Problem Statement - Part II

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 1

We have used Lasso and Ridge Regression for our analysis:

1. In case of Lasso Regression, we generated the model_cv using the list of alphas to tune. Post generating the model_cv, we fetched the best parameters using(model_cv.best_params_) that gave us 0.001 as the best alpha. But we observe that the negative mean absolute error stabilizes at 0.4 and after. So we choose alpha to be low to balance off the tradeoff between the bias variance and thus getting the coefficients of smallest of features. Thus we choose 0.001 which makes most of the insignificant variables coefficients to zero.
2. In case of Ridge Regression, we generated the model_cv using the list of alphas to tune. Post generating the model_cv, we have fetched the best parameters using(model_cv.best_params_) that gave us 10 as the best alpha. But we see Negative mean absolute error stability at around 2 alpha but we use 10 that we got as best param.

Now to answer the question on doubling the alpha value, in case of ridge regression, then this results in increase of shrinkage and hence more coefficients of the variables tend to become zero which results in more error in prediction. Similarly in Lasso if we double the alpha value we will make more variable coefficients to tend to zero resulting in decrease of R2 value

The most significant variables after doubling the alpha values in Lasso are:

['GrLivArea', 'OverallQual', 'OverallCond', 'TotalBsmtSF', 'BsmtFinSF1', 'GarageArea', 'Fireplaces', 'LotArea', 'LotFrontage', 'BsmtFullBath', 'WoodDeckSF', 'ScreenPorch', 'KitchenQual_TA', 'MSSubClass', 'KitchenAbvGr', 'PropAge']

The most significant variables after doubling the alpha values in Ridge are:

OverallQual, Neighborhood_Crawfor, GrLivArea, OverallCond, SaleCondition_Normal, SaleCondition_Partial

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?

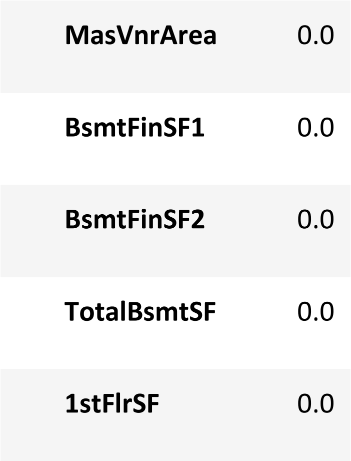
Answer 2

We see that both Lasso and Ridge regression have given us similar values of R2 for both train and test sets, but we can conclude that **Lasso** is better as it gives a zero value to all the insignificant features, making it easier to choose predictive variables/features.

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 3



MasVnrArea	0.0
BsmtFinSF1	0.0
BsmtFinSF2	0.0
TotalBsmtSF	0.0
1stFlrSF	0.0

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 4

In order to maintain accuracy so that it is not less than the training score and it is in consistency with it, we ensure the model to be robust and generalized. Basically the model should work even on other datasets apart from the ones used during training. We will have to utilize the various regularization techniques that help in maintaining the trade-off between bias and model complexity which in turn is related to the model robustness. When we have a complex model, even a slight change in the dataset will cause a lot of variance and will tend to be highly unstable. Whereas, a model that is simple, even with a change in the dataset, will not have a lot of variance. Hence, we should maintain both bias and variance to maintain accuracy in the model, reducing the errors occurring.