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 values of alpha for ridge is 50.0 and Lasso is 0.001.

a. Ridge

Before:

R2 score (test) : 0.9063747171226282

RMSE (test) : 0.1268016826600198

After changing alpha to 100 for ridge:

R2 score (test) : 0.9051751993451124

RMSE (test) : 0.1276113827146637

Most important predictor variables: 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'GarageCars', 'RoofMatl_CompShg', 'LotArea', 'MSZoning_RL', 'BsmtFullBath', 'Neighborhood_NridgHt'

b. Lasso

Before:

R2 score (test) : 0.9044339715549808

RMSE (test) : 0.12810916902768948

After changing alpha to 0.002 for lasso:

R2 score (test) : 0.9077396484666771

RMSE (test) : 0.1258739896860086

Most important predictor variables: 'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'OverallCond', 'GarageCars', 'LotArea', 'RoofMatl_CompShg', 'Neighborhood_NridgHt', 'BsmtFullBath', 'Neighborhood_Crawfor'

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:

Here ridge has performed little better than Lasso. But we do not see much significant difference. Will use Lasso model as final model as it marks insignificance features to zero which we have already analyzed as a part of EDA analysis, and helps to keep our model simple.

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:

The top 5 features are:

2ndFlrSF	0.067528
MSZoning_RL	0.066441
GarageCars	0.057220
MSZoning_RM	0.054915
OverallCond	0.053678

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:

To make sure a model is robust, we need to make sure the model does not overfit. Model overfits means it performs well on train data but does not perform well on test data. An overfit model has high variance and small change in data affects the model performance a lot. A model using too many features means

that model remembers all the data along with noise, which is called complex model. To protect from overfitting the model, we add some kind of biasness to our model. We generally use Regularization techniques using Lasso/Ridge and need to tune hyperparameter to get the optimal values of alpha.