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:

I have got the below optimal values for alpha

Ridge - 0.1
Lasso - 0.0001

Metric	Ridge regression	Lasso regression	
0	R2 Score Train	0.897530	0.895662
1	R2Score Test	0.851409	0.863824
2	RSS Train	10.201532	10.387456
3	RSS Test	6.048088	5.542747
4	MSE Train	0.012816	0.013050
5	MSE Test	0.017684	0.016207

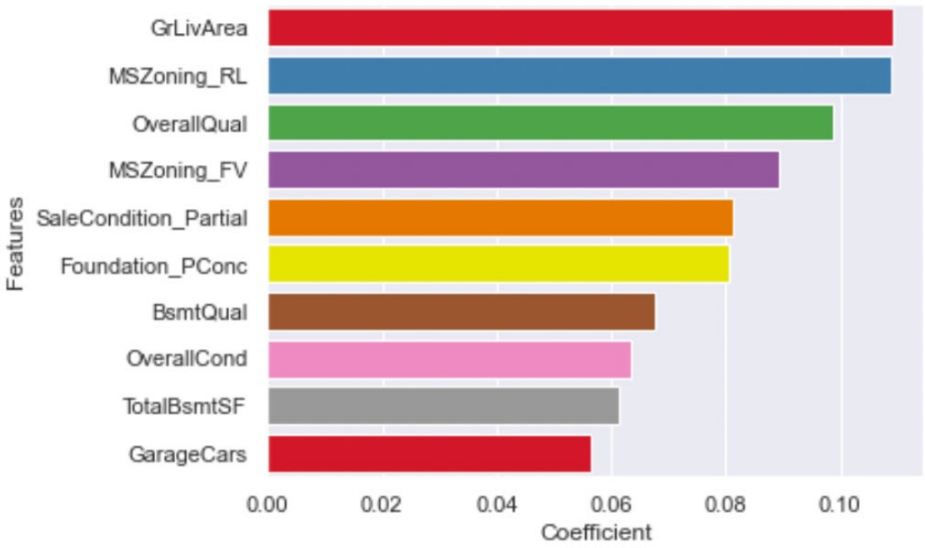
When double the value of alpha

Ridge - 0.2
Lasso - 0.0002

Metric	Ridge regression	Lasso regression	
0	R2 Score Train	0.897151	0.895662
1	R2Score Test	0.853970	0.863824
2	RSS Train	10.239215	10.387456
3	RSS Test	5.943839	5.542747
4	MSE Train	0.012863	0.013050
5	MSE Test	0.017380	0.016207

I got very similar values for the first time for both Lasso and Ridge.
Very slight difference in the R2 score for Ridge regression when we double the value of alpha.
A noticeable difference is there in RSS test in Ridge Regression when we double the value of alpha.

The important variables after this change is implemented are:



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:

I got near identical values for both Ridge and Lasso for the alpha values 0.1 and 0.0001.
As per the alpha/Lambda values I have received, Ridge regression does not zero any of the coefficients.
Lasso zeroed one or two coefficients in the selected features, Lasso is a better option and it also helps in some of the feature elimination. So I would choose 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:
I have excluded the following five most important variable I have got prior.
- MSZoning_FV
- MSZoning_RL
- MSZoning_RM
- MSZoning_RH
- Foundation_Stone

I have created a new model after removing these columns code mentioned in the Python notebook. After the Lass Regression, I have got the other important predictors which are

- MSSubClass
- BldgType_TwnhsE

- Condition1_PosA
- Condition1_Feedr
- EnclosedPorch

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 the model is robust and generalisable, the model must be very simple although its accuracy will reduce. Alternatively, using the Bias-Variance trade-offs this can be well understood. The simpler the model, more the bias, lesser the variance and more generalisable it gets. In terms of accuracy, its implication is that a robust and generalisable model will perform equally well on both training as well as test data, that is, the accuracy does not deviate much for training and test data.

Bias: Error that occurs in model when a model is too weak to learn from the data, we call it a Bias. Higher Bias implies the model is unable to learn the details present in the data. The model performs inadequately on training and testing data.
Variance: A variance is a form of error in the model when a model tries to overlearn from the data. A high variance means that the model performs exceptionally well on the training data as it is very well trained on this data but performs ver badly upon testing data as it would be an unseen data for the model.
It is important to strike a right balance between Bias and Variance to avoid overfitting and under-fitting of data.