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** Regression and **Lasso** Regression are **500** and **0.01** respectively.

Doubling the alphas:

Ridge Regression: When I double the alpha for Ridge Regression to 1000, the training r2 changes from 0.8854 to 0.8629 and the test r2 changes from 0.8531 to 0.8433. RSS on the training data has increased from 116.94 to 139.89 and the same has increased from 66.27 to 70.65 on the test set. The r2 sees a slight dip and the RSS has gone up quite a bit, both on training and test data. The most important predictor variables are **OverallQual, GrLivArea, 1stFlrSF, Neighborhood_NoRidge and Neighborhood_NridgHt**.

Lasso Regression: When I double the alpha for Lasso Regression to 0.02, the training r2 changes from 0.8947 to 0.8761 and test r2 changes from 0.8354 to 0.8301. RSS on the training set has increased from 107.49 to 126.53 and the same has increased from 74.25 to 76.61 on test set. The r2 sees a slight dip and the RSS has gone up quite a bit, both on training and test data. The most important predictors are **GrLivArea, OverallQual, Conditions_PosN and PosN (**Negative connotation, combination of Condition1 = PosN and Condition2 = PosN**), Neighborhood_NridgHt, Neighborhood_NoRidge**

In terms of feature selection by Lasso, the features selected dropped from 102 to 62.

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:

Let us look at the metrics below:

Metric	Linear Regression	Ridge Regression (alpha=500)	Lasso Regression (alpha=0.01)
R2 Score (Train)	0.9429	0.8855	0.8947
R2 Score (Test)	-6.24E+21	0.8531	0.8354
RSS (Train)	58.3016	116.9475	107.4908
RSS (Test)	2.81E+24	66.2793	74.2589
MSE (Train)	0.2390	0.3384	0.3245
MSE (Test)	80156610000.0000	0.3890	0.4118

From the above, its obvious that Linear Regression is not a good fit.

For **Ridge Regression**, as determined by GridSearchCV, the optimal values of **alpha=500**. As we can see in the above table Ridge Regression has a very good r2 train score of 0.8855 and r2 test score of 0.8531. The training RSS score is 116.9475 and test RSS score is 66.2793. There are a total of 264 predictor variables.

For **Lasso Regression**, as determined by GridSearchCV, the optimal values of **alpha=0.01**. As we can see in the above table Ridge Regression has a very good r2 train score of 0.8947 and r2 test score of 0.8354. The training RSS score is 107.4908 and test RSS score is 74.2589. There are a total of 102 predictor variables.

From the above data, it is evident that there is not much to choose from between Ridge and Lasso regression based on Metrics. But, because Lasso Regression has a much simpler model (much lesser predictor variables), I would choose Lasso Regression.

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:

When I remove 5 most important predictor variables in Lasso Regression, the next 5 most important predictors are: **2ndFlrSF, 1stFlrSF, ExterQual, GarageCars, BsmtQual**.

The optimal **alpha** value selected by the GridSearCV algorithm is **0.01**.

The new model has a **train r2** score of **0.8650** and **test r2** score of **0.8475**.

The **train RSS** score is **137.79** and **test RSS** score is **68.79**.

The **train MSE** score **0.135** and **test MSE** score is **0.157**.

When compared to the earlier model, we see that the difference between r2 train and test scores have reduced. However, the train RSS has gone up significantly.

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:

The following steps are taken to make sure that the model is robust and generalizable:

1. **Majority Null Column Removal**: This is done to remove the columns which have more than 40% nan / NA / null values. This process will make sure that data is consistent and the regression on valid values is correctly predicted.
2. **Missing Data Treatment**: The data which have nulls within limits are appropriately imputed. This will help in maintaining data integrity and reduce bias.
3. **Converting Data into correct types**: The data is converted to appropriate types which can help in regression. The Linear Regression does not make accurate predictions when the data type is non-numeric. Therefore, the data is converted into numeric types. For e.g., ordinal data types, creating 'dummy' columns for categorical data.
4. **Exploratory Data Analysis**: Manually explore each data field to analyse data and understand the importance of each field (improve domain knowledge). Also make sure plot the data against the target variable to check the outliers and figure out important values for a given column.
5. **Train / Test data split**: The data is split into training data (70%) and test data 30%). This is done so that we can evaluate our model on unseen data.
6. **Scaling data**: This is an important step because some algorithms use the Euclidean distance between two data points in their computations/derivations, which is sensitive to the scale of the variables. If features are on different scales, features with larger ranges will dominate the coefficients, leading to biased results.
7. **Regularization**: Helps to prevent overfitting by shrinking irrelevant coefficients and helps in achieving bias-variance tradeoff.
8. **Model Validation**: Perform k-fold Cross Validation so that the test data does not influence the model evaluation and we have a real unseen test data set.
9. **Model Evaluation**: Evaluate metrics like r2, RSS, MSE on test and train data sets.
10. **Linear Regression Assumptions**: Making sure all the assumptions of Linear Regression are evaluated.