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:
The optimal value of alpha for

- Ridge: 10.0
- Lasso: 0.001

If using 70:30 ratio of train and test data.

If we double the alpha value for ridge and lasso, I observed the following:
Ridge:

- Train set R2 decreased from 0.94 to 0.9328 & Test set R2 increased from 0.89 to 0.8839
- Train RSS increased from 9.53 to 10.78, Test RSS from 8.26 to 8.36
- No change in RMSE values

Lasso:

- Train set R2 decreased from 0.92 to 0.9049 & Test set R2 increased from 0.88 to 0.8709
- Train RSS increased from 12.65 to 15.25, Test RSS from 8.44 to 9.30
- Train RSME increased from 0.11 to 0.1222, Test RSS from 0.14 to 0.1457

Output from the notebook:

RIDGE	
R-Squared (Train): 0.94 (Test): 0.89 RSS (Train): 9.53 (Test): 8.26 RMSE (Train): 0.10 (Test): 0.14	{'alpha': 20} R-Squared (Train): 0.93 (Test): 0.88 RSS (Train): 10.78 (Test): 8.36 RMSE (Train): 0.10 (Test): 0.14
LASSO	
R-Squared (Train): 0.92 (Test): 0.88 RSS (Train): 12.65 (Test): 8.44 RMSE (Train): 0.11 (Test): 0.14	{'alpha': 0.002} R-Squared (Train): 0.90 (Test): 0.87 RSS (Train): 15.25 (Test): 9.30 RMSE (Train): 0.12 (Test): 0.15

Important predictor variables after doubling the alpha value (from both ridge and lasso models):

Out[88]:

	Ridge
OverallQual_9	0.074194
GrLivArea	0.070013
Neighborhood_Crawfor	0.065620
CentralAir_Y	0.063903
Functional_Typ	0.061403
TotalBsmtSF	0.060931
OverallQual_8	0.054137
2ndFlrSF	0.053105
MSSubClass_70	0.051948
OverallCond_7	0.051555

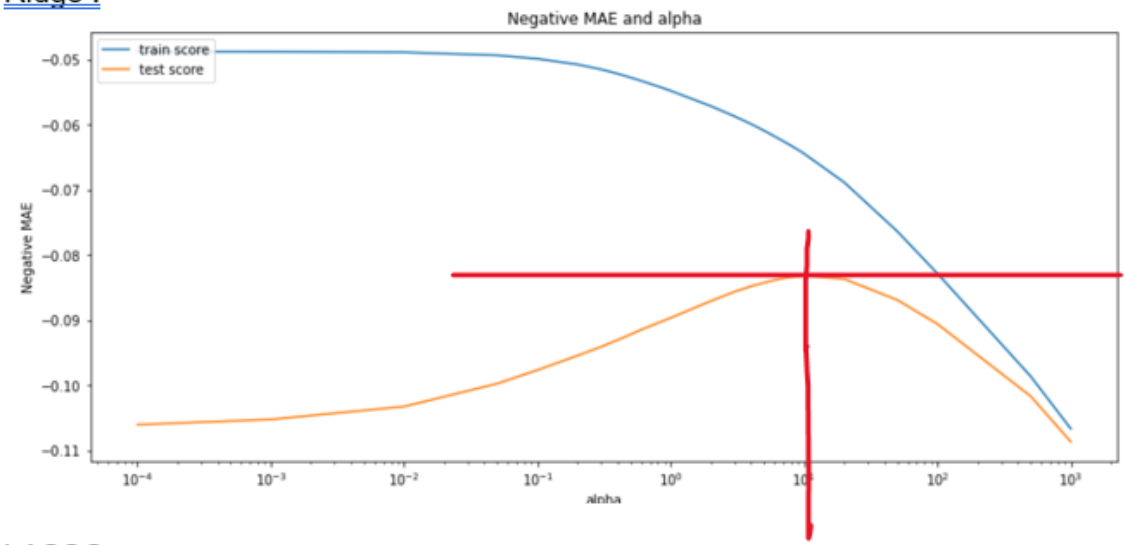
c

Question 2

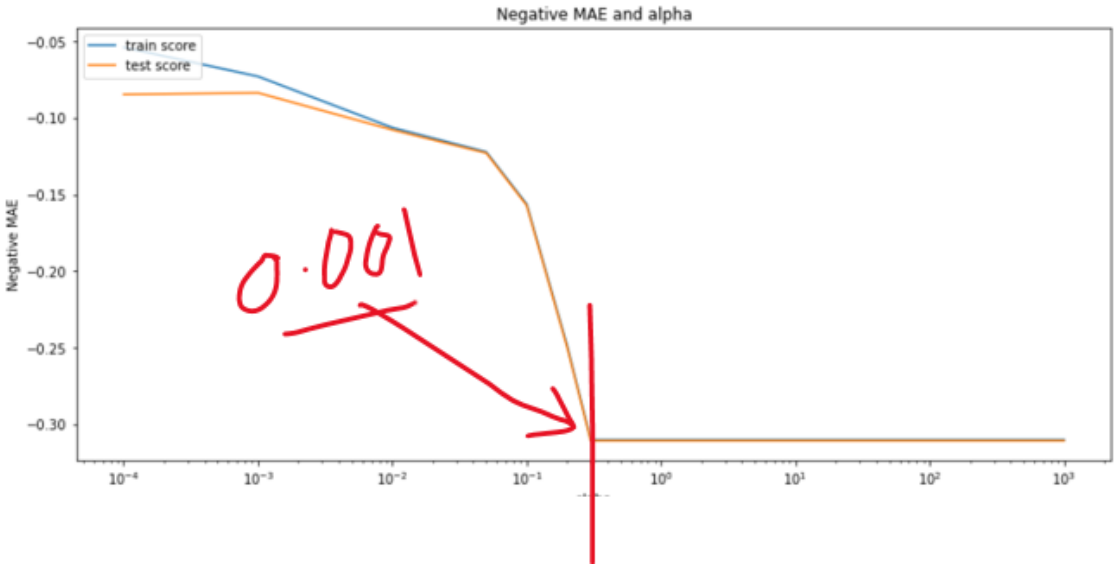
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:
The determined optimal value of lambda for ridge and lasso 10 and 0.001 respectively.

Ridge :



LASSO:



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:

Top 5 Lasso predictors were:

- OverallQual_9
- GrLivArea
- OverallQual_8
- Neighborhood_Crawfor
- CentralAir_Y

After dropping the above 5, another lasso model was created, and the following 5 new predictor variables were identified.

- 2ndFlrSF
- 1stFlrSF
- Neighborhood_Somerst
- MSSubClass_70
- TotalBsmtSF

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 considered as robust and generalised based on its performance with any variation in the data and its ability to properly adapt new and unseen data.

In order to ensure that a model is robust and generalised, we need to perform the following:

1. Outliers treatment for both independent variables and target variables
2. Regularization to avoid overfitting

These are the following implications of the same for the accuracy of the model:

1. An overfit model (complex model) will have high variance and a smallest of change in the data affects the accuracy of the model. An overfit model will have high accuracy on train data but will fail with unseen test data. So a balance between the model complexity and accuracy needs to be maintained. It can be achieved by Ridge/Lasso regression (regularization technique).
2. Outliers are data points which are distant from most of the other data points. As the models are sensitive to the distribution of the data points, outliers can result into less accurate models. We use log transform, capping between percentiles or using standard deviation to treat the outliers.