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. optimal values of alpha(tuning hyper parameter) is:
    - **ridge**: 1
    - **lasso**: 0.0001
2. when alpha is doubled for ridge and lasso regression, there will be slight decrease in r2 score for train & test dataset.
    - **ridge regression** -->
        - **training** --
            - old model: 0.865
            - new model: 0.859
        - **testing** --
            - old model: 0.827
            - new model: 0.816
    - **lasso regression** -->
        - **training** --
            - old model: 0.867
            - new model: 0.859
        - **testing** --
            - old model: 0.842
            - new model: 0.833
3. most important 5 predictor variables after change in alpha are having maximum coefficients value:
    - **ridge**: 'OverallQual', 'TotalSF', 'TotRmsAbvGrd', '2ndFlrSF', 'GarageCars'

	features	coefficient
12	TotalSF	0.238065
15	OverallQual	0.174593
2	TotRmsAbvGrd	0.109740
10	LotArea	0.101952
27	BsmtQual	0.089104

    - **lasso**: 'TotalSF', 'OverallQual', 'Neighborhood_NridgHt', 'GarageCars', 'TotRmsAbvGrd'

	features	coefficient
12	TotalSF	0.395423
15	OverallQual	0.185009
0	GarageCars	0.075310
27	BsmtQual	0.069338
23	Neighborhood_NridgHt	0.065818

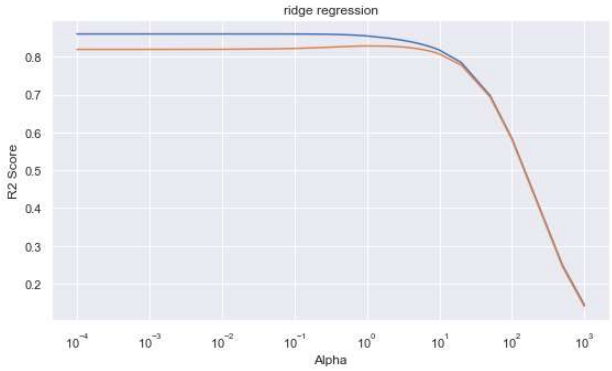
Question 2
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

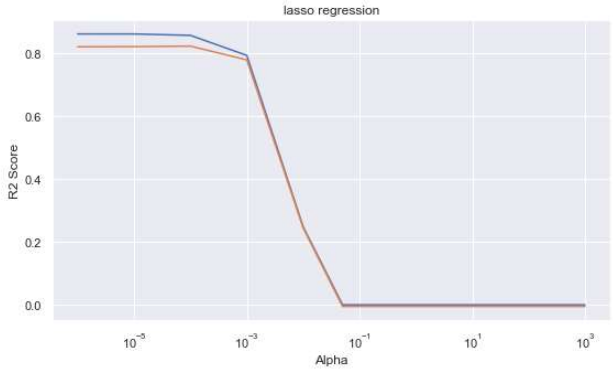
Answer

1. **ridge regression**: maximum R2 score is achieved at alpha between 0 & 10, beyond 10, R2 score decreases. for an alpha between 0 & 5, R2 score hold maximum value and difference between train and test dataset is very less.

2. **lasso regression**: R2 score steadily decreases from the lowest value of alpha, but after 10^-3 value of alpha, R2 score decreases linearly, also at this point, there is very less difference between train and test dataset. thus model is good fit.

when alpha value is doubled, there is a light decrease in R2 score of the model, thus we will god ahead with the previously selected optimal values.





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:
most important predictors before operation:
* **ridge regression model**: ['TotalSF', 'OverallQual', 'TotRmsAbvGrd', 'LotArea', 'BsmtQual']

	features	coefficient
12	TotalSF	0.238065
15	OverallQual	0.174593
2	TotRmsAbvGrd	0.109740
10	LotArea	0.101952
27	BsmtQual	0.089104

* **lasso regression model**: ['TotalSF', 'OverallQual', 'GarageCars', 'houseAge', 'BsmtQual']

	features	coefficient
12	TotalSF	0.395423
15	OverallQual	0.185009
0	GarageCars	0.075310
5	houseAge	-0.070360
27	BsmtQual	0.069338

After removing the predictors received from the incoming dataset, new most important predictors are:
* **ridge regression model**: ['GarageCars', 'houseAge', 'BedroomAbvGr', 'Neighborhood_NridgHt', 'Neighborhood_StoneBr']

	features	coefficient
0	GarageCars	0.209425
7	houseAge	-0.161072
21	BedroomAbvGr	0.132813
16	Neighborhood_NridgHt	0.121283
6	Neighborhood_StoneBr	0.109272

* **lasso regression model**: ['TotRmsAbvGrd', 'KitchenAbvGr', 'LotArea', 'BedroomAbvGr', 'Neighborhood_NridgHt']

	features	coefficient
3	TotRmsAbvGrd	0.396019
18	KitchenAbvGr	-0.242668
15	LotArea	0.213286
21	BedroomAbvGr	-0.165940
16	Neighborhood_NridgHt	0.136034

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?

Asnwer:

For a model to be robust and generalisable, it has to be a good fit:
* less difference between training and testing scores.
* R2 score similar to penalized R2 score.
* all data points are considered with least residual sum of square values.
* match pattern of data points for interpolation preidction.

implication of model accuracy:
* R2 score: score of train and test set more close to 1, more robust is the model.
* mean absolute error: less is the score for train and test set, more good is the prediction ability of the model.
---
* R2 score --
    * training : 0.89875
    * testing : 0.82751

* Mean Absolute Error --
    * training : 0.02181
    * testing : 0.03231