Q1: 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?

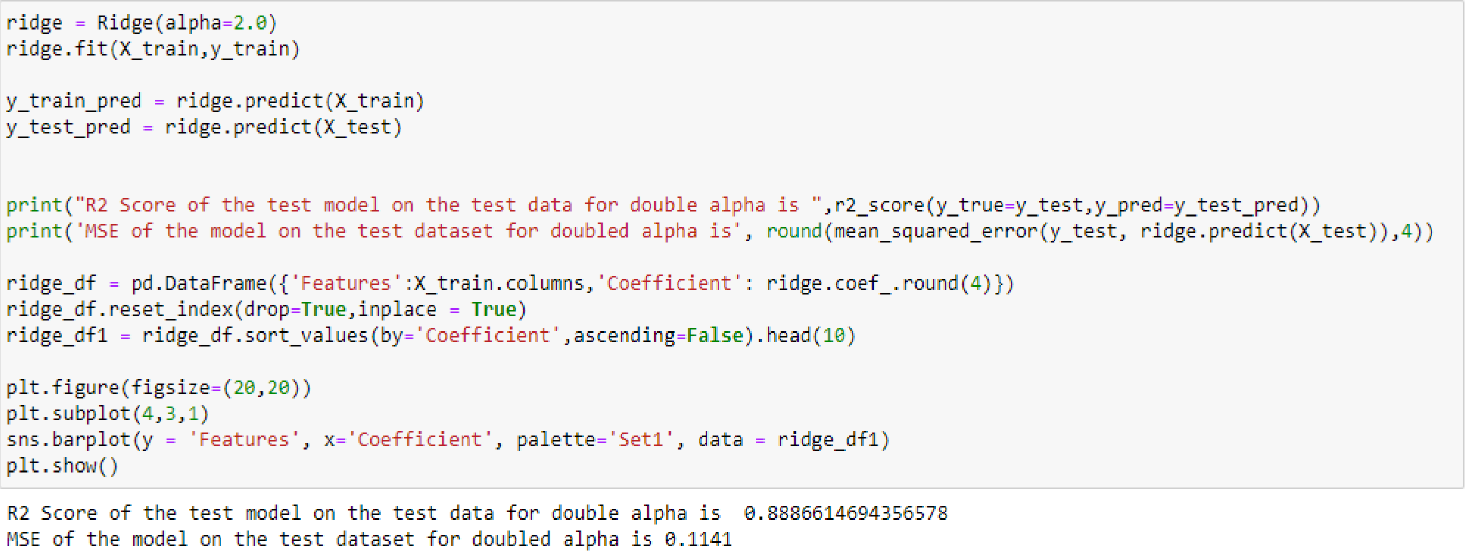
A: The optimal value of alpha for ridge and lasso is as follows:

- Ridge: 1.0
- Lasso: 0.001

If we make the changes in the model by doubling the value of alpha for Ridge and lasso

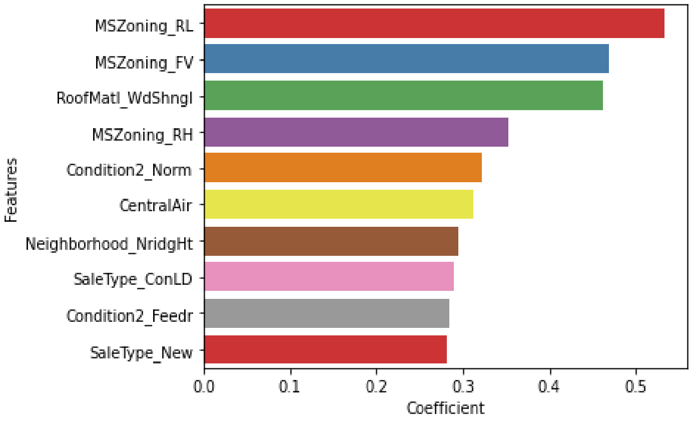
- Observed slight reduction in the Coefficient value when there is an increase in Optimal value
- There is a slight increase in the RSS if we increase the optimal value

We can observe the important predictor variable of Ridge in below Fig. if we double the value of Ridge to **2**

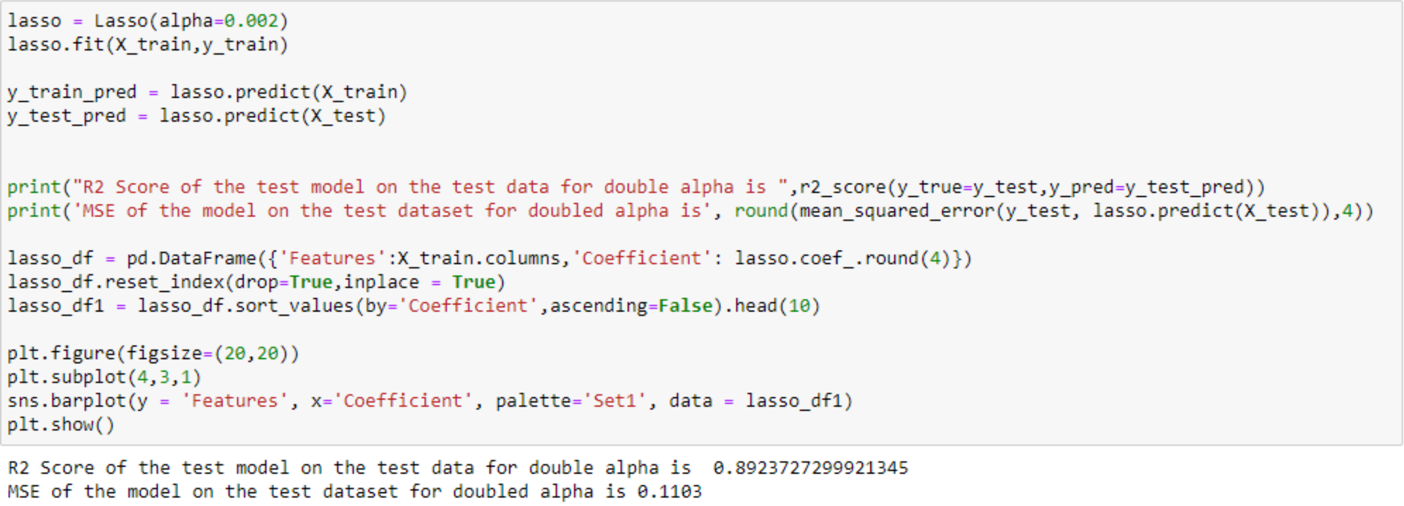


```
ridge = Ridge(alpha=2.0)
ridge.fit(X_train,y_train)

y_train_pred = ridge.predict(X_train)
y_test_pred = ridge.predict(X_test)


print("R2 Score of the test model on the test data for double alpha is ",r2_score(y_true=y_test,y_pred=y_test_pred))
print('MSE of the model on the test dataset for doubled alpha is', round(mean_squared_error(y_test, ridge.predict(X_test)),4))

ridge_df = pd.DataFrame({'Features':X_train.columns,'Coefficient': ridge.coef_.round(4)})
ridge_df.reset_index(drop=True,inplace = True)
ridge_df1 = ridge_df.sort_values(by='Coefficient',ascending=False).head(10)

plt.figure(figsize=(20,20))
plt.subplot(4,3,1)
sns.barplot(y = 'Features', x='Coefficient', palette='Set1', data = ridge_df1)
plt.show()
```

```
R2 Score of the test model on the test data for double alpha is  0.8886614694356578
MSE of the model on the test dataset for doubled alpha is 0.1141
```

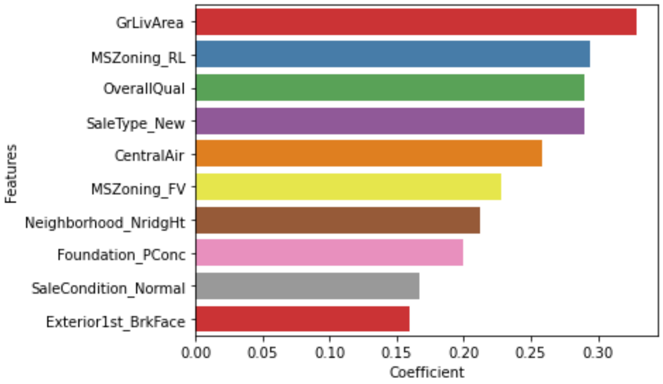


We can observe the important predictor variable of Lasso in below Fig. if we double the value of Lasso to **0.002**



```
lasso = Lasso(alpha=0.002)
lasso.fit(X_train,y_train)

y_train_pred = lasso.predict(X_train)
y_test_pred = lasso.predict(X_test)


print("R2 Score of the test model on the test data for double alpha is ",r2_score(y_true=y_test,y_pred=y_test_pred))
print('MSE of the model on the test dataset for doubled alpha is', round(mean_squared_error(y_test, lasso.predict(X_test)),4))

lasso_df = pd.DataFrame({'Features':X_train.columns,'Coefficient': lasso.coef_.round(4)})
lasso_df.reset_index(drop=True,inplace = True)
lasso_df1 = lasso_df.sort_values(by='Coefficient',ascending=False).head(10)

plt.figure(figsize=(20,20))
plt.subplot(4,3,1)
sns.barplot(y = 'Features', x='Coefficient', palette='Set1', data = lasso_df1)
plt.show()
```

```
R2 Score of the test model on the test data for double alpha is  0.8923727299921345
MSE of the model on the test dataset for doubled alpha is 0.1103
```



Q2: 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?

A: Post analysis, I found the optimal value of lambda for Lasso and Ridge is **1.0** & **0.001**. However the mean squared value of Ridge and Lasso is **0.1149** & **0.1110**

I would be able to choose Lasso model over Ridge model as it helps in feature eliminationand the variables predicted by Lasso are significant variables for predicting the price of a house

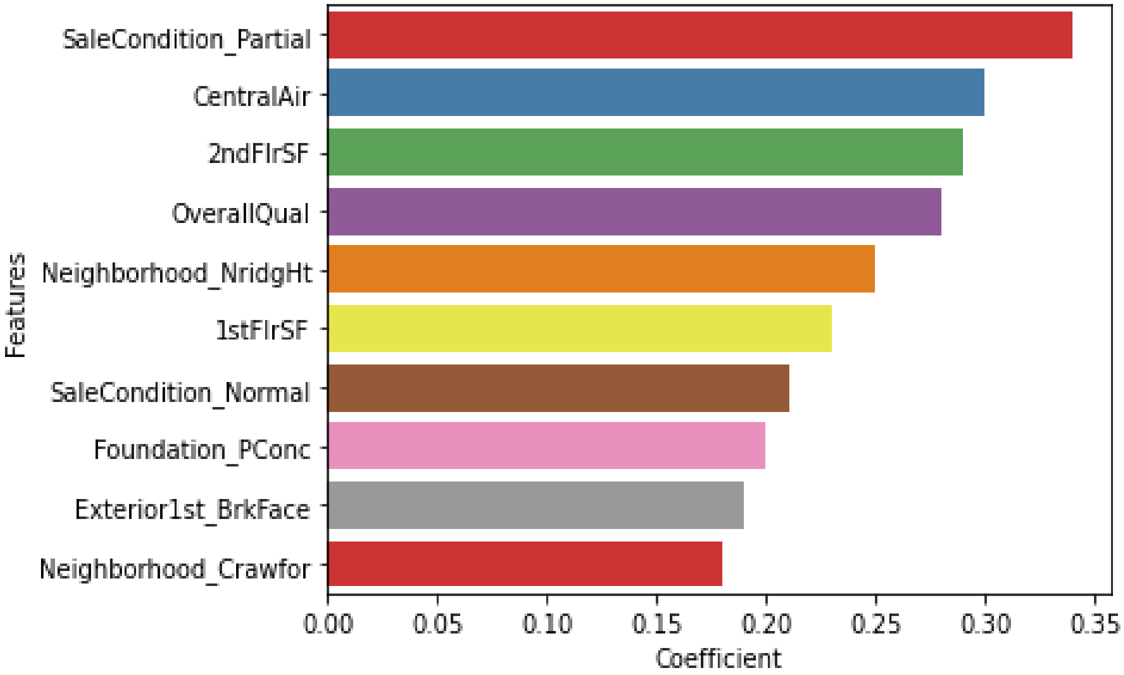
Q3: After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

A: Please find below the top 5 important variable in the Lasso Model

- RoofMatl_WdShngl
- MSZoning_RL
- MSZoning_FV
- SaleType_New
- GrLivArea

After excluding the above variables from the model, I detected the top 5 most important variables as follows:

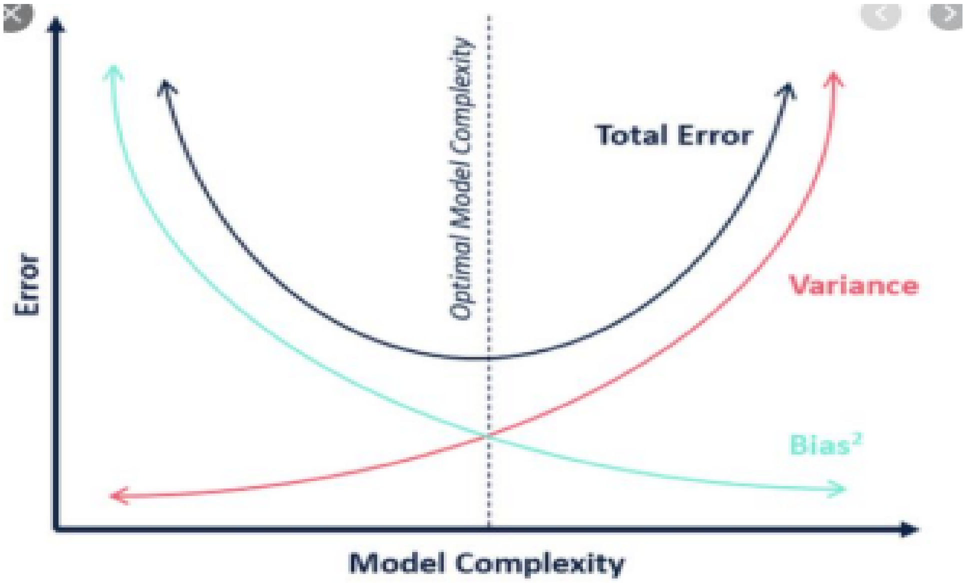
- SaleCondition_Partial
- CentralAir
- 2ndFlrSF
- OverallQual
- Neighborhood_NridgHt



Q4: How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A:

- The model needs to be robust and generalizable as the data are not impacted by outliers and don't affect its performance.
- In addition,the model must generalizable in order for test accuracy to match training score
- The model should not be complex as it tends to change wildly with changes in the training data set and is not appropriate in order to be robust and generalizable
- Complex models led to very high accuracy. In order to have a robust and generalizable model, we have to decrease the variance which will lead to some bias
- The outliers shouldn't be given an excessive amount of weight in order to maintain the high level of model accuracy. Only those outliers that are pertinent to the dataset should be kept once the outlier analysis is completed to ensure that this is not the case. The dataset must be cleaned up of any outliers that don't make sense to preserve. This would improve the model's ability to forecast outcomes accurately.
- Regularization helps to strike the delicate balance between accuracy and complexity
- Ridge Regression and Lasso are the techniques used for Regularization
- Sometimes making the model simple leads to BIAS Variance Trade Off
- Variance refers to the degree of changes in the model itself with respect to changes in the training data.Thus 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


Optimal Model Complexity
Total Error
Error
Variance
Bias²
Model Complexity