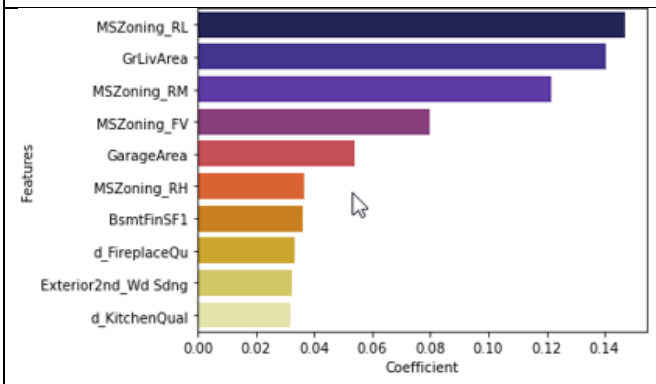
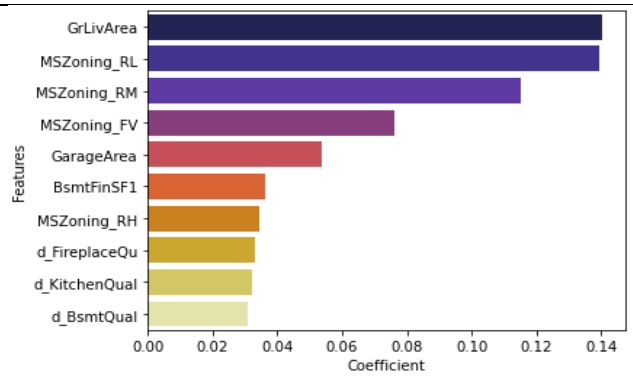
Assignment Part II – Surprise Housing society

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 is **0.0001**. When this alpha is doubled to 0.0002, the following observations were found –

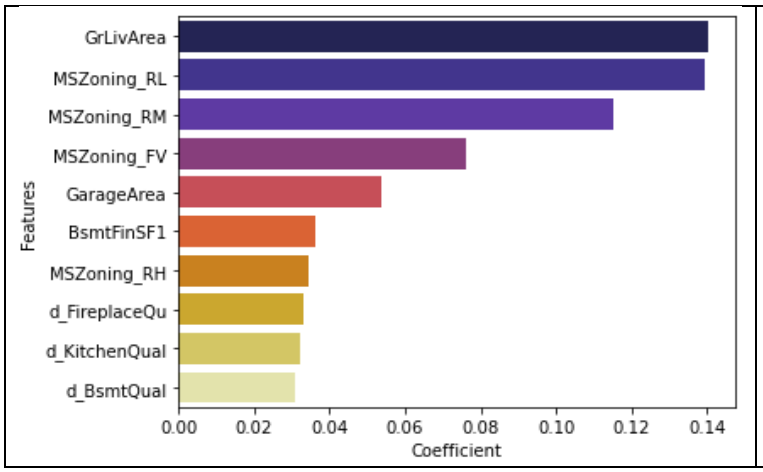
Ridge Regression:

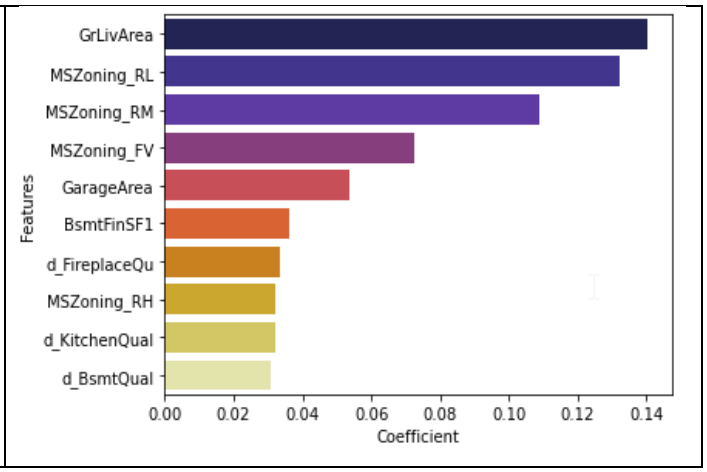
Alpha = 0.0001	**Alpha = 0.0002**
Train R2 score = 0.8932155054326304	Train R2 score = 0.8932155054320406
Test R2 score = 0.8880791858627183	Test R2 score = 0.8880791595330271
Mean Squared error = 0.016644575120509254	Mean Squared error = 0.01664457903619309
	

As observed above, there are **NO** major changes in R2 score or the MSE. With the predictor variables, the variables list is the same but the coefficients of the variables like GrLivArea , MSZoning and KitchenQual has minor changes.

Lasso Regression:

Alpha = 0.0001	**Alpha = 0.0002**
Train R2 score = 0.8931792156134563	Train R2 score = 0.8930715507916066
Test R2 score = 0.8882995159913158	Test R2 score = 0.8884607666117653
Mean Squared error = 0.01661180819145302	Mean Squared error = 0.01658782741463336 2





As observed above, there are **minuscule** changes in R2 score or the MSE. With the predictor variables, the variables list is the completely same and the coefficients of predictor variables remain intact.

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 lambda, Mean Squared error, and R2 score are the same in both Ridge and Lasso Regression. Ridge regression has 46 variables that have non-zero coefficients, but Lasso regression has 43 variables that have non-zero coefficients. Since Lasso helps in feature reduction (as the coefficient value of more of the features became 0), Lasso has a better edge over Ridge. Hence we will choose the Lasso regression.

Question 3

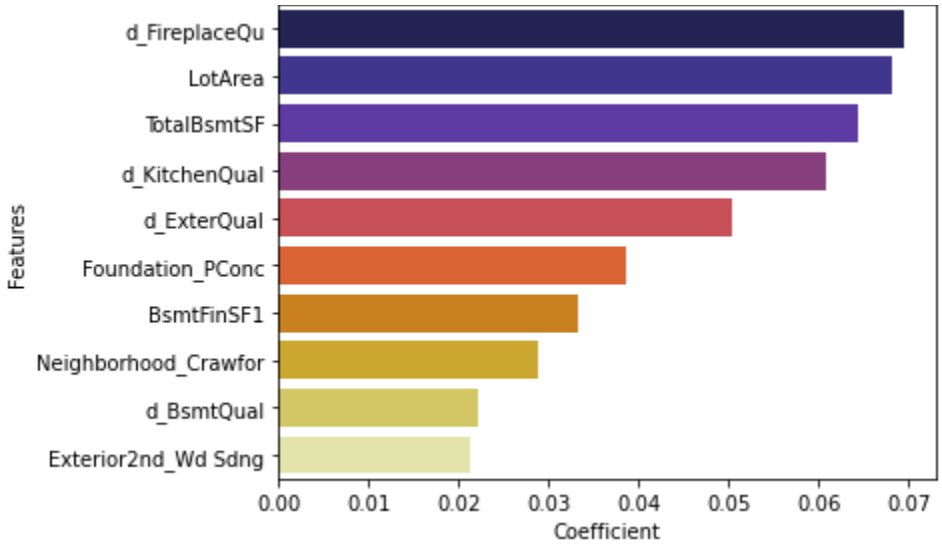
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer:

The top 5 features of Lasso Regression are - 'GrLivArea', 'MSZoning_RL', 'MSZoning_RM', 'MSZoning_FV', 'GarageArea'. After dropping them the Model Train R2 score dropped from

89.3% to 80.4% and Text R2 score from 88.8% to 78.5%. The Mean Square error increased from 0.016 to 0.032.

The next top 5 predictor variables are : d_FireplaceQu, LotArea, TotalBsmtSF, d_KitchenQual and d_ExterQual. Their coefficient values are shown below –



Question 4

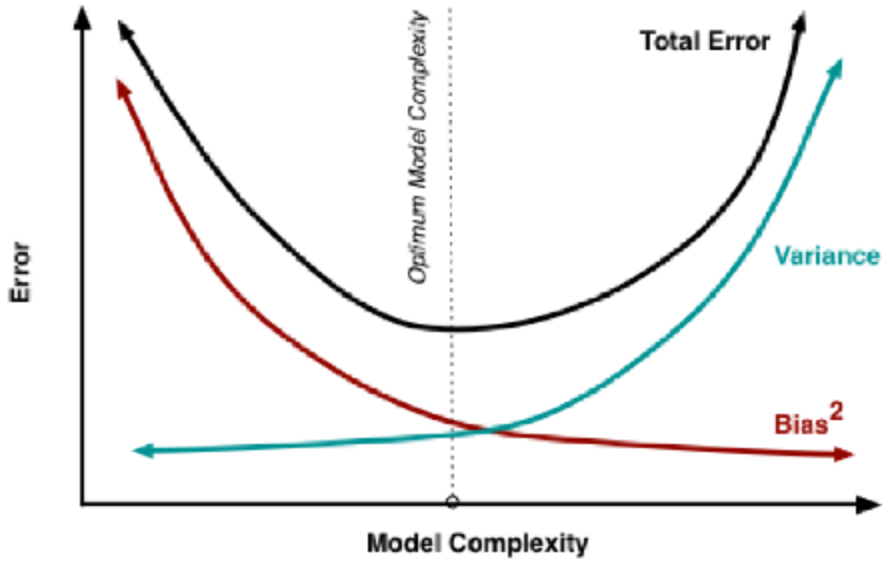
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

A model can be made resistant and generalizable by making a model that is –

1. Simple. Simpler models are usually more generic and more widely applicable
2. Resistant to outliers. It is always better to remove outliers in the data before training the model.
3. More robust error metrics like mean square error and mean absolute error being very low towards zero.
4. Having R2 score in the range of 75 to 80%.

Making a model simple leader to Bias-Variance Trade-off. Bias quantifies how accurate is the model on the test data. Variance refers to the degree of changes in the model itself with respect to changes in the training data. The accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error as shown in the below graph –



Outliers have significant influences on the methods for estimating the accuracy of models, and these influences will become more extreme when sample size and variance are small. Hence, removing outliers increases the accuracy of the model.