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 - 3
- Lasso - 0.0002

When we double the value of alpha for both ridge and Lasso, Ridge-0.8 and lasso-0.0004 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 . 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:-

1. MSZoning_RL
2. GrLivArea
3. MSZoning_RM
4. OverallQual
5. MSZoning_FV
6. TotalBsmtSF
7. Foundation_PConc
8. OverallCond
9. GarageCars
10. MSZoning_RH

The most important variable after the changes has been implemented for Ridge regression are as follows:-

1. MSZoning_RL
2. GrLivArea
3. MSZoning_RM
4. OverallQual
5. TotalBsmtSF
6. MSZoning_FV
7. Foundation_PConc
8. OverallCond
9. GarageCars
10. MSZoning_RH

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:

Lasso helps in feature reduction (as the coefficient value of became 0), Lasso has a better edge over Ridge.

Hence based on Lasso, the factors that generally affect the price are the Zoning classification, Living area square feet, Overall quality and condition of the house, Foundation type of the house, Number of cars that can be accommodated in the garage, Total basement area in square feet and the Basement finished square feet area

Therefore, the variables predicted by Lasso is significant variables for predicting the price of a house.

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 creating another model excluding the five most important predictor variables, Five most important predictor variables now are:

1. MasVnrArea
2. BsmtFinSF1
3. BsmtUnfSF
4. TotalBsmtSF
5. 1stFlrSF

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 should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. 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. 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.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of 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 under-fitting of data.