Assignment 2 - House price prediction

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?

 Solution:

Optimal value of alpha for Ridge regression is **10** as train R2 score in ridge is 0.89 and test R2 score ridge is 0.87 for this value of alpha.

Optimal value of alpha for Lasso regression: **0.001** as train R2 score lasso is 0.87 and test R2 score lasso is 0.86 for this value of alpha.

The changes in the model if you choose double the value of alpha for both ridge and lasso:

R2 score: Upon doubling the values of alpha in both the Ridge and Lasso regressions, the corresponding R2 scores decreases. The difference between training and test R2 scores does not become significant after doubling the alpha hyperparameter though.

Coefficients: We notice the reduction in the coefficient values in both the ridge and lasso regressions, upon doubling the alpha hyperparameter. That clearly indicates that both the regressions are applying regularization more rigorously, upon increasing alpha values.

The most important variables for Ridge regression after doubling the alphas:

1. OverallQual
2. GarageCars
3. GrLivArea
4. FullBath
5. 2ndFlrSF

The most important variables for Lasso regression after doubling the alphas:

1. OverallQual
2. GrLivArea
3. GarageCars
4. FireplaceQu
5. KitchenQual

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?

 Solution:

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance, and making the model interpretably. Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation.

Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. 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 make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the 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?

Solution:

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

1. GrLivArea
2. OverallQual
3. GarageCars
4. FireplaceQu
5. OverallCond

The five most important predictor variables now:

1. FullBath
2. Fireplaces
3. 2ndFlrSF
4. KitchenQual
5. BsmtQual

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?

Solution:

The model should be as simple as possible, though its accuracy will decrease but it will be more 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 and more generalizable. 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.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data