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:

Used an alpha of 0.01. if we double the value then the model will apply more penalty . When we observe graph for alpha 10 we have more errors on both training data and test data too.

The most predictor variables before regression are

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

After regression the most predictor variables are

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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 term with lambda is often called 'Penalty' since it increases RSS. We iterate certain values onto the lambda and evaluate the model with a measurement such as 'Mean Square Error (MSE)'. So, the lambda value that minimizes MSE should be selected as the final model. This **ridge regression model is generally better than the OLS model in prediction**.

In this case Ridge regression includes all variables in final model unlike Lasso Regression.

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:

We shall use the next 5 predictor variables which are

1. GarageArea
2. Fireplaces
3. LotArea
4. LotArea
5. LotFrontage

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 needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable 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 weightage 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 outlier 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. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.