Question1 :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:

Optimum value of alpha for ridge regression is 10

Optimum value of alpha for lasso regression is 0.001

Double the value of alpha

Ridge - alpha =20

Lasso - alpha = 0.002

Ridge:

R2 score on train and test set for ridge regression are 0.94 and 0.90 res pectively for alpha =10

R2 score on train and test set for ridge regression are 0.93 and 0.90 res pectively for alpha =20

Top 5 preditcor variables for alpha=20 are:

('GrLivArea', 0.08820408141015838),

('Neighborhood_Crawfor', 0.06284622367025135),

('TotalBsmtSF', 0.06127110926598349),

('OverallQual', 0.060506236538536456),

('Exterior1st_BrkFace', 0.04991669686566937)

Lasso:

R2 score on train and test set for lasso regression are 0.92 and 0.90 res pectively for alpha=0.001

R2 score on train and test set for lasso regression are 0.91 and 0.90 res pectively for alpha=0.002

Top 5 predictor variables for alpha=0.002 are

('GrLivArea', 0.09621406414387107),

('OverallQual', 0.07955920246096802),

('TotalBsmtSF', 0.0551667075588023),

('Neighborhood_Crawfor', 0.05191754378612108),

('OverallCond', 0.04538489336102607)

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:

As the r2 score on test dataset is same for both lasso and ridge regression,we can choose either,but to interpret the model and understand the significant variables impact on target variable lasso regression would be as it performs automatic feature elimination.

Question3 : 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', 0.24989960939521258),
('Exterior1st_BrkFace', 0.1274435993563277),
('Exterior1st_Stone', 0.1118324734524794),
('2ndFlrSF', 0.10015591957292543),
('BsmtFinSF1', 0.0935971421983545)]

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 is said to be robust if the dependent/target variable doesn't change much with change in input/independent features.

A model is said to be generalisable if the model adapts to new,previously unseen data thereby giving good accuracy on test data.

Both of the above can be explained by Bias-variance trade off.The simpler the model,bias will be more ,variance will be less and the model is more generalisable and robust.Thus there won't be much difference in the accuracy of training and test datasets.

Bias measures how accuractely a model can describe the actual task.High bias means model is unable to learn details in data.Model peforms poor on training and testing dataset.Variance measures how flexible the model is with respect to changes in training data.High variance means model performs extremely well on training data by learning even the noisy details in data.

As complexity increases bias reduces and variance increases.The main aim is to find the optimal point where the total model error is least to avoid over fitting and under fitting.