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 1

The optimal value of alpha for ridge and lasso regression are **10** and **0.001**, respectively.

If we create model select double the value of alpha for both ridge and lasso, following changes are observed:-

1. Coefficient's values are changed for both for ridge and lasso regression.

Table1: Coefficient values on Ridge regression (for alpha 10 and 20).

	Ridge	Ridge Double
MSSubClass	0.000507	0.000010
LotFrontage	0.007967	0.008071
LotArea	0.021092	0.020829
OverallQual	0.076219	0.076313
OverallCond	0.040466	0.039460
YearRemodAdd	0.013280	0.013704
MasVnrArea	0.010826	0.011844
BsmtFinSF1	0.030743	0.030265
BsmtFinSF2	0.001637	0.001277
BsmtUnfSF	0.002682	0.002718

2. R2 score, RSS and RMSE values are almost same.

Table 2: Metric comparison on Ridge regression (for alpha 10 and 20).

	Metric	Ridge Regression	Double Ridge Regression
0	R2 Score (Train)	0.928747	0.928747
1	R2 Score (Test)	0.912982	0.912982
2	RSS (Train)	8.374832	8.374832
3	RSS (Test)	4.329465	4.329465
4	MSE (Train)	0.099731	0.099731
5	MSE (Test)	0.109512	0.109512

Table 3: Metric comparison on Lasso regression (for alpha 0.001 and 0.002).

	Metric	Linear Regression	Ridge Regression	Lasso Regression	Double Lasso Regression
0	R2 Score (Train)	8.883190e-01	0.928747	0.923835	0.923835
1	R2 Score (Test)	-1.197965e+23	0.912982	0.913134	0.913134
2	RSS (Train)	1.312658e+01	8.374832	8.952124	8.952124
3	RSS (Test)	5.960314e+24	4.329465	4.321889	4.321889
4	MSE (Train)	1.248590e-01	0.099731	0.103111	0.103111
5	MSE (Test)	1.284934e+11	0.109512	0.109417	0.109417

Observations

We do not observe any significant changes in the metrics as well as in the features after doubling the optimal lambda values for Ridge and Lasso Regression.

The most important variable before the changes has been implemented for ridge regression are as follows:-

1. Neighborhood_StoneBr
2. Neighborhood_Crawfor
3. OverallQual
4. GrLivArea
5. Exterior1st_BrkFace

The most important variable after the changes has been implemented for ridge regression are as follows: -

1. OverallQual
2. GrLivArea
3. Neighborhood_Crawfor
4. Neighborhood_StoneBr
5. 1stFlrSF

The most important variable before the changes has been implemented for lasso regression are as follows:-

1. GrLivArea
2. Neighborhood_StoneBr
3. Neighborhood_Crawfor
4. OverallQual
5. Neighborhood_NridgHt

The most important variable after the changes has been implemented for lasso regression are as follows:-

1. MSSubClass

2. Exterior1st_Plywood

3. Exterior1st_ImStucc

4. Exterior1st_HdBoard

5. Exterior1st_CemntBd

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?

Question 2

The r2_score of lasso and ridge is almost equal for the test dataset so we can choose any regression to solve this problem.

There are some important points to select lasso regression for this problem.

1. As we have seen that the number of variables is large so lasso regression helps us to perform feature selection.
2. It also helps in model interpretation.
3. It forces some of the coefficients exactly equal to zero in the penalty term of cost function.
4. In order to obtain a simple model, regularization minimizes the error term and penalize the coefficients.

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 3

Earlier, the 5 most important predictor variables were: -

We realized that those five most important predictor variables in the lasso model are not available in the incoming data.

1. GrLivArea

2. Neighborhood_StoneBr

3. Neighborhood_Crawfor

4. OverallQual

5. Neighborhood_NridgHt

Then, we create another model excluding those five most important predictor variables. In this time, we found the five most important predictor variables: -

1. MSSubClass

2. Exterior1st_Plywood

3. Exterior1st_ImStucc

4. Exterior1st_HdBoard

5. Exterior1st_CemntBd

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?

Answer 4

The variance in a robust model is low. It means that changing one or more dependent variables has no significant effect on the predicted variable. On the other hand, a generalized model reduces model complexity. In a generalized model, as the number of dependent variables increases, it becomes more complex, resulting low bias but high variance. Moreover, a generalized model has just enough dependent variables to have the lower possible variance.

The Bias-Variance tradeoff finds balance between robust and generalized model as shown in figure 1.

We get good accuracy in train data and very low accuracy of test data. It means that model can be overfitting and underfitting. Outliers cause large variance and are extremely sensitive to the linear regression model. Therefore, we have proceeded Ridge/Lasso regression model with regularization, which incorporates a penalty term in the cost function. This penalty term shifts the variable coefficients towards 0. Due to this, model's complexity minimizes.

Therefore, regularization gives us large variance with a negligible bias trade-off. Regularization helps with managing model complexity by essentially shrinking the model coefficients toward zero. This discourage the model from becoming too complex, this avoids the risk of overfitting. As a result, it helps in the development of a robust and generalized model. In this type of model, the train and test accuracy will be good and consistent**.**

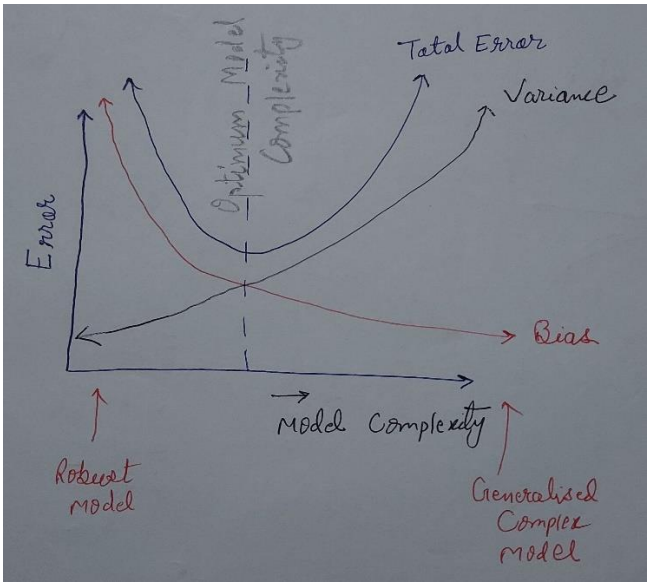

Figure 1: Bias and Variance tradeoff plot.