Q-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: The optimal values for Ridge and Lasso are given below:

Ridge: 5.77154379759519

The top 10 predictors are:

- MSZoning_C (all)
- RoofMatl_ClyTile
- PoolArea_480
- Condition2_PosN
- OverallQual_9
- Neighborhood_Crawfor
- OverallCond_3
- Neighborhood_MeadowV
- Fireplaces_3
- OverallCond_9

If we change the value to twice the original then the top 10 predictors are:

- MSZoning_C (all)
- OverallQual_9
- Neighborhood_Crawfor
- OverallCond_3
- PoolArea_480
- RoofMatl_ClyTile
- Neighborhood_Edwards
- Condition2_PosN
- Neighborhood_MeadowV
- Fireplaces_3

With twice the optimal alpha the R^2 value = 0.889895469085712 which is greater than 0.8562 that was originally obtained.

Lasso: 0.00038152899755399596

The top 10 predictors are:

- RoofMatl_ClyTile
- Condition2_PosN
- MSZoning_C (all)
- PoolArea_480
- OverallCond_3
- OverallQual_9
- Neighborhood_Crawfor

Neighborhood_MeadowV
OverallCond_9
BsmtCond_Po

If we change the value to twice the original then the top 10 predictors are:

RoofMatl_ClyTile
Condition2_PosN
PoolArea_480
MSZoning_C (all)
OverallCond_3
OverallQual_9
Neighborhood_Crawfor
Neighborhood_MeadowV
OverallQual_8
Functional_Typ

With twice the optimal alpha the R^2 value = 0.8893982369045244 which is greater than the obtained optimal value of 0.8664.

It looks like the models become more accurate by doubling the alpha value for both regression schemes.

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: We'll apply Lasso regression, because even though by a small margin the r-squared value is better for Lasso 86.6 vs 85.6. It is perhaps still possible to discard some of the predictor variables (which is evident from the large number of zero coefficients in lasso output).

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: The code snippet below gives the new top 10:

OverallQual_9
Fireplaces_3
Neighborhood_Crawfor
Neighborhood_MeadowV
OverallCond_9
OverallQual_8
Neighborhood_Edwards

Neighborhood_StoneBr
OverallCond_7
Neighborhood_NridgHt

```python
# If we drop the following columns from X_train and then do Lasso again what do we get
X_train = X_train.drop(['RoofMatl_ClyTile', 'Condition2_PosN', 'MSZoning_C (all)','PoolArea_480','OverallCond_3'], axis=1)
X_test = X_test.drop(['RoofMatl_ClyTile', 'Condition2_PosN', 'MSZoning_C (all)','PoolArea_480','OverallCond_3'], axis=1)

# Somehow for lasso when using logspace we get better values for score
alpha_values = np.logspace(-4, np.log10(50), num=50)
parameters = {'alpha': alpha_values}
lasso = Lasso()

lasso_cv = GridSearchCV(lasso, parameters, cv=5)

lasso_cv.fit(X_train, y_train)

# Print the best parameters and the best score
print("Best parameters: ", lasso_cv.best_params_)
print("Best score: ", lasso_cv.best_score_)

# Fitting the Lasso model
lasso = Lasso(alpha=lasso_cv.best_params_['alpha'])
lasso.fit(X_train, y_train)

# Making predictions
y_pred = lasso.predict(X_test)

feature_names = X_train.columns
coefficients = lasso.coef_
#Now we create a dictionary from these two and print it out
coefficients_dictionary = dict(zip(feature_names, coefficients))
nonzero_coefficients_dictionary = {feature: coef for feature, coef in zip(feature_names, coefficients) if coef != 0}
# Now we find the coefficients
sorted_coeffiecient_dictionary = {k: v for k, v in sorted(nonzero_coefficients_dictionary.items(), key=lambda item: abs(item[1]), reverse=True)}
# Let us take the top 10 from this list
top_10_predictors_lasso = list(sorted_coeffiecient_dictionary.keys())[:10]

print("Top 10 predictors in Lasso regression are:")
for predictor in top_10_predictors_lasso:
    print(predictor)


--INSERT--
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_coordinate_descent.py:631: ConvergenceWarning: Objective did not converge. You might want to increa
  model = cd_fast.enet_coordinate_descent(
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_coordinate_descent.py:631: ConvergenceWarning: Objective did not converge. You might want to increa
  model = cd_fast.enet_coordinate_descent(
Best parameters:  {'alpha': 0.0006518363448688389}
Best score:  0.8568726156694556
Top 10 predictors in Lasso regression are:
OverallQual_9
Fireplaces_3
Neighborhood_Crawfor
Neighborhood_MeadowV
OverallCond_9
OverallQual_8
Neighborhood_Edwards
Neighborhood_StoneBr
OverallCond_7
Neighborhood_NridgHt
```

,

It can be argued that since these are one-hot encoded variables we can’t really view them as predictors but it can be argued either way.