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?

Ans:

1. Optimal value of alpha for Ridge is 10
2. Optimal value of alpha for Lasso is 0.01

If I double the value of in Ridge, the penalty increases. This results in decrease in the values of coefficients. In Ridge, it does not set the coefficients to zero which means it does not perform feature selection.

If I double the value of in Lasso, the penalty increases like in Ridge. This results in decrease in the values of coefficients. However in Lasso, it can set the coefficients to 0 which will result in feature selection. As we increase the alpha value, more and more coefficients will become zero and only fewer predictor variables will be left in the model

In Ridge, it lowered the coefficients

Reduced values of top5 coefficients

('GarageFinish_RFn', 0.075),

('GarageFinish_Unf', 0.077),

('SaleCondition_Normal', 0.079),

('SaleCondition_Others', 0.089),

('SaleCondition_Partial', 0.108)]

In Lasso,

('GarageFinish_RFn', 0.059),

('GarageFinish_Unf', 0.066),

('SaleCondition_Normal', 0.07),

('SaleCondition_Others', 0.104),

('SaleCondition_Partial', 0.156)]

More features are eliminated as the coefficients are turned to zero

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?

I chose to apply Lasso because the data set provided has large number of predictors and very few of them are important which will have substantial effect on the outcome. Also, Lasso provides feature selection by making the coefficients zero.

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?

In Lassi, the next best five predictors are

('GarageType_BuiltIn', 0.037),

('GarageType_Detchd', 0.043),

('GarageType_No Garage', 0.047),

('GarageType_Others', 0.05),

('GarageFinish_No Garage', 0.052),

In Ridge, the next best five predictors are

('GarageType_BuiltIn', 0.049),

('GarageType_Detchd', 0.049),

('GarageType_No Garage', 0.05),

('GarageType_Others', 0.051),

('GarageFinish_No Garage', 0.052)

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?

A complex model has low bias and high variance. It fits the training data well but falters with the test data. Whereas a simple model has high bias and lower variance, it may not fit the training data well but tends to better with unseen data. Ridge and Lasso regression are techniques to balance the bais-variance tradeoff. They regularize by adding penalty and reducing the coefficients which will help in reducing the variance. This will help in reducing the overfitting making the model robust and generalizable.