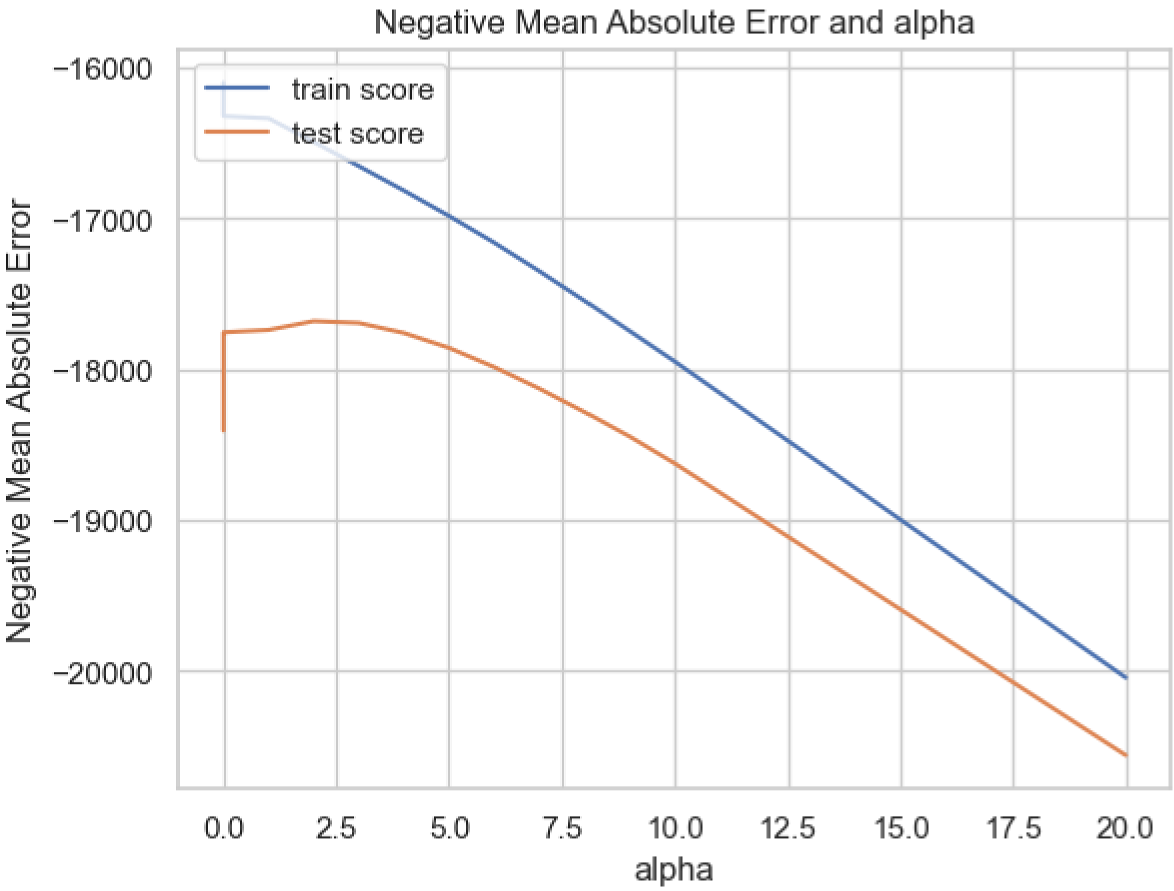
Assignment-Based Subjective Questions:

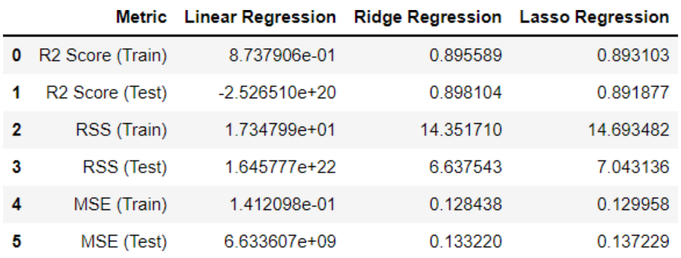
Question 1

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

Ans: In the case of ridge regression, when we plot the curve between negative mean absolute error and alpha, we observe a decreasing trend in the error term as the value of alpha increases from 0. However, the training error shows an increasing trend with the alpha value. Upon reaching an alpha value of 2, the test error is minimized, leading us to choose an alpha value of 2 for our ridge regression.



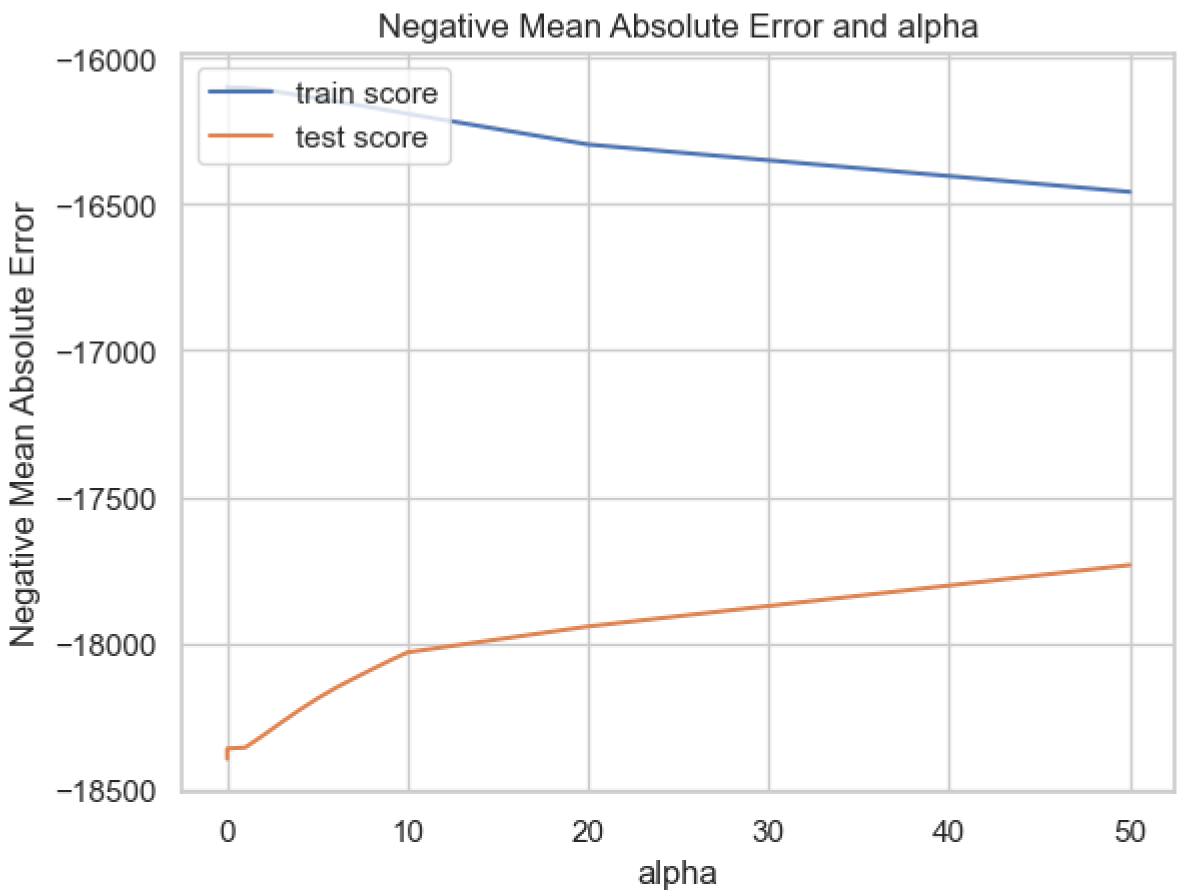
After doubling the optimal alpha values for Ridge and Lasso, Ridge_alpha = 100, Lasso_alpha



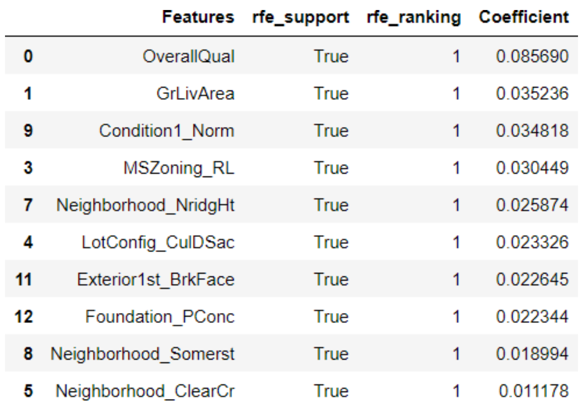
	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.737906e-01	0.895589	0.893103
1	R2 Score (Test)	-2.526510e+20	0.898104	0.891877
2	RSS (Train)	1.734799e+01	14.351710	14.693482
3	RSS (Test)	1.645777e+22	6.637543	7.043136
4	MSE (Train)	1.412098e-01	0.128438	0.129958
5	MSE (Test)	6.633607e+09	0.133220	0.137229

=

0.002

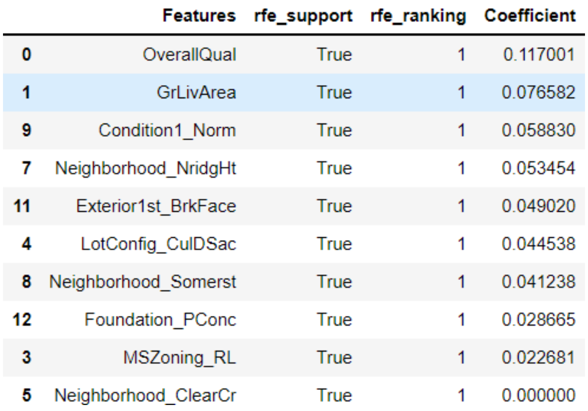


For ridge regression, the important predictors after doubling the value of alpha

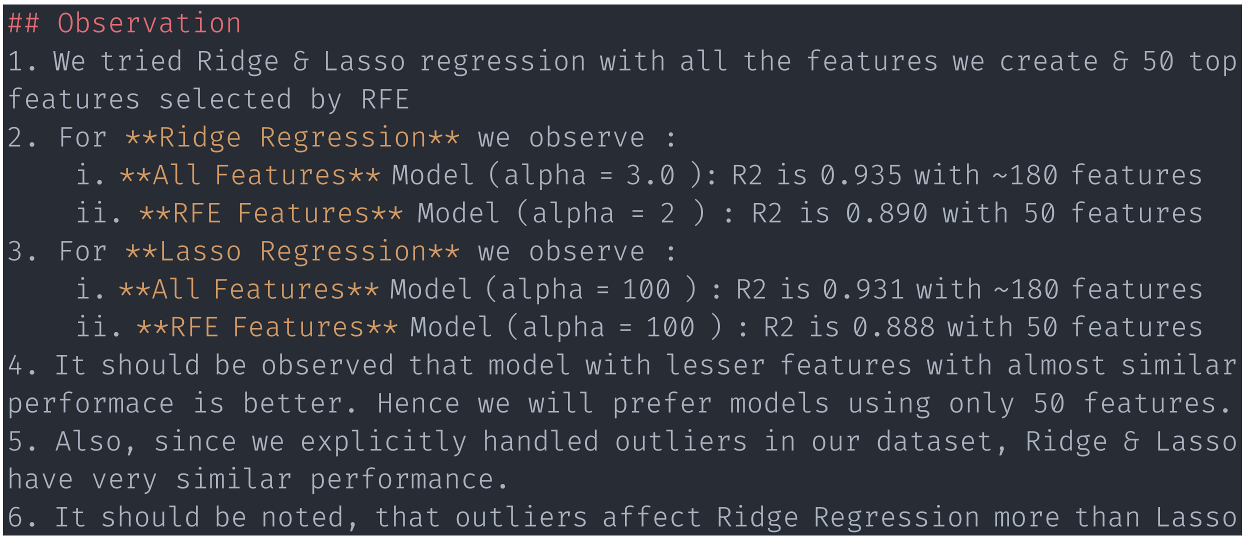


	Features	rfe_support	rfe_ranking	Coefficient
0	OverallQual	True	1	0.085690
1	GrLivArea	True	1	0.035236
9	Condition1_Norm	True	1	0.034818
3	MSZoning_RL	True	1	0.030449
7	Neighborhood_NridgHt	True	1	0.025874
4	LotConfig_CulDSac	True	1	0.023326
11	Exterior1st_BrkFace	True	1	0.022645
12	Foundation_PConc	True	1	0.022344
8	Neighborhood_Somerst	True	1	0.018994
5	Neighborhood_ClearCr	True	1	0.011178

For Lasso regression, the important predictors after doubling the value of alpha



	Features	rfe_support	rfe_ranking	Coefficient
0	OverallQual	True	1	0.117001
1	GrLivArea	True	1	0.076582
9	Condition1_Norm	True	1	0.058830
7	Neighborhood_NridgHt	True	1	0.053454
11	Exterior1st_BrkFace	True	1	0.049020
4	LotConfig_CulDSac	True	1	0.044538
8	Neighborhood_Somerst	True	1	0.041238
12	Foundation_PConc	True	1	0.028665
3	MSZoning_RL	True	1	0.022681
5	Neighborhood_ClearCr	True	1	0.000000



```
## Observation
1. We tried Ridge & Lasso regression with all the features we create & 50 top
features selected by RFE
2. For **Ridge Regression** we observe :
    i. **All Features** Model (alpha = 3.0 ): R2 is 0.935 with ~180 features
    ii. **RFE Features** Model (alpha = 2 ) : R2 is 0.890 with 50 features
3. For **Lasso Regression** we observe :
    i. **All Features** Model (alpha = 100 ) : R2 is 0.931 with ~180 features
    ii. **RFE Features** Model (alpha = 100 ) : R2 is 0.888 with 50 features
4. It should be observed that model with lesser features with almost similar
performace is better. Hence we will prefer models using only 50 features.
5. Also, since we explicitly handled outliers in our dataset, Ridge & Lasso
have very similar performance.
6. It should be noted, that outliers affect Ridge Regression more than Lasso
```

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?

Ans: Opting for regularization is crucial to enhance prediction accuracy while minimizing variance and maintaining model interpretability. Ridge regression, employing a tuning parameter (lambda) to penalize the square of coefficient magnitudes, is chosen through cross-validation.

This penalization helps keep the residual sum of squares small, particularly for coefficients with larger values. Increasing lambda reduces model variance while keeping bias constant. Unlike Lasso Regression, Ridge includes all variables in the final model.

Lasso regression, with its lambda-driven penalty on the absolute value of coefficients, is another option. As lambda increases, Lasso tends to shrink coefficients towards zero, eventually setting some variables exactly to zero. This introduces variable selection. Lasso's performance transitions from simple linear regression for small lambdas to variable neglect for larger lambdas.

Though Ridge Regressor is performing little better compared to Lasso regressor, I prefer to choose Lasso as it helps in feature elimination.

```
# Best model
lasso = Lasso(alpha=100)
lasso.fit(X_train_rfe, y_train)
```

Question 3

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?

Ans: The Variables that positively/directly affect the price of a house are : ['BsmtFinSF1', 'KitchenAbvGr', 'OverallQual', '1stFlrSF', 'PoolArea']

The Variables that negatively/indirectly affect the price of a house are : ['LotArea', 'Condition1_PosA', 'RoofStyle_Hip', 'Exterior1st_WdShing', 'Exterior2nd_ImStucc']

```
1. The top 5 features which affect the prices are:
    i. BedroomAbvGr
    ii. OverallCond
    iii.2ndFlrSF
    iv. LotArea
    v. SaleType_Con


2. Properties of the house which positively affect the prices are:
    i. BedroomAbvGr
    ii. OverallCond
    iii. 2ndFlrSF
    iv. 2ndFlrSF
    v. BsmtUnfSF


3. Properties of the house which negatively affect the prices are:
    i. LotArea
    ii. MSSubClass
    iii. SaleCondition_Partial
    iv. Exterior2nd_CmentBd
    v. Neighborhood_NridgHt
```

Question 4

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?

Ans: Model simplicity is key for robustness and generalizability, even at the cost of reduced accuracy. This aligns with the Bias-Variance trade-off, where a simpler model introduces more bias but less variance, enhancing generalizability. A balanced approach between bias and variance is essential to prevent overfitting or underfitting. A robust model performs consistently well on both training and test data, ensuring minimal accuracy discrepancies between the two.

Bias refers to model errors when it's weak to learn from data, resulting in poor performance on both training and testing. Variance, on the other hand, arises when the model overlearns, excelling on training but failing on unseen testing data.

Criterias:

1. Detecting and treating the outliers in the predictors
2. Imputing the missing values appropriately after outlier treatment with the measures of central tendency based on the nature of predictor (numerical vs categorical)
3. Understand the feature variables in conjunction with the domain before dropping the features. Don’t drop the predictors just because the data in the predictors are skewed (We need to bother only if the target variable is skewed as it might give inclined results). Dropping the predictors unnecessarily will reduce the predictive power of the model.
4. Scaling the predictors and target variable using scaling techniques
5. Perform Cross Validation using K-fold to detect if there is a model

overfit Implications:

1. The model accuracy reduces when we the outliers are not treated
2. The model accuracy reduces when missing values are not properly imputed
3. The model accuracy reduces when we drop the features which has predictive power
4. The model accuracy reduces when the predictors and target variable are not in the same scale. The model coefficients will be very large for few and very small for others which may lead to certain model coefficients (which is very small) to be insignificant
5. Apply Regularization techniques if overfit is detected so that model accuracy improves