**Introduction:**

Machine learning can be used to evaluate real estate market to increase sales and profits. It is less time and resource consuming compared to other evaluation techniques. Machine learning regression for residential real estate evaluation will help improve performance.

Exploratory Data Analysis:

There were two data sets provided. The first data set is train, that is used to create the model and the second data set is to test the model and predict the survival of passengers. The train data has 1460 rows of data and 81 variables. The data is a combination of categorical variables that describe the features of the house. The response variable is SalePrice.

**Data Preparation and Overview of Programming:**

The first step to get the data ready to create the model, is to figure out the missing values and to drop or fill the values based on the percentage missing. The next step was to drop the columns that the response variable is not strongly dependent on. There are 27 columns left that influence the response variable. I converted the string values in columns manually to numeric values, which is the same as LabelEncoder. The histograms below are the distribution plots of the variables.
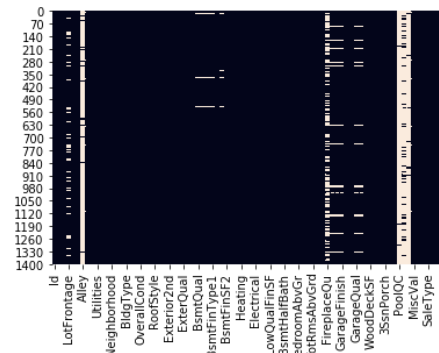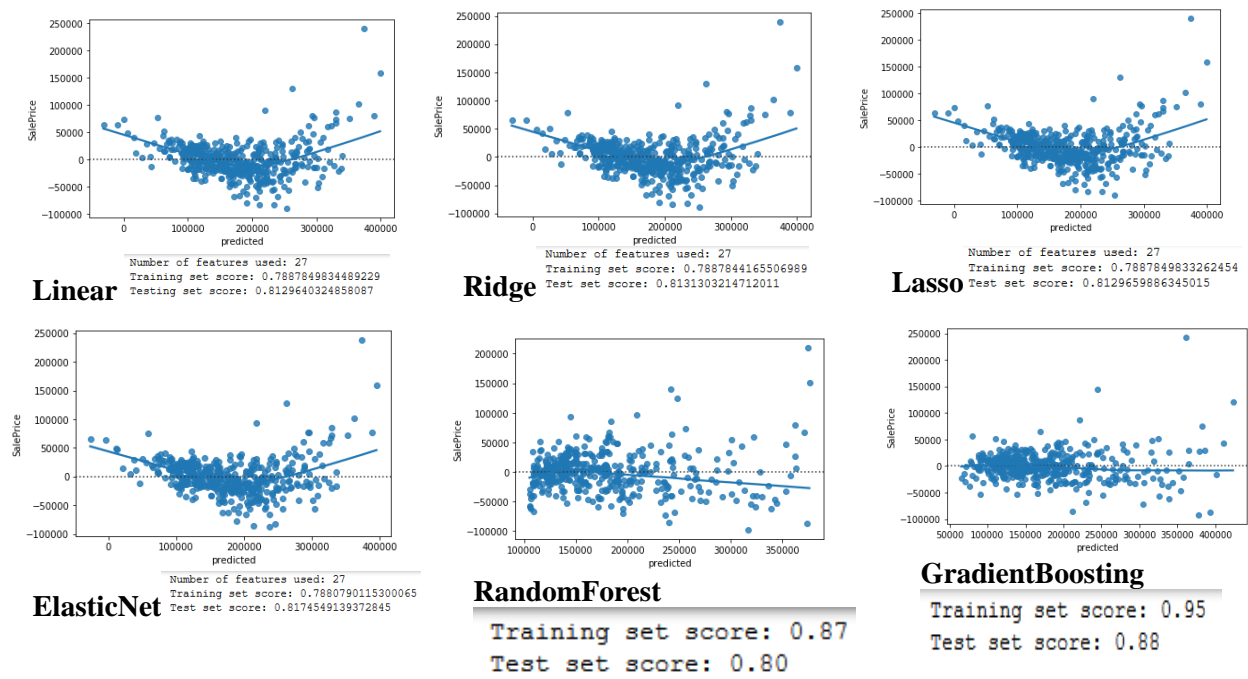

Programming:

The four regression models used for this analysis are Linear, Ridge, Lasso and ElasticNet. The python packages used for data handling are Pandas, Number, SciKitLearn. The variable 'SalePrice' was removed to create a subset of variables and response variable. For every
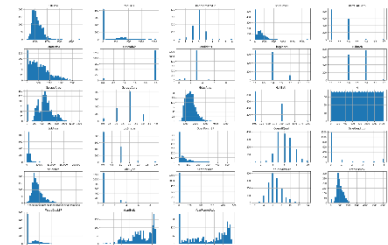
regression model the model score and root mean squared error metric for performance evaluation was calculated.

Linear

Number of features used: 27
Training set score: 0.7887849834489229
Testing set score: 0.8129640324858087

Ridge

Number of features used: 27
Training set score: 0.7887844165506989
Test set score: 0.8131303214712011

Lasso

Number of features used: 27
Training set score: 0.7887849833262454
Test set score: 0.8129659886345015

ElasticNet

Number of features used: 27
Training set score: 0.7880790115300065
Test set score: 0.8174549139372845

RandomForest

Training set score: 0.87
Test set score: 0.80

GradientBoosting

Training set score: 0.95
Test set score: 0.88

Random Forest and Gradient Boosting modeling methods were also implemented on the train and test datasets to predict the response variable.

	Model	RMSE	Score
6	GradientBoost	27561.595794	87.033341
3	ElasticNet	31501.495105	83.061236
1	Ridge	32016.802238	82.502529
2	Lasso	32338.101788	82.149581
0	Linear	32339.243598	82.148320
5	RandomForest	34006.804799	80.259823
4	Decision Tree	37402.100910	76.121262

Root mean squared Error Metric and Model Score

Insights & Conclusions:

Based on the performance of the models with respect to the root mean squared error metric, I recommend that the Gradient Boosting model be used to predict the home values. The variables that influence the price of the home are OverallQual, YearBuilt, TotRmsAbvGrd, FirePlaces, GrarageCars, GarageArea, YearRemodAdd.

Appendix:

The ipynb notebook and an html version of the notebook along with the output and Kaggle submission scores are included in the submission.