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?

Answer:

Optimal value of alpha:

Optimal alpha value for Ridge regression model is: 8

Optimal alpha value for Lasso regression model is : 0.0007

Double the value of alpha for both Ridge and Lasso:

The cost function of Ridge and Lasso are:

$$\text{Ridge Regression cost} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{RSS} + \underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{penalty}$$

$$\text{Lasso Regression cost} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{RSS} + \underbrace{\lambda\sum_{j=1}^{p}|\beta_j|}_{Penalty}$$

y_i - Actual value of ith datapoint

$\hat{y}_i$ - predicted value of ith datapoint

β_j - co-efficient of jth Feature.

In order to do regularization we add up penalty term to the cost function. This penalty term helps the model trade bias for a significant reduction in the variance. In both cases (Ridge and Lasso) penalty term increases with higher value of beta co-efficient. Ridge regression uses sum of square of all beta coefficients as shrinking penalty. Lasso uses sum of absolute values of all beta coefficients as shrinking penalty. In both the equations, the penalty term is multiplied by lambda or alpha. This is a hyperparameter and its optimal value can be obtained by hyperparameter tuning. Value of alpha can be any number greater than or equal to 0. If the value of lambda move towards higher and higher number, the shrinkage penalty increases and it will push the coefficients further and further towards 0. So, the mode will be simpler. It will increase the bias and variance will get reduced.

If we increase the value of alpha to a very large number, then all coefficients of Lasso become zero and Ridge coefficients becomes closer to zero. Then the model will have high bias and low variance and it may result in Underfitting. If we reduce the value of lambda then the shrinking penalty will be lower, so the model bias will reduce and variance will increase. Choosing good lambda value becomes crucial.

We need to find the optimal value of alpha by performing hyperparameter tuning.

Top 10 features with beta coefficient values obtained from Ridge after double the value alpha (16)

```
OverallQual              0.205808
GrLivArea                0.145821
OverallCond              0.111776
GarageArea               0.106464
1stFlrSF                 0.103543
2ndFlrSF                 0.102571
FullBath                 0.091342
Neighborhood_StoneBr     0.087768
MSSubClass_30           -0.083300
Exterior1st_BrkFace      0.078694
```

Top 10 features with beta coefficient values obtained from Lasso after double the value alpha (0.0014)

```
OverallQual              0.377499
GrLivArea                0.372707
GarageArea               0.154041
OverallCond              0.123692
YearRemodAdd             0.078064
CentralAir               0.077008
Neighborhood_NridgHt     0.074724
Exterior1st_BrkFace      0.073594
MSSubClass_30           -0.072364
Neighborhood_StoneBr     0.067855
```

After Doubling value of alpha the most important variables are:

In Ridge model: (OverallQual & GrLivArea)-coefficients (0.205808 & 0.145821)
In Lasso Model: (OverallQual & GrLivArea)-coefficients (0.377499 & 0.372707)

(In Lasso, Variables (OverallQual & GrLivArea) have almost equal coefficients)

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?

Answer:

Regularization helps managing the models complexity by basically shrinking the model coefficients towards zero. Model complexity depends on main two things: **Magnitude of beta coefficients** and **number of independent variables**.

Now, Ridge and Lasso both have similar r2 score and MAE on test dataset. But Lasso has eliminated 115 features and final number of features in Lasso Regression model is 104. Where Ridge model has all 219 features. So, Lasso model is simpler than Ridge with having similar r2 score and MAE.

Ridge:

```
r2 score on test dataset: 0.8900525915673781
MSE on test dataset: 0.018610398565659873
RMSE on test dataset: 0.13641993463442165
MAE on test dataset: 0.09400644622692728
```

Lasso:

```
r2 score on test dataset: 0.8949539666532987
MSE on test dataset: 0.017780760603572966
RMSE on test dataset: 0.13334451846091375
MAE on test dataset: 0.09150936913949212
```

Both Ridge and Lasso models shows almost similar performance on test dataset, we should choose the simpler model. So, I will Choose Lasso model.

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?

Answer:

Initial top 5 features in Lasso model are as below:

GrLivArea	0.373734	(Above grade (ground) living area square feet)
OverallQual	0.319486	(Rates the overall material and finish)
OverallCond	0.142915	(Overall condition of the house)
GarageArea	0.137348	(Rates the overall condition of the house)
Neighborhood_StoneBr	0.127501	(Dummy variable of Neighborhood- Physical locations)

As Neighborhood_StoneBr is a dummy variable, dropping entire Neighborhood feature.
After dropping the top five variables (GrLivArea, OverallQual, OverallCond, GarageArea, Neighborhood) and Rebuilt the Lasso model again with rest of features.

Now, Top 5 features after dropping top 5 features of Previous Lasso model are as below.

1stFlrSF	0.406328	(First Floor square feet)
2ndFlrSF	0.359682	(Second Floor square feet)
KitchenQual_TA	-0.128247	(Dummy variable kitchenQual Typ/Avg)
YearRemodAdd	0.126438	(Remodel date)
Exterior1st_BrkFace	0.125430	(Dummy variable Exterior covering)

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?

<u>Answer:</u>

As per Occam's razor, a model should not be unnecessary complex, given two models that shows similar performance in the finite training and test data, we should pick the one that makes fewer on the test data due to following reasons:-

- Simpler models are usually more generic and are more widely applicable
- Simpler models require fewer training samples for effective training than the complex ones and hence are easier to train.
- Simpler models are more robust.
- Complex models tend to change widely with changes in the training data set.
- Simple models have low variance, high bias and complex models have low bias, high variance.
- Simpler models make more error in the training set. Complex models lead to overfitting, they work very well for the training samples, fail miserably when applied to other test samples.

Therefore to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.
Regularization can be used to make the model simple. Regularization helps to strike the delicate balance between, keeping the model simple and not making it too naive to be any use.
For regression, regularization involves adding a regularization terms to the cost that adds up the absolute values or the squares of the parameters of the model.

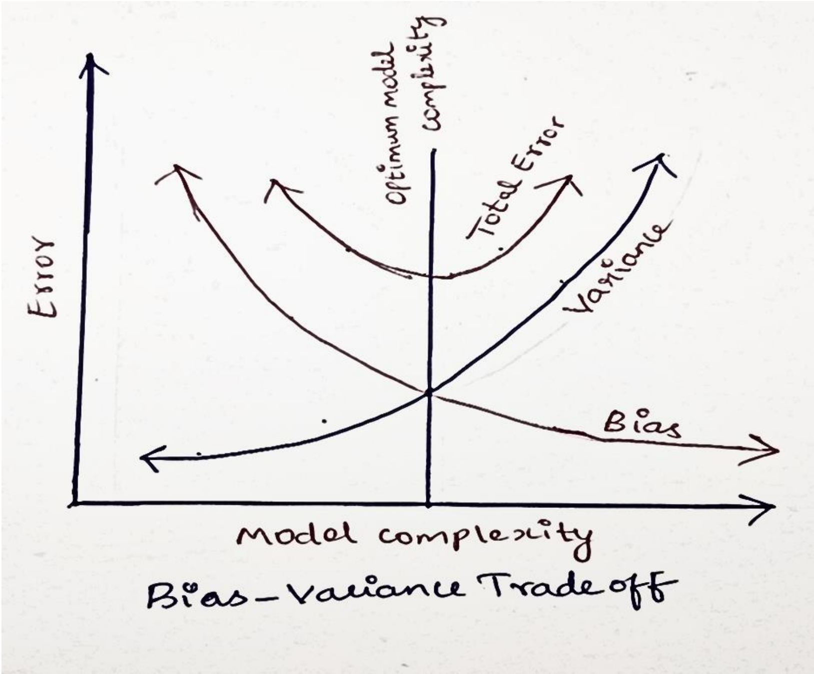
Also, making a model simple leads to Bias-Variance trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change widely even if more points are added or removed.

Bias quantifies now accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data.

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



Optimum model complexity
Total Error
Variance
Error
Bias
Model complexity
Bias-Variance Tradeoff