CDA 541
STATISTICAL DATA MINING - I
PREDICTION OF HOUSE PRICES

Group 3

Team Members:

Venkata Geethik Kumar Rachaputi (UB ID : 50476154)

Sai Sahiti Chittem (UB ID : 50478898)

Sai Kumar Nalla (UB ID : 50481156)

University at Buffalo The State University of New York

Introduction

The opportunity to analyze and forecast the future direction of property prices in real estate markets is intriguing. Property price forecasting is becoming increasingly important and beneficial. Home price predictions can be difficult to make. There are numerous other factors that influence the price of a house or piece of property; buyers are simply unconcerned about the home's square footage. Finding the right set of characteristics that contribute to understanding the buyer's behavior can be difficult.

Dataset

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

The Dataset contains both train and test files with each file having,

1460 observations

81 variables

SalePrice is the response variable.

Models

- Xgboost
- Lasso Regression
- Linear Regression
- Random Forest

Pre – Processing : Dataset

- Train data :

  81 variables with numeric and characters

- Combining both the train and test data together for further processing.

  Dimensions of the new dataframe (total = train + test):

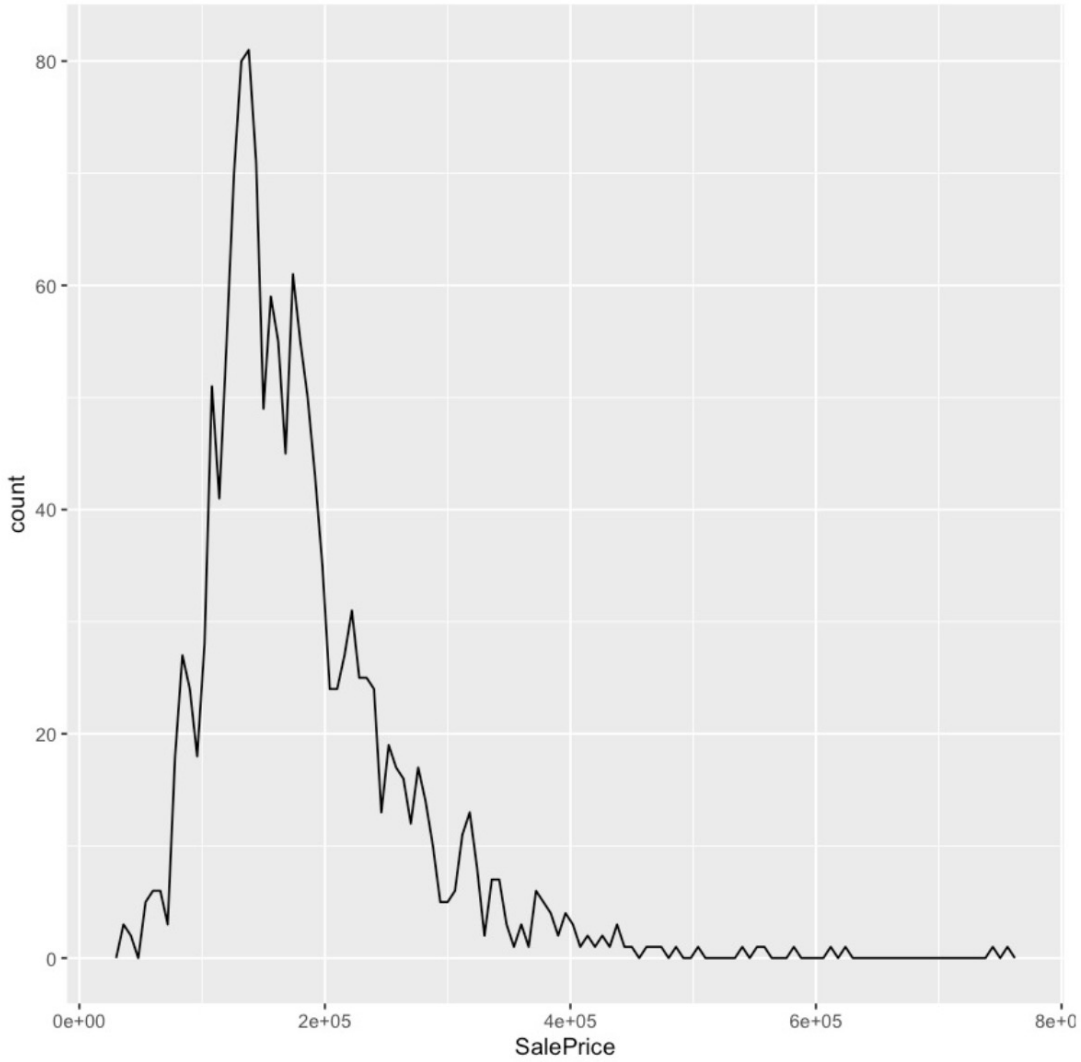
  2919 observations

  80 variables

- There are 37 numerical variables and 44 character variables in total

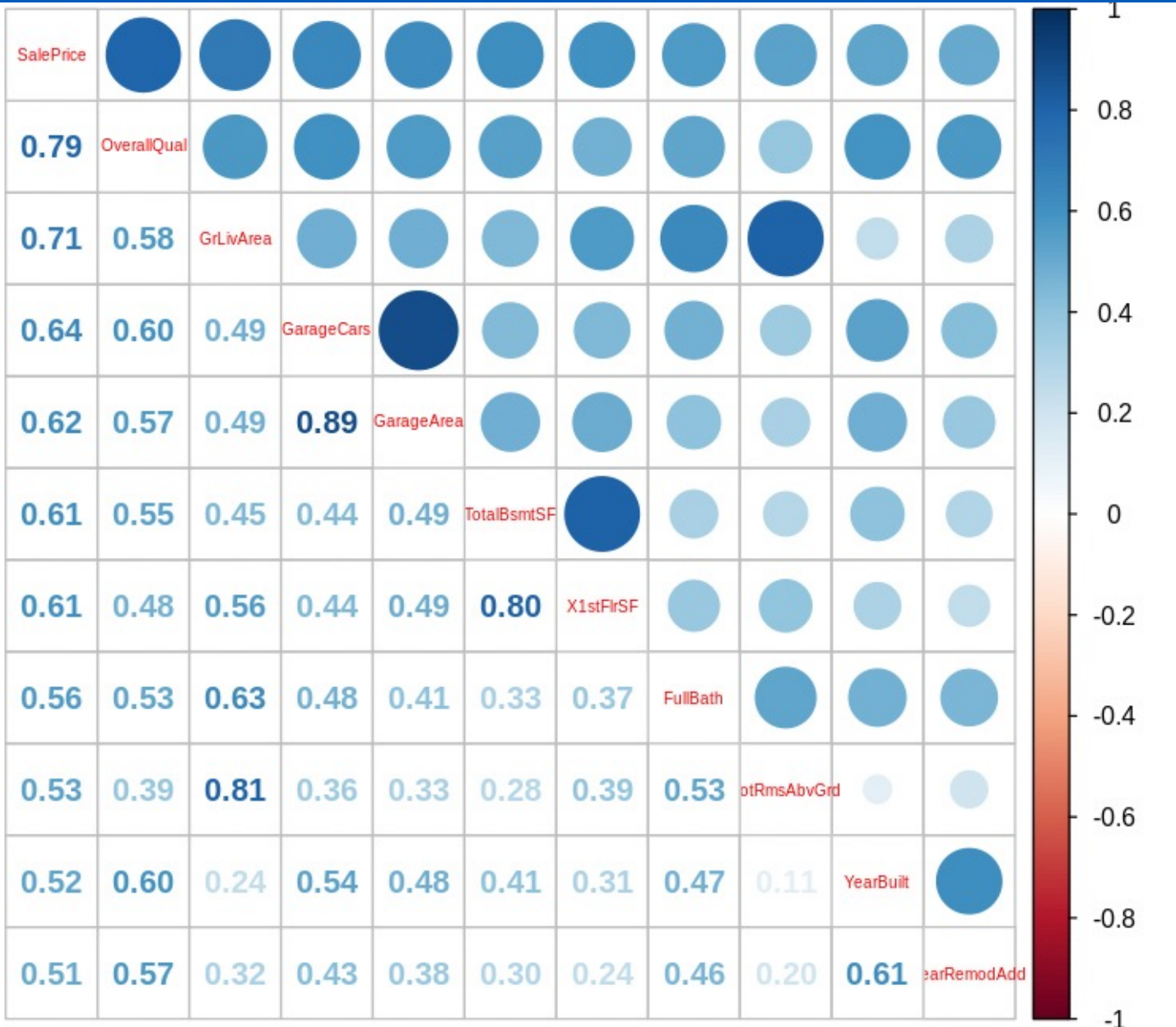
Pre - Processing

The sale prices are skewed towards right as few wealthy people bear enough wealth to acquire luxurious houses.



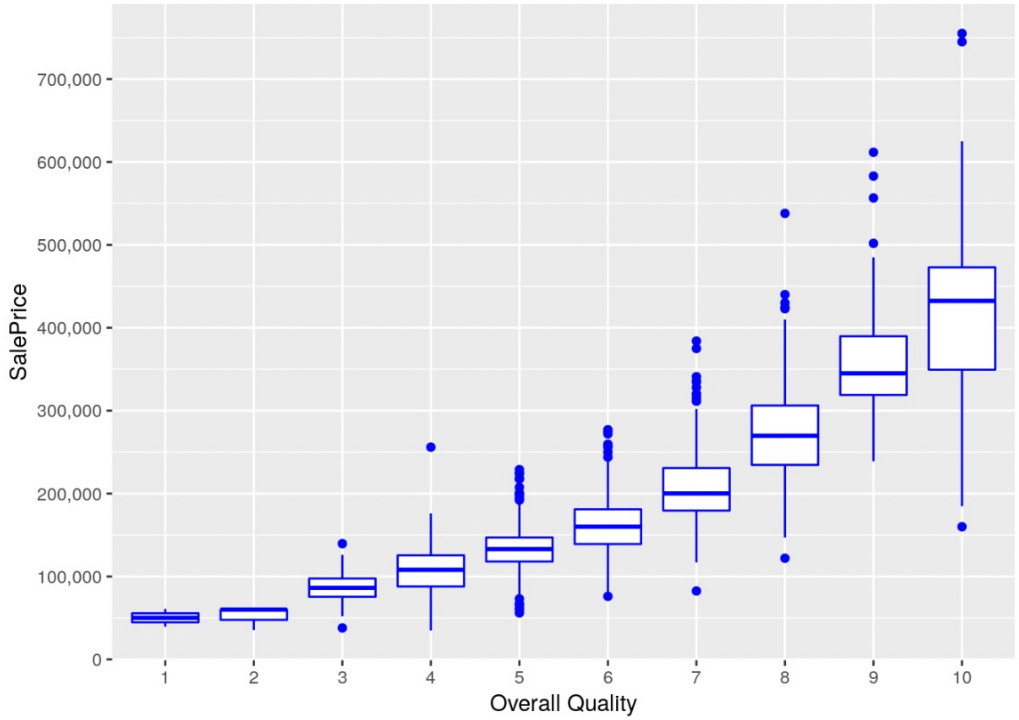
Correlation :

SalePrice is highly correlated with Overall Quality and GrLivArea variables in the data.



Pre – Processing

Overall Quality



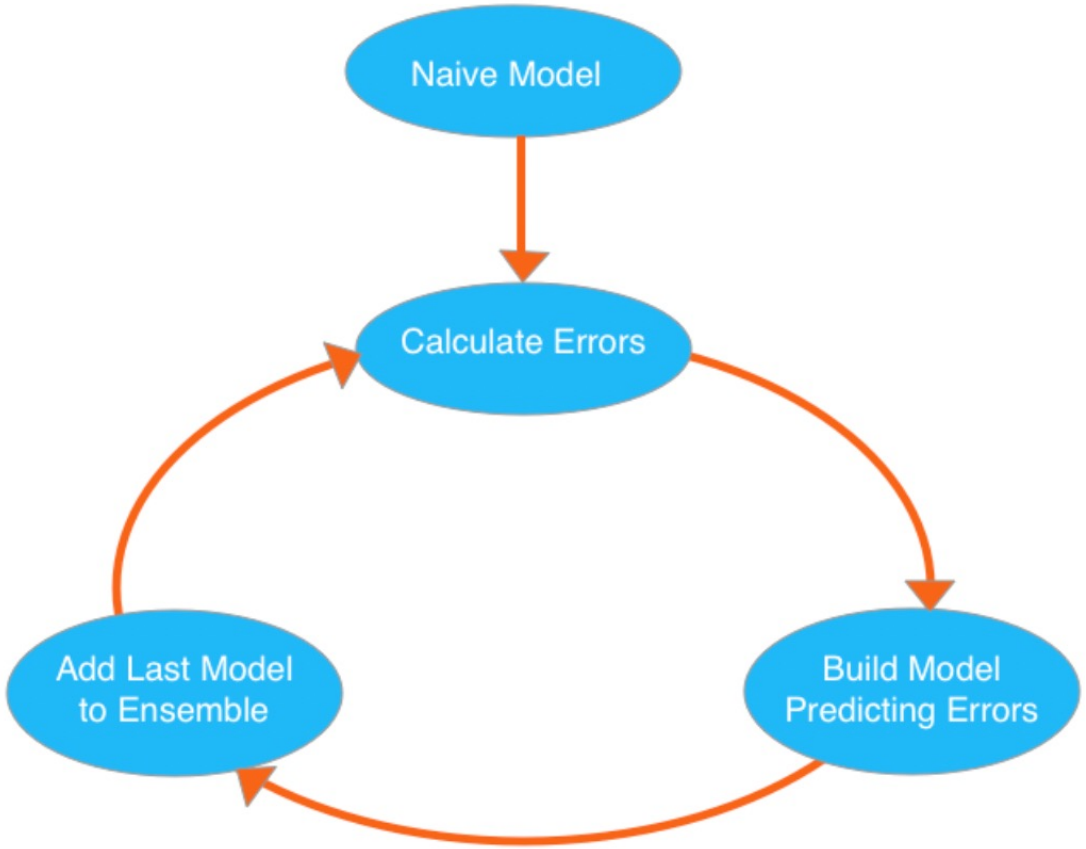
GrLivArea

Pre - Processing

- Converting character variables into ordinal integers or factors
- Dealing character variables with missing values
- Skewness and normalizing numeric predictor values
- One - hot encoding on categorical variables
- Dealing with skewness of SalePrice
- Create new train and test sets for modeling

Xgboost

- XGBoost is an implementation of the **Gradient Boosted Decision Trees** algorithm .]
- It supports various objective functions, including regression, classification and r anking.



Features of Xgboost

- It has several features:
- Speed: it can automatically do parallel computation on *Windows* and *Linux*, with *OpenMP*. It is generally over 10 times faster than the classical gbm.
- Input Type: it takes several types of input data:
  - Dense Matrix: R's dense matrix, i.e. matrix ;
  - Sparse Matrix: R's sparse matrix, i.e. Matrix::dgCMatrix ;
  - Data File: local data files ;
  - xgb.DMatrix: its own class (recommended).
- Sparsity: it accepts *sparse* input for both *tree booster* and *linear booster*, and is optimized for *sparse* input ;
- Customization: it supports customized objective functions and evaluation functions.

Xgboost

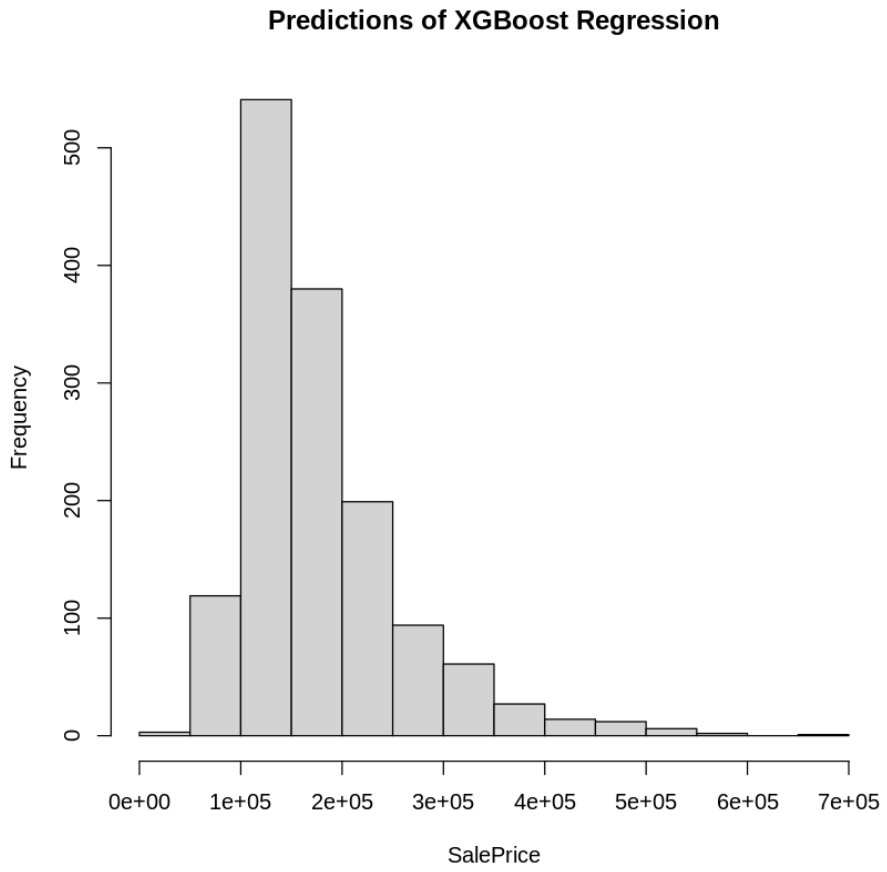
ADVANTAGES

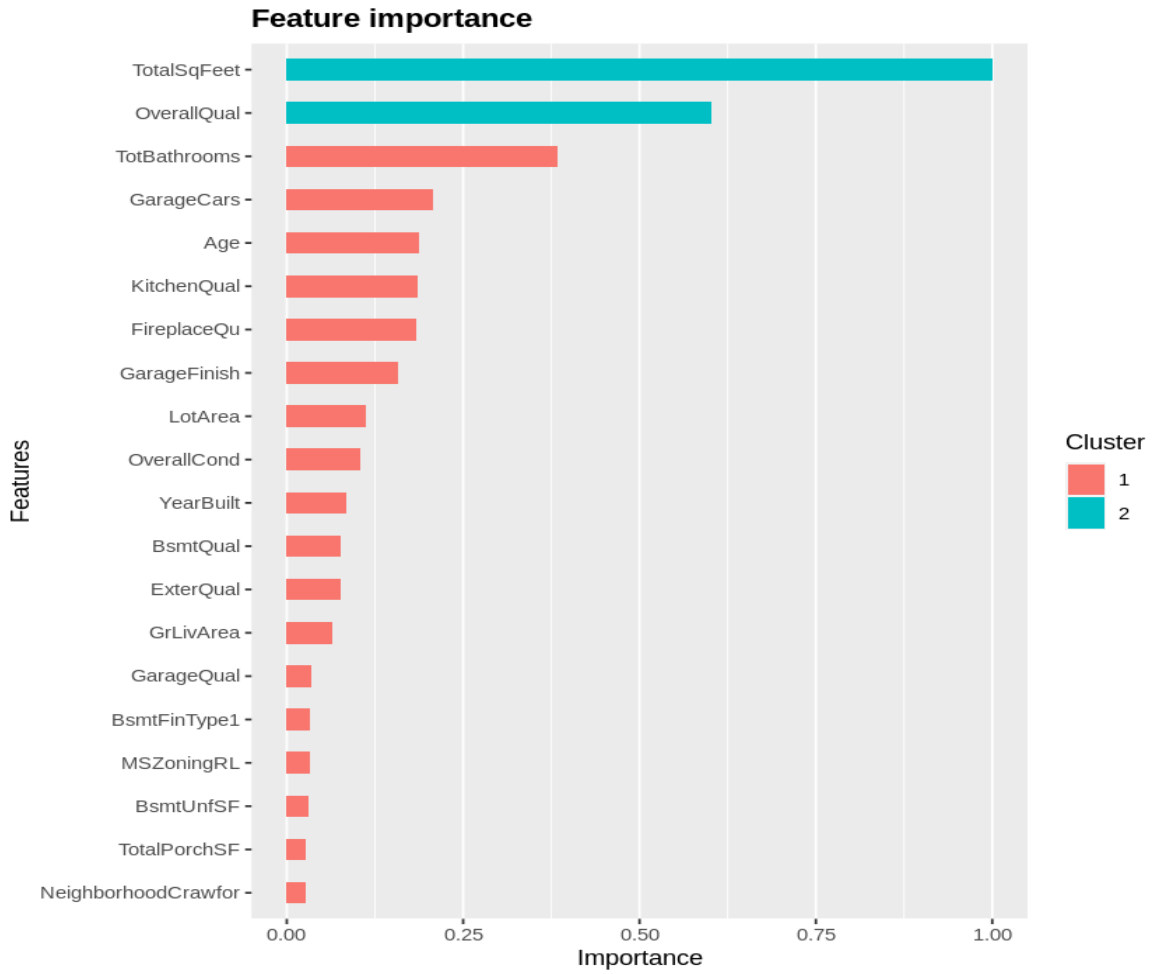
- It is Highly Flexible.
- It uses the power of parallel processing.
- It is faster than Gradient Boosting.
- It supports regularization

DISADVANTAGES

- XGBoost **does not perform so well on sparse and unstructured data**.

XgBoost Prediction and variable importance on Housing Dataset

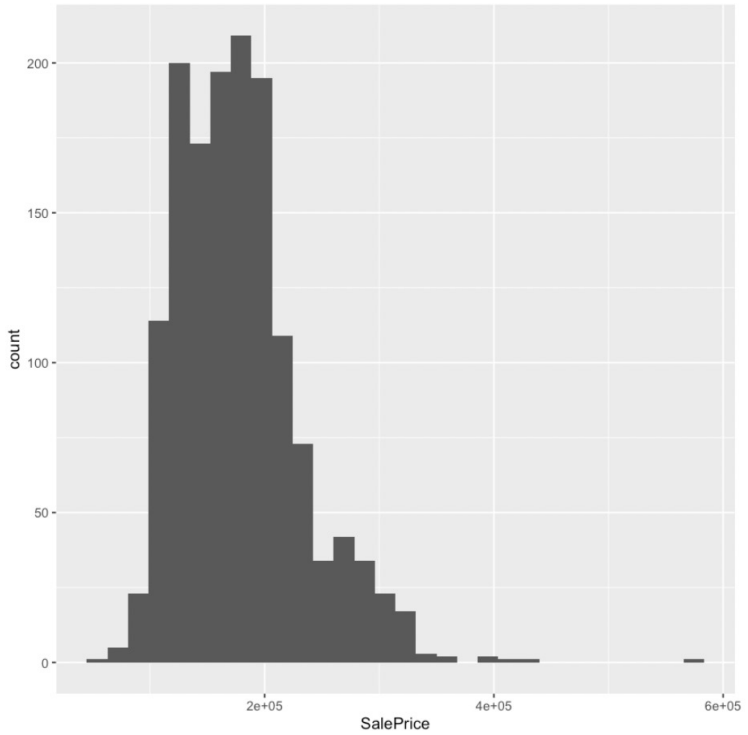




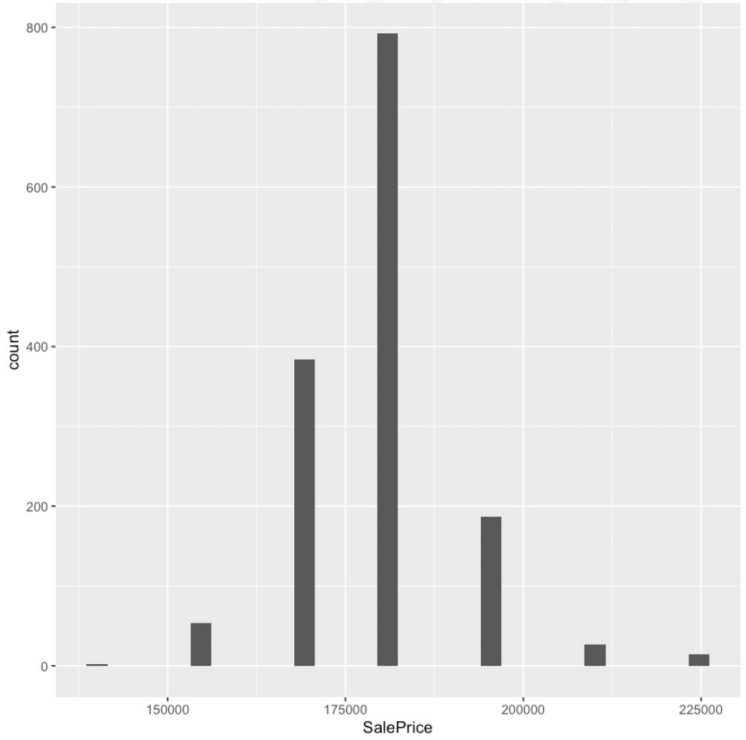
RMSE Score for data = 0.069

Linear Regression

GrLivArea

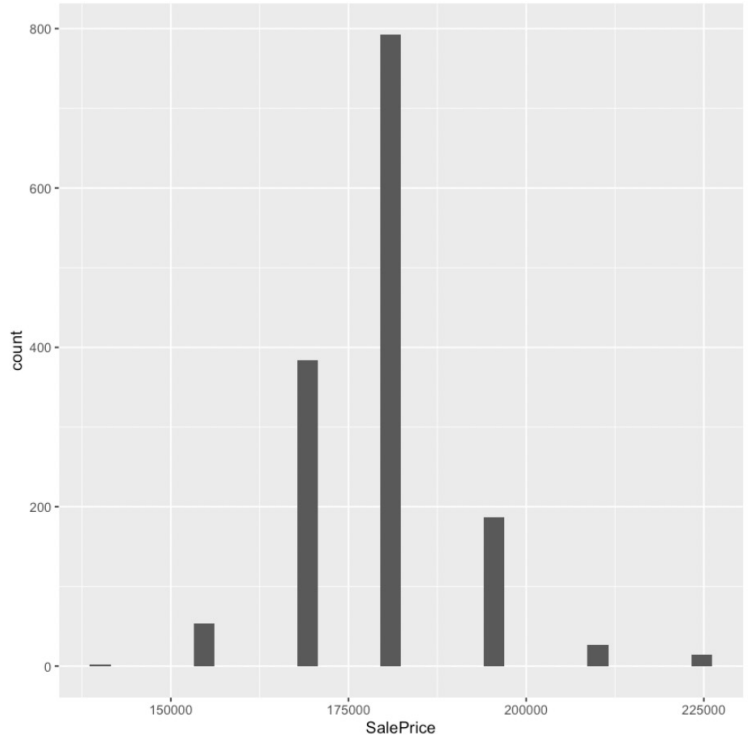


BedroomAbvGr

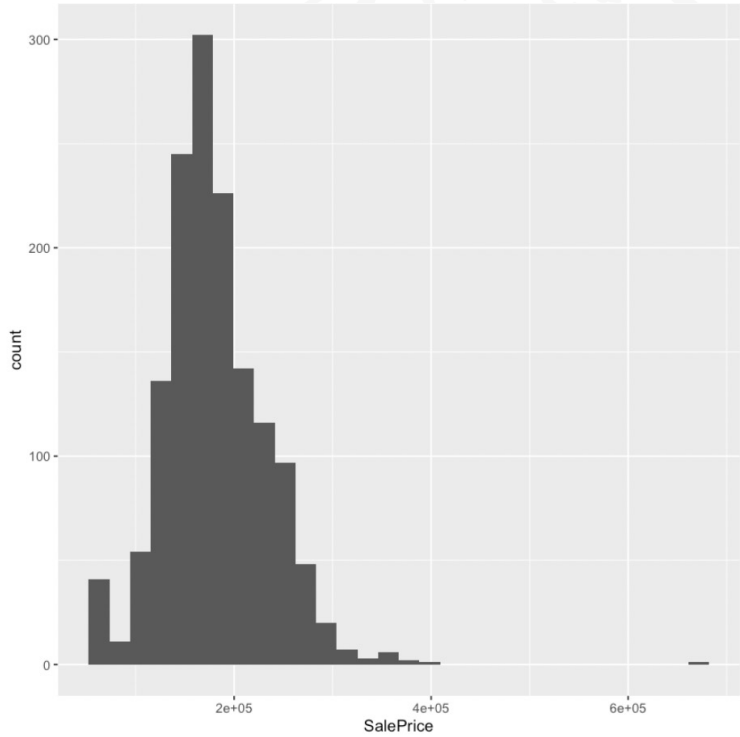


Linear Regression

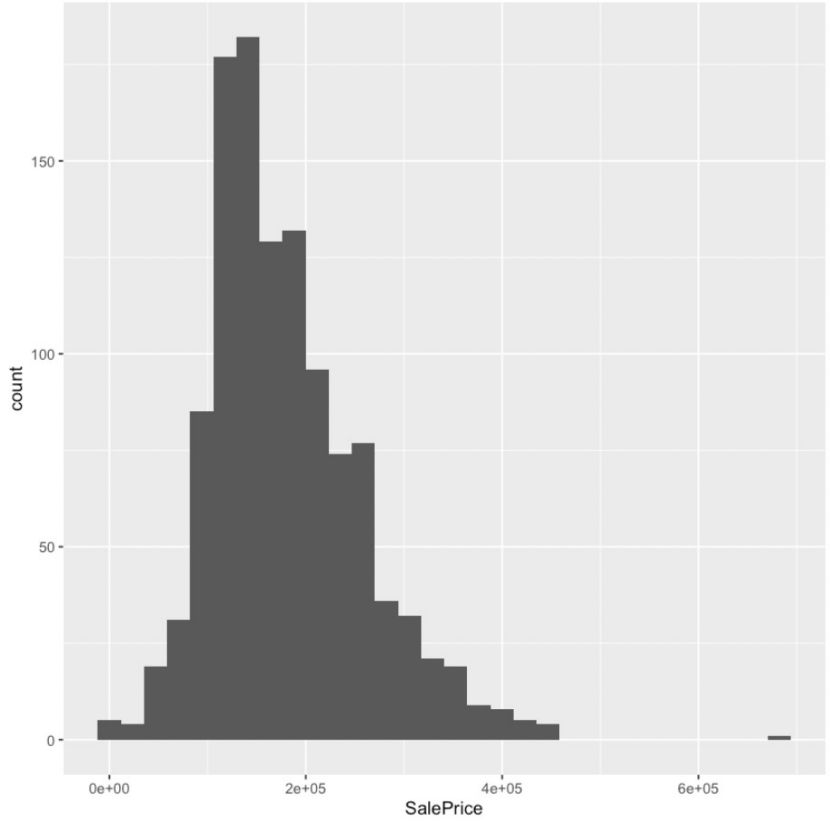
YearBuilt



TotalBsmtSF



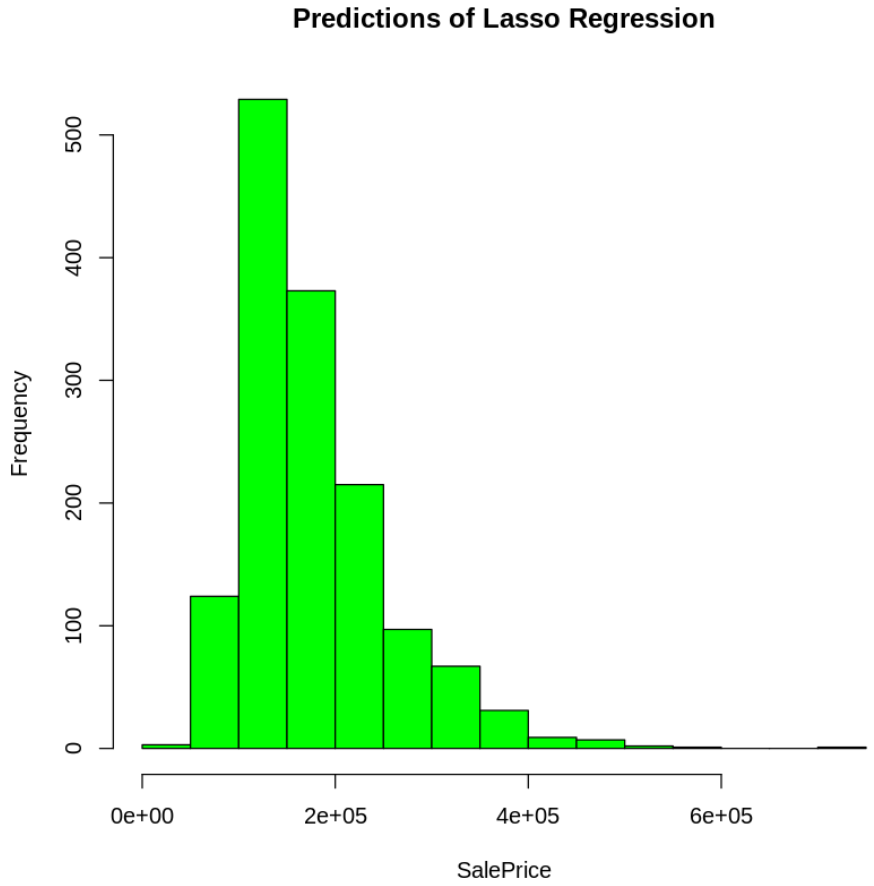
Linear Regression



RMSE Score for the data = 0.144

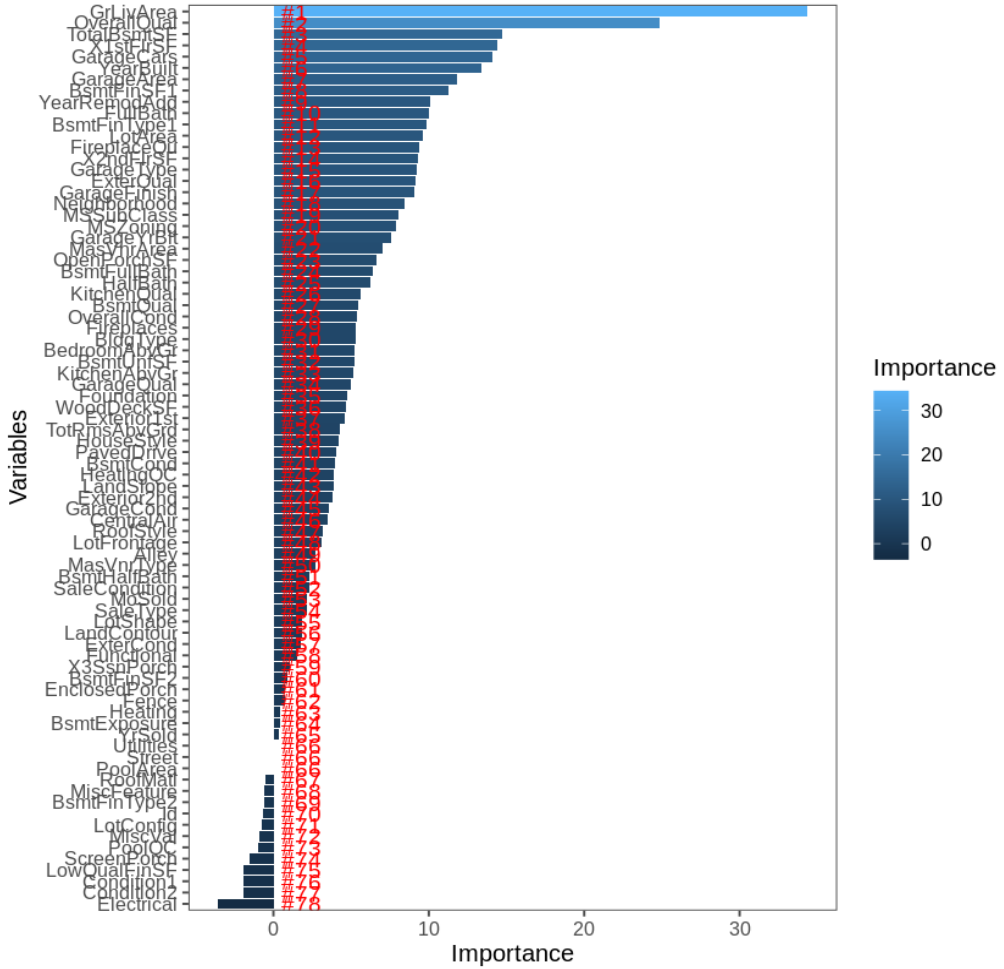
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-924819.9475	1432503.3384	-0.6456	0.5187
Id	0.4417	2.2207	0.1989	0.8424
MSSubClass	-101.3179	29.0123	-3.4922	0.0005
LotFrontage	49.8434	51.3447	0.9708	0.3319
LotArea	0.8271	0.1268	6.5240	0.0000
OverallQual	16264.4433	1220.5737	13.3252	0.0000
OverallCond	3505.3585	1054.0311	3.3257	0.0009
YearRemodAdd	278.3759	70.6033	3.9428	0.0001
MasVnrArea	28.2705	5.8775	4.8099	0.0000
BsmtFinSF1	46.8705	5.0926	9.2037	0.0000
BsmtFinSF2	22.3195	7.5665	2.9498	0.0033
BsmtUnfSF	22.4597	4.4674	5.0275	0.0000
X1stFlrSF	49.8577	6.2885	7.9284	0.0000
X2ndFlrSF	49.7360	5.2914	9.3994	0.0000
LowQualFinSF	8.3423	23.0124	0.3625	0.7170
BsmtFullBath	601.4488	2720.8597	0.2211	0.8251
BsmtHalfBath	-4813.4046	4229.3277	-1.1381	0.2554
FullBath	2138.0464	2868.8040	0.7453	0.4563
HalfBath	2794.9958	2686.3660	1.0404	0.2984
BedroomAbvGr	-12082.3235	1796.2521	-6.7264	0.0000
KitchenAbvGr	-24146.1543	5537.9411	-4.3601	0.0000
TotRmsAbvGrd	5587.6263	1260.4061	4.4332	0.0000
Fireplaces	1222.5826	1850.3733	0.6607	0.5089
GarageYrBlt	191.9431	66.5661	2.8835	0.0040
GarageCars	7193.0229	2930.5867	2.4545	0.0143
GarageArea	18.3861	10.0239	1.8342	0.0669
WoodDeckSF	6.3328	8.4593	0.7486	0.4543
OpenPorchSF	40.1191	16.3632	2.4518	0.0144
EnclosedPorch	-2.9181	16.6453	-0.1753	0.8609
X3SsnPorch	29.0406	34.5424	0.8407	0.4007
ScreenPorch	42.1486	16.5683	2.5439	0.0111
PoolArea	-56.5070	31.9319	-1.7696	0.0771
MiscVal	-5.4349	6.0031	-0.9054	0.3655
MoSold	-253.8719	348.5938	-0.7283	0.4666
YrSold	-34.4613	711.6960	-0.0484	0.9614

Lasso Regression predictions on dataset

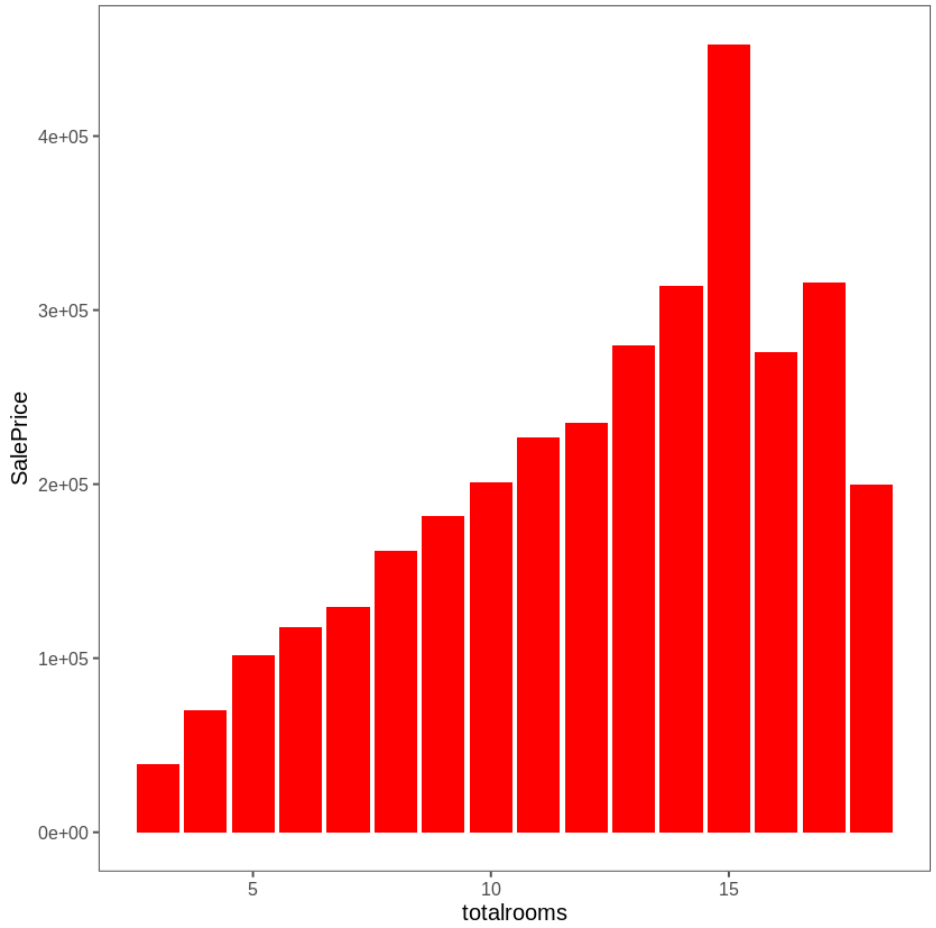


RMSE Score for the data = 0.112

Random Forest Regression variable importance on dataset



Total rooms vs Sales Price



RMSE Score for the data = 0.1445

Root Mean Squared Errors (RMSE)

MODELS	RMSE VALUES
• Xgboost	0.069
• Lasso Regression	0.112
• Linear Regression	0.14407
• Random Forests	0.1445

Conclusion

- Among all the models performed, Xgboost performed very well with good RMSE score, accuracy and least errors. The predicted house prices do help the buyers understand the market and take the decision wisely while the prediction of Xgboost explains that it performs well on the structured data when compared to other models.
- The predicted house prices looked as follows :

```
##        Id SalePrice
## 1461 1461  115030.1
## 1462 1462  162238.9
## 1463 1463  181801.5
## 1464 1464  194189.9
## 1465 1465  199721.3
## 1466 1466  168640.3
```

THANK YOU !!