Question 1

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:-

Below is the Optimal Value of alpha for Ridge and Lasso

Ridge - 0.8

Lasso - 0.0001

Metric of Ridge and Lasso

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.926801	0.934705
1	R2Score Test	0.900364	0.906424
2	RSS Train	1.438550	1.283213
3	RSS Test	0.805244	0.756267
4	MSE Train	0.001452	0.001295
5	MSE Test	0.001890	0.001775

Doubling the value of alpha

Ridge - 0.16

Lasso - 0.0002

Metric of Ridge and Lasso

	Metric	Ridge regression	Lasso regression
0	R2 Score Train	0.927425	0.930591
1	R2Score Test	0.900804	0.904642
2	RSS Train	1.426272	1.364063
3	RSS Test	0.801693	0.770676
4	MSE Train	0.001439	0.001376
5	MSE Test	0.001882	0.001809

There is no significant change in the metric after doubling the alpha value.

The R2 value is almost same after for Lasso and Ridge on train and test set.

The important feature of Lasso model as it has less difference in R2 in comparison to Ridge.

BsmtFullBath

OverallCond

MasVnrArea

HeatingQC

MSSubClass

OverallQual

GarageQual

BsmtFinType2

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:-

Based on the alpha/Lamda values I have got, Ridge regression does not zero any of the coefficient, Lasso zeroed one or more coefficients in the selected features, Lasso is better option and it also helps in the some of the feature elimination.

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:-

Removing the top 5 features

BsmtFullBath, OverallCond, MasVnrArea, HeatingQC, OverallQual which I have derived using lasso.

After removable of these feature again training the model with these top 5 feature yield me following top feature using lasso.

FullBath,

MSSubClass,

2ndFlrSF,

WoodDeckSF,

GarageQual

Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer:-

There different techniques which we need to check while building a robust and generalizable.

1. Treatment of Outlier
   Whenever we have dataset it can have outlier and we should check it using boxplot. Once we visualised the outlier in features we should remove it proper else it will affect the help the model to have make better prediction by using unbiased data.
2. Missing value treatment
   In dataset, some feature has few missing value then we should remove it by using mean, median and mode depending upon the type and nature of the feature.
   If feature has higher % of missing value then we should analyse it and drop those feature.
3. Significance of predicted variable
   Model significance can be determined by P-value, R2 and adjusted R2.

   Implications of accuracy of the mode
   1. We should have more and more data containing different combination should that model should learn from different feature instead of working with small dataset.
   2. Outliers and missing value treatment.
      Outlier present in the dataset for feature can affect the model accuracy and should be removed after visualisation using boxplot and provide standard data to the model.
      Missing value should be treated based on the type of features and amount of missing value else it lead to an inaccurate model.
   3. Feature Selection
      Based on the domain knowledge we need to select the feature which is important and related to target variable.
      We can use VIF value along with p-value for feature selection.
   4. Feature Engineering

If we can derive more important feature from the existing feature which can help us in getting more insight and better relation with target variable then we should do that.

5. Correct algorithm selection
   Choosing the right algorithm for the problem statement is very crucial for accuracy of the model. It comes with knowledge and experience.
6. Cross validation
   More accuracy can sometime lead to over fitting the model.
   In those case we use cross validation to get the model with correct accuracy.