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:

Ridge Regression

The Optimum value for alpha for Ridge Regression is **3**.
The R2 Score for the train data is **0.947** and Mean Squared Error is **0.001.**
The R2 Score for the test data is **0.905** and Mean Squared Error is **0.002.**

When Doubling the alpha there is a variation in the r2 score.

Results after doubling the Alpha(alpha=6)

R2 score for train data: 0.9416
Mean squared error for train data: 0.001
R2 score for test data: 0.9031
Mean squared error for test data: 0.002

The Top predictor variables:

LotArea
GarageArea
OverallCond
Neighborhood_NoRidge
2ndFlrSF
TotalBsmtSF
Neighborhood_StoneBr
BsmtFinSF1
1stFlrSF
OverallQual
GrLivArea

Lasso Regression

The Optimum value for alpha for Lasso Regression is **0.0001**.
The R2 Score for the train data is **0.946** and Mean Squared Error is **0.002.**
The R2 Score for the test data is **0.906** and Mean Squared Error is **0.002.**

When Doubling the alpha there is a variation in the r2 score.

R2 score for train data: 0.9404127598803985
Mean squared error for train data: 0.0013196225566313871
R2 score for test data: 0.9049542043258593
Mean squared error for test data: 0.0024015102594832907

The Top predictor variables:

SaleCondition_Partial
GarageCars
LotArea
GarageArea
Neighborhood_NoRidge
OverallCond
BsmtFinSF1
Neighborhood_StoneBr
TotalBsmtSF
OverallQual
GrLivArea

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 and Ridge algorithm given approximately the same R2 score and Mean Squared Error for the given data, for both train and test split.

The advantage of Lasso over Ridge in the current context is that Lasso helps to provide feature selection by shrinking the insignificant variables completely to zero and removing them from the model.

Since Lasso helps in feature selection we are choosing Lasso over Ridge for this assignment.

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:

OverallCond
BsmtFinSF1
TotalBsmtSF
OverallQual
GrLivArea

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 build a robust and generalizable model

- The model should be simple as possible
- The model should give comparable results for test data with approximately the same R2 score that we get from train data.

To understand the implications of accuracy we have to see bias–variance tradeoff. It describes the relationship between a model's complexity and prediction accuracy.

Bias represents the error in models, difference between prediction and correct value. High bias gives a large error in training and test data.

Variance represents the sensitivity of the model to changes in the training data. High variance means the model performs well on training data but shows high error in test data.

We have to create a model that will perform well in training data and give comparable results for test data or unseen data as well. We need to compromise in bias to achieve a good generalized model.