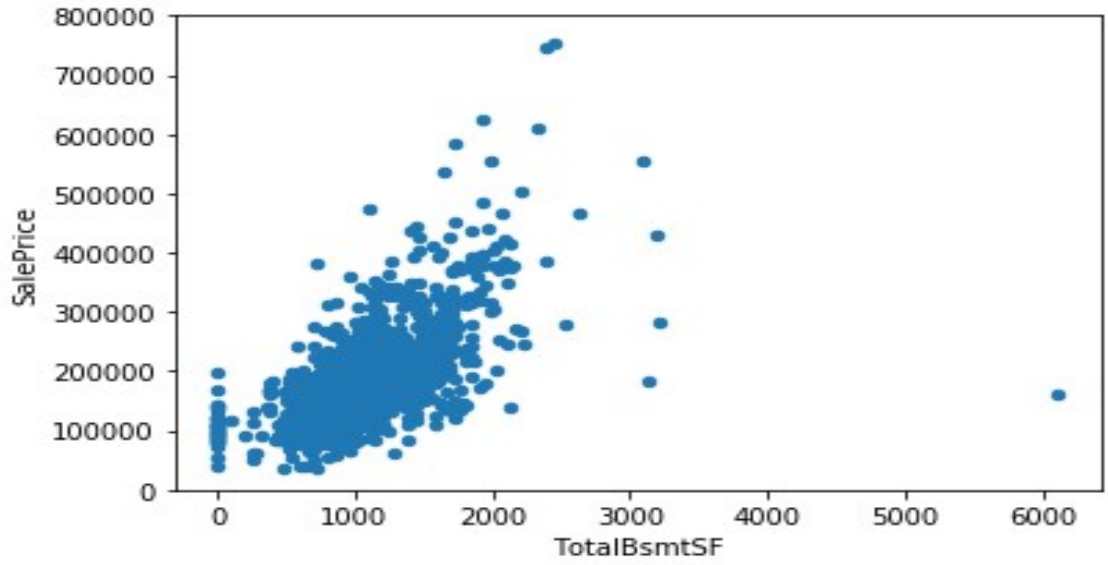
PROBABILITY PLOT

The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line. Departures from this straight line indicate departures from the specified distribution. The correlation coefficient associated with the linear fit to the data in the probability plot is a measure of the goodness of the fit. Estimates of the Ordered Values and Theoretical quantiles of the distribution are given by the intercept and slope.



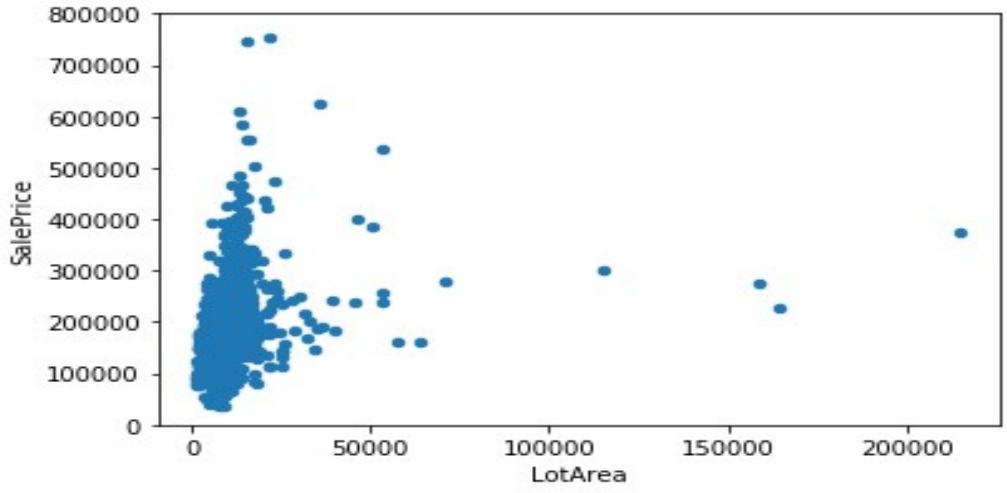
Scatter plot total Bsmtsf/sale price

A **scatter plot** is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. In this case we are Plotting along the Total BsmtSF and Sale price.

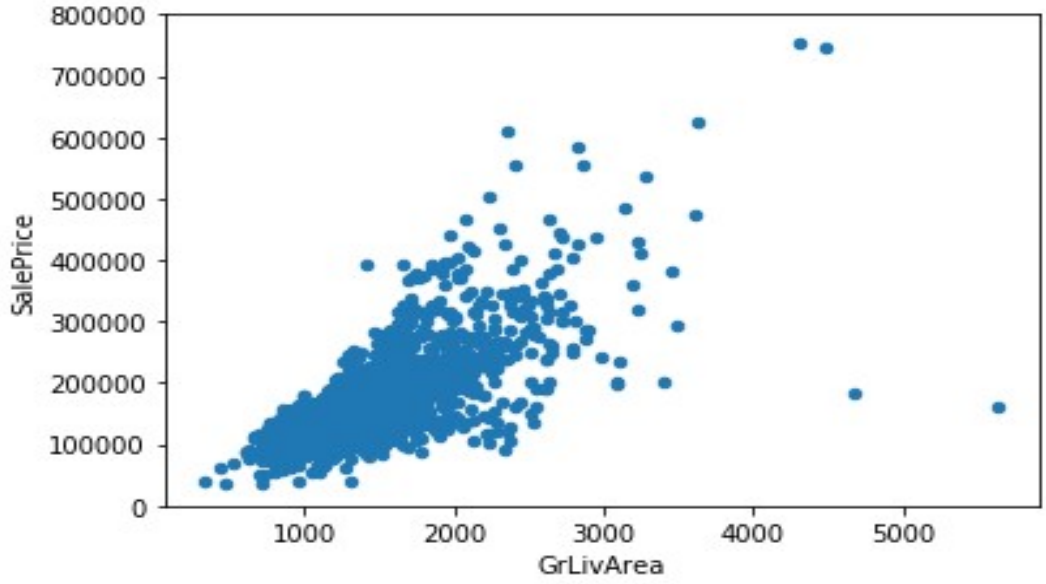


Scatter plot LotArea/sale price

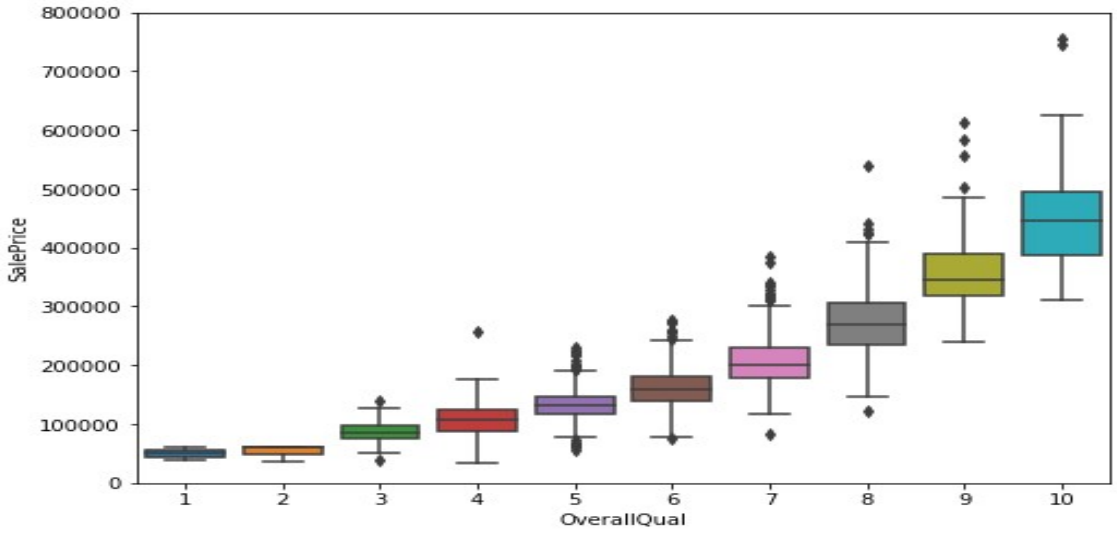
The Lot Area is nothing but the gross/total square unit of measure of every floor of a building, minus whatever the local zoning/building laws define as exceptions. Here, we are using the scatter plot to compare Sale price and Lot Area.



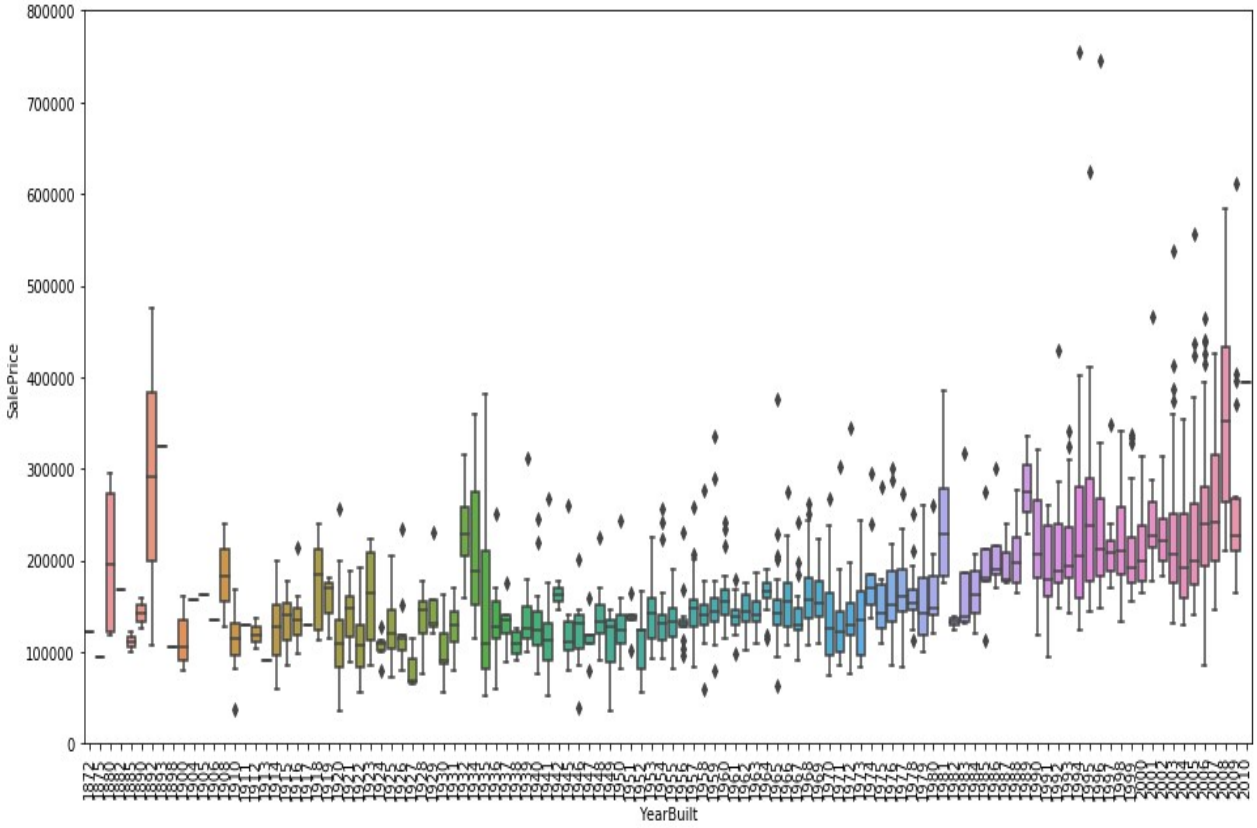
Scatter plot grlivarea/sale price



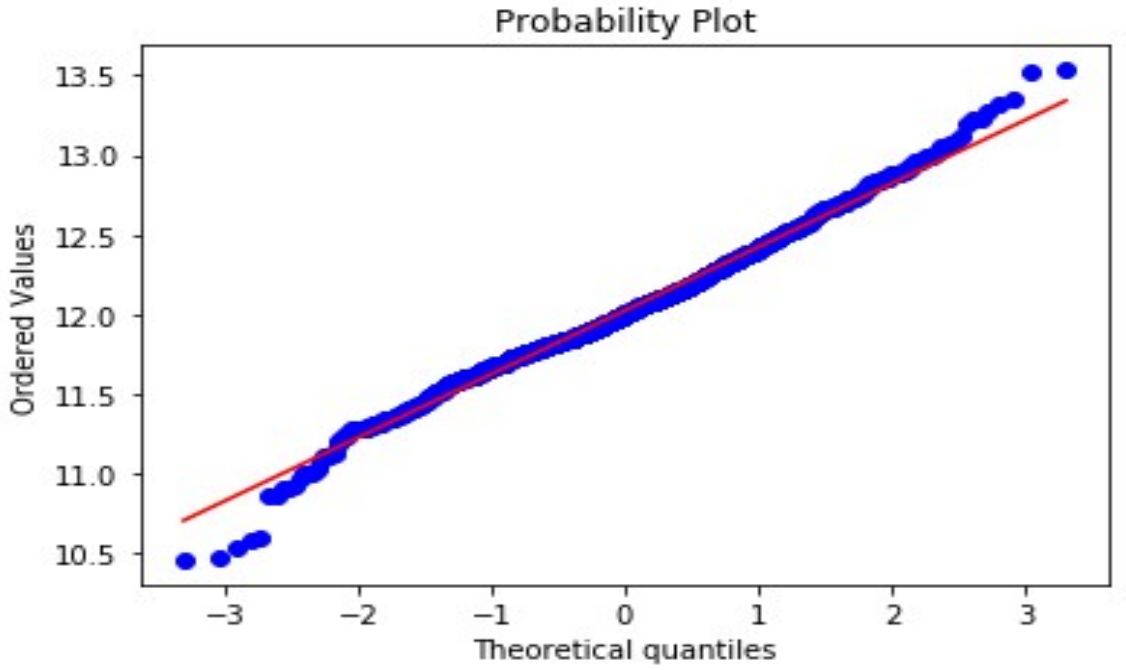
Box plot overallqual/saleprice

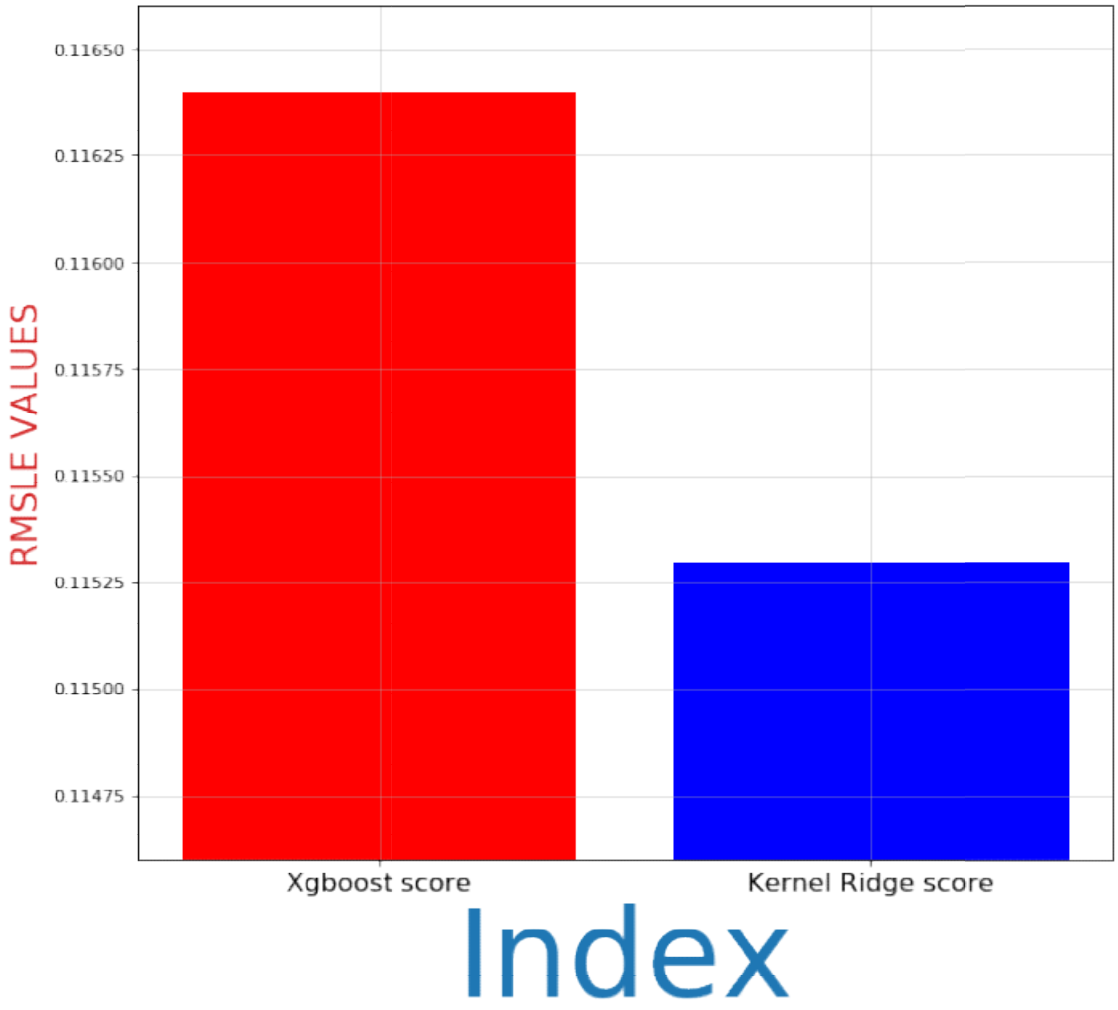


correlation matrix

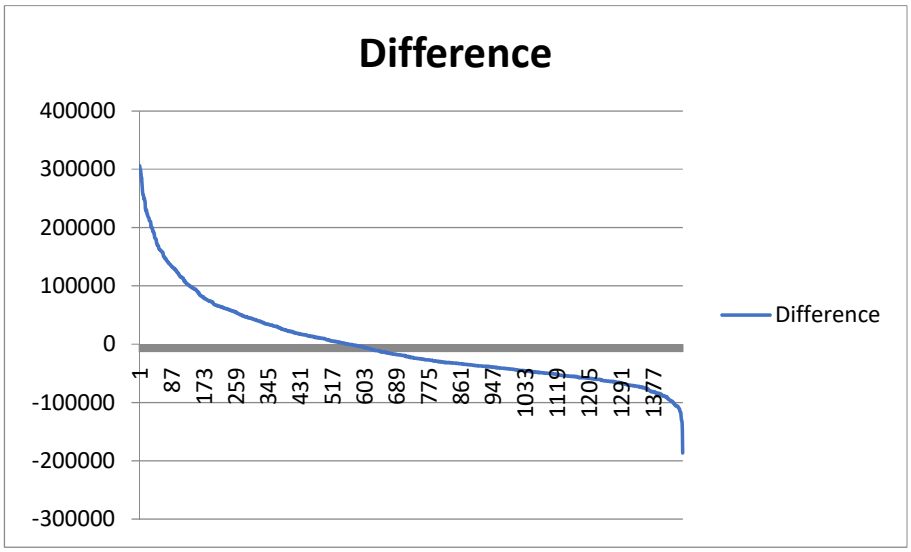


QQ-plot





The above comparison plots XG Boost & Ridge. Since XGBoost performed better, thus for better results, we will go with the predictions provided by XGBoost model.



More than 90% of the values have the variation between + & - 1 lacs dollars.

Compare the model performance

Root Mean Squared Logarithmic Error

Root Mean Squared Logarithmic Error (RMSLE) to evaluate the performance predicting the sale price of a category of equipment. Here we are taking the score values of each algorithm as score **(mean)** and score **(standard deviation)** CALCULATIONS

Lasso score: 0.1115 (0.0074) ---------- .0892=.8*0.1115

Kernel Ridge score: 0.1153 (0.0075) ---------- .02306=.2*.1153

ElasticNet score: 0.1116 (0.0074)

Gradient Boosting score: 0.1167 (0.0083) ---------- .02328=.2*.0.1167

Xgboost score: 0.1164 (0.0070)

LGBM score: 0.1162 (0.0057)

Finally we make predictions using the Lasso and XGBoost model and store results as lasso preds and xgb preds. For our final prediction we use the formula given below:

CALCULATIONS above……..

preds = 0:8 ——————————— lasso preds+0:2 ——————————— xgb preds =.11248 + Standard Deviation
(>>>>>>where preds is directly proportional to rmsle) thus higher the rmsle gives higher preds values

Our final predictions were based on the Ridge Predictions (ridge preds) and Lasso Predictions (lasso preds). We used the following formula to calculate the final predictions.

CALCULATIONS above……..

pred = 0:8 ——————————— lasso preds+0:2 ——————————— ridge preds =.11226 + Standard Deviation

I did a quick analysis on RMSLE in this kernel. My conclusion is that you are better off predicting more than the actual revenue, rather than less than actual revenue

4. Conclusion

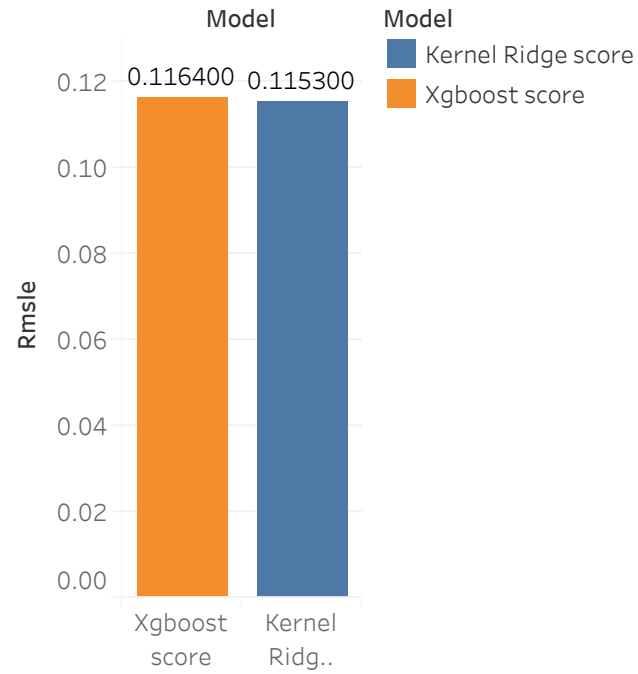
We predicted the SalePrice of the houses for the given Ames Housing dataset using two different methods**. The Prediction include XGBoost/ Lasso & Ridge/Lasso. XGBoost performed better, thus we will go with the predictions provided by XGBoost model.** This prediction data will help eventual buyers to have a better knowledge of the property.

5. Future Work

Our model had a low rmsle score, but there is still room for improvement. In a real world scenario, we can use such a model to predict house prices. This model should check for new data, once in a month, and incorporate them to expand the dataset and produce better results We can try out other advanced regression techniques, like Random Forest and Bayesian Ridge Algorithm, for prediction. Since the data is highly correlated, we should also try Elastic Net regression technique.

Code provided as attachement in the mail.

Sheet 1



Sum of Rmsle for each Model. Color shows details about Model.