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:

Ridge Regression:

The optimal value for Alpha in Ridge refression is 4

Objective = RSS + α * (sum of square of coefficients)

1. In ridge regression, as the value of Alpha increases, the model complexity reduces thus increase in RSS value.
2. The smaller the alpha value, that leads to severe reduction of magnitude of coefficient value of the predictor variables
3. Higher the alpha value will lead to under fitting as the complexity reduces and increase in magnitude of coefficient values for predictor variables

Though the coefficients are **very very small**, they are **NOT zero in ridge regression**.

Lasso Regression:

Objective = RSS + α * (absolute sum of coefficients)

The optimal value for Alpha in Ridge refression is 0.0001

Lasso tells us again that the model complexity decreases with increase in the values of alpha. But notice the straight line at alpha=1.

Apart from the expected inference of higher RSS for higher alphas, we can see the following:

1. For the same values of alpha, the coefficients of lasso regression are much smaller as compared to that of ridge regression.
2. For the same alpha, lasso has higher RSS (poorer fit) as compared to ridge regression
3. Many of the coefficients are zero even for very small values of alpha

We can observe that **even for a small value of alpha, a significant number of coefficients are zero**. This also explains the horizontal line fit for alpha=1 in the lasso plots, its just a baseline model! This phenomenon of most of the coefficients being zero is called '**sparsity**'.

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:

I prefer for the Lasso regression than Ridge regression because Lasso performs an addition work of feature selection(i.e. reducing the predictor variables by making the coefficients to zero) which was not possible in ridge regression. Few more reasons and differences explained below.

The key difference between lasso and ridge regression are

- **Ridge:** It includes all (or none) of the features in the model. Thus, the major advantage of ridge regression is coefficient shrinkage and reducing model complexity.
- **Lasso:** Along with shrinking coefficients, lasso performs feature selection as well. (Remember the '*selection*' in the lasso full-form?) As we observed earlier, some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model.

Applicability:

- **Ridge:** It is majorly used to *prevent overfitting*. Since it includes all the features, it is not very useful in case of exorbitantly high #features, say in millions, as it will pose computational challenges.
- **Lasso:** Since it provides *sparse solutions*, it is generally the model of choice (or some variant of this concept) for modelling cases where the #features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored.

Question 3

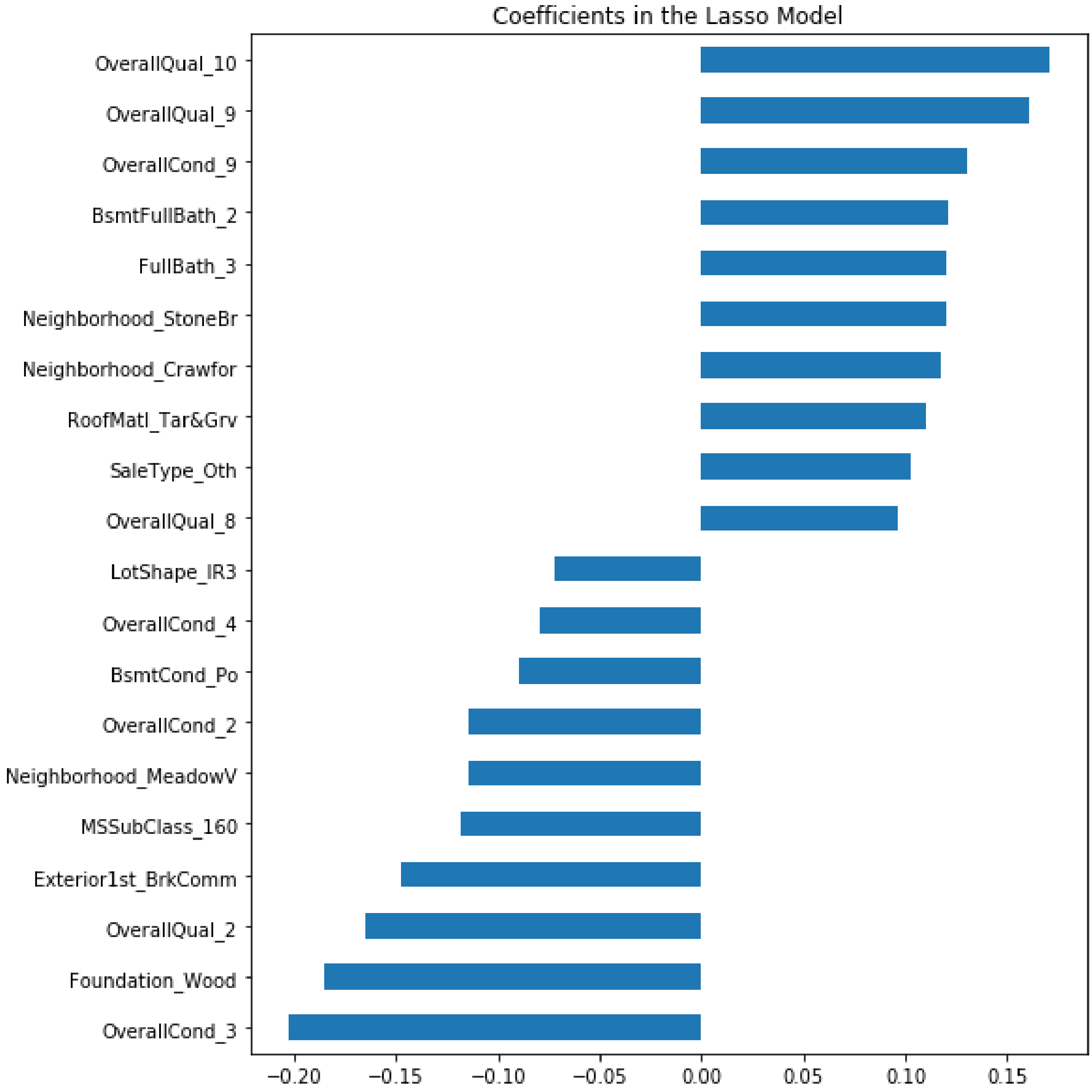
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans:

We have tried to remove the most important variables that comes from the model as shown in below figure.

The top 5 predictor having high coefficients that we removed from model are

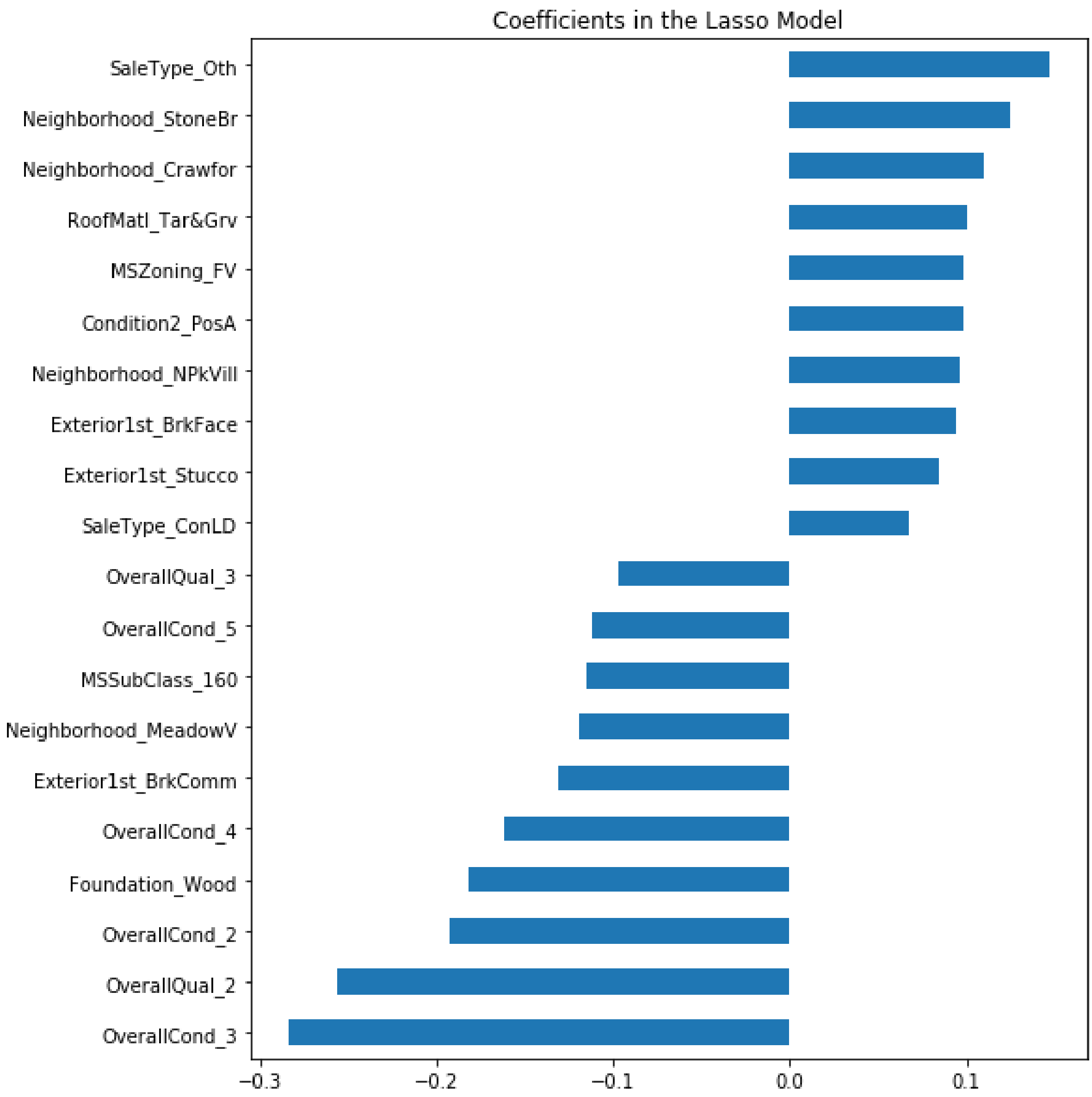
OverallQual_10,OverallQual_9,OverallCond_9,BsmtFullBath_2,FullBath_3



Now we remove the above top 5 predictor variables and rebuild the model and now the new predictor variables (top 5) are going to be as shown in figure.

SaleType_oth, Neighbourhood_stoneBr, Neighbourhood_Crawfor, RoofMatl_Tar&Grv, MSZoning_FV

The model remains same with same optimal values of alpha in both lasso(0.0001) and ridge regression(4) and the r2 and mse as well remains same after removing the tope 5 predictor variables.



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?

Ans:

More accurate model leads to overfitting. A flexible one will learn from data a lot. So usually avoiding accurate model and preferring a robust one is better. Although, in practice, we often cannot have a low bias and low variance model. As the model complexity goes up, the bias reduces while the variance increases, hence the trade-off.

While creating the best model for any problem statement, we end up choosing from a set of models which would give us the least test error. Hence, the test error, and not only the training error, needs to be estimated in order tocselect the best model. This can be done in the following two ways.

1. Use metrics which take into account both model fit and simplicity. They penalise the model for being too complex (i.e. for overfitting), and thus are more representative of the unseen ‘test error’. Some examples of such metrics are Mallow's Cp, Adjusted R&, AIC and BIC

2. Estimate the test error via a validation set or a cross-validation approach.
   In validation set approach, we find the test error by training the model on training set and fitting on an unseen validation set while in n-fold cross-validation approach, we take the mean of errors generated by training the model on all folds except the kth fold and testing the model on the kth fold where k varies from 1 to n

Let's look into these one by one

1. Mallow's Cp

$$C_p = \frac{1}{n}(RSS + 2d\sigma^2)$$

2. AIC (Akaike information criterion)

$$AIC = \frac{1}{n\sigma^2}(RSS + 2d\sigma^2)$$

3. BIC (Bayesian information criterion)

$$BIC = \frac{1}{n}(RSS + ln(n)d\sigma^2)$$

4. Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

AIC and BIC are defined for models fit by maximum likelihood estimator. We can notice that as we increase the number of predictors d, the penalty term in Cp, AIC and BIC all increase while the RSS decreases. Hence, lower the value of Cp, AIC and BIC, better is the fit of the model. Higher the Adjusted $R2$, better is the fit of the model.