PROBLEM STATEMENT – II SUBJECTIVE QUESTIONS

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:

1. The Optimal value of alpha for ridge is 0.01 and for lasso it is 0.0001.
2. R2 score for both Ridge and Lasso Models are 0.89 approx. on test data and 0.90 on train data.
3. On doubling the alpha values of the Ridge and Lasso model, the prediction accuracy remains around 0.90 with minor change in the co-efficient values.

RIDGE COEFFICIENTS

```
print('Ridge Original')
betas['Ridge'].sort_values(ascending=False)[0:5]
```

```
Ridge Original

GrLivArea      0.962081
OverallQual    0.467781
OverallCond    0.345036
LotArea        0.308243
1stFlrSF       0.236247
Name: Ridge, dtype: float64
```

```
print('Ridge Alpha Double')
betas['ridge_alpha_double'].sort_values(ascending=False)[0:5]
```

```
Ridge Alpha Double

GrLivArea      0.955080
OverallQual    0.468087
OverallCond    0.344763
LotArea        0.306605
1stFlrSF       0.237307
Name: ridge_alpha_double, dtype: float64
```

LASSO COEFFICIENTS

```
: print('Lasso Original')
  betas['Lasso'].sort_values(ascending=False)[0:5]

  Lasso Original

: GrLivArea      0.948586
  OverallQual    0.480681
  OverallCond    0.342535
  LotArea        0.252205
  GarageCars     0.218025
  Name: Lasso, dtype: float64

: print('Lasso Alpha Double')
  betas['lasso_alpha_2'].sort_values(ascending=False)[0:5]

  Lasso Alpha Double

: GrLivArea      0.929948
  OverallQual    0.494536
  OverallCond    0.338344
  GarageCars     0.220170
  LotArea        0.198938
  Name: lasso_alpha_2, dtype: float64
```

4. Overall since the alpha values are small, we do not see a huge change in the model after doubling the alpha.
5. The most important predictor are as follows:
    a. Total Garage Area
    b. Overall Quality
    c. Overall Condition
    d. Lot Area
    e. Garage Cars

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?

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Ridge Regression alpha * 2	Lasso Regression alpha * 2
0	R2 Score (Train)	0.905987	0.905987	0.905532	0.905985	0.904731
1	R2 Score (Test)	0.893900	0.893931	0.894821	0.893959	0.894707
2	RSS (Train)	15.088288	15.088375	15.161433	15.088628	15.289935
3	RSS (Test)	7.646455	7.644273	7.580104	7.642227	7.588293
4	MSE (Train)	0.121565	0.121565	0.121859	0.121566	0.122374
5	MSE (Test)	0.132127	0.132109	0.131553	0.132091	0.131624

1. The optimum lambda value of Ridge and Lasso are as follows: -
    a. Ridge – 0.01
    b. Lasso – 0.0001
2. The Mean Squared Error in case of Ridge and Lasso are:
    a. Ridge - 0.121565
    b. Lasso - 0.121859
3. As we can see the Mean Squared Error are almost same for both models.
4. Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used as the final model.

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:

1. The top five important predictor variables in the lasso model are: -
    a. GrLivArea
    b. OverallQual
    c. OverallCond
    d. LotArea
    e. GarageCars
2. After removing these attributes from the dataset, the new Top 5 predictors are: -
    a. 1stFlrSF
    b. TotRmsAbvGrd
    c. FullBath
    d. GarageQual
    e. KitchenQual

```
In [110]: betas_2 = pd.DataFrame(index=X_train_RFE.columns)
          betas_2['lasso_quest_3'] = lasso_quest_3.coef_
          betas_2['lasso_quest_3'].sort_values(ascending=False)[0:5]

Out[110]: 1stFlrSF        0.690035
          TotRmsAbvGrd    0.548737
          FullBath        0.335345
          GarageQual      0.291006
          KitchenQual     0.212871
          Name: lasso_quest_3, dtype: float64
```

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:

1. As these two models show similar 'performance' in the finite training or test data then we should pick the one that makes fewer on the test data due to following reasons:-
    a. Simpler models are usually more 'generic' and are more widely applicable.
    b. Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
    c. Simpler models are more robust.
        i. Complex models tend to change wildly with changes in the training data set.
        ii. Simple models have low variance, high bias and complex models have low bias, high variance.
        iii. Simpler models make more errors in the training set. Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples
2. Therefore, the model should be generalized so that the test accuracy is not lesser than the training score.
3. The model should be accurate for datasets other than the ones which were used during training.
4. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained.
5. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.