Assignment Part - II

Q1.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?

Optimal value of alpha for
Ridge-568.9
Lasso-10
<u>The impact of doubling the alpha is explained below</u>
If the alphas is doubled in Ridge Regression i.e. 1138, following observations are seen in the model
- Bias will be High
- Variance will reduce
- Model will be much more stable
- Slight drop in the value of R-square
- and slight increase in RMSE

If the alpha is doubled in Lasso regression case i.e. 20, following observations are seen in the model
- Each predictor has a coefficient hence when alpha increases , only strong predictors can survive long , has less number of predictors are retained
- Model therefore becomes simpler
- Chances that performance will be weak as probability of important predictors getting removed gets higher

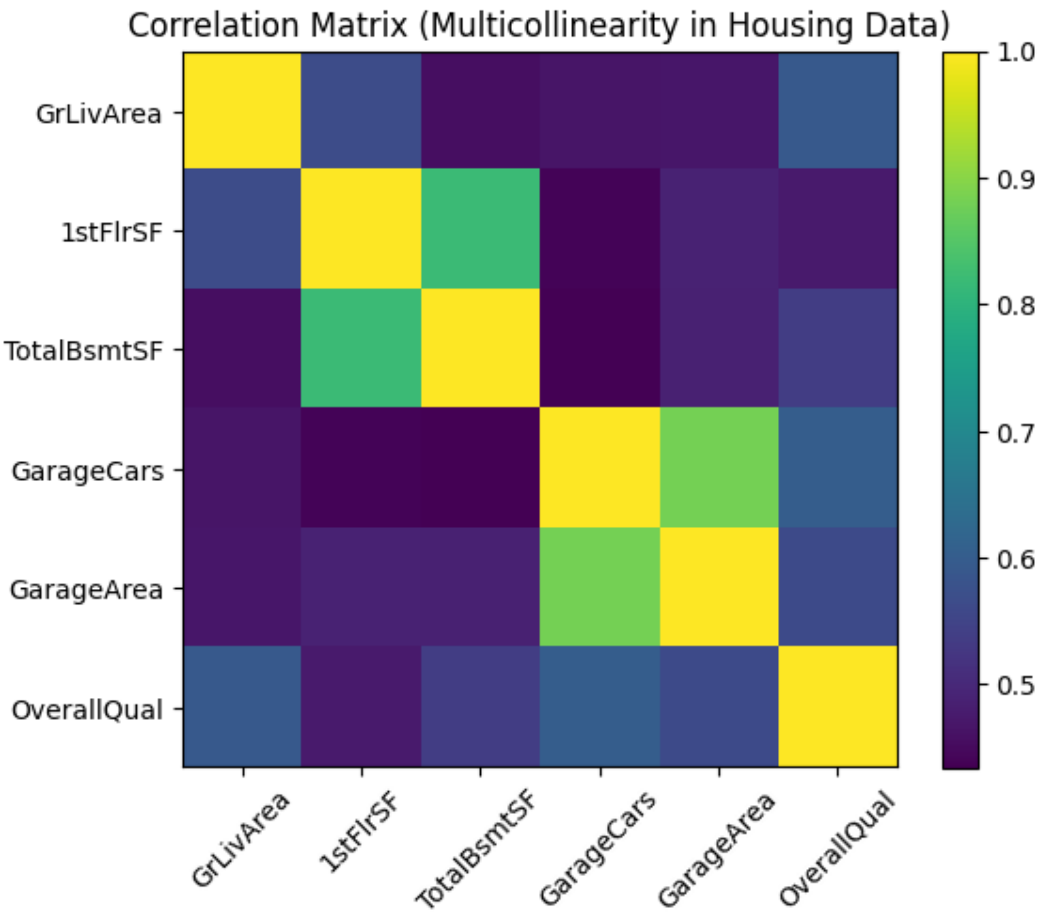
Q2.You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

After evaluating the result from code line <ridge_results = evaluate_model(best_ridge, X_test, y_test)>, we can observe that

- Ridge RMSE is <= Lasso RMSE
- R-square of Ridge is slightly better

Hence we will choose Ridge Regression
Reasons:-
- the housing data has huge multicollinearity as shown in Fig 2.1
- Ridge Regression handles correlated predictors in a much better way
- Provides a stabler estimate on coefficients

As we can see superior performance on unseen data we will choose Ridge Regression



[Fig 2.1]

Q3.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?

In case of lasso, if the strongest predictors are removed or unavailable, it selects the next best options that capture the same information for model building
Lets see which are the original top predictors

- GrLivArea
- OverallQual
- YearBuilt
- TotalBsmtSF
- GarageCars

After removing above, lasso opts the next best variables which are

- 1stFlrSF -> this variable is closely related to GrLivArea which was removed, still depicts the size of the house

- ExterQual -> Represents the quality of the material on the exterior same as OverallQual

- YearRemodAdd -> This tells the construction date of the house which is similar to the meaning of 3rd top predictor

- TotRmsAbvGrd -> This also depicts the size of house hence an important predictor

- GarageArea -> This provides the same information as variable GarageCars hence the removal of GarageCars gives to way to this predictor which is closer in the meaning

Q4.How can you make sure that a model is robust and generalisable? What are the implications of the same

To ensure model is robust, we can do following :

- Cross validation, training the model multiple times
- Train and test split
- Using regularisation methods i.e Ridge or Lasso. This adds a penalty to the loss function in large coefficients hence prevents the model from fitting and reduces variance too
- Feature Scaling. This handles missing values, any outliers. Prevents any dominance of larger valued features
- Perform tests on unseen data. Final evaluation done on a completely untouched data simulates real world performance

Implications of above methods on model accuracy can be explained as:

- Training accuracy may decrease
- Overfitting is reduced to a good extent
- Test accuracy will improve

In conclusion, a robust model should not fit training data completely and perfectly, but must perform well on any unseen data