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 alpha using 70/30 ratio of train/test data:

- Ridge: 9.0
- Lasso: 0.001

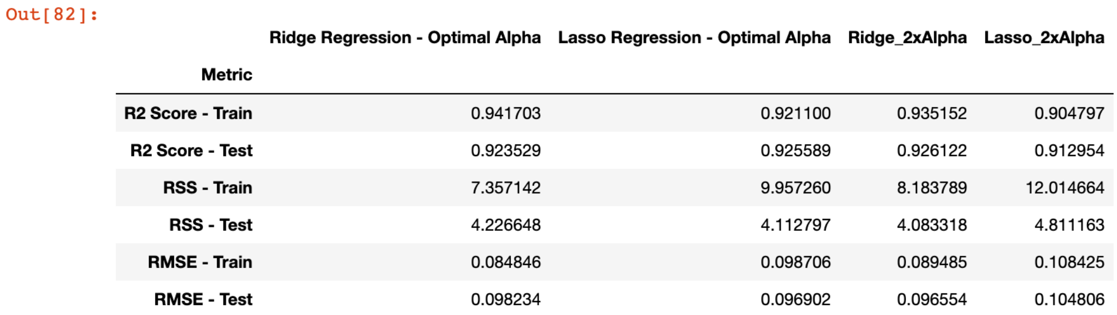
After doubling the alpha value for both ridge and lasso, following changes were observed:
Changes in Ridge metrics:

- Train set R2 decreased from 0.941703 to 0.935152
- Test set R2 increased from 0.923529 to 0.926122
- Train RSS increased, Test RSS slightly decreased
- Train RMSE increased, Test RMSE slightly increased

Changes in Lasso metrics

- Train set R2 decreased from 0.921100 to 0.904797
- Test set R2 decreased from 0.925589 to 0.912954
- Train and Test RSS slightly increased
- Train and Test RMSE slightly increased

Output from the notebook:



Out[82]:

	Ridge Regression - Optimal Alpha	Lasso Regression - Optimal Alpha	Ridge_2xAlpha	Lasso_2xAlpha
Metric				
R2 Score - Train	0.941703	0.921100	0.935152	0.904797
R2 Score - Test	0.923529	0.925589	0.926122	0.912954
RSS - Train	7.357142	9.957260	8.183789	12.014664
RSS - Test	4.226648	4.112797	4.083318	4.811163
RMSE - Train	0.084846	0.098706	0.089485	0.108425
RMSE - Test	0.098234	0.096902	0.096554	0.104806

Important predictor variables after doubling the alpha value (from both ridge and lasso models):

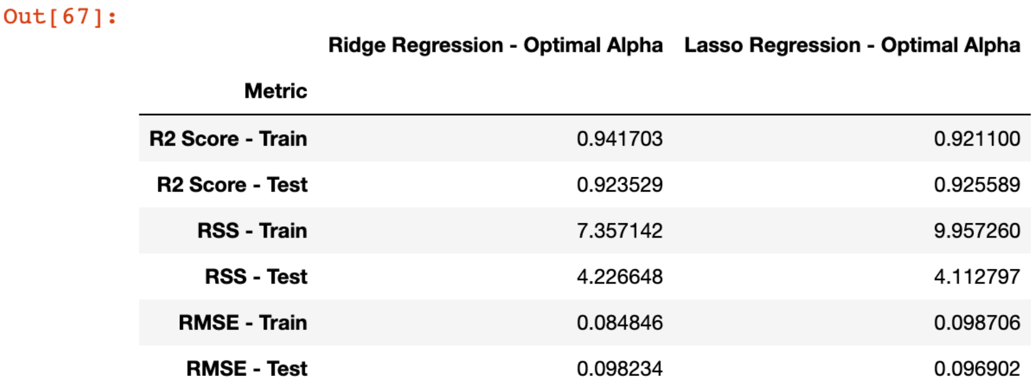
- GrLivArea
- OverallQual_8 (overall material and finish of the house in “Very Good” condition)
- OverallQual_9 (overall material and finish of the house in “Excellent” condition)
- Functional_Typ (Home with typical functionality)
- Neighborhood_Crawfor (Crawford in the neighbourhood)
- Exterior1st_BrkFace (Brick face exterior)
- TotalBsmtSF
- CentralAir_Y (home with central air conditioning)

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:

With the optimal value of lambda determined for ridge and lasso, R2 score, RSS, MSE and RMSE on the test set are almost the same for both ridge and lasso models. Most of the top predictor fields are very closely matching as well.



Out[67]:

Metric	Ridge Regression - Optimal Alpha	Lasso Regression - Optimal Alpha
R2 Score - Train	0.941703	0.921100
R2 Score - Test	0.923529	0.925589
RSS - Train	7.357142	9.957260
RSS - Test	4.226648	4.112797
RMSE - Train	0.084846	0.098706
RMSE - Test	0.098234	0.096902

Based on metric comparison of Ridge and Lasso models using the optimal value of lambda, Lasso appears to be performing better than Ridge. R2 score for Lasso model is greater than R2 score for Ridge model on test data. RSS and RMSE on test set are lower for Lasso model than Ridge model as well.

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 Lasso predictors were:

- OverallQual_8
- OverallQual_9
- GrLivArea
- Neighborhood_Crawfor
- Functional_Typ

After dropping the above 5, another lasso model was created and the following 5 new predictor variables were identified.

- 2ndFlrSF
- Exterior1st_BrkFace

- MSSubClass_70
- 1stFlrSF
- Neighborhood_Somerst

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 model is considered as robust and generalised based on its performance with any variation in the data and its ability to properly adapt new and unseen data.

In order to ensure that a model is robust and generalised, we need to perform the following:

1. Outliers treatment for both independent variables and target variables
2. Regularization to avoid overfitting

Following are the implications of the same for the accuracy of the model:

1. Outliers are data points which are distant from most of the other data points. As the models are sensitive to the distribution of the data points, outliers can result into less accurate models. We use log transform, capping between percentiles or using standard deviation to treat the outliers.
2. An overfit model (complex model) will have high variance and a smallest of change in the data affects the accuracy of the model. An overfit model will have high accuracy on train data but will fail with unseen test data. So a balance between the model complexity and accuracy needs to be maintained. It can be achieved by Ridge/Lasso regression (regularization technique).