1:

What is the optimal value of alpha for ridge and lasso regression? What will change the model if you choose to double the alpha value for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: Optimal value of alpha for both lasso & ridge regression is

Ridge alpha value: 8.0

and Lasso alpha value: 0.001

Test & Train score before & after doubling the optimal value of alpha is given below-

	estimator_name	train_r2_score	test_r2_score	train_mse_score	test_mse_score
0	Linear Regression	0.942079	-3.997702e+17	0.057921	3.432624e+17
1	Ridge Regression with alpha 8.0	0.914768	8.579496e-01	0.085232	1.219715e-01
2	Ridge Regression_with alpha 16	0.904935	8.628343e-01	0.095065	1.177773e-01
3	Lasso Regression with alpha_0.001	0.928180	8.334368e-01	0.071820	1.430193e-01
4	Lass Regression_with alpha 0.002	0.915253	8.561725e-01	0.084747	1.234974e-01

As per the given table, I observe increment in train_mse_score for both Lasso as well as Ridge Regression after doubling the optimal value and but test score became better. It may be model is more generalizing when we double the alpha value.

After making changes below are the most important predictors.

Important Predictors for Lasso

```
[('Condition2_PosN', -2.101730470022199),
 ('OverallQual_10', 1.0667070700595709),
 ('RoofMatl_WdShngl', 0.65150913762827),
 ('Neighborhood_StoneBr', 0.5095927634600115),
 ('OverallQual_9', 0.4493458092339428),
 ('Neighborhood_NoRidge', 0.3980357551010922),
 ('2ndFlrSF', 0.3270636553868147),
 ('KitchenQual_Ex', 0.2810687069710022),
 ('1stFlrSF', 0.2583810413544561),
 ('BsmtQual_Ex', 0.24801239306764264),
 ('BsmtExposure_Gd', 0.24335518047087842),
 ('Neighborhood_Crawfor', 0.2103686959172543),
 ('Neighborhood_NridgHt', 0.20228423569811313),
 ('ExterQual_Ex', 0.19828733304600327),
 ('PoolQC_Ex', 0.19422336752005953),
 ('Functional_Typ', 0.1935494715373101),
 ('OverallQual_8', 0.16404894894839064),
 ('OverallCond_4', -0.1563516274455205),
 ('Exterior1st_BrkFace', 0.14871433473002102),
 ('MSSubClass_90', -0.14219290433804427)]
```

Important Predictors for Ridge

```
[('OverallQual_10', 0.36272625255000324),
 ('Neighborhood_StoneBr', 0.2971098082339202),
 ('2ndFlrSF', 0.28823435969727035),
 ('Neighborhood_NoRidge', 0.2723762692272622),
 ('RoofMatl_WdShngl', 0.229973691760992),
 ('KitchenQual_Ex', 0.22871041535632755),
 ('BsmtExposure_Gd', 0.21356484817194943),
 ('ExterQual_Ex', 0.21193389218617623),
 ('1stFlrSF', 0.20862485717437704),
 ('Condition2_PosN', -0.20085919151947643),
 ('BsmtQual_Ex', 0.1760635471327914),
 ('Neighborhood_NridgHt', 0.16118715471818),
 ('Functional_Typ', 0.15953193995480522),
 ('OverallQual_9', 0.15804951724556598),
 ('RoofMatl_CompShg', -0.15583103168179321),
 ('Condition2_Norm', 0.1542096524524302),
 ('PoolQC_Ex', 0.14823435419150752),
 ('Neighborhood_Edwards', -0.14747568757878934),
 ('OverallCond_9', 0.13656382992190882),
 ('OverallCond_4', -0.13141187885306838)]
```

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:

Optimal value of alpha for both lasso & ridge regression is

Ridge alpha value: 8.0, Test R2 score: 0.85, and Test MSE score: 0.121,

and Lasso alpha value: 0.001, Test R2 score: 0.84 and test MSE score: 0.123

I chose Lasso Regression. Because I observed that obtained model is not complex and more generalized. Lasso does feature selection also. I also see there is no big score gap b/w Lasso and Ridge models. Since Lasso provides a better edge than Ridge so many non-important coefficients become zero. Which results more generalized model.

I also observed that the model becomes stable (low vif and low p) quickly after getting features from lasso regression.

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?**

Ans:

So if I remove the 5 most important predictors and build a new lasso model with the same alpha 0.001, then the 5 most important predictors

RoofMatl_Tar&Grv
RoofMatl_CompShg
Neighborhood_NoRidge
KitchenQual_Ex
ExterQual_Ex

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?**

Ans

We can make sure the model is robust & generalizable by checking the following

1. As a variance bias trades off, the model should not be complex. The complex model behaves well in the train datasets but gives a bad score for the test dataset.
2. But also model should not be simple enough, that model is not able to capture all patterns.
3. To make the model generalize, we keep the model robust by penalizing the weight parameter.
4. L1 & L2 regularization technique used to make the model more generalized.
5. There are different techniques to prevent the model to memorize all data points like Lasso & Ridge Regression.
6. For Neural networks, dropping technique, weight capping, minimizing hidden layers etc.
7. In the tree model, we make a generalized model by tuning Hyperparametes.
8. For Hyperparameter tuning, we use cross-fold validation to avoid data leakage and get model statistics with different hyperparameters.

Implication

9. We finally compare test and train scores. if there is not much gap b/w training and test scores, we can say the model achieving towards robust.
10. Model should also follow assumptions to behave well.