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:

The optimal lambda value in case of ridge and lasso is as below:

- Ridge - 8.0
- Lasso - 0.0003

I choose these values based on best fit determined by algorithm.

When we double the value of alpha for our ridge regression no we will take the value of alpha equal to 16, the model will apply more penalty on the curve and try to make the model more generalized that is making model more simpler and no thinking to fit every data of the data set. The most important variable after the changes has been implemented for ridge regression are as follows:-

- OverallQual
- MSZoning
- BsmtExposure
- Neightborhood
- RoofStyle
- Foundation
- HouseStyle

Similarly when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases. The most important variable after the changes has been implemented for lasso regression are as follows:-

- SaleCondition
- GrLivArea
- Neighborhood
- OverallQual
- LotConfig
- BsmtExposure
- KitchenQual
- RoofStyle

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:

The Mean Squared error in case of Ridge and Lasso are:

- Ridge - 0.016025
- Lasso - 0.016022

The Mean Squared Error of Lasso is slightly lower than that of Ridge. Also, since Lasso helps in feature reduction Lasso has a better edge over Ridge. So I will choose lasso.

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:

- MSZoning
- Neightborhood
- GrLivArea
- OverallQual
- BsmtExposure

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 model has to be as simple as possible, although its accuracy will decrease a bit but it will be more robust and generalisable. It can be also understood using the Bias Variance trade-off. The simpler the model the more the bias but less variance and will be more generalizable.

The implications for the accuracy is that robust and generalisable model will perform equally well on both training and test data. The accuracy doesn't change much for training and test data.

High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

High variance means model performs exceptionally well on training data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and underfitting of data