Regularization - 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?

Ans:
The optimal value of alpha for Ridge: 14
The optimal value of alpha for Lasso: 0.0001

With doubling of the alpha we found followings:
For Ridge: That the R2 has got decreased for both test and train sets. RSS for the train set has reduced a bit but for test set the RSS has increased. RMSE has also increased for both train and test datasets.
For Lasso: that the R2 has not change much, infact it is almost the same. RSS has increased little for train and test sets. RSME increased for train set while it is almost same for test set.

Following are top 5 predictors (We consider Lasso):
1. GrLivArea,
2. 'VeryExcellent' (10) and 'Excellent' (9) values in column named 'OverallQual',
3. The 'WdShngl' value in the column named 'RoofMatl,
4. The value 3 in the column named - 'FullBath'

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?

Ans: We choose lambda (alpha) with Lasso regression as the test data is performing better with Lasso and also because Lasso has eliminated mant predictors during it's internal processing of feature selection.

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?

Ans: After droping the five most importand predictors and re-modling the top five predictors now are:
1. 1stFlrSF,
2. 2ndFlrSF,

3. LotArea,
4. The value ‘1892’ of the column ‘YearBuilt’,
5. The value ‘3’ of the column ‘GarageCars’

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?

Ans: Following aspects are considered while model building to make it robust and generalisable:
1. The Feature selection: The model should be able to do selection of importand features while ignoring the others. With regularization and with selection of appropriate hypertunning parameter this can be achieved.
2, Cross-Validation: With k-folds it can be cross-validated to see how well the model generalizes the different sub-sets of the data. We used 5 folds cross-validation in the assignment.
3. R2 and RSME value: The train and test data R squared values should be able to describe the relevant data well. RSME should not be high.

Apart from above the ‘Outlier Handling’ should also be done effectively to increase the robustness of the model.

With use of Lasso regression we get most of the requirements except data preparation (remove outlier, removing nulls, feature enguneering etc) are achieved.