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:

The optimum value of alpha for ridge regression is 500 and for lasso regression is also 500.

Alpha = 500					Alpha = 1000				
	Metric	Linear Regression	Ridge Regression	Lasso Regression		Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.57E-01	8.94E-01	9.32E-01	0	R2 Score (Train)	9.57E-01	8.70E-01	8.99E-01
1	R2 Score (Test)	-8.06E+21	8.63E-01	8.48E-01	1	R2 Score (Test)	-8.06E+21	8.52E-01	8.43E-01
2	RSS (Train)	2.75E+11	6.78E+11	4.32E+11	2	RSS (Train)	2.75E+11	8.29E+11	6.42E+11
3	RSS (Test)	2.27E+34	3.87E+11	4.30E+11	3	RSS (Test)	2.27E+34	4.18E+11	4.42E+11
4	MSE (Train)	1.64E+04	2.58E+04	2.06E+04	4	MSE (Train)	1.64E+04	2.85E+04	2.51E+04
5	MSE (Test)	7.20E+15	2.97E+04	3.13E+04	5	MSE (Test)	7.20E+15	3.09E+04	3.18E+04

From above chart we can see that, with alpha = 500 we get values of R2 score as follow:

For Ridge Regression:

On training set R2 score is 0.8936641677004936 and on testing set 0.862558947239897.

With these R2 score values we can say model is performing good enough on both the sets. There for it is acceptable. The change in the R2 score values is 3%.

All 278 feature coefficients are present in the model.

For Lasso Regression:

On training set R2 score is 0.9322771190403039 and on testing set 0.8475354701397759.

With these R2 score values we can say model is not performing good. We can clearly see the overfitting problem in the training set. The change in the R2 score values is 9%.

As lasso does model selection, 136 features coefficients are 0. So, 142 features are present in the model.

If we choose double the value of alpha i.e., 1000 then R2 score value changes as follows:

For Ridge Regression:

On training set R2 score is 0.870132090410631 and on testing set 0.8515556261582596.

With these R2 score values we can say model is performing even better on both the sets as compared to previous alpha value of 500. There for it is also acceptable. The change in the R2 score values is 2%.

All 278 feature coefficients are present in the model.

For Lasso Regression:

On training set R2 score is 0.8993171045075061 and on testing set 0.8432697516215422.

With these R2 score values we can say model is performing good on both the data sets. We can clearly see the overfitting problem in the training set has been reduced by adding more penalty to RSS value. The change in the R2 score values is 6%.

As lasso does model selection, 179 features coefficients are 0. So, 99 features are present in the model.

Most Important Predictors after change are as follows:

Positive important predictors: GrLivArea, OverallQual_10, OverallQual_9, OverallQual_8, GarageCars, YearBuilt, Neighborhood_NoRidge

Negative important predictors: Condition2_PosN, KitchenAbvGr, FireplaceQu_NA, ExterQual_TA, Neighborhood_Edwards

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:

As we have seen with alpha = 500 we have:

For Ridge Regression:

On training set R2 score is 0.8936641677004936 and on testing set 0.862558947239897.

The change in the R2 score values is 3%.

All 278 feature coefficients are present in the model. So, **It is complex model.**

For Lasso Regression:

On training set R2 score is 0.9322771190403039 and on testing set 0.8475354701397759.

The change in the R2 score values is 9%. **Overfitting Problem.**

As lasso does model selection, 136 features coefficients are 0. So, 142 features are present in the model. Less complex as compared ridge with alpha = 500.

With alpha = 1000 we have:

For Ridge Regression:

On training set R2 score is 0.870132090410631 and on testing set 0.8515556261582596.

The change in the R2 score values is 2%.

All 278 feature coefficients are present in the model.

For Lasso Regression:

On training set R2 score is 0.8993171045075061 and on testing set 0.8432697516215422.

The change in the R2 score values is 6%. **Overfitting problem reduced**.

As lasso does model selection, 179 features coefficients are 0. So, 99 features are present in the model. **Complexity reduced even more.**

Conclusion: Since, Lasso regression model with alpha value as 1000 is performing good on both training and test data set and also it has done model selection as only 99 features are there in the final model, I will choose this model to apply.

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:

After removing the 5 most important predictors such as GrLivArea, OverallQual_10, OverallQual_9, OverallQual_8, GarageCars, we have 273 predictors in the training and test data sets.

Five most important predictors in the updated model are: 2ndFlrSF, 1stFlrSF, Neighborhood_NridgHt, Neighborhood_NoRidge, YearBuilt.

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:

Models need to be made more robust and generalisable so that they are not impacted by the outliers present in the training data set. It should be generalised enough to provide equal test accuracy as of train accuracy. The model should be accurate for all the unseen datasets. Outlier analysis need to be done and only those features should be retained which are relevant to dataset. Weightage given to outliers should be low to increase the accuracy prediction of the model.

Simpler models have low variance and high bias. And complex models have low bias and high variance. We need to find the such a minimum RSS value for which we will get low bias with significantly low variance for unseen data. To achieve this, we used regularization techniques like ridge and lasso which add the penalty for the use of more predictors in the model. This penalty increases the RSS value which results in compromising the bias to get significantly reduction in variance.

The model may have non-linearity in the data. In generalised models the non-linearity in the data is handled with the help of data transformation techniques. We transform the response variable if error terms are not normal of if the residuals exhibit non constant variance. We transform the predictors when we observe non-linear trend in the residual plot.

To make the predictive models more robust we can follow several methods such as:

Removing outliers: It will improve the accuracy of the model.

Use of robust statistics: Use of median and interquartile range (IQR) are more robust to outliers than the mean and standard deviation.

Use of robust regression: Use of least absolute deviations (LAD) regression and the least squares regression with L1 regularization (Lasso) are more robust to outliers than the least squares regression.

Using ensemble methods: Combining multiple models can make the final model more robust to outliers because it will not be affected by a single outlier.

Data pre-processing: Techniques such as data normalization, data scaling and data transformation can also make a predictive model more robust to outliers.