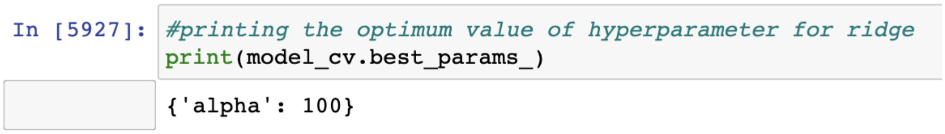
Advanced Regression Assignment - Subjective Questions

Question 1

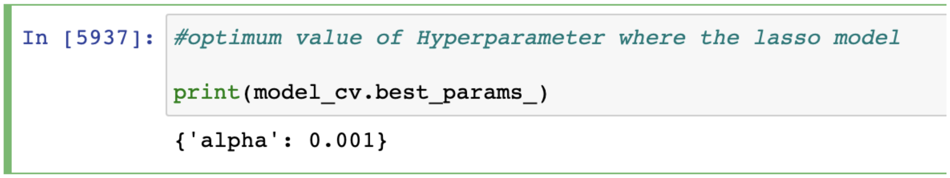
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer :

The optimum value of alpha for ridge and lass obtained as as follows :

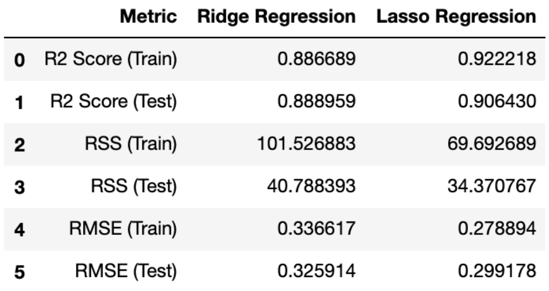


```
In [5927]: #printing the optimum value of hyperparameter for ridge
           print(model_cv.best_params_)

           {'alpha': 100}
```



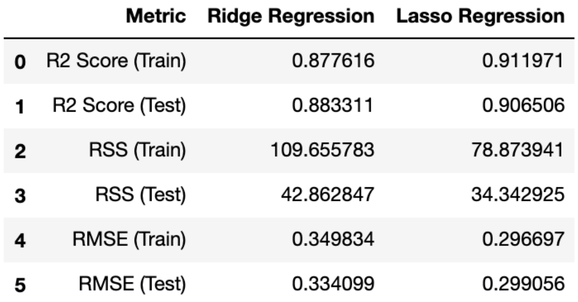
```
In [5937]: #optimum value of Hyperparameter where the lasso model

           print(model_cv.best_params_)

           {'alpha': 0.001}
```

The statistics R-squared, RSS and RMSE obtained for the Ridge and Lasso models with the above alpha values are as follows :



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.886689	0.922218
1	R2 Score (Test)	0.888959	0.906430
2	RSS (Train)	101.526883	69.692689
3	RSS (Test)	40.788393	34.370767
4	RMSE (Train)	0.336617	0.278894
5	RMSE (Test)	0.325914	0.299178

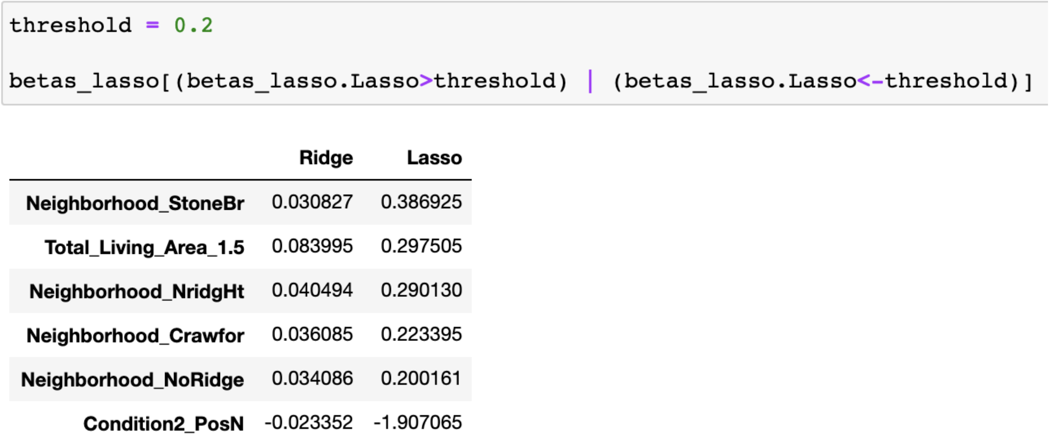
The statistics R-squared, RSS and RMSE obtained for the Ridge and Lasso models with the alpha values 2x the optimum values are as follows:



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.877616	0.911971
1	R2 Score (Test)	0.883311	0.906506
2	RSS (Train)	109.655783	78.873941
3	RSS (Test)	42.862847	34.342925
4	RMSE (Train)	0.349834	0.296697
5	RMSE (Test)	0.334099	0.299056

We can clearly see that when the 2x alpha is being used rather than the optimum value of alpha, we seem to get a slight degradation in the model performance as in both Ridge and Lasso the R2-score decreases for train, test data. Also the RSS increases marginally which in turn has a slight increase in the RMSE.

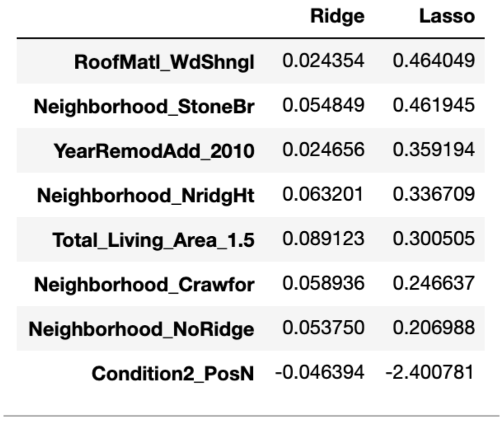
The top variables when using the optimum alpha for Lasso Regression (with a coefficient higher than |0.2|) are :



```
threshold = 0.2

betas_lasso[(betas_lasso.Lasso>threshold) | (betas_lasso.Lasso<-threshold)]
```

	Ridge	Lasso
Neighborhood_StoneBr	0.030827	0.386925
Total_Living_Area_1.5	0.083995	0.297505
Neighborhood_NridgHt	0.040494	0.290130
Neighborhood_Crawfor	0.036085	0.223395
Neighborhood_NoRidge	0.034086	0.200161
Condition2_PosN	-0.023352	-1.907065

The top variables when using 2x the optimum alpha for Lasso Regression (with a coefficient higher than |0.2|) are :



	Ridge	Lasso
RoofMatl_WdShngl	0.024354	0.464049
Neighborhood_StoneBr	0.054849	0.461945
YearRemodAdd_2010	0.024656	0.359194
Neighborhood_NridgHt	0.063201	0.336709
Total_Living_Area_1.5	0.089123	0.300505
Neighborhood_Crawfor	0.058936	0.246637
Neighborhood_NoRidge	0.053750	0.206988
Condition2_PosN	-0.046394	-2.400781

Lasso just selects **97 variables out of 284** when 2X alpha is chosen where as it selects **135 variables** when modelling with optimum alpha.

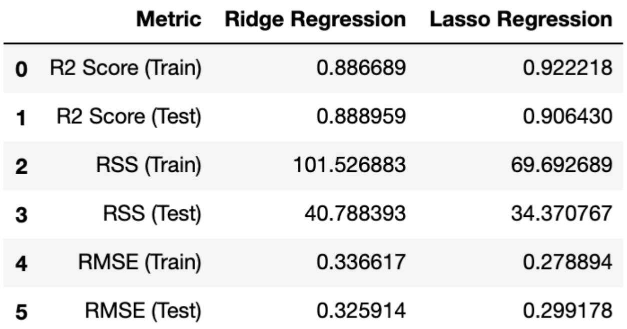
Question 2

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 :

The optimum values for lambda obtained for Ridge = 100 and that for Lasso = 0.001.

As per the train-test statistics the following are the metrics :



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.886689	0.922218
1	R2 Score (Test)	0.888959	0.906430
2	RSS (Train)	101.526883	69.692689
3	RSS (Test)	40.788393	34.370767
4	RMSE (Train)	0.336617	0.278894
5	RMSE (Test)	0.325914	0.299178

The above metrics suggest that Lasso model is performing better as R2 score of train test is better than ridge, RSS of train-test data is also lesser than that obtained by Ridge followed by the RMSE.

The residual plot and distplot of residuals for Lasso on train-test shows better randomness and normal behavior than the Ridge which also is a sign for a better model as Lasso justifies the assumptions of Linear Regression better in this case than the Ridge.

In terms of complexity of the model, Ridge model includes 274 out of 285 variables to be a part of the model. However, the Lass model includes only 135 out of 285 variables which makes it a simpler yet robust model comparatively.

Hence considering the above factors it is wise to choose the lambda obtained by the Lasso model with the 135 variables obtained.

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:

The top 5 most important predictor variables with their betas are as follows :

‘Condition2_PosN’	**-**	**-2.400781**
'RoofMatl_WdShngl',	**-**	**0.464049**
'Neighborhood_StoneBr',	**-**	**0.461945**
'YearRemodAdd_2010',	**-**	**0.359194**
'Neighborhood_NridgHt',	**-**	**0.336709**

After removing these variables from the dataset and performing the Lasso Regression, the top 5 variables are as follows :

Condition2_Norm	**-**	**0.774292**
Condition1_PosN	**-**	**-0.381092**
Total_Living_Area_1.5	**-**	**0.242500**
Neighborhood_Timber	**-**	**-0.196109**
Neighborhood_CollgCr	**-**	**-0.199547**

NOTE : Total_Living_Area_1.5 is the power (1stFlrSF + 2ndFlrSF, 1.5)

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 :

To ensure that a model is generalisable and robust we need to ensure the following :

1. Remove any outliers present in the data to make sure your model is not learning the highly unlikely data and that's affecting the test results.

2. Transform the data appropriately to ensure that all the predictors being used have a linear or almost linear relationship with the target variable.

3. Scale the data so that the model so that the results are visually understandable and to avoid misinterpretation of a particular predictors significance having a coefficient too high or too low.

We can say that this model is robust and generalisable by the following :

1. The model was built by using a 5 fold cross validation method so it is not biased towards any particular feature.

2. The R-squared on the training set and test test doesn't vary too much and hence it performs considerably good on unseen data as well.

3. The RMSE of this model is also low both for train and test which is a sign that the model's accuracy is considerably good both for train and test data and hence it can perform good on any other data that is unseen.