Subjective Questions

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: The optimal value for ridge and lasso regression is 100 and 0.001 respectively.

Model details for ridge and lasso regression ***before DOUBLING*** the ALPHA:

Ridge: R-Squared for train and test data are 0.947625787783842 and 0.8992550377187744

Top 10 features:- GrLivArea , OverallQual, OverallCond , 1stFlrSF , TotalBsmtSF , 2ndFlrSF , BsmtFinSF1 , SaleCondition_Normal , CentralAir_Y , YearBuilt

Lasso: R-Squared for train and test data are 0.9477991224203927 and 0.9071193096175336

Top 10 features:- GrLivArea ,OverallQual ,YearBuilt ,OverallCond ,TotalBsmtSF ,2ndFlrSF ,BsmtFinSF1 ,SaleCondition_Normal ,MSZoning_RL,LotArea

Model details for ridge and lasso regression ***After DOUBLING*** the ALPHA:

Ridge: R-Squared for train and test data are 0.943583823233938 and 0.8992766698307472.

Top 10 important features:- OverallQual ,GrLivArea,1stFlrSF ,OverallCond ,TotalBsmtSF ,2ndFlrSF ,BsmtFinSF1 ,GarageArea ,CentralAir_Y ,GarageCars .

Lasso: R-Squared for train and test data are 0.9432618738805428 and 0.9110727627094882.

Top 10 important features:- GrLivArea ,OverallQual ,YearBuilt ,OverallCond ,TotalBsmtSF ,BsmtFinSF1 ,GarageArea ,MSZoning_RL ,LotArea ,SaleCondition_Normal .

Hence as evident from the above details there is not much change in the R-Squared value but there are some changes in the features.

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 found the optimal values of lambda for lasso and ridge to be 100 and 0.001 respectively.

TOP 5 FEAUTURES FROM ABOVE MODELS ARE:

1) GrLivArea

2) MSZononig_RL

3) OverallQual

4) YearBuilt

5) MSZoning_RM

1. For the same values of alpha, the coefficients of lasso regression are much smaller as compared to that of ridge regression.

2. For the same alpha, lasso has higher RSS (poorer fit) as compared to ridge regression

3. Many of the coefficients are zero even for very small values of alpha.

Typical Use Cases

- Ridge: It is majorly used to prevent overfitting. Since it includes all the features, it is not very useful in case of exorbitantly high #features, say in millions, as it will pose computational challenges.

- Lasso: Since it provides sparse solutions, it is generally the model of choice (or some variant of this concept) for modelling cases where the #features are in millions or more. In such a case, getting a sparse solution is of great computational advantage as the features with zero coefficients can simply be ignored.

Its not hard to see why the stepwise selection techniques become practically very cumbersome to implement in high dimensionality cases. Thus, lasso provides a significant advantage.

We will select lasso regression with this dataset.

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: The first five most important variable for the lasso regression are **GrLivArea', 'OverallQual' , 'YearBuilt' , 'MSZoning_RL' ,'MSZoning_RM' .**

Since these are not available as per the question hence they have been dropped.

So the top five features of the new model are as follows:-

1:-1stFlrSF

2:-2ndFlrSF

3:-TotalBsmtSF

4:-OverallCond

5:-SaleCondition_Partial

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: While creating the best model for any problem statement, we end up choosing from a set of models that would give us the least test error. Hence, the test error, and not only the training error, needs to be estimated in order to select the best model. This can be done in the following two ways:

Use metrics that take into account both the model fit and its simplicity. They penalise the model for being too complex (i.e., for overfitting) and, consequently, more representative of the unseen 'test error'. Some examples of such metrics are adjusted R2, AIC and BIC.

Estimate the test error via a validation set or a cross-validation approach. In the validation set approach, we find the test error by training the model on a training set and fitting on an unseen validation set, while the in n-fold cross-validation approach, we take the mean of errors generated by training the model on all folds except the kth fold and testing the model on the kth fold, where k varies from 1 to n.

So far, you have understood that MSE of the training might not be a good estimate of the test error. The aforementioned parameters are a manipulation of the RSS (residual sum of squares), wherein a penalty term is introduced to compensate for the increase in complexity due to the increase in the number of predictors.