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

Best alpha Value for Lasso: 0.001

Best alpha value for Ridge: 1.0

If we double the value of alpha for ridge and lasso:

Lasso Alpha – 0.001 x 2 = 0.002

Ridge Alpha – 1.0 x 2 = 2.0

For Lasso Regression:

Before doubling –

train r2: 0.8663072943057095

test r2: 0.748275574339103

After doubling –

train r2: 0.820154767812517

test r2: 0.8440154496909013

There is a decrease in the r2 scores on train and test set indicating this is not a good alpha value.

More features can also be removed if we increase alpha in a lasso regression model.

For Ridge Regression:

Before doubling –

train r2: 0.8784936256234364

test r2: 0.8001645517702862

After doubling –

train r2: 0.8630065049357419

test r2: 0.8329158487186958

Coefficients are increasing as shown in the notebook. R2 score of training data has decreased.

Top Features: OverallQual, BsmtFinSF1, MSZoning_RM, MSZoning_FV , MSZoning_RH, PoolQC_Gd, SaleCondition_Partial, BsmtFinType1_No_Basement, GarageType_No_Garage, GarageCond_Fa

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 would choose Lasso Regression model as it gives us feature selection capabilities. It penalizes unnecessary variables without affecting model accuracy. This makes for a more generalized model with better accuracy on unseen test data.

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

BsmtFinSF2, MSZoning_RL, SaleCondition_Partial, GarageType_No_Garage, PoolQC_Gd

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

For making a model more robust and generalisable:

- Model accuracy should be > 70-75%. Lower than this indicates poor accuracy and we should continue to tune our variables or hyper parameters. Both our ridge and lasso model accuracies fall in this range.
- P-values of all the features should be close to 0 or at lease < 0.05 to indicate that the independent variable is significant.
- VIF values of all selected features should be less than 10. Ideally they should be less than 5. But greater than 10 means those variables should definitely be dropped or refined.