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=8

The optimal value of alpha for lasso= 0.0001

If the value of alpha is doubled for both ridge and lasso, the changes in the model would be as follows:

For Ridge Regression:

- The optimal value of alpha would be 16 (double of 8).
- The R2 Score (Train) would decrease to 0.878 from 0.893.
- The R2 Score (Test) would become NaN (not a number) from 0.851.
- The Difference in R2 test and train would become NaN from 0.042.
- The RSS (Train) would increase to 1.503 from 1.322.
- The RSS (Test) would become NaN from 6069.503.
- The MSE (Train) and MSE (Test) would remain the same at 0.001 and NaN respectively.
- The RMSE (Train) would increase to 0.038 from 0.036.
- The RMSE (Test) would become NaN from 0.054.

For Lasso Regression:

- The optimal value of alpha would be 0.0002 (double of 0.0001).
- The R2 Score (Train) would increase to 0.897 from 0.912.
- The R2 Score (Test) would become NaN from 0.848.
- The Difference in R2 test and train would become NaN from 0.064.
- The RSS (Train) would increase to 23116.923 from 23559.134.
- The RSS (Test) would become NaN from 1.303.
- The MSE (Train) and MSE (Test) would remain the same at 0.001 and NaN respectively.
- The RMSE (Train) would decrease to 0.035 from 0.033.
- The RMSE (Test) would become NaN from 0.055.

After the change is implemented, the most important predictor variables for each regression model would be as follows:

For Ridge Regression:

1. GrLivArea with a coefficient of 0.0037
2. OverallQual_10 with a coefficient of 0.0037

3. OverallQual_9 with a coefficient of 0.0037
4. RoofMatl_WdShngl with a coefficient of 0.0037
5. Neighborhood_NoRidge with a coefficient of 0.0037
6. GarageCars with a coefficient of 0.0037
7. OverallQual_8 with a coefficient of 0.0037
8. 2ndFlrSF with a coefficient of 0.0037
9. Neighborhood_Crawfor with a coefficient of 0.0037
10. FullBath with a coefficient of 0.0037

For Lasso Regression:

1. GrLivArea with a coefficient of 0.2954
2. OverallQual_10 with a coefficient of 0.1286
3. OverallQual_9 with a coefficient of 0.0992
4. RoofMatl_WdShngl with a coefficient of 0.0780
5. Neighborhood_NoRidge with a coefficient of 0.0572
6. GarageCars with a coefficient of 0.0562
7. OverallQual_8 with a coefficient of 0.0462
8. 2ndFlrSF with a coefficient of 0.0351
9. Neighborhood_Crawfor with a coefficient of 0.0348
10. FullBath with a coefficient of 0.0312

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 will choose the optimal value of lambda for lasso regression because:

1. R2 Score (Test): Lasso Regression has a higher R2 score on the test set (0.847765) compared to Ridge Regression (0.850747). A higher R2 score indicates a better fit of the model to the data.
2. R2 Score (Test): Lasso Regression has a higher R2 score on the test set (0.847765) compared to Ridge Regression (0.850747). A higher R2 score indicates a better fit of the model to the data.
3. RSS (Test): Lasso Regression has a lower residual sum of squares (RSS) on the test set (1.303491) compared to Ridge Regression (6069.503029). A lower RSS indicates better prediction accuracy.
4. MSE (Test): Lasso Regression has a lower mean squared error (MSE) on the test set (0.002976) compared to Ridge Regression (0.002918). A lower MSE indicates better prediction accuracy.
5. RMSE (Test): Lasso Regression has a slightly higher root mean squared error (RMSE) on the test set (0.054553) compared to Ridge Regression (0.054016). However, the difference is very small and may not significantly impact the model's performance.

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 five most important predictor variables after removed excluding the five most important predictor variables are:

Lasso - New	
1stFlrSF	0.290141
2ndFlrSF	0.200863
LotArea	0.073609
GarageCars	0.061325
MasVnrArea	0.046905

```
best_cols = pd.DataFrame(index=X_train.columns)
pd.set_option('display.max_rows', None)
best_cols['Lasso'] = lasso_model.coef_
best_cols.sort_values(by=['Lasso'],ascending=False).head(5)
```

	Lasso
GrLivArea	0.322699

	Lasso
OverallQual_10	0.132449
RoofMatl_WdShngl	0.119158
OverallQual_9	0.093789
Neighborhood_NoRidge	0.054337

```
# Dropping the columns
X_train.drop(['GrLivArea','OverallQual_10','RoofMatl_WdShngl','OverallQual_9','Neighborhood_NoRidge'],axis=1,i
nplace=True)
```

```
# Building the model again with new data set for Lasso Regression
lasso_model_new = Lasso(alpha=lasso_model_cv.best_estimator_.alpha)
lasso_model_new.fit(X_train, y_train)
y_train_pred_new = lasso_model_new.predict(X_train)
```

```
cols_new = pd.DataFrame(index=X_train.columns)
cols_new['Lasso - New'] = lasso_model_new.coef_
cols_new.sort_values(by=['Lasso - New'],ascending=False).head(5)
```

	Lasso - New
1stFlrSF	0.290141
2ndFlrSF	0.200863
LotArea	0.073609
GarageCars	0.061325
MasVnrArea	0.046905

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:

Metric	Linear Regression with RFE	Ridge Regression	Lasso Regression	
0	R2 Score (Train)	0.736554	0.892730	0.911511
1	R2 Score (Test)	0.665627	0.850747	0.847765
2	Diffrence in R2 test and train	0.070927	0.041982	0.063745
3	RSS (Train)	3.245765	1.321613	23559.134369
4	RSS (Test)	2.863030	6069.503029	1.303491
5	MSE (Train)	0.003176	0.001293	0.001067
6	MSE (Test)	0.006537	0.002918	0.002976
7	RMSE (Train)	0.056355	0.035961	0.032661
8	RMSE (Test)	0.080849	0.054016	0.054553

The model metrics give an indication of the performance of three different regression models: Linear Regression with RFE (Recursive Feature Elimination), Ridge Regression, and Lasso Regression.

The R2 score measures how well the model explains the variance in the data, with values closer to 1 indicating a better fit. The Ridge Regression and Lasso Regression models have higher R2 scores (0.893 and 0.912 respectively) compared to Linear Regression with RFE (0.737), suggesting better overall performance.

The difference between the R2 scores of the test and train sets (R2 test - R2 train) indicates the model's generalization capability. Smaller differences imply better generalization.

The RSS and MSE metrics provide information about the model's residuals (the difference between predicted and actual values). Lower values indicate better accuracy, with Ridge Regression having the lowest RSS for both train and test sets.

Similarly, RMSE measures the average magnitude of residuals. Lower values imply better accuracy, with Lasso Regression having the lowest RMSE for both train and test sets.

Overall, the higher R2 scores, lower RSS, MSE, and RMSE values for Ridge Regression and Lasso Regression suggest that these models are more accurate and have better generalization capability compared to Linear Regression with RFE.