Advanced Linear Regression Assignment

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?

Ans:

The optimum value of alpha is

1. ridge regression - 0.2

```
In [3676]:
1 # getting the best estimator for lambda
2 ridge_model_cv.best_estimator_

Out[3676]: Ridge(alpha=0.2)
```

2. lasso regression – 50

```
In [3683]:
1 lasso_model_cv.best_estimator_

Out[3683]: Lasso(alpha=50)
```

When the value of alpha is doubled ideally it means that we are applying high regularization which results in a simple model which is underfitting of the data.

Upon applying double alpha, r2_score of both ridge and lasso went down and the coefficients of the predictor variables shrunk.

The most important predictor variables after the implementation of the change are

1. BsmtUnfSF
2. BsmtFinSF1
3. BsmtFinSF2
4. 2ndFlrSF
5. OverallQual_Very Excellent
6. KitchenAbvGr
7. Age
8. ExterCond_Fa

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:

The ideal choice is to apply Lasso regression due to the following reasons

1. r2_score of Lasso regression is slightly higher than that of ridge regression.
2. The number of predictors in Lasso regression are less are it performs feature selection.

Even though Lasso is computationally intensive it is ideal to lasso regression as in the test data set the noise would ignored by Lasso as the coefficients would have become 0 for them due to feature selection. This helps in increase of adjusted r2 square score as well.

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:

Initially the strong predictor variables are

BsmtUnfSF

BsmtFinSF1

BsmtFinSF2

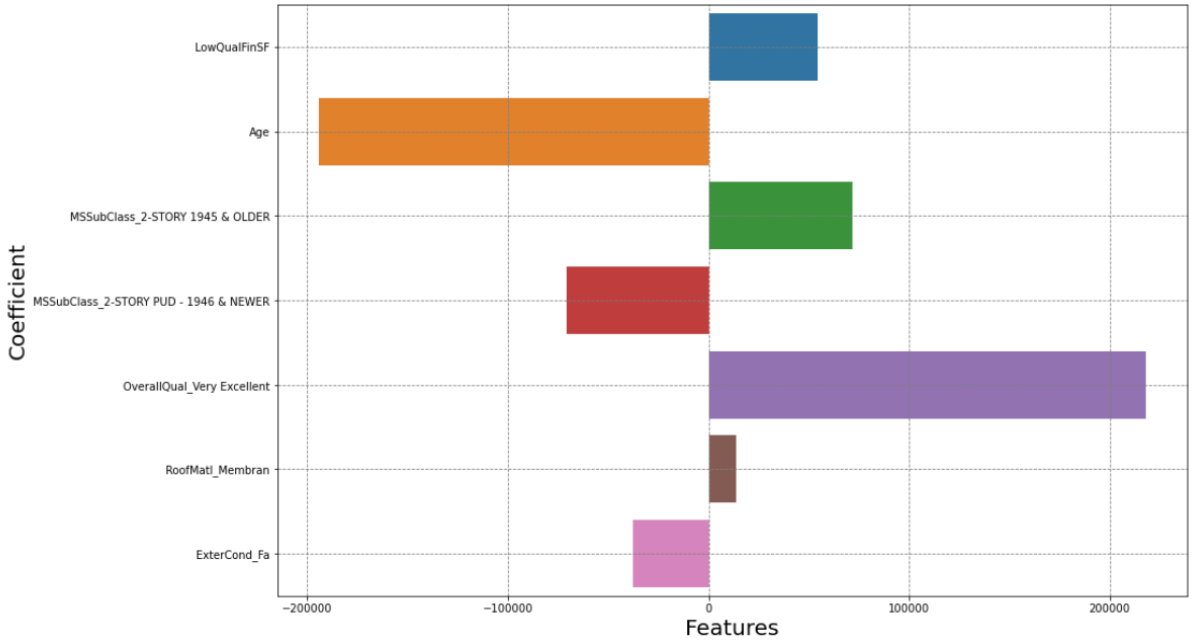
2ndFlrSF

KitchenAbvGr

The above predictor variables were removed, lasso regression was applied which gave the following coefficients.

	Features	Coefficient
4	OverallQual_Very Excellent	218005.75
2	MSSubClass_2-STORY 1945 & OLDER	71532.93
0	LowQualFinSF	54339.67
5	RoofMatl_Membran	13913.94
6	ExterCond_Fa	-37605.25
3	MSSubClass_2-STORY PUD - 1946 & NEWER	-70887.67
1	Age	-194159.72

When a plot was made features w.r.t to their coefficients



The top predictor variables as per the new model are

OverallQual_Very Excellent

Age

MSSubClass_2-STORY 1945 & OLDER

MSSubClass_2-STORY PUD - 1946 & NEWER

LowQualFinSF

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:

To make sure we are building a robust and generalizable model we need to make sure that our model is simple enough for generalization and is not underfitted. Regularization helps to strike a balance between keeping model simple, yet not making it too naive to be any use.

In Regularization we add an extra Regularization term to the error term which makes sure that this balance exists.

There are mainly two kinds of Regularization

Lasso (L1):

The Regularization term is made of absolute values of the parameters of the model.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Cost function

Ridge (L2):

The Regularization term is made of squares of the parameters of the model.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Cost function

We use hyperparameter named lambda to apply the significance of regularization.

Regularization also helps to balance bias-variance trade off.

Bias is the error made on test data. Variance is refers to change of output w.r.t to input.