**Final Report - House Prices: Advanced Regression Techniques**

Group Members: Akansha Toppo, Abhishek Chandran, Rajshree Nanaware

**Introduction:** The Kaggle Competition "House Prices: Advanced Regression Techniques" is to predict House Price in Ames, IA. Dataset has been split into test and train dataset 50% each and there are total 79 predictor variables in this dataset representing different aspects of residential homes.

**Objective:** The objective of this project is to fit a parsimonious model predicting the Sale Price of Homes using the most significant factors by reducing the RMSE value and increasing the R-squared value.

**Exploratory Data Analysis :**

1. By analyzing the structure of data we found that Alley, Fence, MiscFeature, PoolQC are having NA in almost every observation. So, we dropped above 4 columns.

2. Using MissForest we imputed the NA's of the dataset.

3. Same set of operations has been performed on test dataset as well.

3. We used StepAIC (Forward, Backward, Both) methods for automatics predictors selection.

a) Forward method selected 76 variables

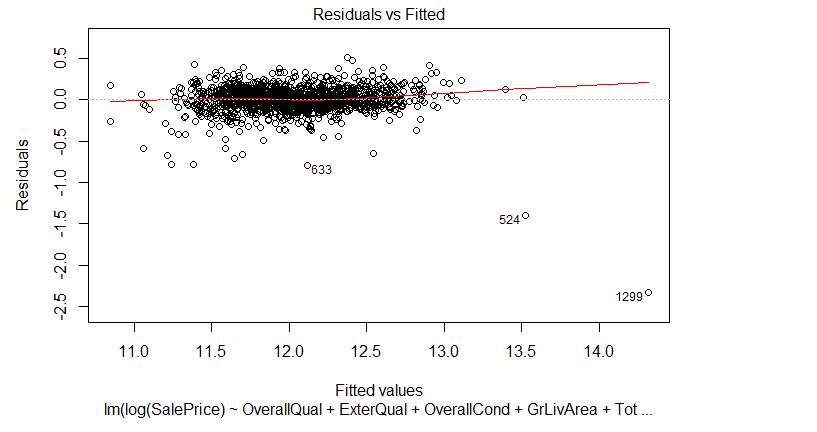
b) Backward method selected 76 variables

c) When both were used, model selected 47 variables

**Models:**

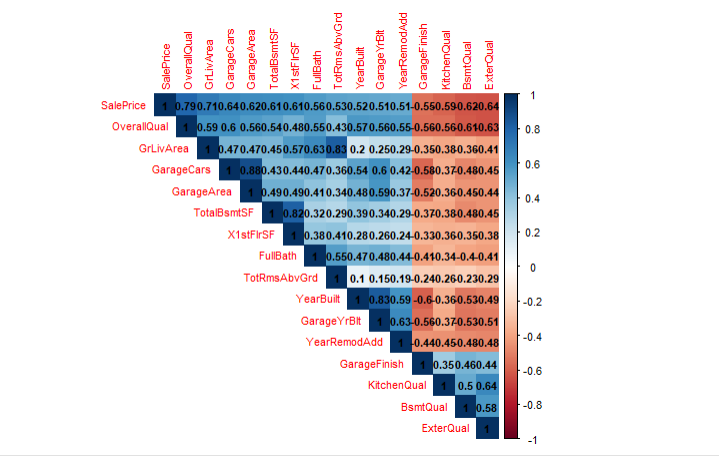
**LM MODEL:**

1. We ran LM model for all the variables. It gave a Rsquared value of 0.88
2. We ran LM model with the selected predictors and Rsquared value dropped to 0.85
3. Got an in-sample RMSE of 49164.74



Residual vs Fitted values Plot

1. We did cross-validation to find the expected RMSE
   1. Expected out of sample RMSE of train data: 0.1569
   2. Expected Rsquared of train data: 0.8470
2. We cross checked LM analysis by comparing it with correlation. For that we found most correlated numeric predictor which are having correlation > 0.5



Correlation Matrix of Outcome variable and Predictors

**GLMNET MODEL:**

We ran Ridge, Lasso and mix models with 10-fold cross validation and identified best lambda values for which we received smallest out of sample RMSE and high Accuracy

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Alpha** | **Lambda** | **RMSE(In sample)** | **R-squared** | **MAE** | **RMSE(Out of Sample)** |
| Ridge | 0 | 0 | 42399.23 | 0.748 | 19084.64 | 0.1632 |
| Lasso | 1 | 0.01 | 0.14650 | 0.8667 | 0.0941 | 0.1509 |
| Mix | 0.55 | 0.0065 | 0.1425968 | 0.8719 | 0.0898 | 0.1504 |

Out of all three models Mix had the best R-squared and RMSE values, 0.8719 and 0.1504 respectively with an alpha value of 0.55

**K-NN MODEL:**

We ran K-NN model on all predictors using carret library

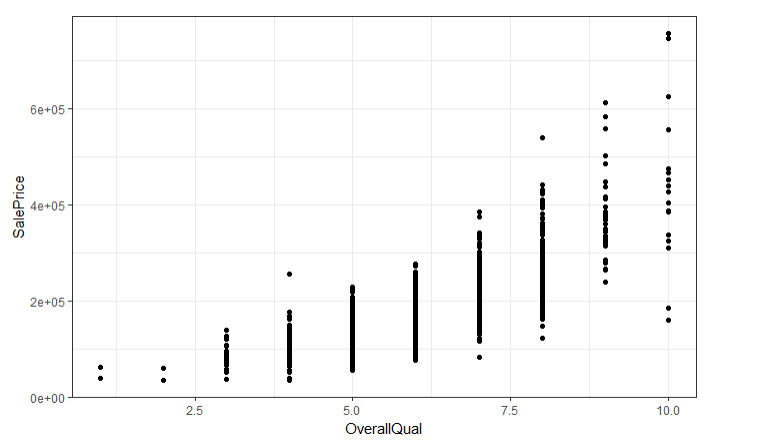
We got an out of sample RMSE value of 0.190087 and Rsquared value of 0.7779 at K equal to 9

**Visualizations:**

Analyzing Data and effect of different most Significant Predictors that were commonly identified through LM, KNN, GLMNET using ggplot library.

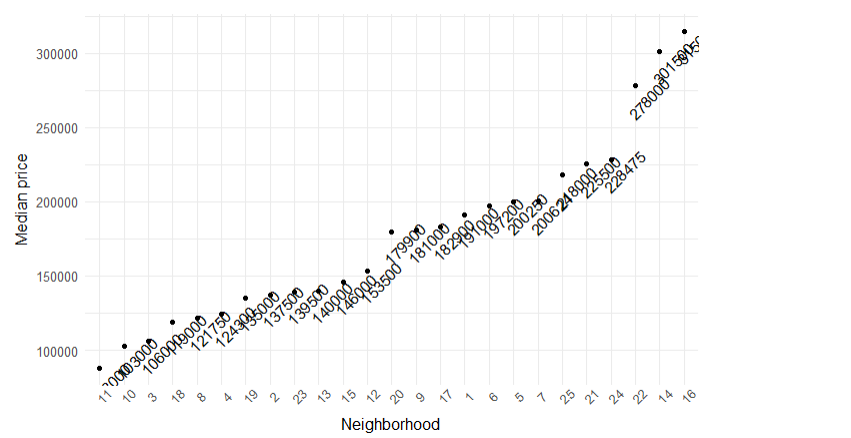
1. **Impact of most significant factor, Overall Quality on Sale Price:**

The coefficient for overall quality is .79 which seems to be the highest from rest all predictors. The graph clearly shows that better the quality higher is the sales price.



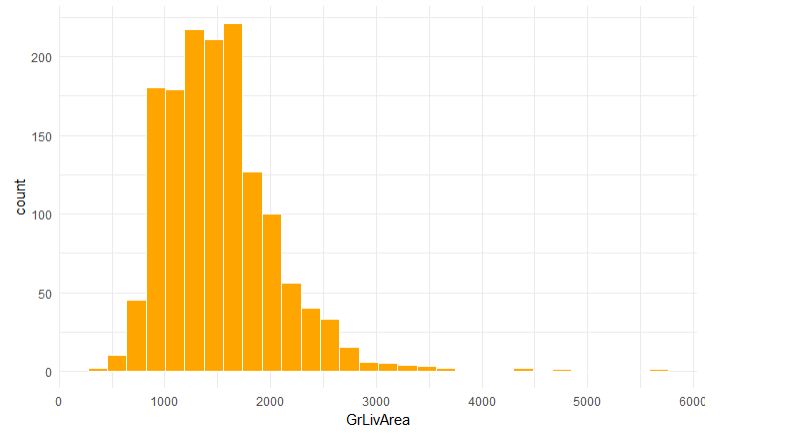
1. **Neighborhood with most expensive homes:**

In the below graph it a representation with median price and neighborhood we can see that for certain neighborhood there are more expensive homes than others concluding it to be significant predictor.



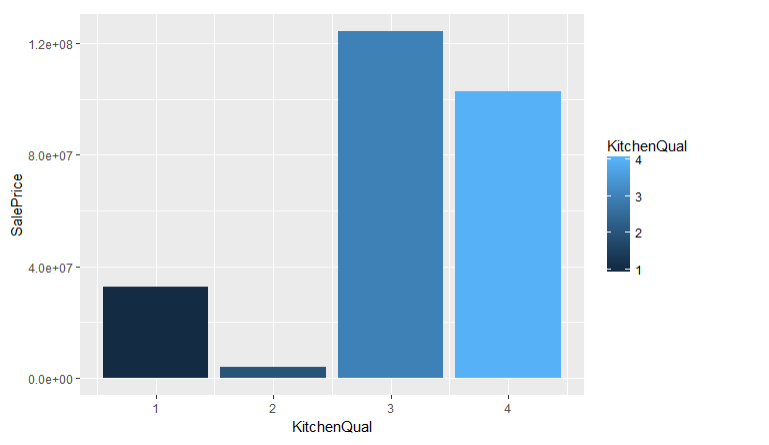
1. GrLivingArea Impact on Housing Price:

In the below graph it can be seen that the count is pretty high for around houses having ground living area between 1000 to 2000 sqft, hence making it a great predictor influencing the price.



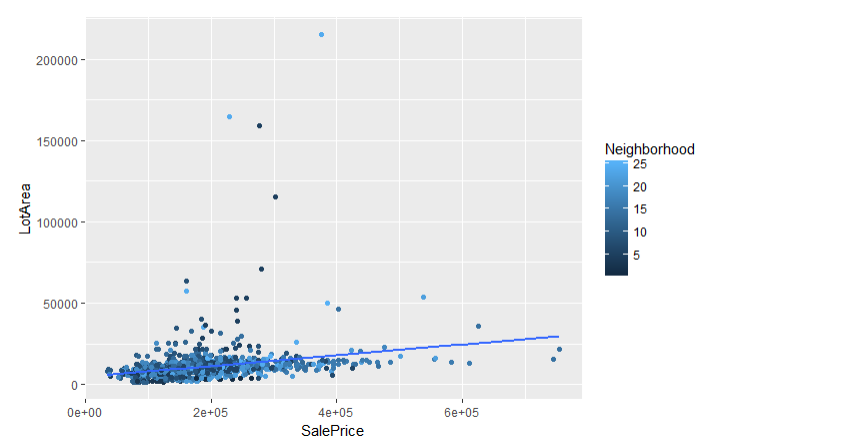
1. **Effect of Kitchen Quality on Sale Price:**

A good kitchen potentially always makes the house look better, as we can see in the below graph as well quality is segregated into four labels 1 being the least and 4 the best, the sales prices is highly corelated.



1. LotArea is contributing to higher Sale Price:

In some areas LotArea contributes to high sale price, while in other areas it is not that significant. We have displayed in regards with neighborhood making it easier to determine the correlation.



**Kaggle Result:**

So based on comparison of different models, we found that GLMNET(mixed) model performance was best based on out of sample RMSE and Rsquared values.

Kaggle score: 0.13575

Kaggle Rank :1822