Question 1

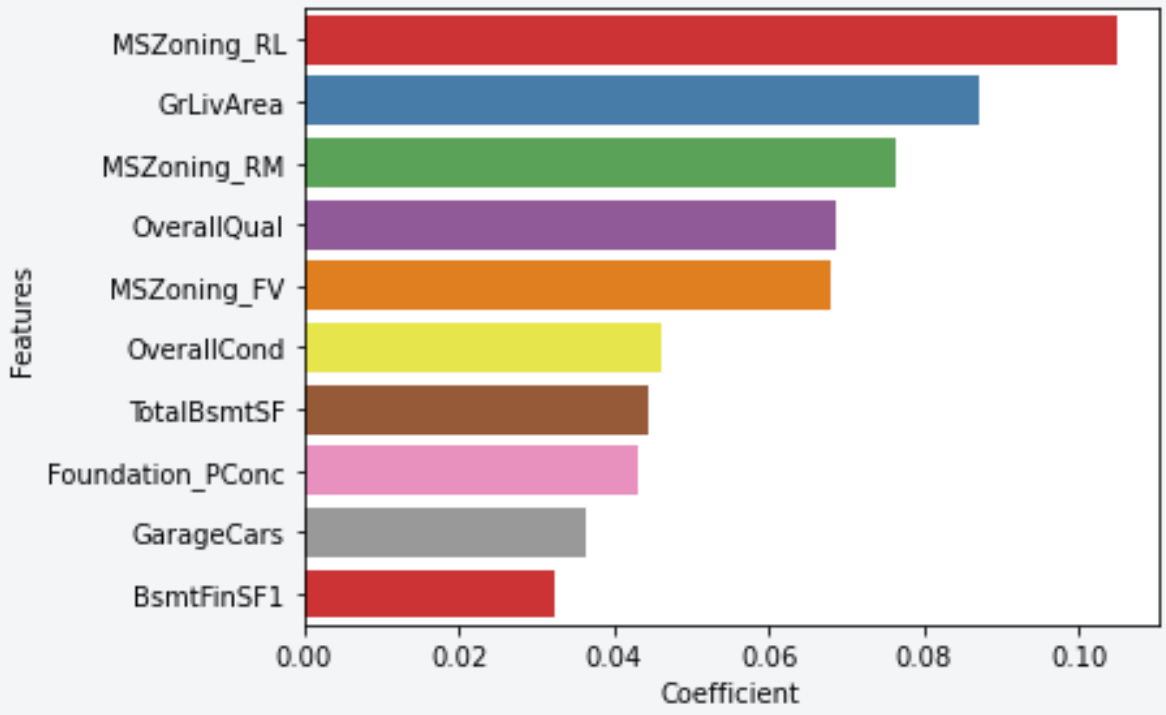
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer : The optimal value of alpha for Ridge and Lasso Regression are 1 and 0.0001.

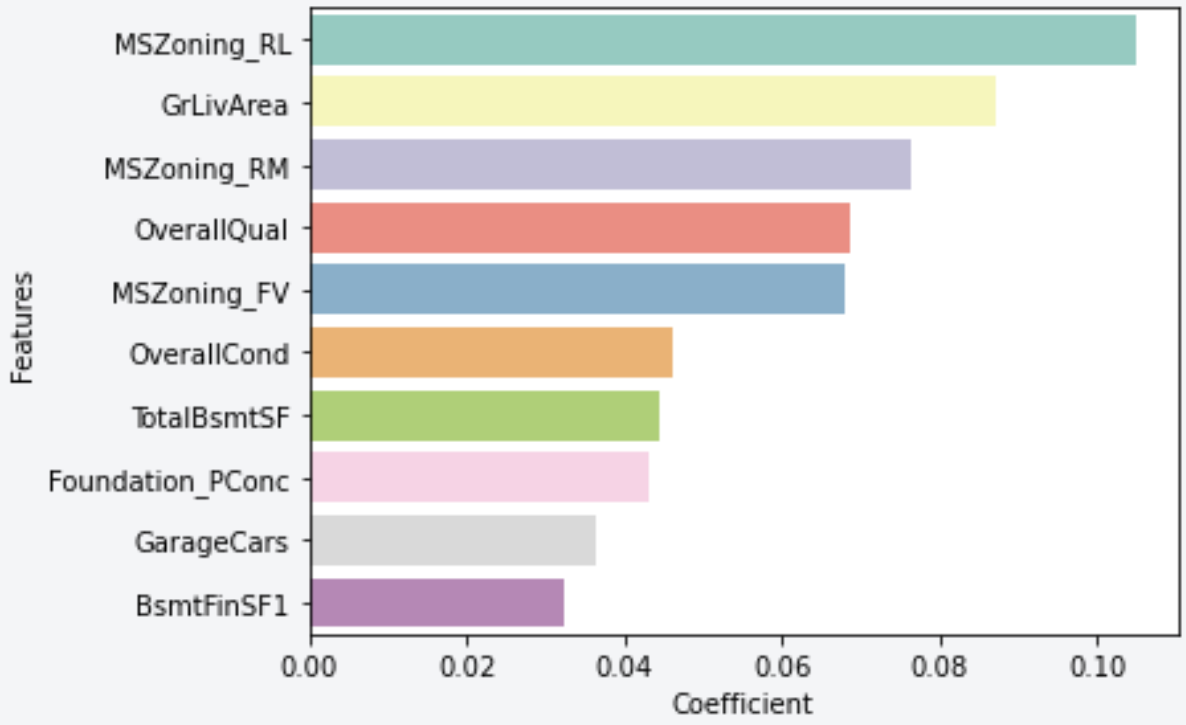
In both the Regression model the effects of doubling the value of alpha is that the value of coefficients has moved towards zero for both the coefficients i.e. positive and negative. Some predictor variables have also changed the order of importance . But the first 10 variables have not shown much changes in both the regression models (Ridge and Lasso).

For RIDGE Regression:

When Alpha = 5.0



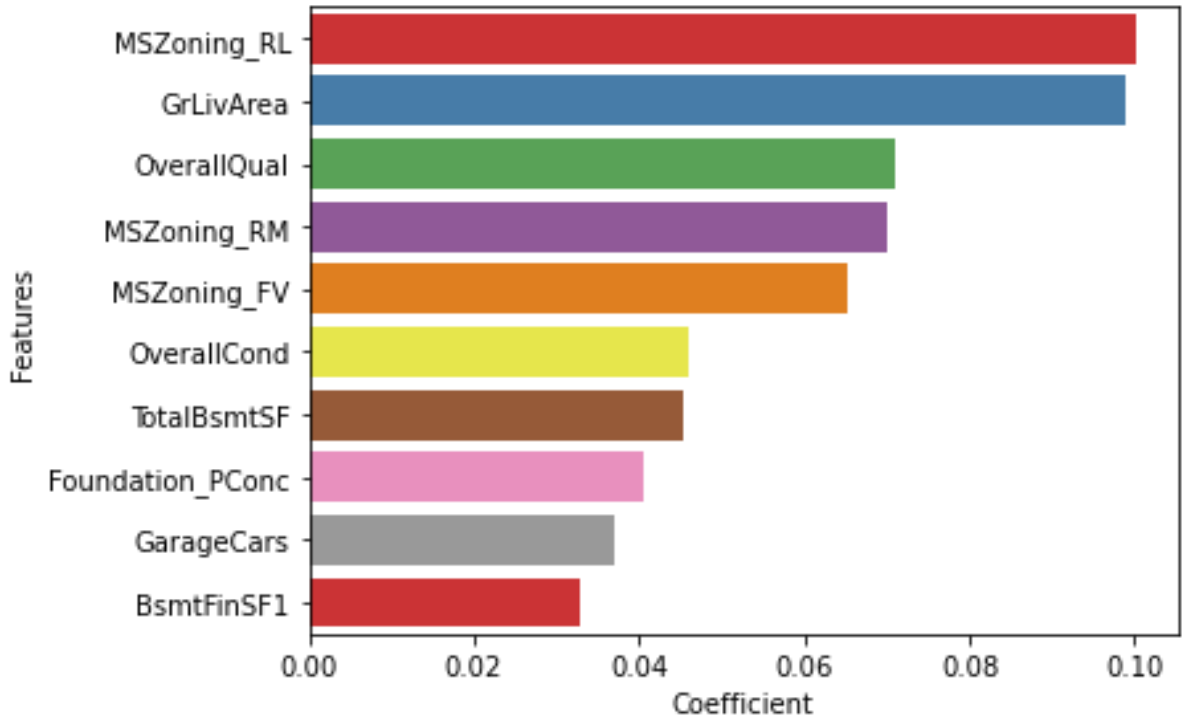
When Alpha = 10.0



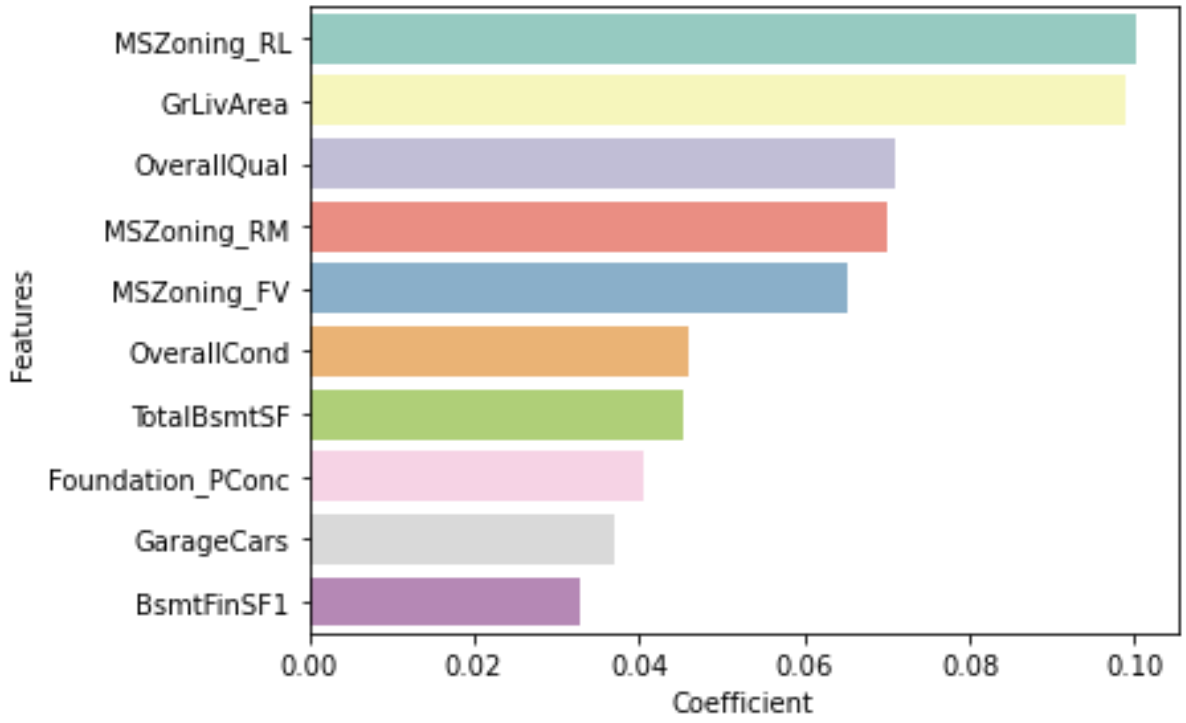
In Both the graphs it is clearly visible the order is not changed

For LASSO Regression :

When Alpha = 0.0005



When Alpha = 0.001



There is not much change in the value of coefficients .

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 optimal value of Ridge and Lasso Regression calculated as per the model are:

```
Ridge :          Alpha = 5.0
                 Mean Squared Error = 0.013714
                 Training Accuracy = 91.97%
                 Test Accuracy = 91.37%

Lasso :          Alpha = 0.0005
                 Mean Squared Error = 0.013464
                 Training Accuracy = 91.67%
                 Test Accuracy = 90.4%
```

We can clearly observe from the above values that Mean Squared Error of Lasso is less than the Ridge and Lasso also helps in feature reduction i.e. the coefficient value of one or more the feature can become zero thus , Lasso has edge over the Ridge Regression . So we clearly choose Lasso over Ridge.

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 : After excluding the top five most important predictor variables namely

'MSZoning_RL', 'GrLivArea', 'OverallQual', 'MSZoning_RM' , 'MSZoning_FV' we get alpha = 0.001 . Thus , the top five predictor models after dropping these variables with their respective coefficient values are :

Features	Coefficient
MSZoning_RL	0.100283
GrLivArea	0.098793
OverallQual	0.070912
MSZoning_RM	0.070136
MSZoning_FV	0.065277

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?

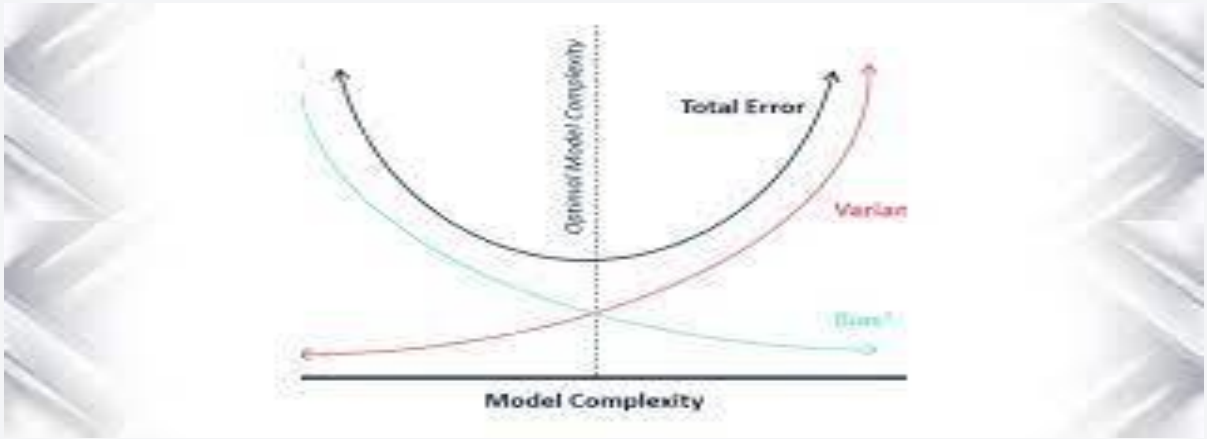
Answer : A model is said to generalisable model if it does not overfits the training data that means the training data should not be completely memorised . So when the model is tested on an unseen data it should either give the test accuracy or some possible acceptable errors.

The model is considered to be Robust if the results remains the same the same or somewhat same even after some of the values are changed.

Simpler models tends to Robust and Generalisable . Hence we should make the model simpler but not simpler as it has its own cons.

A model can be made Generalisable by striking a balance between Overfitting , Underfitting and Accuracy . Regularization is one such way to solve this problem as it punishes those coefficients which are very large in comparison with other values .

The model should have enough complexity that can solve the underlying patterns and not not be too complex . Hence we compromise on the complexity of the model to increase the accuracy of test data .



Here , The optimal value is when the model has enogh bias to generalised and enough variance to give the least errors.