Subject Question and Answer – Rohit Keshari

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:

The optimal value of alpha for ridge and lasso regression are:

Alpha for Ridge: **20**

Alpha for Lasso: **0.001**

If we double the value of alpha. Then,

Alpha for Ridge: **40**

Alpha for Ridge: **0.002**

In the model if you choose to double the value of alpha for Ridge and Lasso below effect will be seen.

Ridge - The larger is the alpha, the higher is the smoothness constraint. So, the smaller the value of alpha, the higher would be the magnitude of the coefficients.

Lasso - The alpha value increases, more features have a coefficient of 0.

Below are the five most important predictor variables after the change is implemented:

Ridge	
MSZoning_C (all)	-0.231103
Neighborhood_Edwards	-0.056137
MSZoning_RM	-0.055723
SaleCondition_Abnorml	-0.046907
LandContour_Bnk	-0.044877

Lasso	
OverallQual	0.061406
Neighborhood_Crawfor	0.053635
Condition1_Norm	0.049430
OverallCond	0.045502
Neighborhood_Somerst	0.039296

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:

After determining the value of alpha for both lambda for ridge and lasso.

I will go for Lasso as :

1. It helps in feature elimination (Lasso shrinks some of the variables coefficients to 0 thus performing variable selection)
2. The model will be more robust
3. It's simpler while balancing the 'bias-variance' trade-off

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 building the model below are five most important predictor variables in the lasso model:

1. Neighborhood
2. Condition1
3. Exterior1st
4. OverallQual
5. Functional

I have removed above columns data from train data set and ran the model again and below are the five most important predictor variables now:

1. BsmtQual_Ex
2. KitchenQual_Ex
3. MSZoning_FV
4. OverallCond
5. GarageCars

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 and generalizable when there is no impact by outliers in training data set. Means that if any new data point is incoming so model should predict correctly, if model is not predicted correctly then means it impacted by outliers and model is overfitted.

The model should also be generalisable so that the test accuracy is not less than training score also model should be accurate new unseen data point which were not used during training the model.

Too much weightage should not give to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset.

This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations).

This would help standardize the predictions made by the model. If the model is not robust , it cannot trust for predictive analysis.

--------End---------