Problem Statement - Part II Question

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?

Answer1:

Optimal value of alpha for Lasso: 0.001

Optimal value of alpha for ridge: 100

After double of alpha for ridge and lasso i.e. 200 and 0.002

For Ridge Regression:

- Model coefficients is decreased when we use double alpha value in Ridge Regression model.
- There is a drop in R2 score also of train and test data
- R2 score train : 0.90489 to 0.90114
- R2 score test : 0.91003 to 0.90864

For Lasso :

- Model coefficients for 'GrLivArea', 'OverallQual', 'Fireplaces' decreased.
- Model coefficients for 'OverallCond', 'SaleCondition_Partial', 'TotalBsmtSF', 'SaleCondition_Normal', 'Foundation_PConc', 'LotArea', 'Exterior2nd_Wd Sdng' is increased.
- R2 score is decresed for tarin data but increased for test data.
- R2 score train : 0.9075 to 0.9068
- R2 score test : 0.9085 to 0.9106

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 2:

We have determined optimul value of lambda for ridge and lasso regression i.e. (100 for ridge and 0.001 for lasso). So I am using ridge regression for my final model as it gives good R2 score on train and test data i.e. (90.48% for train data and 91.0 % for test data) and it has lower rmse value for test data.

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 3:

Top five features are : 'GrLivArea', 'OverallQual', 'OverallCond', 'SaleCondition_Partial', 'TotalBsmtSF'.

After dropping them the lasso model's accuracy reduced from 90.75% to 88.6% for train data and 90.85% to 87.7% for test data.

Now top 5 features are: '2ndFlrSF', '1stFlrSF', 'BsmtFinSF1', 'KitchenQual', 'Fireplaces'.

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 4:

To make model robust and generalisable we check model has good accuracy score and handled Asssumptions properly:

Accuracy score :-

Our Model accuracy is coming 90.48%(Train) and 91.0%(Test) which indicate our model has good accuracy score.

Assumptions Analysis :-

1. Residual analysis proof :- After building ridge model with alpha value 100 we have plot a graphs between error term and residual. Which shows our error distribution is normally distributed across 0. Which indicate our model has handled the assumption of error normal distribution properly.
2. Residual vs Prediction plot :- We have plot this plot which shows that our predictions has randomly scattered around the 0 horizontal line. Which indicate our model has handled the assumption properly.

Thus we are sure that model is robust and generalisable.