Advanced Linear Regression Assignment

Question1. 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?

Optimal value of alpha in ridge regression and lasso regression is 20 and 0.001 respectively.

- After doubling alpha in ridge (alpha=40), the evaluation metrics are as below:

```
r square training score: 0.91
r square test score: 0.9
rss training score: 87.32000000000001
rss test score: 35.53
mse training score: 0.09
mse test score: 0.09
```

Not much change to r square training score and r square test score for ridge regression.

After doubling alpha in ridge, top 5 most important predictor variables are :

Coeff Values	Variables	Absolute Coefficent
0.197513	OverallQual	0.197513
0.185436	GrLivArea	0.185436
0.131211	1stFlrSF	0.131211
0.127301	Neighborhood_StoneBr	0.127301
0.120601	SaleType_New	0.120601

- After doubling alpha in lasso (alpha = 0.002), the evaluation metrics are:

```
r square training score: 0.91
r square test score: 0.91
rss training score: 84.51
rss test score: 34.31
mse training score: 0.09
mse test score: 0.08

Training r square score reduces from 0.92 to 0.91.
```

After doubling alpha in lasso, top 5 most important predictor variables are :

Coeff Values	Variables	Absolute Coefficent
0.418175	Neighborhood_StoneBr	0.418175
0.337145	SaleType_New	0.337145
0.322534	GrLivArea	0.322534
0.212594	Neighborhood_Crawfor	0.212594
0.195404	OverallQual	0.195404

Question2. 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?

Metrics for ridge regression when alpha = 20 and metrics for lasso regression when alpha = 0.001:

Metric	Linear Regression	Ridge Regression	Lasso Regression
R2 Score (Train)	0.865052	0.914574	0.917888
R2 Score (Test)	0.827203	0.904061	0.905282
RSS (Train)	130.629210	82.692226	79.483941
RSS (Test)	62.943383	34.946968	34.502165
RMSE (Train)	0.367352	0.292277	0.286551
RMSE (Test)	0.388981	0.289840	0.287989

The training and test r square for Lasso is more than ridge regression and also the RMSE for training and testing data is lesser in lasso which means better model.
As can be seen, Lasso regression performs better. Hence, I will choose Lasso Regression.

Also, Lasso regression does feature selection by default (making less significant coefficient values equal to 0). Hence making the model more generic.

Question3. 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. Top 5 predictor variables in lasso are :

1. Neighborhood_StoneBr

2. SaleType_New
3. GrLivArea
4. Neighborhood_Crawfor
5. Exterior2nd_CmentBd

After removing these 5 variables, trained the lasso model again and the resultant model had below top 5 predictor variables along with corresponding coefficient values:

Coeff Values	Variables	Absolute Coefficent
0.297066	SaleCondition_Partial	0.297066
0.293535	1stFlrSF	0.293535
0.282948	2ndFlrSF	0.282948
0.219799	Exterior1st_CemntBd	0.219799
-0.218242	BldgType_Twnhs	0.218242

Question4. 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?

1. Model can be considered as robust and generalizable based on Occam's Razor principle. Occam's Razor basically states that given competing theories and explanations, the simplest ones should be preferred.
2. In Machine learning, it can be implemented as given two models that show similar 'performance' in the finite training or test data, the thumb rule is to choose the simpler model.
3. Benefits of using simpler model:

    - A simpler model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen data sets.
    - A simpler model requires fewer training data points. This becomes extremely important because in many cases, one has to work with limited data points.
    - A simple model may make more errors in the training phase but is bound to outperform complex models when it views new data. This happens because of **overfitting**.
    - A simple model is more robust and does not change significantly if the training data points undergo small changes.

4. For making a model simpler, the technique used is called Regularization. Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
   The commonly used regularization techniques are ridge regression and lasso regression.

A simpler model will perform better on the unseen test set as it prevents overfitting. Overfitting is a phenomenon wherein a model becomes highly specific to the data on which it is trained and fails to generalise to other unseen data points in a larger domain. A model that has become highly specific to a training data set has 'learnt' not only the hidden patterns in the data but also the noise and the inconsistencies in it. In a typical case of overfitting, a model performs quite well on the training data but fails miserably on the test data.
As simpler models are more robust and generalize better, hence, it increases the accuracy on test sets.

For having a trade-off between simple and complex models, bias-variance trade-off concept is introduced.

The 'variance' of a model is the **variance in its output** on some test data with respect to the changes in the training data. In other words, variance here refers to the **degree of changes in the model itself** with respect to changes in the training data.

Bias quantifies how **accurate the model is likely to be** on future (test) data. Extremely simple models are likely to fail in predicting complex real-world phenomena. Simplicity has its own disadvantages.

Hence, accuracy of the model can be maintained by making a model complexity to optimum Model complexity line as shown as below.