Questisson 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 for Ridge and Lasso regression are **100** and **0.01** respectively.*

Changes in the Ridge and Lasso Model on doubling the value of alpha :-

- ***Ridge Regression***

  *After doubling the alpha from **100** to **200**:*
  - *Train R2 score decreased to 0.9383 from 0.94265*
  - *Test R2 Score remains the same.*
  - *There is a reduction in the coefficient of the features.*

  The following are the top 5 predicted variables and their coefficients:

FEATURE	COFFICIENT
BsmtFullBath	0.1065
OverallCond	0.1031
2ndFlrSF	0.0869
YearBuilt	0.0767
1stFlrSF	0.0734

- ***Lasso Regression***

  *After doubling the alpha from **0.01** to **0.02**:*
  - *Train R2 score decreased to 0.9224 from 0.9327*
  - *Test R2 Score decreased to 0.8845 from 0.8918.*
  - *The no of coefficient reduced to 49 from 86.*
  - *There is change in the coefficient value of the features.*

  The following are the top 5 predicted variables and their coefficients:

FEATURE	COFFICIENT
BsmtFullBath	0.3148
OverallCond	0.2384
1stFlrSF	0.1208
OverallQual	0.1081
YearBuilt	0.0939

***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 #2

We know that both Ridge and Lasso regularize the coefficients by reducing them in value, essentially causing shrinkage of the coefficients.

However, I will go with Lasso model. Because, while model building, it is observed that the r2_scores are almost the same for both Ridge and Lasso Regression models. But as lasso penalizes more on the dataset and can also help in feature elimination (as evident that no of non-zero features in the lasso model with alpha = 0.01, reduced to 86), it is best to go with Lasso as a 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 # 3

Following are the five most important predictor variables and their coefficients, after removing the original five:

FEATURE	COFFICIENT
BsmtHalfBath	0.3602
2ndFlrSF	0.1331
YearRemodAdd	0.0831
SaleType_Oth	0.0674
GarageArea	0.0658

***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?

We can make a robust and generalisable model by making it as simple as possible. Extremely simpler model could highly biased and is likely to fail in predicting complex real world scenarios. On one hand, simplicity is generalizable and robust and on the other hand, some problems are inherently complex in nature.

Implication of simpler model is that - machine learning does not simply involve building models to fit the available data. The real challenge is to learn patterns which can be used to explain the behaviour of similar unseen data, which a simple model could not achieve.

Imagine solving digital image processing problems using simple linear regression when much more complex models like neural networks are typically successful in these problems. We say that the linear model has a high bias since it is way too simple to be able to learn the complexity involved in the task.

There is a trade-off between the bias and variance, which is known as the bias-variance tradeoff in machine learning.

A process called Regularization is used to create an optimally complex model, i.e. a model which is as simple as possible while performing well on the training data.Through regularization, we try to strike the delicate balance between keeping the model simple, yet not making it too naive to be of any use.

There are two commonly used regularised regression methods - Ridge regression and Lasso regression. Both these methods are used to make the regression model simpler while balancing the 'bias-variance' trade-off.