## SUBJECTIVE QUESTIONS – ADVANCED LINEAR REGRESSION

***Question 1 - What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?***

***Answer –*** Optimal value for alpha for ridge and lasso are 10 and 100 respectively.
Post doubling, ridge and lasso alphas became 20 and 100 respectively.
With this increase, we were able to see that there is a change in the coefficient values for both ridge and lasso. Some coefficients increased and some dropped.
For Ridge, R2 values dropped from 81.7% to 81%
For Lasso, R2 values dropped from 82% to 81.4%
Top 5 Features are as follows: *Neighborhood_NoRidge, Neighborhood_NridgHt, OverallQual, Neighborhood_Veenker, Neighborhood_Somerst*

***Question 2 - You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?***
***Answer –*** When you look at the final grid as show below, Lasso model has removed one of the unwanted feature *(Exterior1st_ImStucc).* Removing features helps to make the model simple and become more generalised. Hence I would prefer Lasso regression.

Changes oberserved in the coefficients post regularization

```
[114]: betas = pd.DataFrame(index=X_train_new.columns)
       betas.rows = X_train_new.columns
       betas['Linear'] = lm.params
       betas['Ridge'] = ridge.coef_
       betas['Lasso'] = lasso.coef_
       pd.set_option('display.max_rows', None)
       betas.head(30)
```

[114]:

| | Linear | Ridge | Lasso |
|---|---|---|---|
| OverallQual | 27647.360170 | 30197.260074 | 28863.132242 |
| BsmtUnfSF | -8003.662858 | -7971.701458 | -7854.130565 |
| TotalBsmtSF | 9279.749866 | 10302.799528 | 9535.679861 |
| 1stFlrSF | 20441.708979 | 21067.339051 | 20794.986319 |
| 2ndFlrSF | 22293.776139 | 22419.811189 | 21973.715801 |
| MSSubClass_90 | -12705.404824 | -10782.423534 | -10189.787524 |
| MSSubClass_120 | -24969.485209 | -15070.880943 | -19073.213566 |
| MSSubClass_160 | -26146.878217 | -17777.077213 | -21720.993797 |
| MSZoning_RM | -7424.650337 | -10404.453028 | -8617.806883 |
| LotShape_2 | -29970.479199 | -16247.709454 | -19443.269157 |
| LotConfig_CulDSac | 16063.088674 | 14769.833497 | 15087.340123 |
| Neighborhood_ClearCr | 28400.224142 | 17099.303242 | 22784.217815 |
| Neighborhood_Crawfor | 29790.365972 | 18767.423629 | 24175.728540 |
| Neighborhood_NoRidge | 61624.658164 | 41983.467170 | 55702.634357 |
| Neighborhood_NridgHt | 61178.037715 | 43836.597241 | 55191.242645 |
| Neighborhood_Somerst | 29486.801447 | 19166.268752 | 24819.180450 |
| Neighborhood_StoneBr | 37862.435338 | 14492.149873 | 24382.535145 |
| Neighborhood_Veenker | 36174.260024 | 16529.826352 | 25793.530567 |
| HouseStyle_1Story | 14254.162243 | 12590.964563 | 12800.873608 |
| Exterior1st_ImStucc | -74891.157565 | -3431.025860 | -0.000000 |
| Exterior1st_Stucco | -19994.363639 | -14938.387798 | -16133.664803 |
| Exterior2nd_BrkFace | 17952.954449 | 10171.457492 | 10475.783466 |
| Exterior2nd_CmentBd | 18373.753150 | 14262.888747 | 15316.266930 |
| Exterior2nd_ImStucc | 37854.979510 | 12482.649342 | 14301.043421 |
| Exterior2nd_VinylSd | 13420.637152 | 12245.338185 | 12316.894992 |
| GarageType_NA | -23369.994554 | -17204.332013 | -19993.679852 |
| GarageFinish_1 | -17571.892770 | -14927.980264 | -15981.049623 |
| GarageFinish_2 | -13500.522442 | -11445.738902 | -12116.236185 |

***Question 3 - After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?***

***Answer -*** Post removal of top 5 predictors, now the new top 5 predictors in Lasso model are - *GarageType_NA, HouseStyle_1Story, GarageFinish_1, 1stFlrSF, Exterior2nd_VinylSd*

**Question 4 - How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?**

**Answer –** The generalised aspects that we need to focus on are the p-value, the VIF value and the R2 & Adj. R2 values. For our model,

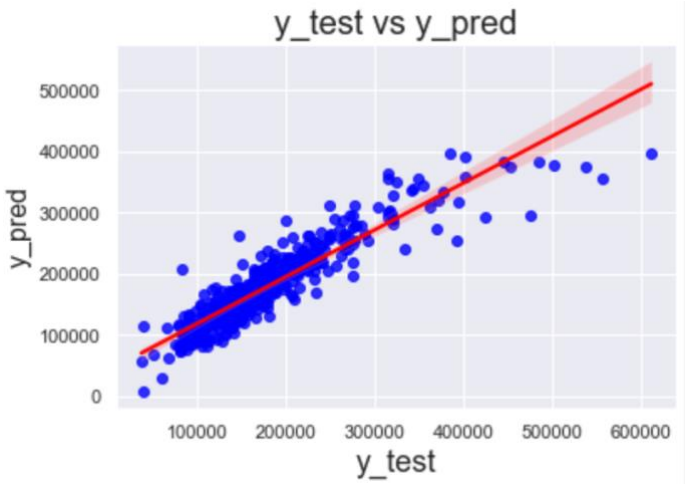
- We have attained a good VIF value of less than 5
- The p values are also less than 0.05
- Accuracy scores are consistent with training and test dataset ranging from 80%-82%.

Visualising through a table

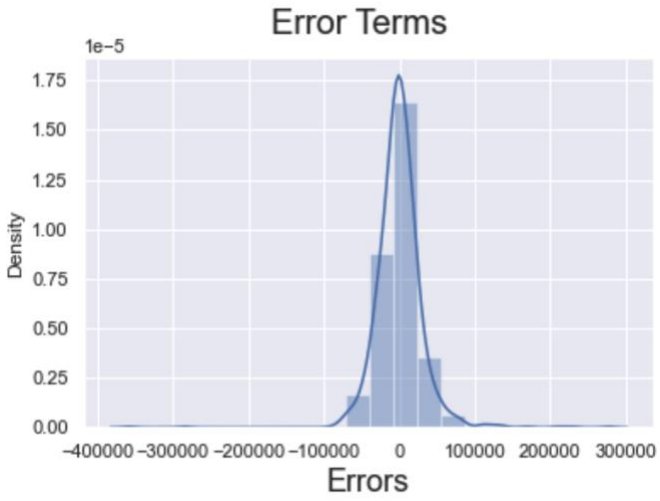
```
# Master table with all metrics for comparison

lr_table = {'Metric': ['R2 Score (Train)','R2 Score (Test)','RSS (Train)','RSS (Test)',
                       'MSE (Train)','MSE (Test)'],
        'Lasso Regression': metric1
        }

lr_metric = pd.DataFrame(lr_table ,columns = ['Metric', 'Lasso Regression'] )
rg_metric = pd.Series(metric2, name = 'Ridge Regression')
ls_metric = pd.Series(metric3, name = 'Linear Regression')

master_metric = pd.concat([lr_metric, rg_metric, ls_metric], axis = 1)
master_metric
```

| | Metric | Lasso Regression | Ridge Regression | Linear Regression |
|---|---|---|---|---|
| 0 | R2 Score (Train) | 8.213612e-01 | 8.171958e-01 | 8.238588e-01 |
| 1 | R2 Score (Test) | 8.208627e-01 | 8.172023e-01 | 8.238645e-01 |
| 2 | RSS (Train) | 1.140047e+12 | 1.166630e+12 | 1.124108e+12 |
| 3 | RSS (Test) | 5.061927e+11 | 5.165360e+11 | 4.977104e+11 |
| 4 | MSE (Train) | 3.341555e+04 | 3.380288e+04 | 3.318113e+04 |
| 5 | MSE (Test) | 3.395672e+04 | 3.430189e+04 | 3.367101e+04 |

- The y_test vs y_pred graph also seems to follow a linear pattern and the points are closer to the fitted line.



- Error terms are normally distributed as shown below




All these proves that that model that we have come up with is a much generalised and robust one.