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:

For my models the optimal values for Lambda were:-

- Ridge: 2.0 :

       ridge_train_r2:0.9415187383849262

       ridge_test_r2:0.9089636180621271

       Difference : 0.03255512032279906

- Lasso: 0.001 :

       lasso_train_r2:0.8871738764695702

       lasso_test_r2:0.8830078244206411

       Difference:0.00416605204892917

On Doubling the Lamda:

       The R2 score decreases and error increases for both train and test datasets. Basically increase in lambda cause higher regularizations and the model becomes simpler but we use very high values the model will underfit and will give huge errors

After implementing the change(Exactly doubling the lambdas), the most important predictors:

- Ridge: 4.0 :

GrLivArea
1stFlrSF
TotalBsmtSF
BsmtFinSF1
OverallQual_9
Neighborhood_Crawfor
2ndFlrSF
MSZoning_FV
LotArea
Neighborhood_StoneBr

- Lasso: 0.002 :

GrLivArea
TotalBsmtSF
CentralAir_Y
BsmtFinSF1
GarageCars_3
BsmtExposure_Gd
Foundation_PConc
MSZoning_RL
GarageQual_TA
GarageType_Attchd

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 prefer Lasso Regression (Lambda=0.001) instead of Ridge Regression(Lambda=2.0) because:-

- Even though the R2 was better for Ridge regression, Lasso regression had much lower difference between Train and test r2 scores, which highlights that there is no overfitting and it's a robust model
- And in Lasso we have much lower number of features which again highlights that its robust

Question-3:

After building the model, you realized 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:

As per our current model the 5 most important predictor variables are:-

- GrLivArea

- TotalBsmtSF
- BsmtFinSF1
- CentralAir_Y
- SaleType_New

Incase these 5 variables are not present in our test set, we drop these variables and recreate the model. In such case our top features become:

OpenPorchSF

LowQualFinSF

2ndFlrSF

BsmtFinSF2

OverallCond_3

Question-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?

Answer:

Model can be made robust and generalizable by keeping it simple and less complex. Some of the following methods can be used in keeping the model simple:-

- Using proper regularization techniques
- Not having too many features
- Not having huge number of decimal places in weights of features
- Using lower degrees of features

Keeping the model simple leads to lower R2 scores & higher errors for the training set but it is better for predictions of test data because it can handle new patterns by kepping the model simple. So keeping the model simple is the thumb rule but its should be taken care that it should not underfit, basically there should be a proper balance between bias & variance.