Advanced Regression Assignment Questions

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:
Subjective to the assignment dataset, the optimal alpha value for Ridge is **10** and Lasso regression came out to be **100**. And the accuracy of Lasso model was relatively higher than Ridge after looking at all the evaluation metrics like *R2 score, RSS and RMSE.*

If we try to use double value of alpha, which would be 20 and 200 in this case, then the model will have underfitting as a side effect as all the features/columns/fields will get very less coefficient values due to very high alpha value.

In the assignment, the observed changes are

1. R2 score of Ridge model dropped from *0.94 to 0.93*
2. R2 score of Lasso model dropped from *0.93 to 0.92*

The most important predictor variables after the alpha value change are as follows:

1. *OverallQual_9*
2. *OverallQual_8*
3. *GrLivArea*
4. *Neighborhood_Crawfor*
5. *Functional_Typ*

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:

I will choose *Lasso Regression Model* to apply on the dataset and predict the SalePrice as

1. Lasso model had better evaluation scores (*R2 score, RSS and RMSE scores*) when compared with Ridge Regression Model
2. Lasso model eliminates features/fields while determining coef because the coef value in Lasso model can tend to be zero(0) it is a much simpler Regression Model.

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:

After removing the 5 top most predictors from the dataset, the Lasso Regression Model is fitted again and new coefficients are gathered. The top 5 of these new remaining predictors are as follows:

1. Co*ndition2_PosA*
2. *RoofMatl_WdShngl*
3. *RoofMatl_CompShg*
4. *RoofMatl_WdShake*
5. *2ndFlrSF*

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 robust if any variation in data does not effect the efficiency or accuracy performance of the model.

You can make sure that the model in question is Generalised by validating that all overfitting is handled. In Ridge/Lasso Regression model, we use alpha value to achieve this. But overdoing this also has side-effects, you can end up underfitting the dataset

Accuracy of a model increases as we try to resolve overfitting using alpha generalization in our model. And eventually if you use very high alpha values, the accuracy again starts dropping, which can also be observed in Question-1’s Answer when we doubles the alpha value of our models. The R2-score dropped when the Alpha value was doubled indicating that the relation between generalization (alpha value in this context) and accuracy is a *parabolic* curve.