House Pricing Prediction

DS502 Final Project

Yufei Lin, Jingfeng Xia, Jinhong Yu, Shijing Yang, Yanze Wang

Nov 29 2020

Introduction

Description of the Problem

Being able to predict the price of a house tends to be an important skill for both the seller and consumers. For the seller, they could make better sales and consumers could have better understanding when they try to make a purchase. Therefore, in this project, we are planning to make prediction of house price based on the 79 different predictors provided by Kaggle dataset to determine values of residential homes in Ames, Iowa. We have noticed Sale Price has a typical right-skewed distribution, and decided to process it using two ways, logrithmic with base e and square root for a distribution that is much closer to the shape of a Gaussian distribution for better performance in models like linear regression. In this analysis, we will perform random forest on original y-value for importance of variables and all the rest of models on both absolute and processed y-values to see which way would each model do better and provide an ensemble of models at the end of our study.

Description of the Dataset

In terms of the dataset, the entire data set consists of two pieces of data organized as training data set and test data set respectively. Whereas for each of the dataset, approximately 80 columns corresponding parameters would be evaluated with the prediction of house price. Some noteworthy predictors include the location classification, utilities, environment of neighborhood, house style and condition, area, year of built, and number of functioning rooms. There are over 1400 row of data points in both the training data set and the test data set. The sale prices in the train dataset are given as a parameter in the form of five or six figure full flat integers. The test data set will be applied to different regression models in order to distinguish the disparities of different model performances.

Approaches

Given that our data is aimed at predicting Sale Price of a house, it is unreasonable to require a model to fit the exact value of the dataset but only to reach an estimation within a certain range. Therefore, we have decided to use both regression and classification approaches to look at the problem on both the original and processed value. For regression method, we are going to look at if a prediction is within the range of the actual price ±5%, we will say it is an accurate prediction. For classification prediction, we will be tagging the data into several different groups, and would be fitting the threshold accordingly with models like SVM and K-Means clustering.

Data Processing

Read in Data

We have chosen to eliminate the Id column from this dataset because Id has nothing to do with our prediction and would mess up our prediction. We save data in “train.csv”" from Kaggle into a variable named **HousePricing** for further processing and we will separate it into training and testing set. For each model Bootstrapping will be performed before each model’s training process.

Data Exploration

```
## [1] "Original training data set has 1460 rows and 80 columns"
```

```
## [1] "The percentage of data missing in the original training data set is 5.96%"
```

```
## [1] "The number of duplicated rows are 0"
```

```
## [1] "Number of Factors:"
```

```
## [1] 43
```

```
## [1] "Number of Numeric:"
```

```
## [1] 37
```

Table 1: Table continues below

MSSubClass	LotFrontage	LotArea	OverallQual
Min. : 20.0	Min. : 21.00	Min. : 1300	Min. : 1.000
1st Qu.: 20.0	1st Qu.: 59.00	1st Qu.: 7554	1st Qu.: 5.000
Median : 50.0	Median : 69.00	Median : 9478	Median : 6.000
Mean : 56.9	Mean : 70.05	Mean : 10517	Mean : 6.099
3rd Qu.: 70.0	3rd Qu.: 80.00	3rd Qu.: 11602	3rd Qu.: 7.000
Max. :190.0	Max. :313.00	Max. :215245	Max. :10.000
NA	NA’s :259	NA	NA

Table 2: Table continues below

OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1
Min. :1.000	Min. :1872	Min. :1950	Min. : 0.0	Min. : 0.0
1st Qu.:5.000	1st Qu.:1954	1st Qu.:1967	1st Qu.: 0.0	1st Qu.: 0.0
Median :5.000	Median :1973	Median :1994	Median : 0.0	Median : 383.5
Mean :5.575	Mean :1971	Mean :1985	Mean : 103.7	Mean : 443.6
3rd Qu.:6.000	3rd Qu.:2000	3rd Qu.:2004	3rd Qu.: 166.0	3rd Qu.: 712.2
Max. :9.000	Max. :2010	Max. :2010	Max. :1600.0	Max. :5644.0
NA	NA	NA	NA’s :8	NA

Table 3: Table continues below

BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	X1stFlrSF	X2ndFlrSF
Min. : 0.00	Min. : 0.0	Min. : 0.0	Min. : 334	Min. : 0
1st Qu.: 0.00	1st Qu.: 223.0	1st Qu.: 795.8	1st Qu.: 882	1st Qu.: 0
Median : 0.00	Median : 477.5	Median : 991.5	Median :1087	Median : 0

BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	X1stFlrSF	X2ndFlrSF
Mean : 46.55	Mean : 567.2	Mean :1057.4	Mean :1163	Mean : 347
3rd Qu.: 0.00	3rd Qu.: 808.0	3rd Qu.:1298.2	3rd Qu.:1391	3rd Qu.: 728
Max. :1474.00	Max. :2336.0	Max. :6110.0	Max. :4692	Max. :2065
NA	NA	NA	NA	NA

Table 4: Table continues below

LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath
Min. : 0.000	Min. : 334	Min. :0.0000	Min. :0.00000
1st Qu.: 0.000	1st Qu.:1130	1st Qu.:0.0000	1st Qu.:0.00000
Median : 0.000	Median :1464	Median :0.0000	Median :0.00000
Mean : 5.845	Mean :1515	Mean :0.4253	Mean :0.05753
3rd Qu.: 0.000	3rd Qu.:1777	3rd Qu.:1.0000	3rd Qu.:0.00000
Max. :572.000	Max. :5642	Max. :3.0000	Max. :2.00000
NA	NA	NA	NA

Table 5: Table continues below

FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr
Min. :0.000	Min. :0.0000	Min. :0.000	Min. :0.000
1st Qu.:1.000	1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:1.000
Median :2.000	Median :0.0000	Median :3.000	Median :1.000
Mean :1.565	Mean :0.3829	Mean :2.866	Mean :1.047
3rd Qu.:2.000	3rd Qu.:1.0000	3rd Qu.:3.000	3rd Qu.:1.000
Max. :3.000	Max. :2.0000	Max. :8.000	Max. :3.000
NA	NA	NA	NA

Table 6: Table continues below

TotRmsAbvGrd	Fireplaces	GarageYrBlt	GarageCars	GarageArea
Min. : 2.000	Min. :0.000	Min. :1900	Min. :0.000	Min. : 0.0
1st Qu.: 5.000	1st Qu.:0.000	1st Qu.:1961	1st Qu.:1.000	1st Qu.: 334.5
Median : 6.000	Median :1.000	Median :1980	Median :2.000	Median : 480.0
Mean : 6.518	Mean :0.613	Mean :1979	Mean :1.767	Mean : 473.0
3rd Qu.: 7.000	3rd Qu.:1.000	3rd Qu.:2002	3rd Qu.:2.000	3rd Qu.: 576.0
Max. :14.000	Max. :3.000	Max. :2010	Max. :4.000	Max. :1418.0
NA	NA	NA's :81	NA	NA

Table 7: Table continues below

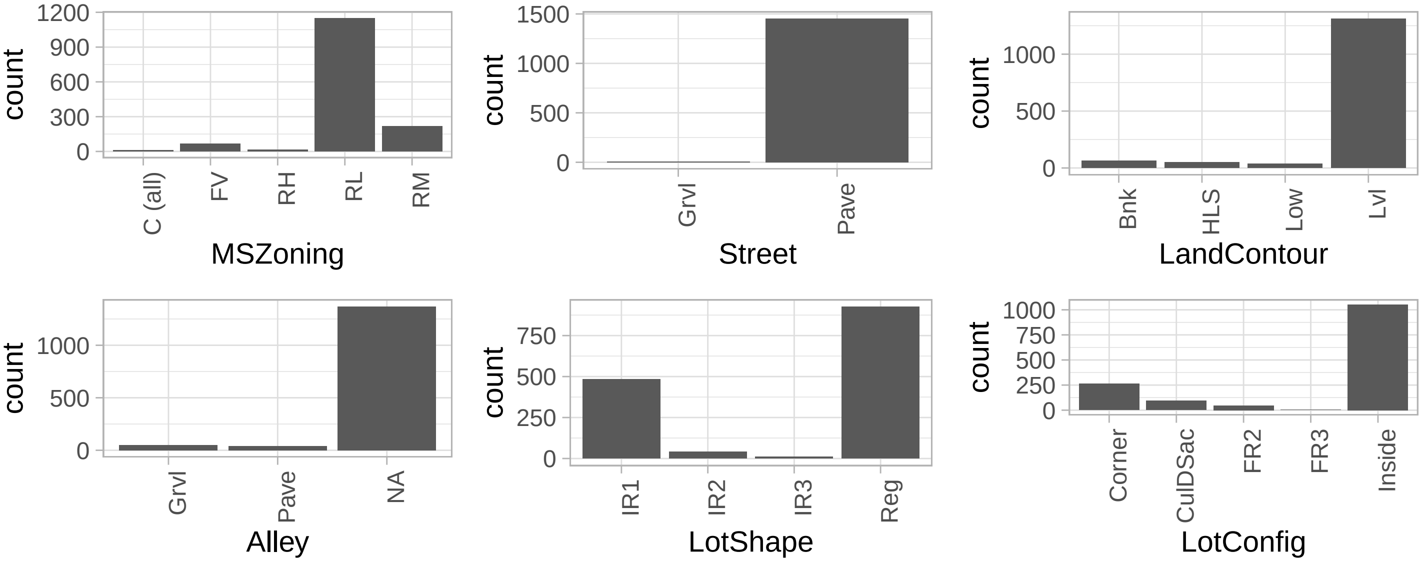
WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00
1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.00
Median : 0.00	Median : 25.00	Median : 0.00	Median : 0.00
Mean : 94.24	Mean : 46.66	Mean : 21.95	Mean : 3.41
3rd Qu.:168.00	3rd Qu.: 68.00	3rd Qu.: 0.00	3rd Qu.: 0.00

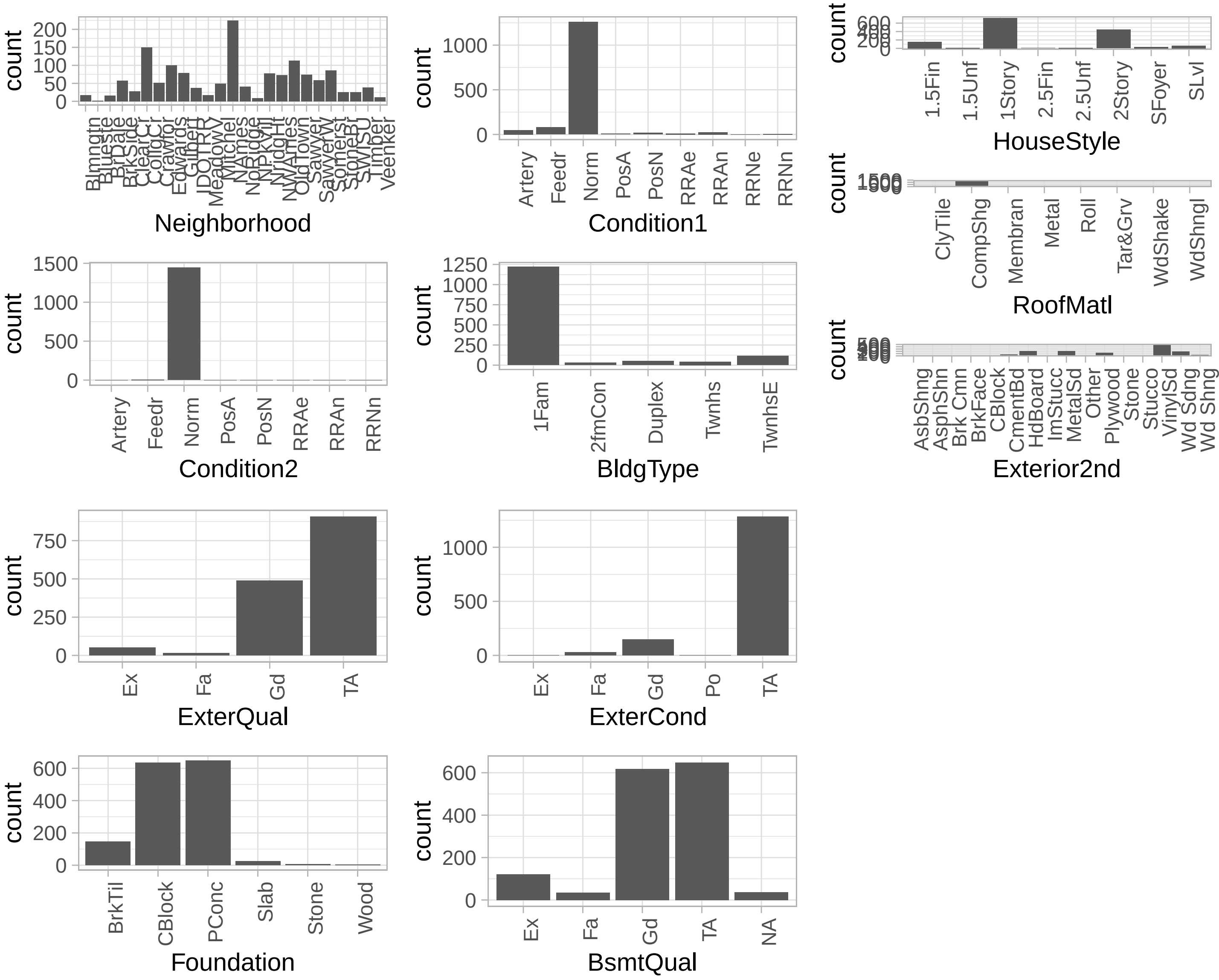
WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch
Max. :857.00	Max. :547.00	Max. :552.00	Max. :508.00
NA	NA	NA	NA

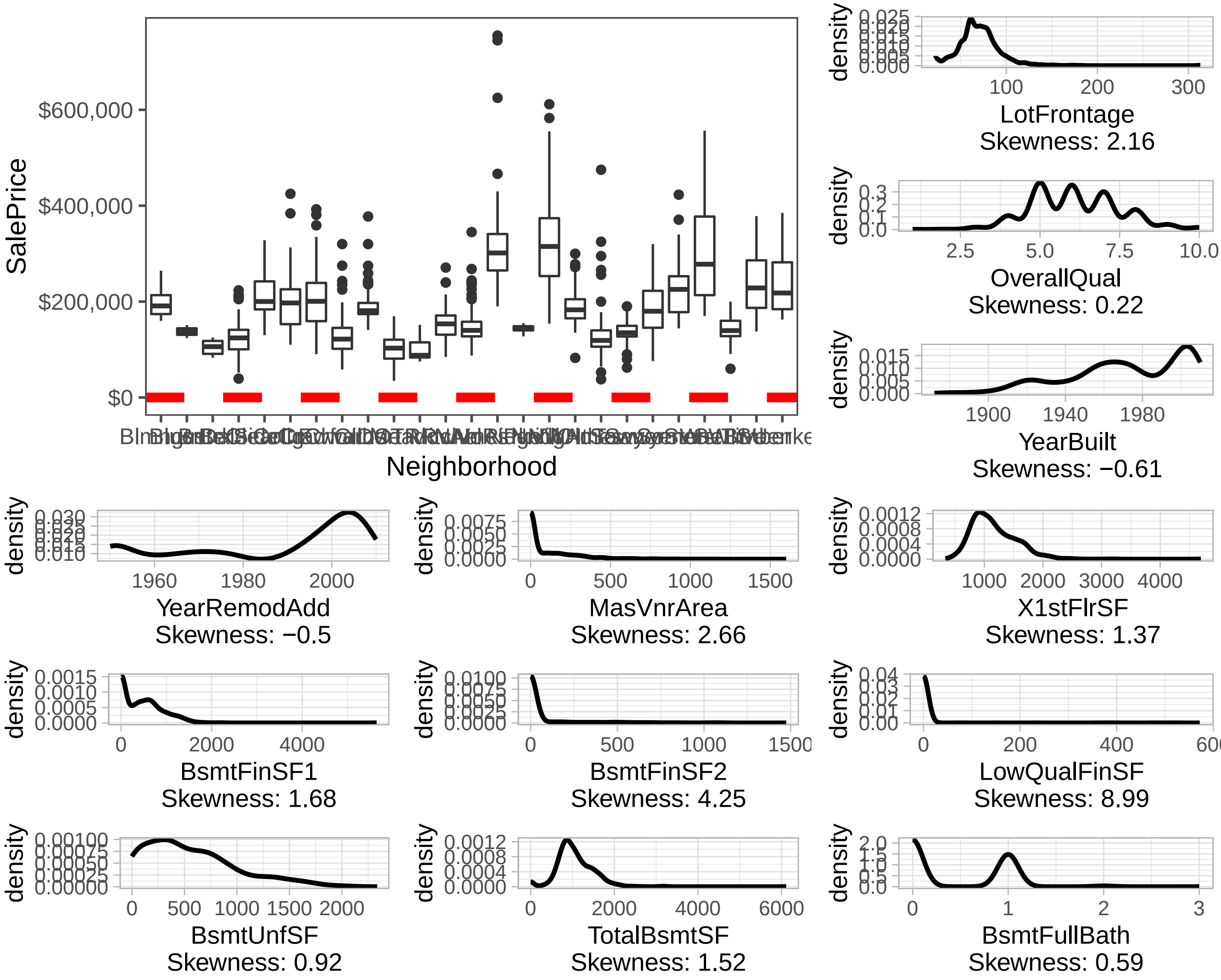
Table 8: Table continues below

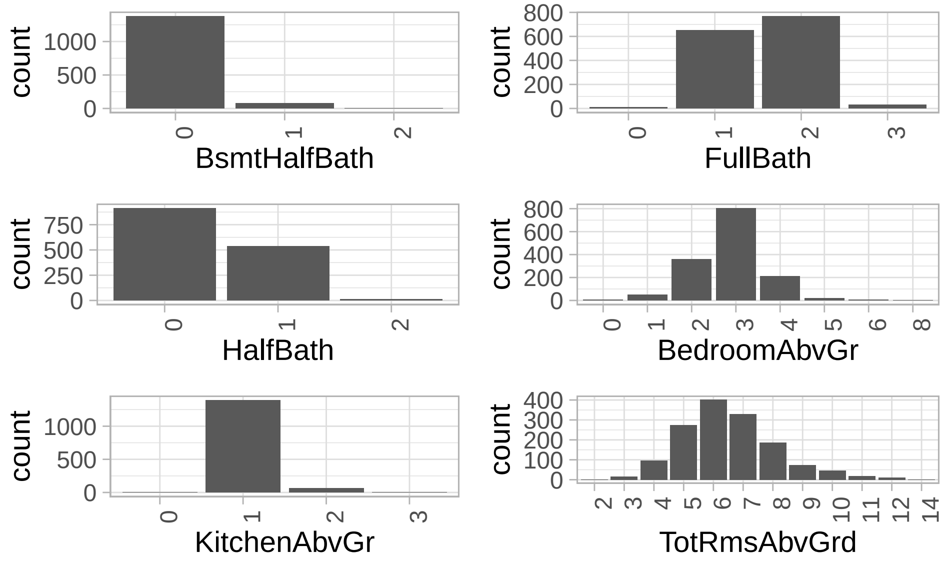
ScreenPorch	PoolArea	MiscVal	MoSold
Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. : 1.000
1st Qu.: 0.00	1st Qu.: 0.000	1st Qu.: 0.00	1st Qu.: 5.000
Median : 0.00	Median : 0.000	Median : 0.00	Median : 6.000
Mean : 15.06	Mean : 2.759	Mean : 43.49	Mean : 6.322
3rd Qu.: 0.00	3rd Qu.: 0.000	3rd Qu.: 0.00	3rd Qu.: 8.000
Max. :480.00	Max. :738.000	Max. :15500.00	Max. :12.000
NA	NA	NA	NA

YrSold	SalePrice
Min. :2006	Min. : 34900
1st Qu.:2007	1st Qu.:129975
Median :2008	Median :163000
Mean :2008	Mean :180921
3rd Qu.:2009	3rd Qu.:214000
Max. :2010	Max. :755000
NA	NA




count
200
150
100
50
0
Blmngtn
Blueste
BrDale
BrkSide
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
MeadowV
Mitchel
NAmes
NoRidge
NPkVill
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU
Timber
Veenker
Neighborhood
count
1000
500
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNe
RRNn
Condition1
count
600
400
200
0
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle
count
1500
ClyTile
CompShg
Membran
Metal
Roll
Tar&Grv
WdShake
WdShngl
RoofMatl
count
1500
1000
500
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNn
Condition2
count
1250
1000
750
500
250
0
1Fam
2fmCon
Duplex
Twnhs
TwnhsE
BldgType
count
500
AsbShng
AsphShn
Brk Cmn
BrkFace
CBlock
CmentBd
HdBoard
ImStucc
MetalSd
Other
Plywood
Stone
Stucco
VinylSd
Wd Sdng
Wd Shng
Exterior2nd
count
750
500
250
0
Ex
Fa
Gd
TA
ExterQual
count
1000
500
0
Ex
Fa
Gd
Po
TA
ExterCond
count
600
400
200
0
BrkTil
CBlock
PConc
Slab
Stone
Wood
Foundation
count
600
400
200
0
Ex
Fa
Gd
TA
NA
BsmtQual

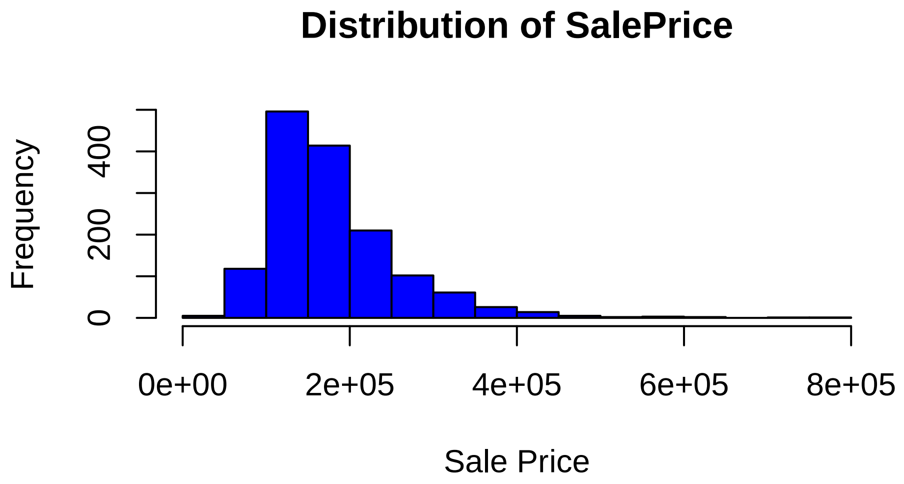

SalePrice
$600,000
$400,000
$200,000
$0
Neighborhood
density
LotFrontage
Skewness: 2.16
OverallQual
Skewness: 0.22
YearBuilt
Skewness: −0.61
YearRemodAdd
Skewness: −0.5
MasVnrArea
Skewness: 2.66
X1stFlrSF
Skewness: 1.37
BsmtFinSF1
Skewness: 1.68
BsmtFinSF2
Skewness: 4.25
LowQualFinSF
Skewness: 8.99
BsmtUnfSF
Skewness: 0.92
TotalBsmtSF
Skewness: 1.52
BsmtFullBath
Skewness: 0.59



target varaible vs. predictors

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   34900  129975  163000  180921  214000  755000
```

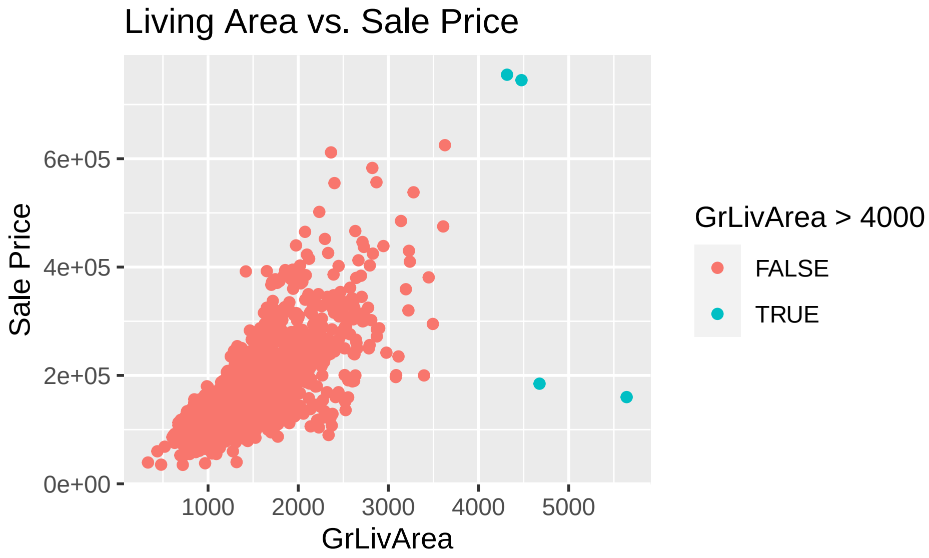
```
##     0%    25%    50%    75%   100% 
##  34900 129975 163000 214000 755000
```



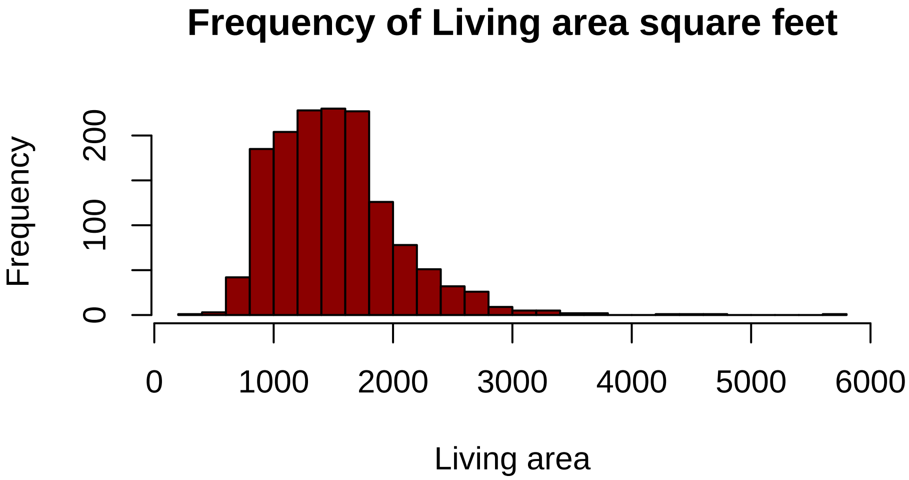
Conclusion

It deviates from normal distribution and it is right skewed

Plotting 'GrLivArea' too see if there are any outliers



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334    1130    1464    1515    1777    5642
```

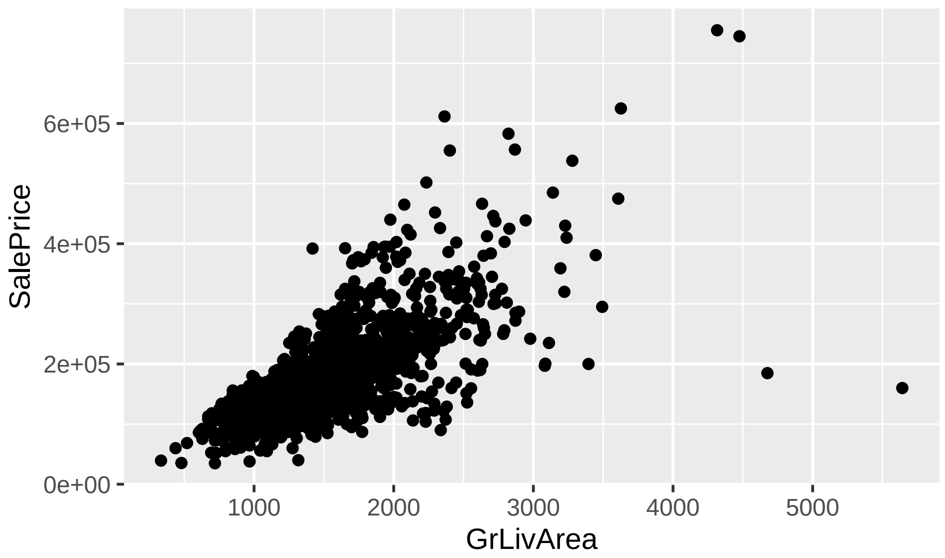


Feature Engineering

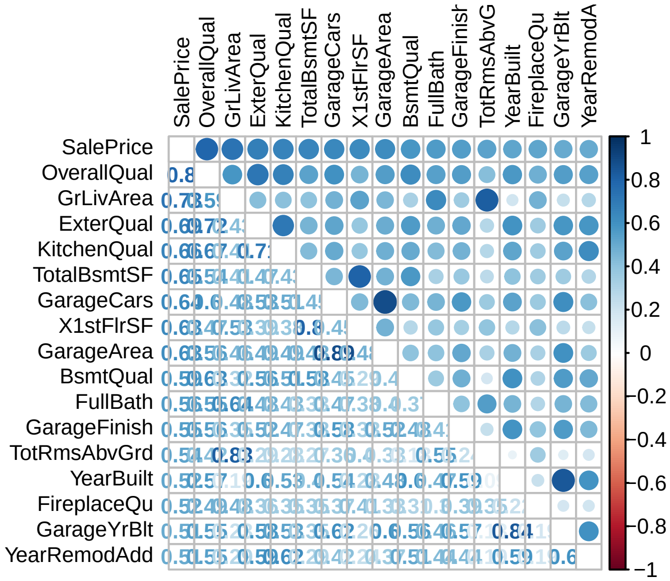
In this section, we convert all missing value based on the following rules:

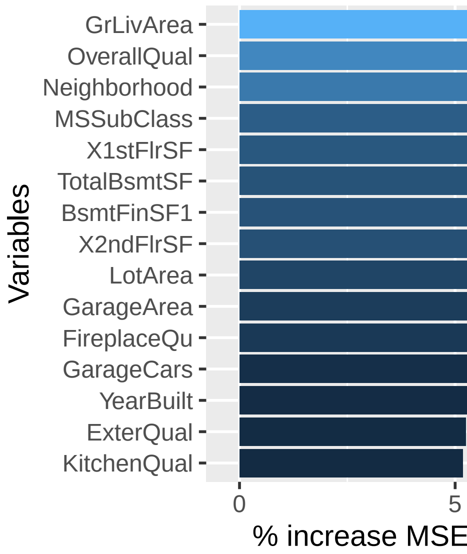
1. Categorical: fill in most common
2. Numeric: fill in median/average

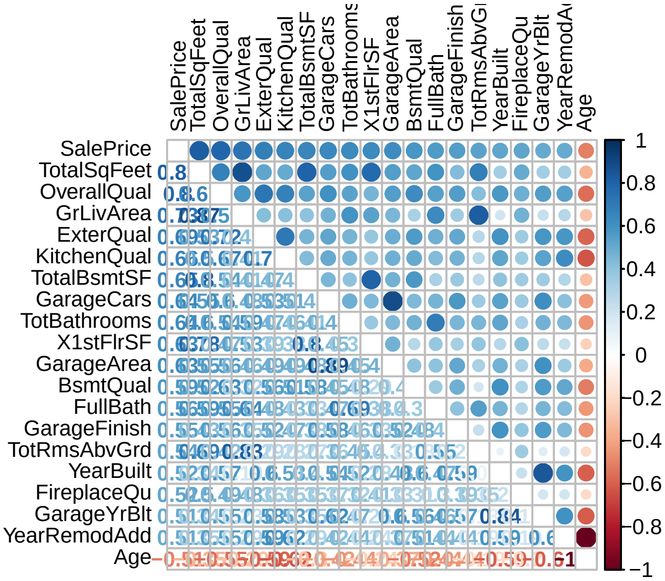
Convert all train to HousePricing



Correlation between the numerical variables

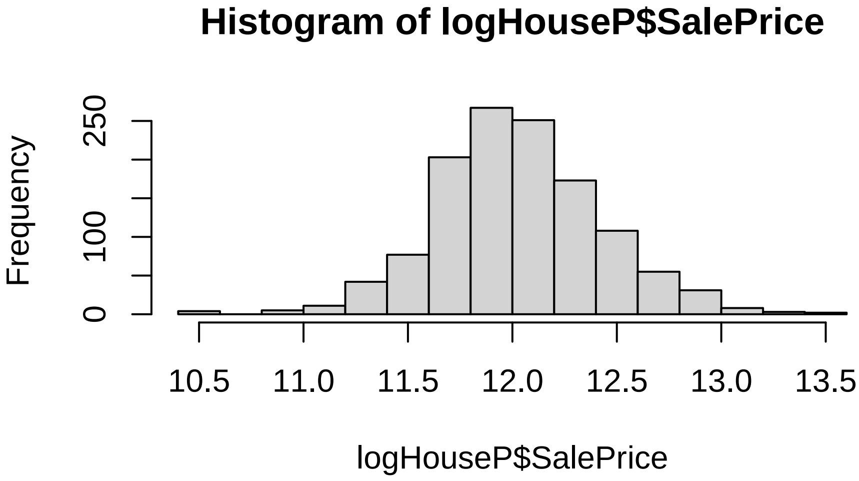




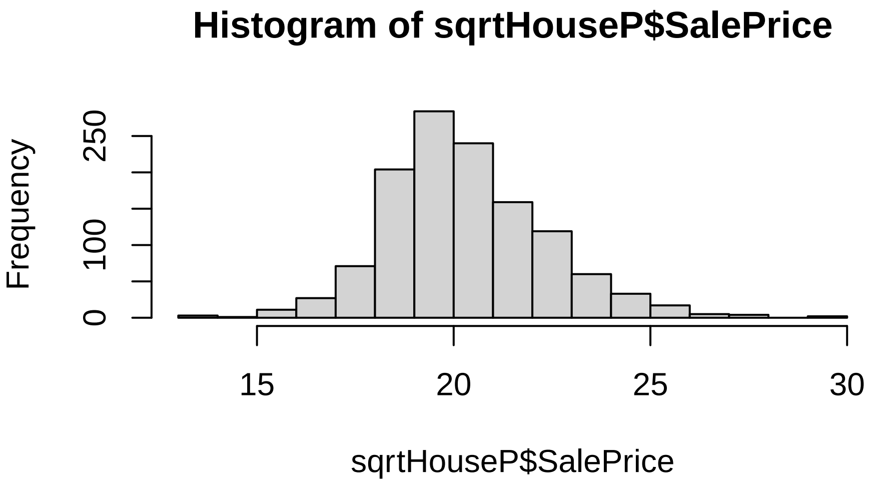


As discussed before, we have decided to use logrithmic with base e and square root to process the data. We have also saved 15% of our data into a variable named vault for the final test of each model.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  130000  164250  181442  214925  755000
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.46   11.78   12.01   12.03   12.28   13.53
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.67   18.99   20.13   20.32   21.53   29.48
```

```
##   0   1   2
## 316 614 310
```

```
##   0   1   2
## 316 614 310
```

```
##   0   1   2
## 316 614 310
```

Seperate into Test and Training Set

Spearate by 70% train, 30% test.

Prediction Algorithms

We choose to use PCR, Random Forest, GAM, Lasso and Ridge, Splines and Linear Regression to look at how each model would be suitable for our regression analysis.

Each model needs a cross validation algorithm Remember to report RMSE

Regression Methods

1. Linear Regression

Explanation

We have chosen this model to understand how each numeric variable is linear related to our House Price prediction.

Check Accuracy

```
## [1] 0.3563218 0.3831418 0.3524904 0.3295019 0.4214559
```

```
## [1] 0.3685824
```

```
## [1] "We have the accuracy of the linear model approximately 36.86%"
```

```
## [1] 0.4674330 0.4980843 0.3946360 0.4367816 0.4329502
```

```
## [1] "We have the accuracy of the linear model after log transformation approximately 44.60%"
```

```
## [1] 0.4750958 0.4291188 0.4559387 0.4827586 0.4597701
```

```
## [1] "We have the accuracy of the linear model after sqrt transformation approximately 46.05%"
```

2. Random Forest

Explanation

We have chosen this model because random forest is based on a collection of decision trees that could help us get better understanding of which tree and division contribute to which section such that we could have a better picture of the overall importance of each different factor in the prediction.

Prepare Model

We have 199 independent variables in the data set, therefore we have set mtry(Number of randomly selected variables for each split) to be the square root of that number for maximum performance of the model.

The following is the result from Random Forest algorithm:

Call: randomForest(formula = SalePrice ~ ., data = train_ori, mtry = sqrt(totalIV), importance = TRUE) Type of random forest: regression Number of trees: 500 No. of variables tried at each split: 15

Mean of squared residuals: 977461686 % Var explained: 83.77

Check Accuracy

We then need to check accuracy, as assumed before, we would look at whether the predicted data is within the ±5% range. The following is the result.

```
## [1] "We have the accuracy of the model approximately 37.90%"
```

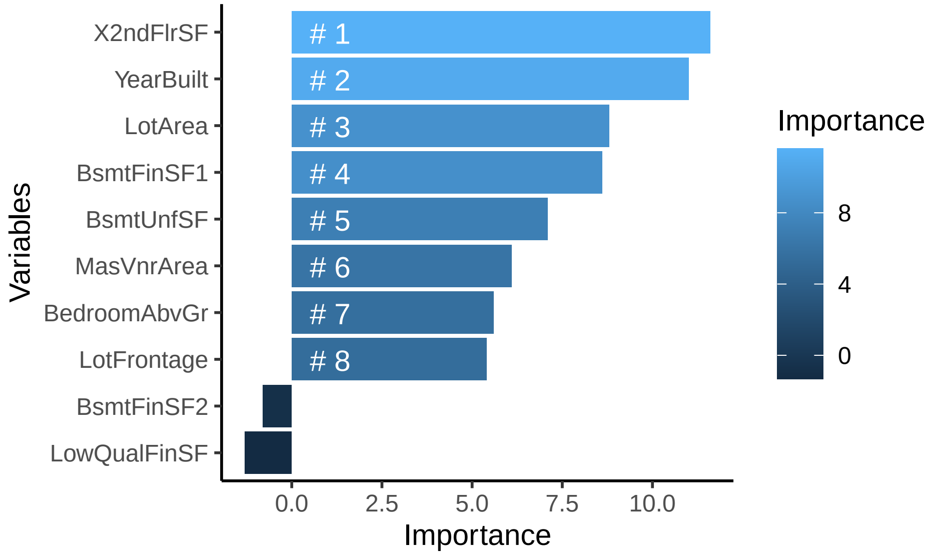
Error Metrics

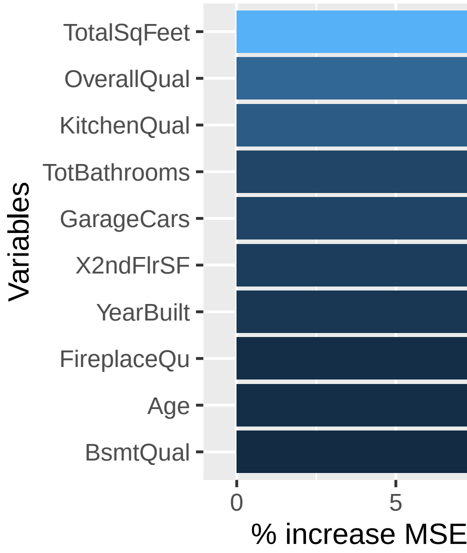
Then let us take a look at the MSE of this model:

```
## [1] "We have the MSE of the model approximately equal to 794429344.85177"
```

Variable Importance

Here we are going to show the top 10 most important variables in predicting sale price of a house.





From the random forest analysis, we have discovered that the top three most important factors for predicting sale price are the following:

1. OverallQual (Overall Quality of the building)
2. ExterQual (Evaluates the quality of the material on the exterior)
3. YearBuilt (The year the house is built)

Cross Validation

In the cross validation, we have chosen to look at R^2, RMSE and MAE.

R2	RMSE	MAE
0.9145	28186	18385

3. PCR (Iris)

Cross Validation

4. Ridge Regression

Explanation

The reason we choose Ridge regression model is Ridge regression is very similar to linear regression, both try to minimize the RSS, but ridge regression has a penalty term, this could help us to prevent overfitting when add more predictors.

Prepare Model

1. Bootstrap Training Data
2. First, we set initial alpha to 1 to fit the ridge regression,and set the values of initial lambda ranging from 10ˆ10 to 10ˆ(-2), essentially covering the full range of scenarios from the null model containing only the intercept, to the least squares fit.

```
## [1] 17096.54
```

1. Then we use cross validation to choose the optimal lambda for Ridge Regression, as the following:

```
## [1]  6752.752 15720.045 19379.844  6978.871 10783.559

## [1] 11923.01

## [1] 0.03497927 0.03301763 0.07009450 0.08046841 0.05678258

## [1] 0.05506848

## [1] 0.1795723 0.1699684 0.2090431 0.3454549 0.2904304

## [1] 0.2388938

## [1] 11923.01

## [1] 0.05506848

## [1] 0.2388938
```

Check Accuracy

We then need to check accuracy, as assumed before, we would look at whether the predicted data is within ±5%. The following is the result.

```
## [1] 0.3908046 0.3908046 0.4137931 0.3984674 0.3678161

## [1] 0.3923372

## [1] "Accuracy of Ridge is approximately 44.90%"

## [1] 0.3946360 0.4789272 0.4444444 0.4559387 0.4712644

## [1] "Accuracy of Ridge with Log Transformation is approximately 44.90%"

## [1] 0.4291188 0.4636015 0.4750958 0.4482759 0.4559387

## [1] 0.4544061

## [1] "Accuracy of Ridge with Sqrt Transformation is approximately 45.44%"
```

Cross Validation

Then let us take a look at the MSE of this model:

```
## [1] 27872.15 28882.61 28645.82 26032.97 29049.92
```

R2	RMSE	MAE
0.8726	28097	17097

```
## [1] 18968.43 21804.18 28115.83 24305.01 31838.40
```

R2	RMSE	MAE
0.902	25006	14649

```
## [1] 664590291469 367770416577 371720863358  87440599536  90945256117
```

R2	RMSE	MAE
0.2273	3.165e+11	3.37e+10

5. Lasso Regression

Explanation

Lasso regression is pretty similar to Ridge regression. But compare to ridge, Lasso is more interpretable. It will make some predictors' coefficient to be exactly 0, which could help us find out which predictor is what Lasso thinks is important.

Prepare Model

1. Set the initial alpha is equal to 1 (Ridge regression is 0), and also use the same initial lambda, then try to use cross validation to choose the optimal lambda for Lasso.
2. With cross validation, we find out the optimal lambda as following:

```
## [1]  277.1055  342.1663 1197.8487  120.2052 1859.1555

## [1] 759.2962

## [1] 0.0030537439 0.0035278000 0.0034186671 0.0009344964 0.0019449767

## [1] 0.002575937

## [1] 0.017113706 0.007467627 0.008349228 0.016389336 0.008561830

## [1] 0.01157635

## [1] 759.2962

## [1] 0.002575937

## [1] 0.01157635
```

Coefficient From Lasso Regression

Here we are going to show the predictors lasso choosed.

```
##  [1] "RoofMatlWdShngl"      "(Intercept)"          "Exterior2ndImStucc"
##  [4] "NeighborhoodStoneBr"  "TotalSqFeet"          "NeighborhoodNridgHt"
##  [7] "NeighborhoodNoRidge"  "SaleConditionPartial" "GarageTypeNone"
## [10] "HeatingWall"

##  [1] "(Intercept)"          "TotalSqFeet"          "OverallQual"
##  [4] "RoofMatlTar.Grv"      "NeighborhoodCrawfor"  "SaleConditionPartial"
##  [7] "NeighborhoodNridgHt"  "NeighborhoodStoneBr"  "GarageCars"
## [10] "TotBathrooms"

##  [1] "(Intercept)"         "RoofMatlTar.Grv"     "TotalSqFeet"
##  [4] "NeighborhoodStoneBr" "NeighborhoodCrawfor" "RoofMatlWdShngl"
##  [7] "NeighborhoodNridgHt" "NeighborhoodVeenker" "NeighborhoodSomerst"
## [10] "Exterior1stBrkFace"
```

Check accuracy

We then need to check accuracy, as assumed before, we would look at whether the predicted data is within the ±5% range. The following is the result.

```
## [1] 0.3295019 0.3639847 0.3678161 0.4521073 0.4022989

## [1] 0.3831418

## [1] "Accuracy of Lasso is approximately 41.46%"
## [1] 0.4252874 0.3524904 0.4406130 0.4176245 0.4367816

## [1] "Accuracy of Lasso with Log Transformation is approximately 41.46%"

## [1] 0.4865900 0.4980843 0.4482759 0.4367816 0.4482759

## [1] 0.4636015

## [1] "Accuracy of Lasso with Sqrt Transformation is approximately 46.36%"
```

Cross Validation

Then let us take a look at the MSE of this model:

```
## [1] 28586.42 23764.23 27996.58 25571.32 24561.55
```

R2	RMSE	MAE
0.8791	26096	16810

```
## [1] 22082.73 20823.16 23653.41 21220.36 22880.75
```

R2	RMSE	MAE
0.9204	22132	14937

```
## [1] 21506.99 19147.44 19245.56 20400.12 20948.25
```

R2	RMSE	MAE
0.9199	20250	12923

6. GAM

Explanation

We have chosen GAM as one of our models because it produces an analysis on those factors that have less linear relationship with the result, for instance LotFrontage, YearRemodAdd, and MasVnrArea that are having relatively high importance but also high p-value that makes them not very linear related to SalePrice.

1) GAM1

In this model, we have LotFrontage, YearRemodAdd and MasVnrArea as predictors, with YearRemodAdd having a degree of freedom 2. We obtain the following result:

2) GAM2

In this model, we have LotFrontage, YearRemodAdd and MasVnrArea as predictors. None of them has a degree of freedom in the fit. We obtain the following result:

3) GAM3

In this model, we have LotFrontage, YearRemodAdd and MasVnrArea as predictorswith LotFrontage having a degree of freedom of 3. We obtain the following result:

GAM Summary

We then take an ANOVA test to understand which model is the best and we have the following result:

We can see that from the anova test that P-value for the second model is the smallest, therefore, it is the most preferred.

Cross Validation

Then, we conduct a cross-validation on the second model only.

Evaluation of different models

Root MSE

Choose best fit model

Conclusion

1. Classfication

Discussion & Future Development

Resources