Assignment Part -II - Subjective Questions and Answers

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 we got in case of Ridge and Lasso is:

1. Ridge - 7.0
2. Lasso - 0.0001

If we double the value of alpha for both ridge and lasso, following changes are observed:

- For Ridge: there is a slight increase in the mean squared error and slight decrease in r2 score of train and test set.
- For Lasso: there is a slight increase in the mean squared error for train set while there is slight decrease in test set. Similarly, for r2 score of train set slightly decreases whereas there is slight increase in r2 score of test set.

The most important predictor variables after the changes are implemented are:

For Ridge:

1. GrLivArea: Above grade (ground) living area square feet
2. OverallQual: Rates the overall material and finish of the house
3. Neighborhood: Physical locations within Ames city limits
4. TotalBsmtSF: Total square feet of basement area
5. OverallCond: Rates the overall condition of the house

For Lasso:

1. MSZoning: Identifies the general zoning classification of the sale.
2. OverallQual: Rates the overall material and finish of the house
3. OverallCond: Rates the overall condition of the house
4. RoofMatl: Roof material
5. SaleType: Type of sale

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

Ridge regression has better Test R2 score overall as compared to Lasso Regression and also there is less difference between R2 score of Train and Test set; Hence we will go ahead with Ridge 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

We dropped the top 5 most important predictor variables in the lasso model and again created again model and got the below five most important predictor variables:

1. GrLivArea: Above grade (ground) living area square feet
2. Neighborhood: Physical locations within Ames city limits
3. BsmtCond: Evaluates the general condition of the basement
4. Exterior1st: Exterior covering on house
5. MSSubClass: Identifies the type of dwelling involved in the sale

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

- The model is expected to be as simple as possible and simpler models are considered as more 'generic', though its accuracy will be decreased but it will be more robust.
- This can be understood from the Bias-Variance trade-off. The simpler the model the more the bias but less variance becoming generalizable. Whereas the complex model will have high variance and low bias.
- Sometimes underfitting and overfitting are the problems associated with the model. Hence, it is important to have balance in Bias and Variance to avoid such problems. This is possible with "Regularization".
- Regularization helps in managing the model complexity by essentially shrinking the coefficients towards zero. This avoids the model becoming too complex, thus reducing the risk of overfitting. • Regularization method should be used to keep the model optimum simpler. It penalizes the model if it becomes more complex.
- Regularization method helps to achieve the Bias-Variance trade off. It compromises by increasing bias to a optimum position where Total Error is minimum.
- This point also known as Optimum Model Complexity where Model is sufficient simpler to be generalisable and also complex enough to be robust.
- Making a model simple lead to Bias-Variance trade off.