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

There is no such optimal value for alpha for ridge and lasso regression, It depends on the dataset and the problem we are trying to solve, the value of alpha can range from 0.0001 to 1000 normally. The optimal Alpha value is chosen for each solution using cross validation, which tries out different values and gives out the best one which minimises or maximises the given parameter. For Eg :- Mean Squared Error should be minimized where as R2 score should be maximised.

If we double the value of alpha, the coefficients tend to move to zero faster, or in other words the regularisation power increases. In case of lasso, increase in alpha, make more coefficients reaching zero. In overall doubling alpha will make the coefficients to shrink more towards zero or become equal to zero, in case of lasso regression.

The most important predictor variable will be the ones with larger coefficients.

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

After finding the optimal value of lambda (hyperparameter) using Cross Validation for ridge and lasso, the decision on choosing depends on multiple factors :-

- Whichever model gives better results with the unseen test set would be a factor.
- If there are large number of predictor variables and needed to analyse each influence of each predictor variable, then lasso would be a better option as it implements feature elimination as well, reducing the number of predictor variables, which makes the analysis easier
- If analysis of individual predictor variable is not important and only the prediction accuracy is important, Ridge would be a better option, as it takes in consideration of all the predictor variables without eliminating them.
- Ridge might be a better option if multicollinearity is present in the dataset, as it does not eliminate any variables
- Also based on computational simplicity, Ridge might be preffered.

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

If the 5 most important variables in lasso model are not available, then the lasso model has to be retrained again excluding these variables. After retraining with the new dataset, there will be another set of important predictor variables arising.

- The 5 most important predictor variables at the initial stage –

```
In [84]: lasso_coeff = pd.Series(lasso.coef_, index= X_train.columns)
         lasso_coeff.sort_values(ascending= False).head(10)

Out[84]: GrLivArea              0.188058
         TotalBsmtSF            0.145294
         OverallQual_9          0.098189
         Neighborhood_NoRidge   0.072559
         OverallQual_8          0.066038
         BsmtFullBath           0.057968
         GarageCars             0.056963
         OverallQual_10         0.053190
         OverallCond_9          0.051827
         Neighborhood_Crawfor   0.050493
         dtype: float64
```

- The 5 most important predictor variable after removing –

```
In [82]: lasso_coeff = pd.Series(lasso.coef_, index= X_train.columns)
         lasso_coeff.sort_values(ascending= False).head(10)

Out[82]: FullBath               0.104497
         LotFrontage            0.087860
         BsmtFullBath           0.066223
         LotArea                0.059392
         BedroomAbvGr           0.059106
         YearLastModified_1997  0.050789
         MasVnrArea             0.049535
         WoodDeckSF             0.048822
         Neighborhood_Crawfor   0.046267
         HalfBath               0.045204
         dtype: float64
```

It can be observed that another set of variables took place when retrained on remaining predictive variables. The R2 score had a small reduction in both train and test but was good enough.

It can also be observed that the coefficient has reduced after retraining with remaining variables.

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

- Test your model with another unseen data, if available
- Measure different model evaluation metrics
- Applying Regularisation techniques like Ridge and Lasso, which add penalty and avoids the model to be complex
- Difference of R2 score of train and test datasets should be minimum
- Feature selection
- The focus should be to have better results on test data than train data, as the model might be overfitting.