Problem Statement – Part II

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:
1.a. The optimal value for alpha for ridge is 5.0 and for lasso regression is 0.0001.
1.b.
With alpha 5.0 and 0.0001 for ridge and lasso

	Ridge	Lasso
Train R2 score	0.914111	0.913117
Test R2 score	0.875989	0.887094
Train RSS	1.005053	1.016693
Test RSS	0.644393	0.586687
Train MSE	0.031592	0.031775
Test MSE	0.038622	0.036852

With alpha 10.0 and 0.0002 for ridge and lasso

	Ridge	Lasso
Train R2 score	0.905765	0.905325
Test R2 score	0.872269	0.887088
Train RSS	1.102718	1.107867
Test RSS	0.663722	0.586718
Train MSE	0.033092	0.033169
Test MSE	0.039197	0.036853

So - on doubling the alphas for Lasso and Ridge regression

- Train R2 score went down by 1 % however test R2 score is almost the same
- MSE went very slightly up.

1.c.i. Most important predictors after doubling the alpha will be for lasso regression

BsmtFullBath, OverallCond, 1stFlrSF, BsmtFinSF1, MasVnrArea, Neighborhood_Timber, Neighborhood_NridgHt, BsmtFinSF2

1.c.ii. Most important predictors after doubling the alpha will be for Ridge regression

MSSubClass, OverallCond, BsmtFullBath, 2ndFlrSF, LowQualFinSF, Neighborhood_NridgHt, BsmtFinSF2, 1stFlrSF, BsmtFinSF1.

Notes: Details on how I got these values are present in Subjective Question section at https://github.com/kc11381/Housing_Price_Prediction_Case_Study/blob/main/house_price_prediction.ipynb

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:
The optimal value of alpha for Ridge and Lasso regression is 5.0 & 0.0001 respectively.
If we see the comparison below table

	Ridge	Lasso
Train R2 score	0.914111	0.913117
Test R2 score	0.875989	0.887094
Train RSS	1.005053	1.016693
Test RSS	0.644393	0.586687
Train MSE	0.031592	0.031775
Test MSE	0.038622	0.036852

R2 score for test score is better in Lasso. Also other values as RSS and MSE are smaller in Lasso.
Another important point is in this assignment we had 200+ features (after dummy variable creation for categorical variables), if we go ahead with Ridge model it will consider all features with penalty but none will become zero. However Lasso regression will make model simpler by penalizing multiple features to become zero.

So because Lasso regression is performing better in terms of R2 score for test dataset, RSS and MSE metrics and even making the model simpler in this case by making multiple features to become zero, **I will choose Lasso Regression here.**

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:
Top 5 new predictors are:

	Feature	Coef
10	BsmtHalfBath	0.226811
4	MasVnrArea	0.114846
5	BsmtFinSF1	0.061820
6	BsmtFinSF2	0.057782
58	Neighborhood_NridgHt	0.057296

We have dropped following previous predictors: 'OverallCond', 'BsmtFullBath', '1stFlrSF', 'LowQualFinSF', '2ndFlrSF'.

Note: Code for this can be found at:
https://github.com/kc11381/Housing_Price_Prediction_Case_Study/blob/main/house_price_prediction.ipynb **at Subjective Q.3 answer.**

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:
A model is generalized when it's not complex in terms of number of features. We always make a trade off between Variance and Bias. Ridge and Lasso regression allow some bias to get a significant decrease in variance, thereby pushing the model coefficients towards 0.
Few points we should always consider for this while building the model:

a. It should not be complex in terms of number of features being too high.
b. We make trade off between Bias and variance to get optimum value of R2 score.
c. There should not be much difference between R2 score of model on train and test data.
d. Never use test data for training the model
e. Always look if the model is overfitting – model should not mug up the train data.
f. Ridge and Lasso regression regression can be used to make model simpler.