Subjective Question/Answers

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 both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

   The optimal value of alpha for ridge regression is 4 and for lasso regression is 50.
   If we choose double the value of both ridge and lasso.
   If we choose to double the alpha, The value of coefficients will decrease.

   After the value of alpha is double, below are the top 5 predictor variables-

   1. **Ridge regression (8)**
      GarageType_Attchd
      GarageFinish_RFn
      GarageFinish_Unf
      SaleType_CWD
      SaleType_Con
   2. **Lasso regression (100)**
      GarageType_Attchd
      GarageFinish_RFn
      GarageFinish_Unf
      SaleType_CWD
      SaleType_Con

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

   The optimal value for ridge regression is 4 and for lasso regression is 50. The test score od ridge regression is 0.9152898297301237 and the test score of lasso regression is 0.9145348792017312.
   The score of ridge regression is more than lasso regression.
   Ridge regression attempts to improve generalization to the testing set, by reducing overfit.
   Lasso regression will reduce the number of non-zero coefficients, even if this penalizes performance on both training and test sets.
   Some of the coefficients in lasso regression are absolutely 0 which reduces the computation by reducing number of features.
   Therefore, I will choose lasso regression

3. **After building the model, you realized that the five most important predictor variables in 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?**

   The important predictor variables after removing the earlier important variables GarageType_Attchd, GarageFinish_Rfn, GarageFinish_Unf, SaleType_CWD and SaleType_Con are as follows.

Subjective Question/Answers

1. Ridge regression (4)
   ‘KitchenQual_TA’, 3585.276
   ‘GarageType_BuiltIn’, 4859.285
   ‘GarageType Detchd’, 5077.973
   ‘GarageType_Attchd’, 4659.589
   ‘GarageType_Carport’, 4953.538

2. Lasso Regression (50)
   ‘KitchenQual_TA’, 1272.858
   ‘GarageType_BuiltIn’, 2332.534
   ‘GarageType Detchd’, 2521.892
   ‘GarageType_Attchd’, 2223.008
   ‘GarageType_Carport’, 2375.183

4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

- The Skewed columns are removed and processed to make normal
- Null or empty values are replaced with median for columns having outliers and mean for other continuous variables.
- The columns having less correlation to target variable have been removed
- The best model out of 5fold cross validated model using Ridge and Lasso regularization has been chosen as final model. This shows the best model out of the available training set.
- On plotting mean test and train scores with alpha we could observe similar curve for both train and test data.

As a result of the above points, I could observe a decent score for Ridge and Lasso Regression which are as follows-

1. Ridge Regression Score:
   Train Score: 0.9460249894962848
   Test Score: 0.9152898297301237
2. Lasso Regression Score:
   Train Score: 0.9451804944627856
   Test Score: 0.9145348792017312

The values of train and test scores are very near which shows the model is properly fit on the data and is supposed to predict target values with higher confidence.