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 to 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 for ridge and lasso regression is 10 and 0.01 respectively.

With increase in value of alpha, variance reduces with a slight adjustment in bias.

If we double the value of alpha for lasso to 0.02, the R2 for train and test will become 0.83 and 0.837. Increase in alpha leads to decrease in r2 value, while coefficient of more predictor variables will become zero, which leads to decrease in Predictor variables to 16 from 19. Most important predictor variables will be OverallQual, GrLivArea, GarageArea and Fireplaces, OverallCond, BsmtFullBath, TotalBsmtSF, 1stFlrSF, WoodDeckSF, LotArea.

If we double the value of alpha for ridge to 20, the R2 for train and test will become 0.908 and 0.883, which is just a slight decrease. Increase in alpha leads to slight decrease in r2 value.Most important predictor variables will be 'OverallQual', 'Neighborhood_Crawfor', 'GrLivArea', 'Neighborhood_NridgHt', 'Condition1_Norm', 'Neighborhood_Somerst', '2ndFlrSF', 'OverallCond', 'Exterior1st_BrkFace', 'MSZoning_RL'.

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:

It is important to have good accuracy of model along with significant feature selection. In this house price prediction model, I would prefer to choose Lasso.

Ridge regression uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it makes the variables exactly equal to 0, to handle high variance.

Hence, lasso regression also serves as a variable shrinkage method, whereas ridge regression does not.

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:

Those 5 most important predictor variables that will be excluded are :-

1.OverallQual 2. GrLivArea 3. GarageArea 4. OverallCond 5. Fireplaces

After excluding these 5 variables, if model has to be re-built, then other significant variables will be : -

1stFlrSF, 2ndFlrSF, TotalBsmtSF, GarageType_Attchd, BsmtFullBath

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 should be simple, so that it can be robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e., the accuracy does not change much for training and test data.