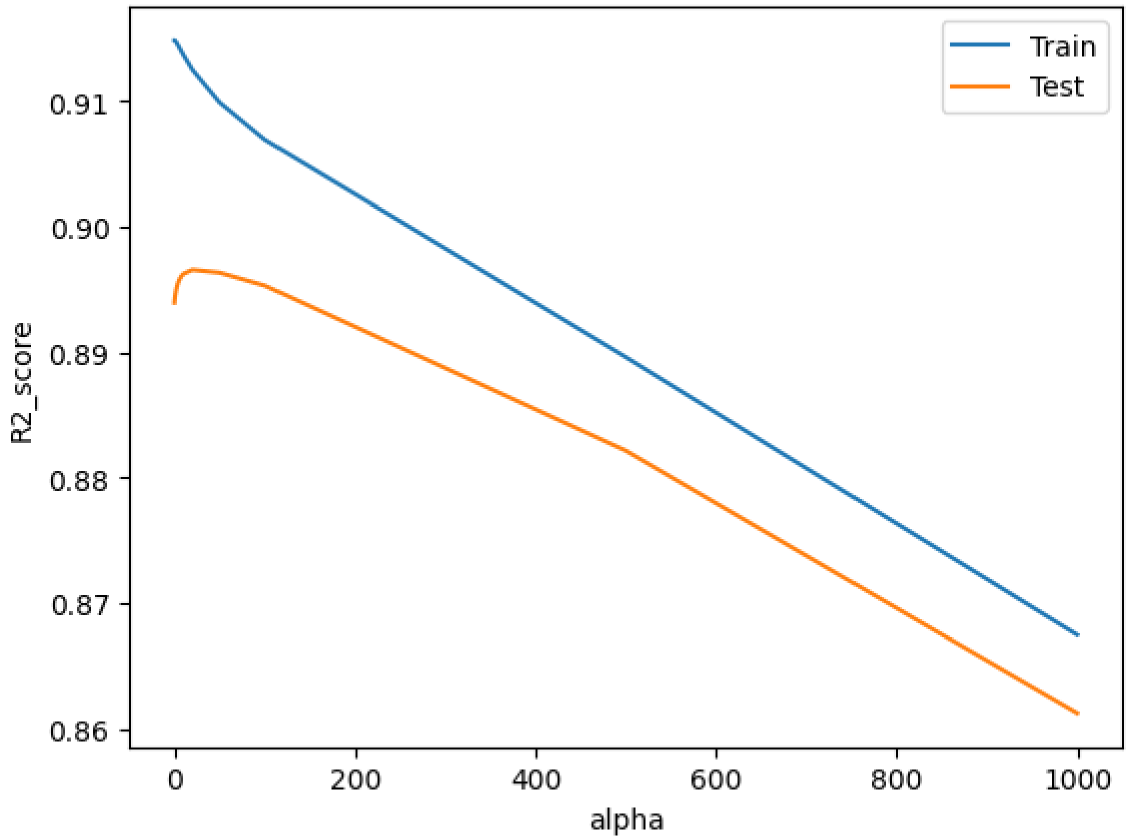
Question 1

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?

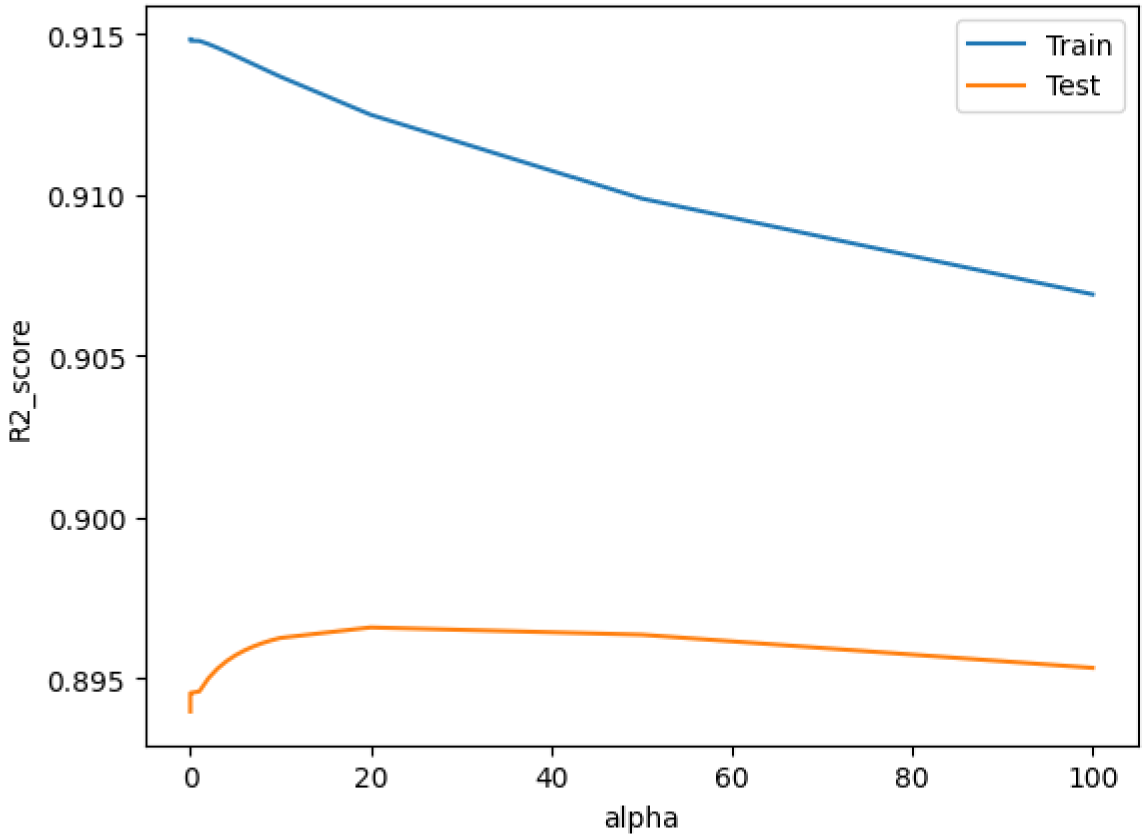
Answer:

- Optimal Value of lamda for Ridge : 20
- Optimal Value of lamda for Lasso : 0.001

Mean R2 score for train and test data for Ridge:



Mean R2 score for train and test data for Lasso:



Evaluation metrics for optimal values of alpha:

	Ridge	Lasso
R-Squared (Train)	0.911932	0.909664
R-Squared (Test)	0.888148	0.886552
RSS (Train)	11.498830	11.795021
RSS (Test)	5.699459	5.780742
MSE (Train)	0.011262	0.011552
MSE (Test)	0.013012	0.013198
RMSE (Train)	0.106124	0.107482
RMSE (Test)	0.114072	0.114883

Coefficient for Lasso model:

```
PavedDrive_Y              0.086
GarageFinish_Unf          0.075
GarageFinish_RFn          0.072
GarageFinish_No Garage    0.048
GarageType_No Garage      0.043
Name: Lasso_Coeff, dtype: float64
```

If we choose to double the value of alpha for both ridge and lasso, below are the results:

	Ridge	Lasso
R-Squared (Train)	0.910250	0.906836
R-Squared (Test)	0.887716	0.885208
RSS (Train)	11.718442	12.164195
RSS (Test)	5.721457	5.849266
MSE (Train)	0.011477	0.011914
MSE (Test)	0.013063	0.013354
RMSE (Train)	0.107133	0.109151
RMSE (Test)	0.114292	0.115562

In case of ridge it will lower the coefficients and in case of Lasso there would be more less important features coefficients turning 0.

The most important predictor variable after the change is implemented are those which are significant, except the values of coeff. has increased.

```
PavedDrive_Y              0.090
GarageFinish_Unf          0.076
GarageFinish_RFn          0.066
GarageFinish_No Garage    0.042
GarageType_No Garage      0.037
Name: Lasso_Coeff, dtype: float64
```

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:

Metrics for both Ridge and Lasso are comparable, hence the model we will choose to apply will depend on the use case.

- If we must include feature selection, then we will use **Lasso**.
- If we don't want to get too large coefficients and reduction of coefficient magnitude is the objective, then we should use **Ridge Regression**.

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:

Top 5 predictor variables for Ridge model:

GarageType_Detchd	0.082
GarageType_BuiltIn	0.073
GarageType_Attchd	0.069
KitchenQual_TA	0.051
KitchenQual_Gd	0.045

Top 5 predictor variables for Lasso model:

GarageType_Detchd	0.087
GarageType_BuiltIn	0.074
GarageType_Attchd	0.072
KitchenQual_TA	0.051
KitchenQual_Gd	0.044

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:

A robust model remains stable despite variations in the data, ensuring consistent performance.

Generalizability implies a model's ability to adapt well to new, unseen data from the same distribution used for training.

To ensure robustness and generalizability, it's crucial to prevent overfitting. Overfitting occurs when a model captures training data patterns excessively but fails to generalize to new data. Also, a model should strike a balance between complexity and simplicity, excessive model complexity leads to high accuracy on training data but may hinder performance on unseen data.

Impact on Accuracy:

- High complexity models may exhibit impressive accuracy on training data but struggle with new data due to overfitting.
- Striking a balance involves reducing variance (complexity) to enhance generalizability, introducing some bias, which may reduce accuracy.
- Achieving a balanced model involves optimizing accuracy without compromising robustness and generalizability.
- Regularization techniques provide a means to fine-tune this balance for optimal model performance.