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:

You can check in the code that, the optimal value of alpha in Ridge is 0.8392 and Lasso is 0.8453.

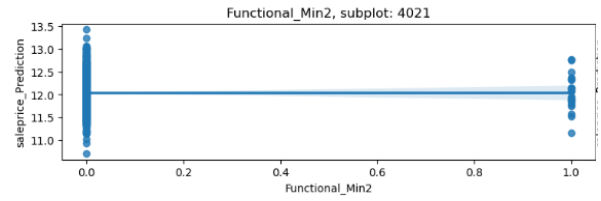
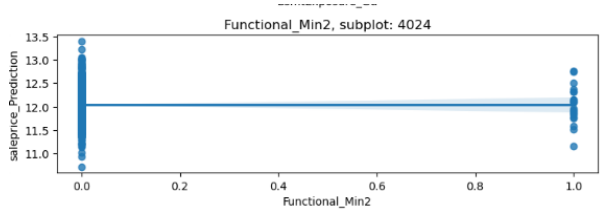
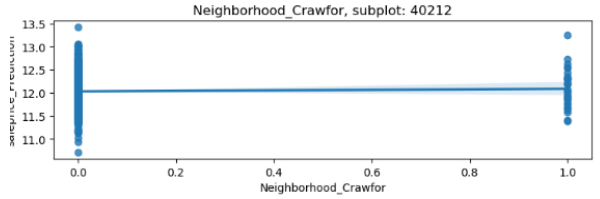
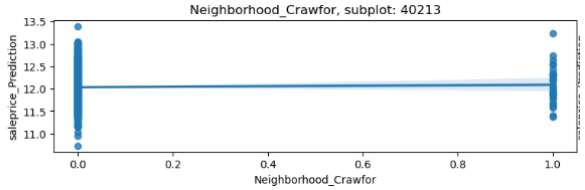
The important predictors are,
'Functional_Min2','GarageCond_Po','Utilities_AllPub''MSZoning_RH','Condition2_Norm','Condition2_PosA','RoofMatl_Membran', 'Condition2_Feedr','Functional_Typ', 'Neighborhood_NridgHt','GarageQual_Ex','Neighborhood_Crawfor', 'RoofMatl_CompShg','Neighborhood_StoneBr', 'RoofMatl_WdShngl'

When changed the alpha values to double (the same you can check at the end of my code) the alpha value changed for Ridge is 0.8455 from 0.8392.

The important predictors now are, 'SaleType_Oth', 'BsmtExposure_Gd', 'Exterior1st_BrkFace', 'Functional_Min2', 'Condition2_PosA','MSZoning_RH','Condition2_Feedr','Condition2_Norm','GarageQual_Ex', 'RoofMatl_CompShg', 'Functional_Typ','Neighborhood_NridgHt', 'Neighborhood_Crawfor', 'RoofMatl_WdShngl', 'Neighborhood_StoneBr'.

If we observe the important predictors have changed. But there are few common like, Neighborhood_Crawfor, Neighborhood_NridgHt, Neighborhood_StoneBr, Condition2_Feedr, Condition2_Norm, Condition2_PosA, Functional_Typ, GarageQual_Ex, RoofMatl_CompShg, RoofMatl_WdShngl

Also if we observe the correlation of these variables for before and after change in alpha as below,

Before doubling in Ridge	After doubling in ridge
	
	

Changed the alpha and built Lasso model. Then alpha changed to 0.8453 from 0.8453.

So if we observe above, even if we change the alpha and the optimal alpha didn't change much or the correlation of the predicted values in both times remains similar.

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:

The optimal value of alpha in Ridge is 0.8392 and Lasso is 0.8453.

Ridge	Lasso
Root Mean Square Error train = 0.09775164407086612 Root Mean Square Error test = 0.14577257400058782 R-Square for training data 0.9347968815077364 R-Square for training data 0.8865350474108582	Root Mean Square Error train = 0.1172244152537997 Root Mean Square Error test = 0.1515412625160463 R-Square for training data 0.9062316278596105 R-Square for training data 0.8773770123541852

I choose the Lass model. Because the errors in predicted values is less when compared to Ridge.

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?

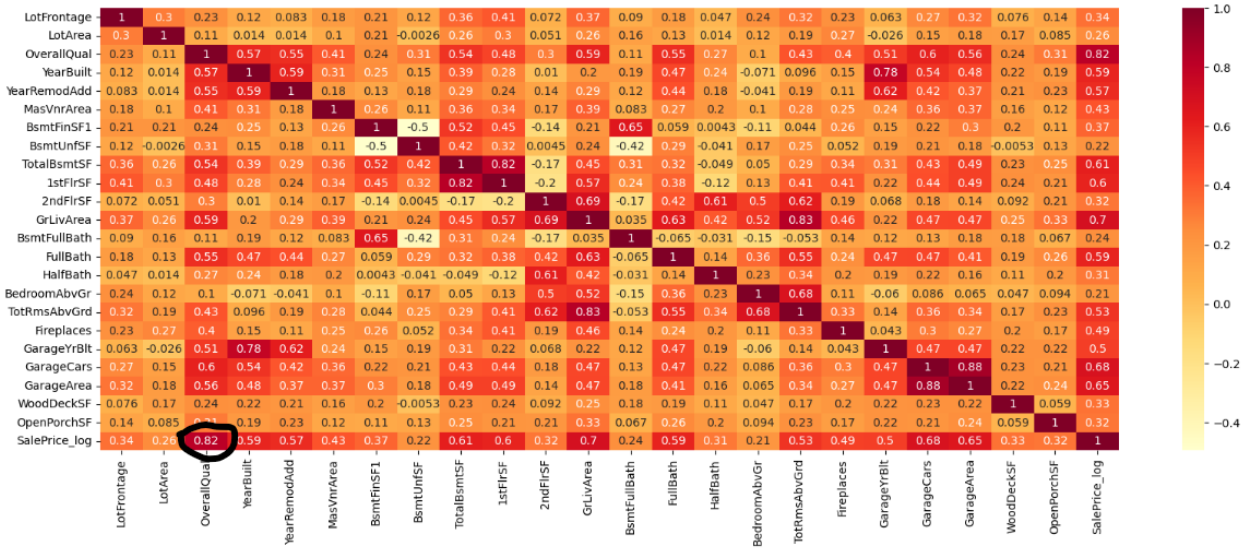
Answer:

The top 15 important predictors and their coefficient values in Lasso are:

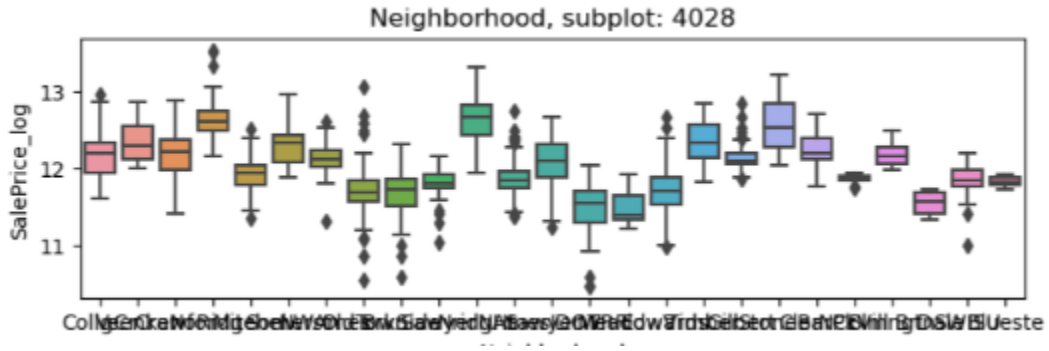
[('HeatingQC_Ex', 0.027053904267803776), ('BsmtFullBath', 0.032135579981689345), ('KitchenQual_Ex', 0.035004183294911125), ('BsmtCond_TA', 0.042263112020467626), ('Condition1_Norm', 0.044811675762985614), ('BsmtCond_Gd', 0.046425038427234666), ('BsmtExposure_Gd', 0.0496653657361928), ('Exterior1st_BrkFace', 0.06157616709332128), ('GarageCars', 0.06271569550649164), ('OverallQual', 0.06950731730733331), ('BldgType_1Fam', 0.0705093539513192), ('Functional_Typ', 0.07327506235301694), ('Neighborhood_NridgHt', 0.08347206323842636), ('Neighborhood_StoneBr', 0.0999830859402537), ('Neighborhood_Crawfor', 0.14215457887291072)]

From the incoming data, Neighborhood, Functional, OverallQual, BldgType and BsmtCond_Gd are the 5 most important predictors as per the coefficient values.

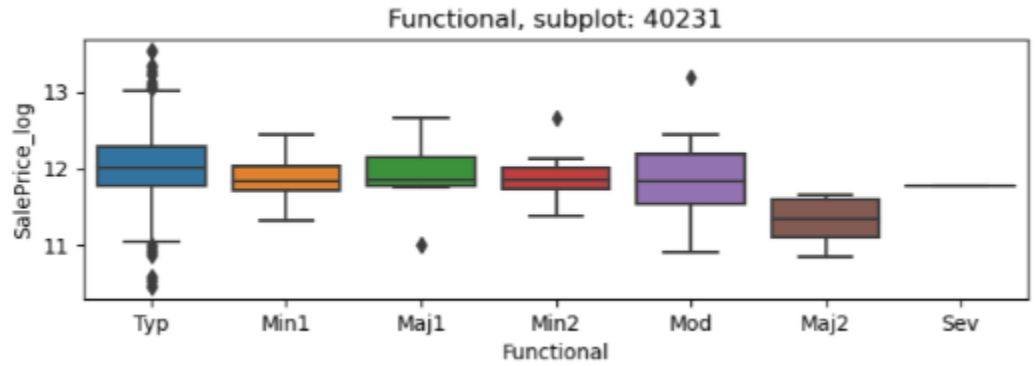
The below heat map shows the same for OverallQual variable.



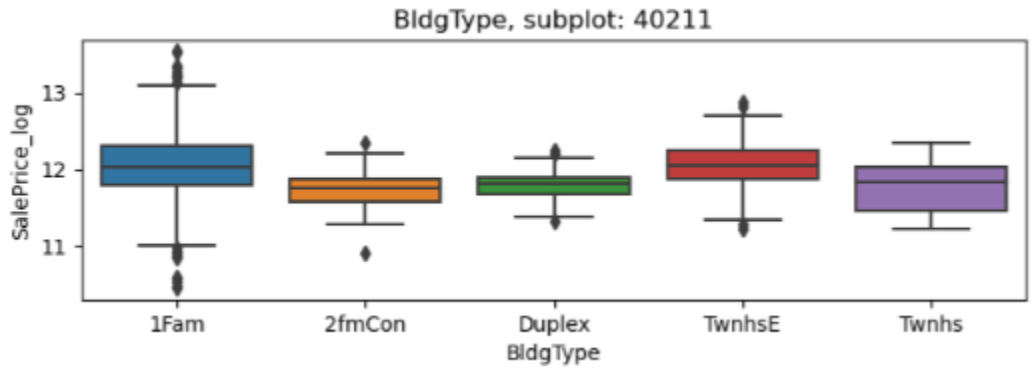
Variance in the Neighborhood variable in the incoming data,



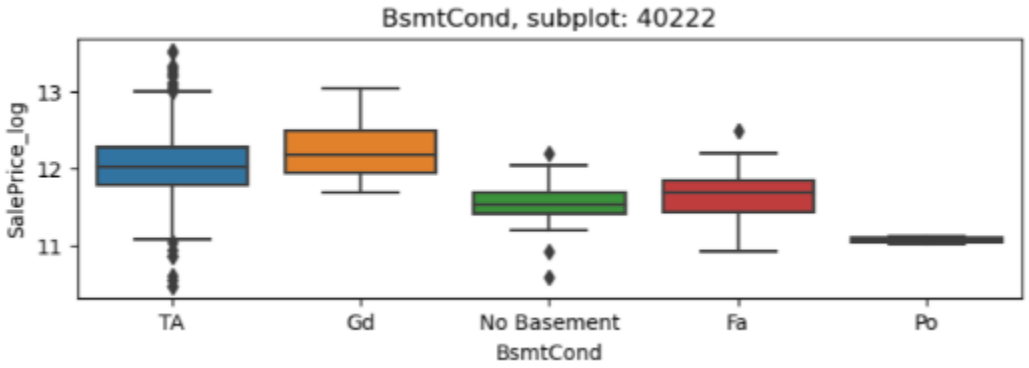
Variance of the Functional variable in the incoming data,



Variance of the BldgType variable in the incoming data,



Variance of the BsmtCond variable in the incoming data,



With the coefficient values generated and the variance in the data in these variables shows above variables are most important for prediction.

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:

We should be building the model on the train data and evaluate the accuracy on the test data (unseen data to model). If the accuracy is not good/accepatable in test vs train data then there is problem of overfitting or model is more complex. In other words RMSE has huge difference between the train and test data predictions then it is an indication of model is not generilised. Model cannot learn from seen data and cannot apply them in the unseen data. Model can work only on the seen data better.

In practice, we should avoid this issue of overfitting and make the model generalisable. This we can achieve by regularization. Regularization is a technique of hyperparmeter tunning while building the model. We choose some random values, mostly 0.0005 to 1. Add them a hyper parameters and find the best value of hyper parameter. The algorithm achieves this by compromising on the bias and not much effecting the variance. This optimal values called alpha is applied on the model.

In our current case, as we saw, the RMSE value in the train and test data are as below,

Root Mean Square Error train = 0.09037520471433773

Root Mean Square Error test = 0.14111528425664943

R-Square for training data 0.9442661794995015
R-Square for test data 0.8936694155726671

You can observe that errors increased in the test data that in the train data. This means the model could perform well in the same way in the test data (unseen). So it is clear problem of overfitting.

After we did the regularization, this problem reduced. After the Ridge regularization,
Root Mean Square Error train = 0.09775164407086612
Root Mean Square Error test = 0.14577257400058782
R-Square for training data 0.9347968815077364
R-Square for training data 0.8865350474108582

After the Lasso regularization,
Root Mean Square Error train = 0.1172244152537997
Root Mean Square Error test = 0.1515412625160463
R-Square for training data 0.9062316278596105
R-Square for training data 0.8773770123541852

The above metrics shows that the errors in Trains data increased after regularisation. And the errors on the test data remain similar. That means the model is generalised.