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?

Answer:

The Optimal value of alpha for ridge and lasso regression is the hyperparameter lambda, which is chosen rightly, so as to reduce only the variance of the model, without compromising on the underlying pattern of the data (which is Bias). the Lambda is chosen rightly such that penalty term helps in regularization. Higher the value of lambda, the magnitude of the model coefficients tends to become lower (leading to 0), denoting more regularization. More regularization will tend to make model coefficients equal to 0 (in Lasso regression) or almost near to 0 (in Ridge regression), causing underfitting.

If the value of lambda/alpha is doubled, few of the predictor variables coefficients tends to 0 (in ridge regression) or becomes 0 (in Lasso). It may also lead to high RSS , low R2_score leading to underfitting eventually.

The most important predictor variables after ridge or lasso regression are the ones who coefficients are not tending to 0 or have not become 0.

Rather the higher magnitude of coefficients are the most important predictor variables. That is –

GrLivArea for ridge :6.822747e+03 for Lasso : 29396.594435

Followed by :

	ridge_doub	lasso_doub
GrLivArea	6.822747e+03	29396.594435
OverallQual_10	6.222225e+03	12591.525794
OverallQual_9	6.036691e+03	13077.585418
Neighborhood_NoRidge	5.273148e+03	5358.010930

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?

Answer:

I would choose to apply Lasso regression with optimal lambda of 300 , as the model performance is reliable and fits for generic underlying data and also provides the feature selections.

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?

Answer :

The top 5 predictors and its coefficient value after removing top 5 predictors from Lasso reg model:

```
2ndFlrSF                26475.900429
1stFlrSF                18058.054093
OverallQual_9           12740.833673
OverallQual_10          12508.777842
OverallQual_8           10130.681219
```

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?

Answer:

Regularization allows to make the model Robust and generalisable. Which can be analysed by checking error terms analysis for heteroscedasticity and homoscedasticity.

Regularization helps in reducing the overfitting of the model but adding a penalty on the error terms, which in turn makes the model understand the underlying data pattern ad makes it more generic enough to such that total error is less along with model being robust and generalisable.

Also , Outliers can degrade the fit of linear regression models when the estimation is done using ordinary least squares. Making sure the outliers are handled as needed makes model more robust.

Error distribution analysis to follow all the regression assumptions another way to understand overfitting and maintain the model robust.

The Model may accuracy may reduce due to some constraint in the error terms while regularization, which is because to fit the more general underlying data pattern, by adding penalty in the error terms.