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 value of alpha(hyperparameter) for Ridge Regression is 11.5

Optimal value of alpha(hyperparameter) for Lasso Regression is 0.001

When we double the values of the alpha (Ridge alpha=14 and Lasso alpha =0.002)

- There is a slight reduction of R2 value
- We see the MSE for Ridge has went
- We can see in both Ridge and Lasso the R2 scores on train reduced with double of alpha.

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 :

- R2 score of lasso model is better
- MSE and RSS on lasso is lowest

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:

SaleType_cwd

GrLivArea

Functional_Type

Exterior1st_BrkFace

LandContour_HLS

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 ensure that our model is robust and capable of generalizing to new data, we employ techniques such as regularization and cross-validation using grid search.

Regularization helps prevent overfitting by adding a penalty term to the model's loss function. This encourages the model to find a balance between accuracy on the training set and simplicity, reducing the risk of overfitting.

Cross-validation is used to evaluate the model's performance. It involves splitting the data into train, validation, and test sets. By using separate validation and test sets, we can ensure that the model does not "sneak-peek" at the test data during training. This allows us to assess the model's performance on unseen data.

Backtesting is another technique employed to evaluate the model's performance. It involves testing the model on historical data to simulate how it would have performed in the past. This helps us gain confidence in the model's ability to generalize to new, unseen data.

The implication of building a robust and generalizable model is that it may result in a slight decrease in accuracy on the training set. However, this trade-off is worthwhile because it leads to an increase in accuracy on the test set. A robust model is less likely to overfit the training data and is more likely to generalize well to new, unseen data.