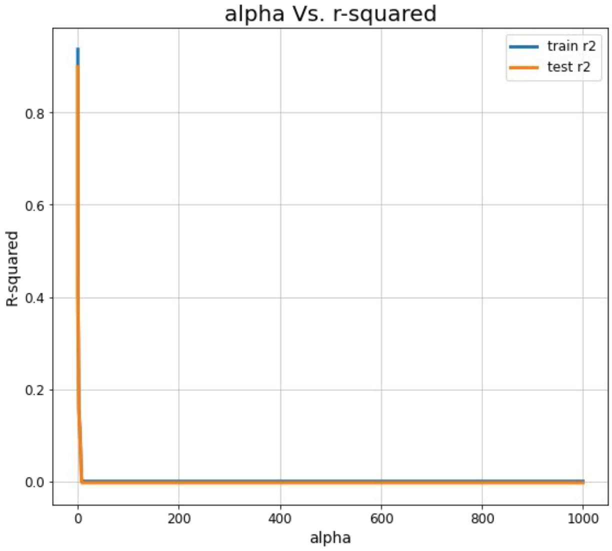
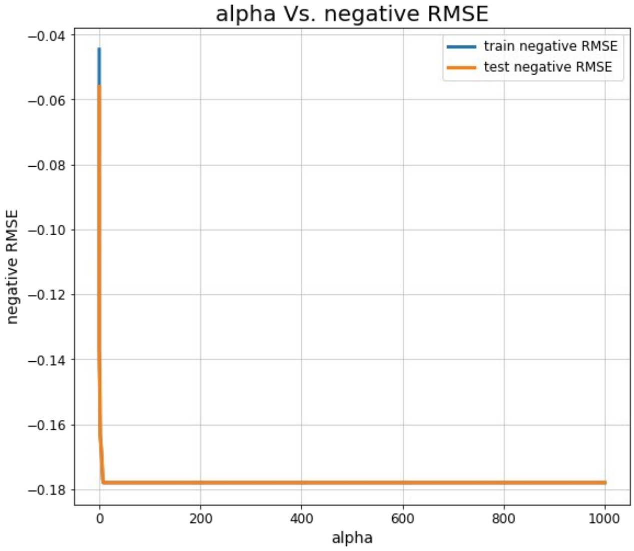
Subjective Questions and Answers

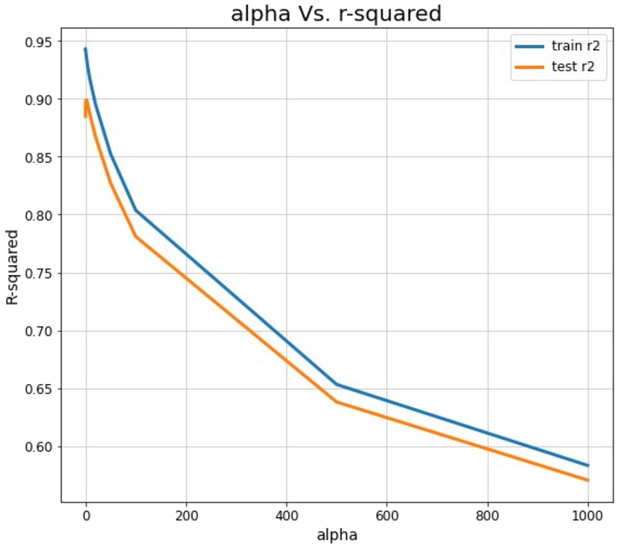
Question-1:
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

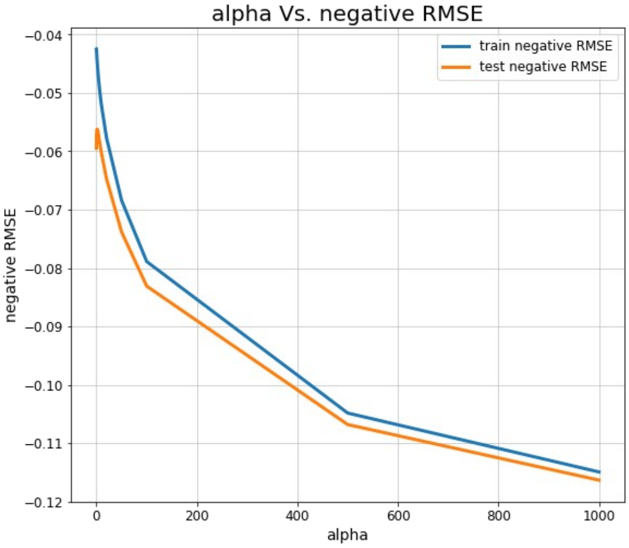
Answer:
The optimal value of alpha for ridge regression is 2.0 and for lasso regression in 0.0001.
As a general trend, we can see that in case of Lasso regression, the r-squared decreases and RMSE increases if alpha is increased, given that value of alpha is very small, in the order of 0.0001. This trend can be clearly visualized from the below two graphs :-



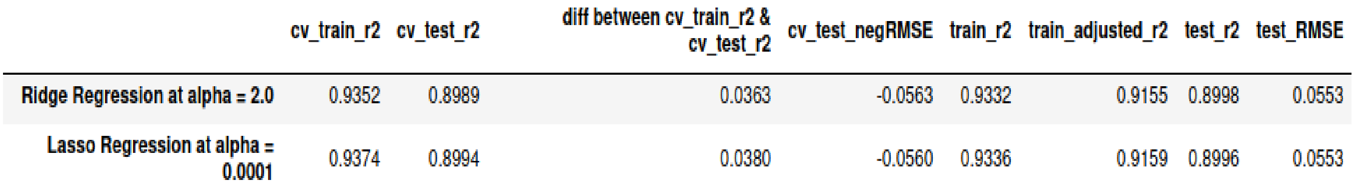


Unlike Lasso regression in case of Ridge regression, the r-squared decreases and RMSE increases with a less steep slope as well but throughout. Following two graphs depict this clearly :-





At optimal values of alpha, the statistical observations for the models are as follows :-

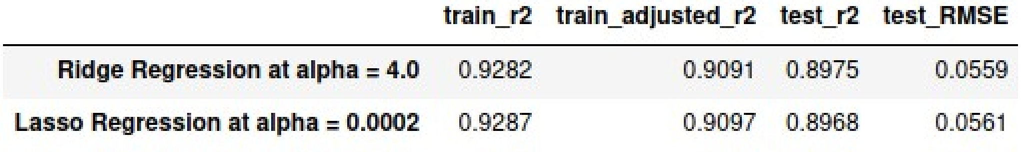


	cv_train_r2	cv_test_r2	diff between cv_train_r2 & cv_test_r2	cv_test_negRMSE	train_r2	train_adjusted_r2	test_r2	test_RMSE
Ridge Regression at alpha = 2.0	0.9352	0.8989	0.0363	-0.0563	0.9332	0.9155	0.8998	0.0553
Lasso Regression at alpha = 0.0001	0.9374	0.8994	0.0380	-0.0560	0.9336	0.9159	0.8996	0.0553

Top 5 important features for Ridge Regression model at alpha = 2.0 are :-
'GrLivArea', 'OverallQual', '2ndFlrSF', 'YearBuilt', 'TotalBsmtSF'

Top 5 important features for Lasso Regression model at alpha = 0.0001 are :-
'GrLivArea', 'YearBuilt', 'OverallQual', 'TotalBsmtSF', 'OverallCond'

But, after doubling the alpha values, the observations are as follows :-



	train_r2	train_adjusted_r2	test_r2	test_RMSE
Ridge Regression at alpha = 4.0	0.9282	0.9091	0.8975	0.0559
Lasso Regression at alpha = 0.0002	0.9287	0.9097	0.8968	0.0561

From the above discussed trends, it is clearly visible that since the doubled alpha value of Lasso regression (0.0002) is very small, the r-squared and adjusted r-squared have decreased and value of RSME has increased for Lasso regression.
Whereas in case of Ridge regression, r-squared decreased and RMSE increased by a certain small amount as well.
Eventually if we see the R-squared for test data set, it has has decreased for both the models as well.

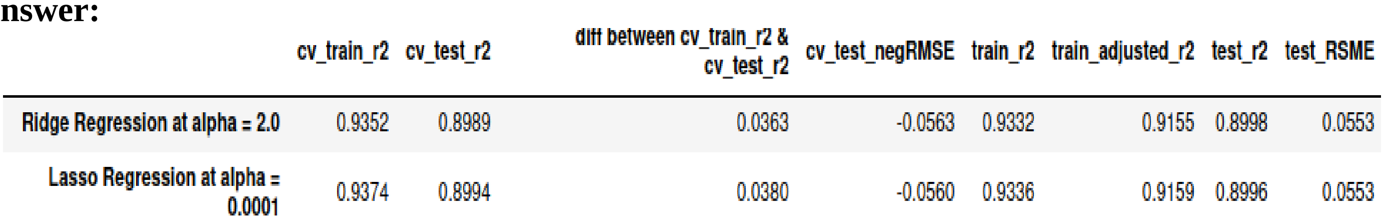
Top 5 important predictor variables for Ridge Regression model if the value of alpha is doubled at 4.0 are :-
'GrLivArea', 'OverallQual', '2ndFlrSF', '1stFlrSF', 'TotalBsmtSF'

Top 5 important predictor variables for Lasso Regression model if the value of alpha is doubled at 0.0002 are :-
'GrLivArea', 'OverallQual', 'YearBuilt', 'TotalBsmtSF', 'BsmtQual'

Question-2:

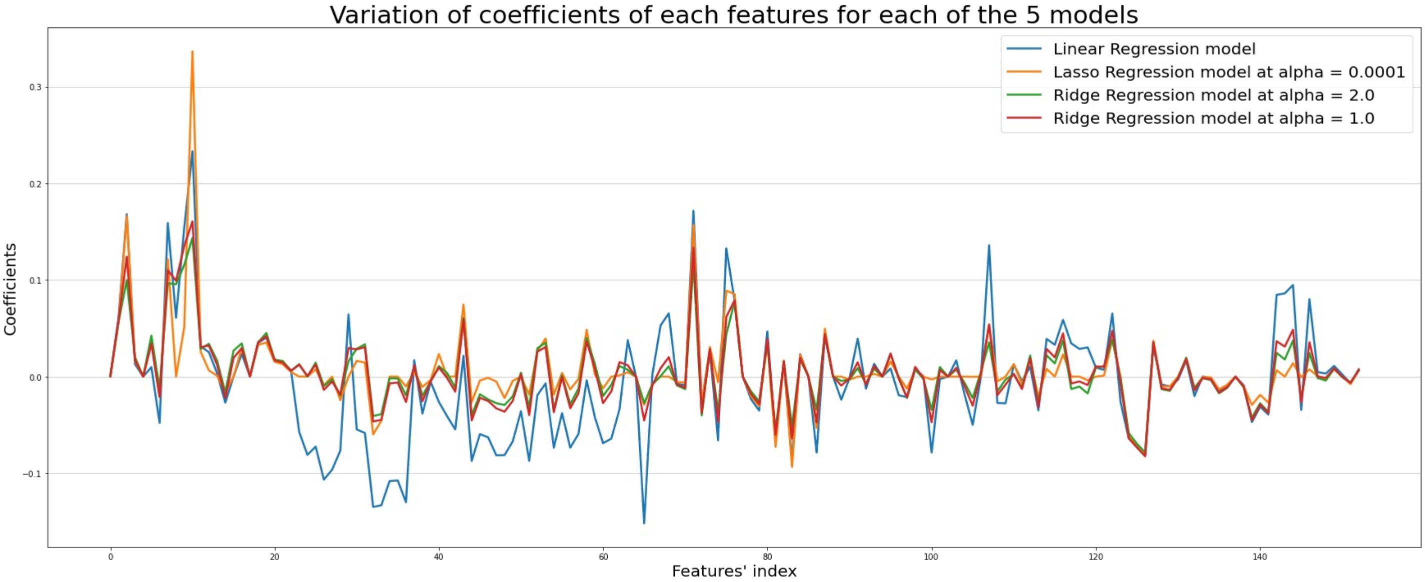
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:



	cv_train_r2	cv_test_r2	diff between cv_train_r2 & cv_test_r2	cv_test_negRMSE	train_r2	train_adjusted_r2	test_r2	test_RSME
Ridge Regression at alpha = 2.0	0.9352	0.8989	0.0363	-0.0563	0.9332	0.9155	0.8998	0.0553
Lasso Regression at alpha = 0.0001	0.9374	0.8994	0.0380	-0.0560	0.9336	0.9159	0.8996	0.0553

Any of the two models from the above table can be chosen and would give almost similar outcome. Moreover, from the following graph, it can be seen that unlike the linear regression model, these two models' coefficients are very much comparable to each other.



I would prefer to choose the Lasso regression model between the two models.
The reasons behind this are discussed below which is nothing but the description of the above table:
The difference between cross validation r-squared values of train and test part is not much for both the models. Even other values don't have any significant difference between the two models.
The main reason is the number of features considered by model for prediction. From the codes in the jupyter notebook, it is clear that Lasso model requires 2 less features to work than other models making it comparatively simpler, which is much needed. Moreover, the adjusted r-squared of Lasso model is 0.9159 and of Ridge model is 0.9155 which is a bit better for the 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:

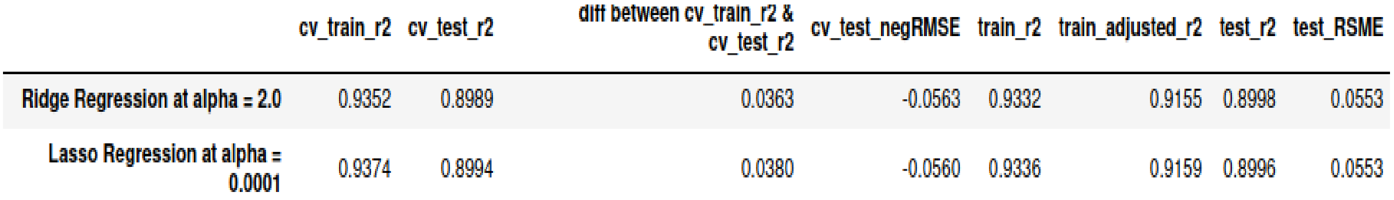
After building the Lasso model, the five most important predictor variables are:
'GrLivArea', 'YearBuilt', 'OverallQual', 'TotalBsmtSF', 'OverallCond'

But, if these features are absent in the incoming data then on building a new model with all the other available features, the most five important predictor variables would be as follows:
'1stFlrSF', '2ndFlrSF', 'BsmtQual', 'Exterior1st', 'KitchenQual'

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:
Let’s discuss with a practical example of the model used in this assignment:-

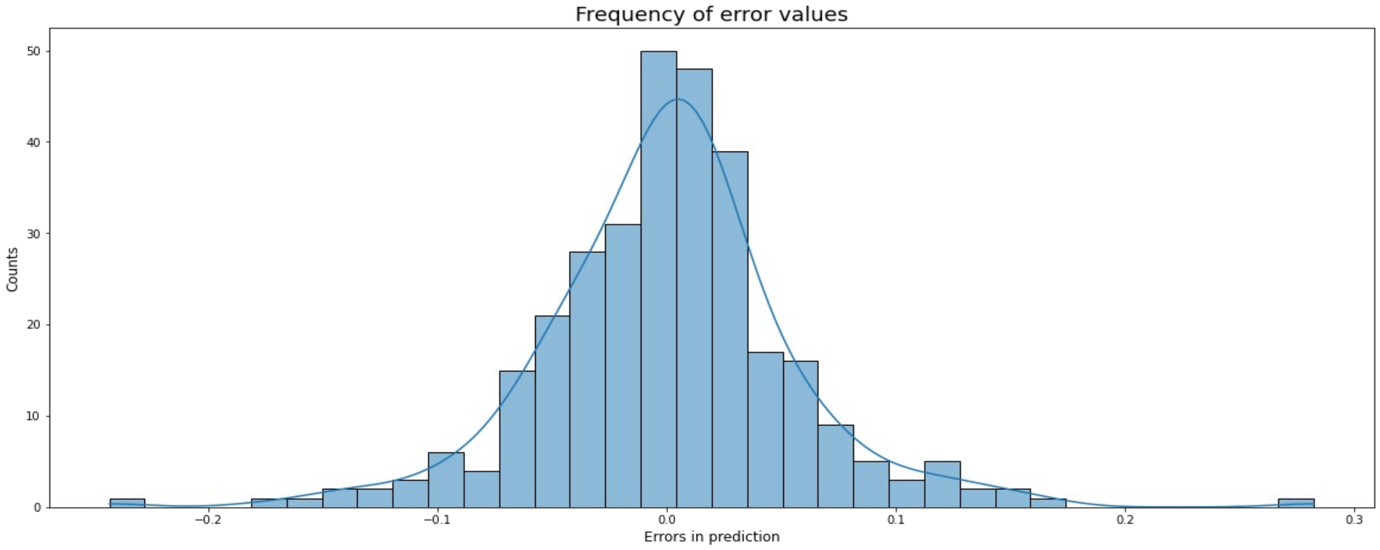
From the following table, it is clearly visible that there is not much difference between cross validation r-squared values of train and test part of the model.



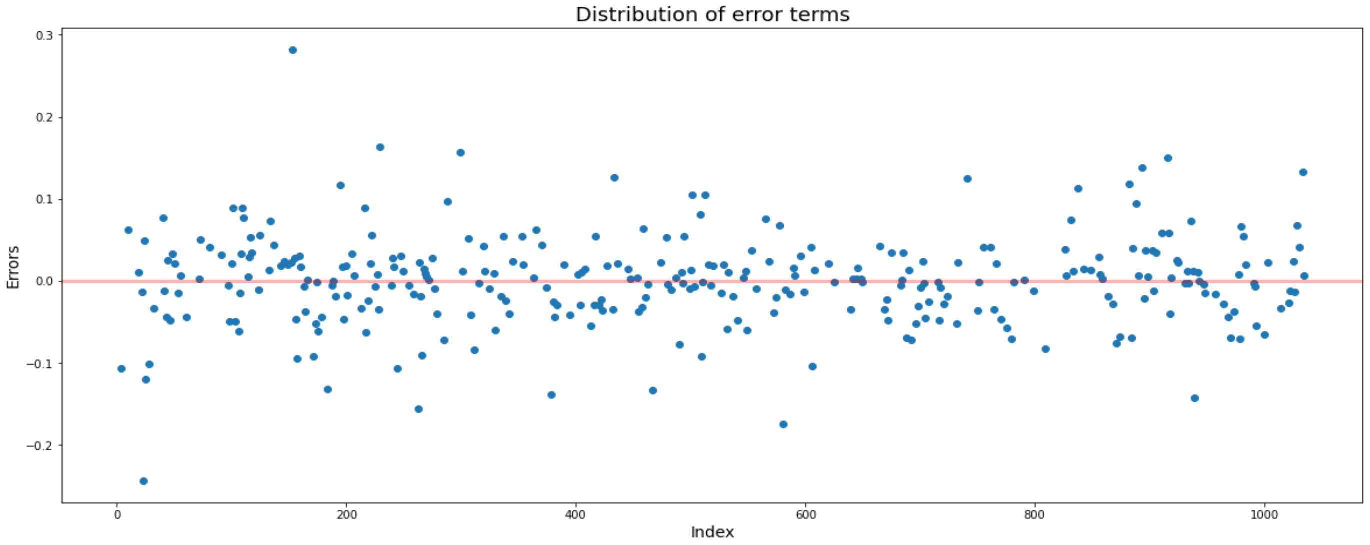
	cv_train_r2	cv_test_r2	diff between cv_train_r2 & cv_test_r2	cv_test_negRMSE	train_r2	train_adjusted_r2	test_r2	test_RSME
Ridge Regression at alpha = 2.0	0.9352	0.8989	0.0363	-0.0563	0.9332	0.9155	0.8998	0.0553
Lasso Regression at alpha = 0.0001	0.9374	0.8994	0.0380	-0.0560	0.9336	0.9159	0.8996	0.0553

The difference between cross validation r-squared values of train and test part is just 0.038 which states that both the bias and variance are low. Therefore the model is pretty much accurate. We can also see that the adjusted r-squared of Lasso model is 0.9159 which is high enough to capture the patterns of data points.
Moreover, the following graphs also give us some useful insights:

The error terms are normally distributed.



The error terms are homoscedastic.


Distribution of error terms
Errors
Index