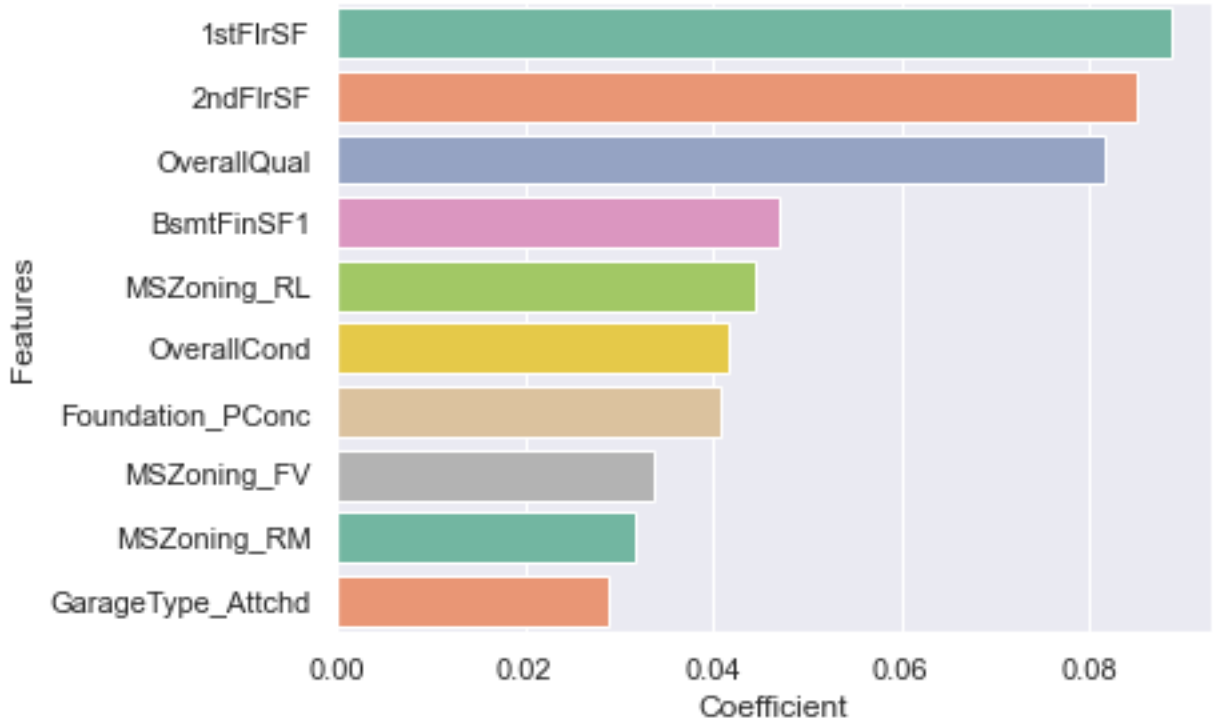
Question 1

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 1

As per the model the optimal value of the alpha that came out was :

#Ridge – 20(original)

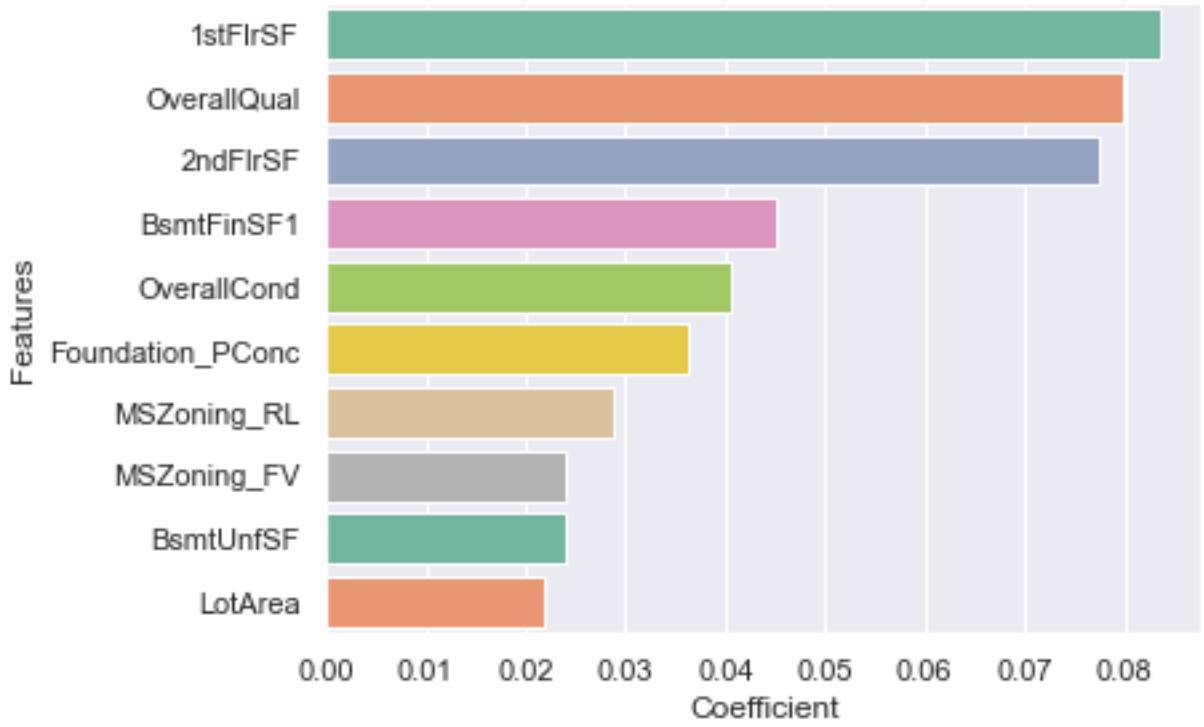


With the new Aplha values :

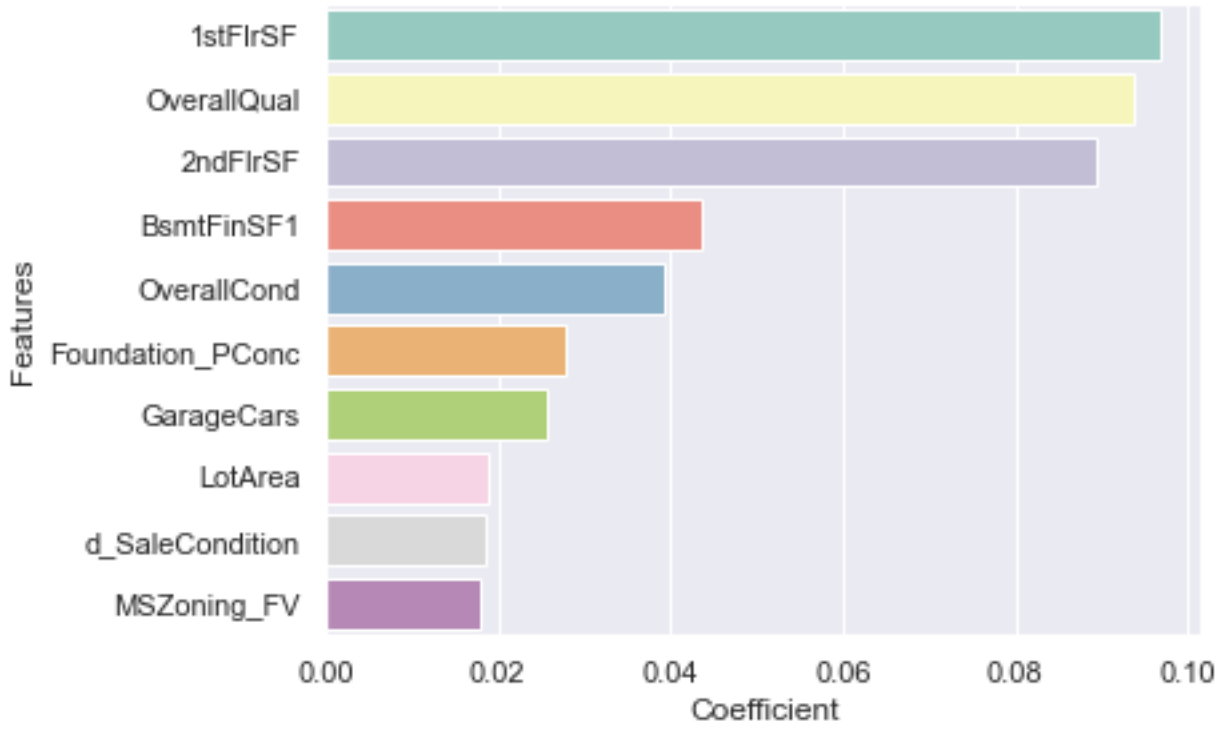
#Ridge - 40

2 new variables came into the picture with almost lowest coefficients : Lot area and bsmntUnsf , replaced Garage type attached and MS: ZoningRM ,

Among them some coefficients became more strong than others which were strong earlier:



#Lasso - 0.0003(original)

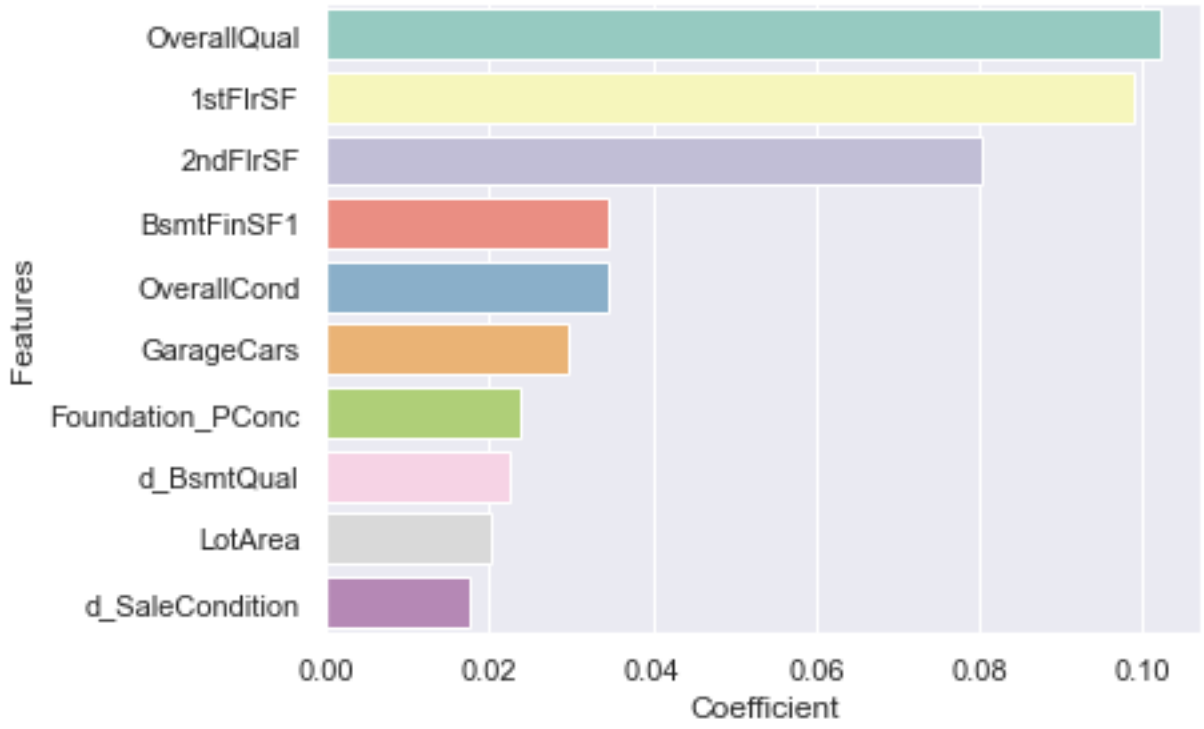


#Lasso - 0.0006

In [432]:

Here d_bsmntqual is added instead of MSZoning_FV , the top variable has also changed and there are more changes among each order.

Generally as the lambda/ alpha increases the coeffecient value decreases and the variance deacreases and bias increases .



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?

#The Mean Squared Error of Lasso is slightly lower than that of Ridge

#Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge.

Since we performed feature selection and Lasso works better on feature selection and also has lower mean sq error value , we would pick Lasso.

In Ridge reg we take lambda as penalty of the square of magnitude of coeff while in lasso we take the magnitude only

#Ridge - 0.016724219916660496

#Lasso - 0.0159117674071282

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?

Excluded variables :

1stFrSf

Overallqual

2ndFlrSf

BsmntFinSF1

OverallCond

Next variables :

Foundation_Pconc

Garagecars

Lotarea

D_Salecondition

MsZoning_FV

Question 4

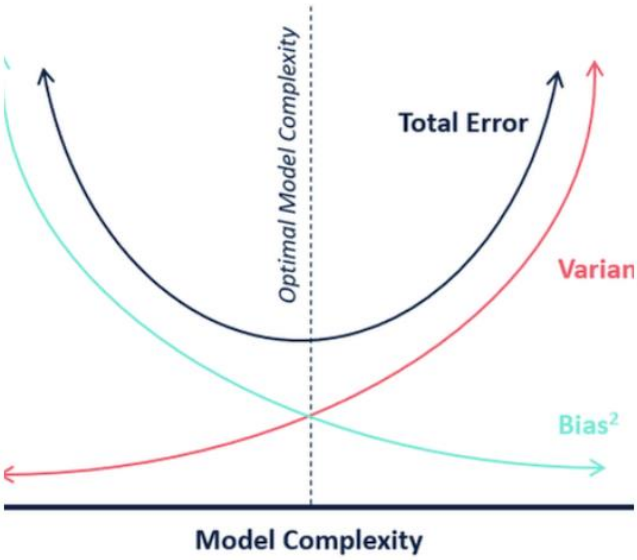
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

A model needs to be made robust and generalizable so that there is a balance between the variance and bias of the model.

The model should not be over fitting as it could lead to poor performance on test data since overfitting causes memorization of training data and also the model should not be far too general as the variance would be low but the bias would be high

The test accuracy should not be less than the training score. The model should be built with optimal lamda value and we should make sure the data is properly analyzed and clean so that there are no outliers and also missing values are taken care of. If a model is too general then accuracy would be low and if the model overfits the training data still there are chances on unseen data it would be perform poorly.

The below graph explains the entire theory .


Optimal Model Complexity
Total Error
Varian
Bias²
Model Complexity