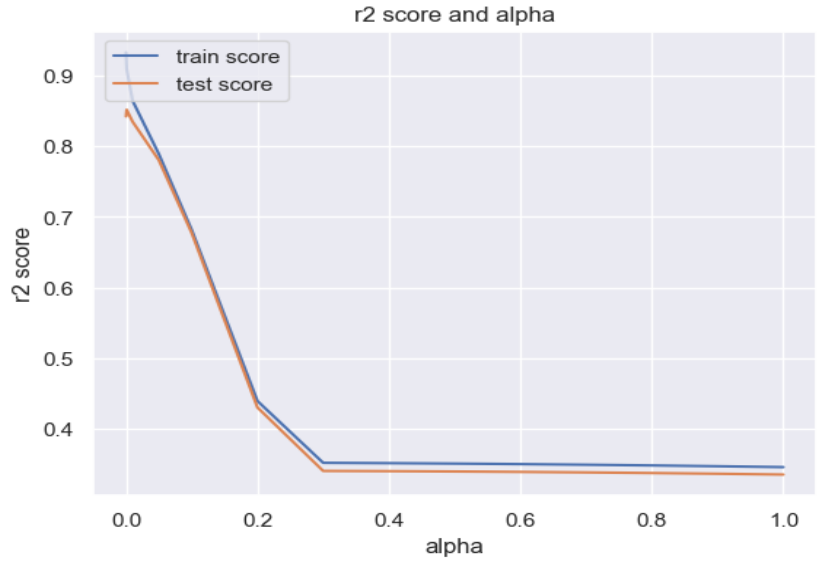
Question 1

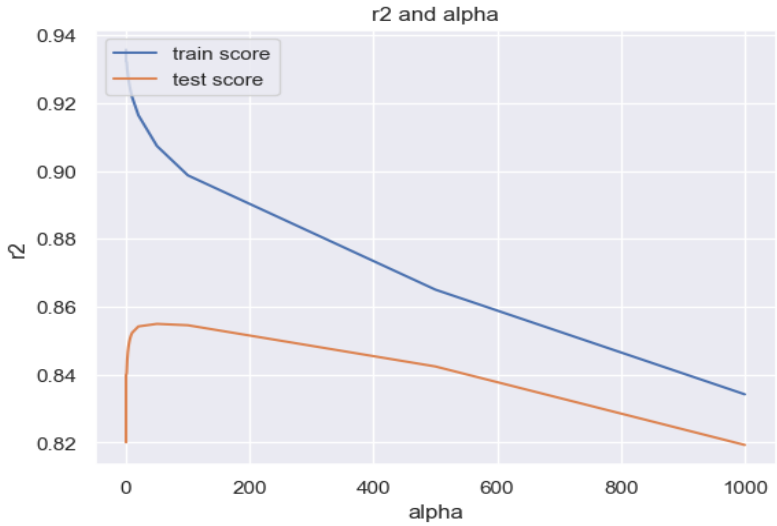
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer ::

For Lasso we have taken optimal alpha value as .01. With which we are able to strike a balance between Train and Test score.



For Ridge we have decided to go for a higher alpha value which is 500. As we can see in below diagram Train and Test Score stabilises after 500.



If we double the value of alpha in Lasso and Ridge regression it will decrease r2 scores further. Hence we can conclude that if we go beyond the optimal value model will become simpler but it may cause underfitting.

Important predictor variable in Lasso model after the change ::

GrLivArea

OverallQual
OverallCond

GarageArea

Fireplaces

Important predictor variables in Ridge model after the change ::

OverallQual

GrLivArea

OverallCond

GarageArea

1stFlrSF

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 ::

Model optimization can be considered as a complex problem and we need to keep Occam's Razor principle whenever we find a solution to a problem. As the principle says, among all the solutions to a problem the simplest one will be the best solution. Same principle can be used while choosing the best model. Though both Lasso and Ridge regularizes a model to avoid overfitting. Advantage of Lasso regression is it would ignore insignificant variables by making coefficient as zero. Thereby it helps to keep the model simple with minimum variables.

There are a lot of features present in Housing Sales Price business problem. It is clear that some of those variables are not having direct impact on Sale Price. However Ridge Regression considers all those variables hence it will be difficult to explain / understand the model. Whereas Lasso Regression considers only important features and it helps Analyst to clearly explain the model functioning to business. **So in our case Lasso regression model would be a better choice than Ridge Regression considering both are giving almost same performance.**

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 ::

5 most important variables as per Lasso regression before exclusion are:

GrLivArea

OverallQual

OverallCond

GarageArea

BsmtFullBath

	Variable	Coeff
0	constant	12.158
13	GrLivArea	0.128
4	OverallQual	0.122
5	OverallCond	0.055
21	GarageArea	0.040
14	BsmtFullBath	0.028

These variables have been removed using below code.

```
]: X_train_new = X_train.drop(['GrLivArea','OverallQual','OverallCond','GarageArea','BsmtFullBath'],axis = 1)
   X_test_new = X_test.drop(['GrLivArea','OverallQual','OverallCond','GarageArea','BsmtFullBath'],axis = 1)
```

We observed a decrease in R2 Score for both train and test. We got r2 scores for train and test as 79.5 and 78 respectively.

A new model has been created using lasso regression and below features turned out to be important.

BsmtFinSF2

TotalBsmtSF

BsmtFinSF1

FullBath

Electrical_SBrkr

	Variable	Coeff
0	constant	12.192
8	BsmtFinSF2	0.128
9	TotalBsmtSF	0.128
7	BsmtFinSF1	0.040
16	FullBath	0.032
170	Electrical_SBrkr	0.028

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 ::

Quite often real world application of a machine learning model is on unseen data. So it is very important that model performance on Test data should be in line with performance on Training data. We should refrain model learning from Outliers so that the right weightage can be given to right variables. This can be achieved only if the model is simple enough to be flexible to avoid overfitting and underfitting. In other words it is imperative to maintain bias vs variance trade off.

- A model is robust if the variation in the data does not affect its performance much.
- A generalizable model can adapt properly to new and unseen data which has similar properties of the training data. Only generalizable models can be used in real world scenarios.

High accuracy can be attained by a complex model but it will be at the expense of decrease in variance. Bias may increase if the model is simple so is accuracy but variance will be less for a simple model. So it is important to strike a balance between bias and variance.