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 of lambda for Ridge Regression is 10 and optimal value of lambda for Lasso Regression is 0.001.
- If we double the value of alpha for both ridge and lasso -
    - Ridge Regression R2 score of train set decreased from 0.943 to 0.937and R2 score of test set remained same at 0.911.
    - Lasso Regression R2 score of train set decreased from 0.925 to 0.910 and score of test set decreased from 0.93 to 0.892.
- The most important predictor variables after we double the alpha values are Neighborhood_Crawfor, GrLivArea, OverallQual_8, OverallQual_9, Functional_Typ, TotalBsmtSF, OverallCond_9, Exterior1st_BrkFace, OverallCond_7, SaleCondition_Normal, OverallQual_7, YearRemodAdd, Condition1_Norm, CentralAir_Y.

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 :

- Since the objective of the exercise is to identify which variables are significant in predicting the price of a house, and how well those variables

describe the price of a house, we will choose lasso regression for this assignment.

- Lasso regression will enable us to eliminate variables that do not contribute to predicting house prices and only get the variables which contribute towards it. Moreover, it can also help us in gauging the impact of those remaining variables which is specifically the aim of our exercise.
- In our case, even though ridge regression outperforms lasso regression by slightest of margins, the performance of both the models can be considered good and as per the expectations. But lasso regression serves our purpose better as it eliminated 305 features out of 383 which helps in avoiding confusion of going through the impact of each and every variable.

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 :

After dropping our top 5 lasso predictors, we get the following new top 5 predictors :

- 2ndFlrSF
- 1stFlrSF
- TotalBsmtSF
- Exterior1st_BrkFace
- Neighborhood_Somerst

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 :

- A model can be considered robust when sudden variations in data doesn't impact it significantly.
- A model can be considered generalisable when it is able to adapt well to some unseen data drawn from the same or similar distribution as that of the data it was trained on.
- We have to make sure the model doesn't overfit for it to be robust and generalisable as an overfitted model has high variance and thus, slight changes in data can affect model predictions. In that case, the model might not perform well on unseen data.
- The model should not be too complex and thus, should not learn all the patterns in the data but only get a generic sense of patterns.
- The accuracy of an overfitted model will mostly be quite high, but it might underperform on unseen data since it has learned the training patterns too well and not generalized those patterns.
- So to counter that, we need to decrease the variance which will lead to some increased bias. That increased bias will lead to less accuracy, but then the model will perform well on unseen data.
- A good model will try to find some balance between accuracy and complexity which can be achieved through techniques such as Ridge and Lasso Regression.