Advaced Linear Regression

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?

Optimal value of Lambda for ridge is 10 and Lasso is 0.001. R2 score and Adjusted R square decreases in both Lasso and ridge for both test and train data when I doubled the lambda value where as RSS, MSE, and RMSE increases.

Important predictor for Ridge:

('Neighborhood_Timber', 0.08155435622589315)

('MSZoning_Residential Low Density', 0.09649426457214931)

('BsmtFinSF2', 0.1083431809279315)

('BsmtUnfSF', 0.12357794992932569)

('LandSlope_Moderate', 0.0787330714942879)

Important Predictor for Lasso:

('BsmtFinSF1', 0.06958354761718198)

('2ndFlrSF', 0.0957252426190483)

('Functional_Typical Functionality', 0.09620235557476964)

('1stFlrSF', 0.09826401594397077)

('OverallQual_Excellent', 0.11211696030613343)

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?

Optimal value of Lambda for ridge is 10 and Lasso is 0.001. The model will work well with both regressions but I will choose Ridge due to multicollinearity of data and use RFE for feature selection.

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?

For Lasso:

- ('CentralAir_Y', 0.06561384677854372)
- ('BsmtFinSF1', 0.06708896685152717)
- ('OverallQual_Very Good', 0.07812259276200326)
- ('Neighborhood_Crawfor', 0.0805977344098445)
- ('Neighborhood_StoneBr', 0.0831007935128889)

For Ridge:

- ('YearRemodAdd', 0.07661019433554986),
- ('Alley_Paved', 0.07797298758601016),
- ('YrSold', 0.07968040887088344),
- ('LandContour_Depression', 0.0797572097899202),
- ('Neighborhood_Timber', 0.08256525274464806)

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?

- Remove outliers
- Check with different values for number of feature for RFE
- Handling missing values
- Standardize the data
- Perform log transform on dependent variable
- Feature Selection
- Check the model is simple to avoid overfitting and at the same time take care we shouldn't it too simple which lead to underfitting. This can be done hyperparameter tuning.

It will have positive implication on the accuracy because factor mentioned above have negative impact on model which lead to low value of training data and lower value of test accuracy.