Assignment-based Subjective Questions

1Question)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)In Ridge Regression:The optimal value of alpha taken is 50 by checking best estimator inbuilt function of sklearn.linear_model python library for ridge.This is verified by plotting a graph betweenNegative mean absolute error and alpha as we see that if the alpha value is increasing the error term is decreasing and train error is showing a increasing trend.
In Lasso Regression: The optimal value of alpha taken is 0.0003 by checking best estimator inbuilt function of sklearn.linear_model python library for Lasso. This is verified by plotting a graph betweenNegative mean absolute error and alpha as the alpha value is increasing the model is trying to penalize the variables and trying to shrink the variables to 0.
The changes in the model if we choose to double the value for both ridge and lasso are :
The r2 after doubling for ridge for train set is 0.927 and test set is 0.887 .The r2 before doubling the ridge for train set is 0.927 and for test is 0.887 which is remaining same with values after doubling.
The r2 after doubling for lasso for train is 0.927 and test is 0.888 which was before doubling the ridge for train set is 0.927 and for test is 0.887 which is remaining same with values after doubling.
The rss is changing significantly:
Ridge: Before doubling->train set:9.746 and test set->6.85
Ridge: After doubling->train set:9.975and test set->6.949
Lasso: Before doubling->train set:9.426and test set->6.624
Lasso: After doubling->train set:9.583and test set->6.710
The most important predicted values after change implemented are:
For Ridge:
1.MSZoning_RL
2. GrLivArea
3. OverallQual
4. MSZoning_RM
5. TotalBsmtSF
6. MSZoning_FV
7. Foundation_PConc
8. OverallCond
9. GarageCars
10. RoofStyle_Gable

For Lasso:
1.MSZoning_RL
2. GrLivArea
3. MSZoning_RM
4. OverallQual
5. MSZoning_FV
6. TotalBsmtSF
7. Foundation_PConc
8. OverallCond
9. GarageCars
10. RoofStyle_Gable

2Question)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)I will choose lasso as the primary goal of the case study is to determine the predictors and from the analysis the lasso will penalize and reduce the variables to 0 so that this efficiently help in analysing the features resulting in preventing overfitting.

If the goal is to avoid large coefficients then I would have choosen ridge.
Also from rss and r2 values of both ridge and lasso the lasso has slightly better values than ridge.

3Question)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)From the graph before dropping the top variables :

The top 5 variables are :
1.MSZoning_RL
2. GrLivArea
3. MSZoning_FV
4. OverallQual
5.MSZoning_RM
Now Steps performed after removing top 5 variables are:
1.2ndFlrSF
2.1stFlrSF
3.TotalBsmtSF
4.OverallCond
5.GarageCars

4Question)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)In order to make sure that a model is robust and generalizable then the model should not overfit or underfit with test data.Generally if the model is overfit it is tend to give more accuracy and as it will tend to learn the patterns in train data more accurately and will not able to predict the patterns of the test data thus leading to more variance and less bais.So to handle this kind of implications we can use lasso and ridge regressions.Thus, doing this we can get more bias and less variance.