Advance Linear Regression

Subjective Questions:

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:

Optimal value of alpha –

Ridge - 4.0

Lasso – 0.01

After make the double alpha for ridge and lasso, the model will be more heavily regularized.

As a result, regularization on the coefficients of the features will increase, leading to potentially smaller coefficients and a simpler model.

the coefficients of both Ridge and Lasso regression models will likely shrink further compared to their original values. This will result in a more pronounced effect of regularization on the models, potentially leading to increased bias but reduced variance.

Lasso -

In [204]:
```
#Doubling alpha

params = {'alpha': [0.2]}
lasso = Lasso()
lasso_cv = GridSearchCV(estimator = lasso,
                        param_grid = params,
                        scoring= 'neg_mean_absolute_error',
                        cv = folds,
                        return_train_score=True,
                        verbose = 1)
lasso_cv.fit(x_train, y_train)
```

```
Fitting 5 folds for each of 1 candidates, totalling 5 fits
```

Out[204]:

In [205]:
```
print(lasso_cv.best_params_)
print(lasso_cv.best_score_)
```

```
{'alpha': 0.2}
-24423.919448996527
```

```
In [214]: lasso = Lasso(alpha=0.2)
          lasso.fit(x_train, y_train)

Out[214]: Lasso
          Lasso(alpha=0.2)

In [215]: y_train_pred = lasso.predict(x_train)
          print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))

          0.8874945319186921

In [216]: y_test_pred = lasso.predict(x_test)
          print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))

          0.8669057921493499

In [ ]:
```

Ridge –

```
]: print(model_cv.best_params_)
   print(model_cv.best_score_)

   {'alpha': 8.0}
   -20914.714581488362

]: ridge = Ridge(alpha=8.0)

   ridge.fit(x_train, y_train)

]: Ridge
   Ridge(alpha=8.0)

]: y_train_pred = ridge.predict(x_train)
   print(metrics.r2_score(y_true=y_train, y_pred=y_train_pred))

   0.8319885606000128

]: y_test_pred = ridge.predict(x_test)
   print(metrics.r2_score(y_true=y_test, y_pred=y_test_pred))

   0.8401899902701019

]:
```

Predictor Variables –

- OverallCond
- BsmtFullBath
- ExterQual
- BsmtCond
- GarageArea
- Functiona

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 -

We will choose Lasso technique as it is giving feature selection option with it and it has removed unwanted features from model without affecting the model accuracy. Which makes are model generalized and simple and accurate.

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-

Top 5 features are –

Neighborhood_NoRidge, Neighborhood_NridgHt, 2ndFlrSF, OverallQual, Neighborhood_Veenker. After dropping them model accuracy reduced from 80 and 81% to 55% and 58%. Now top most features are: Next top 5 features after dropping 5 main predictors

MasVnrType_BrkFace

RoofMatl_Metal

RoofMatl_Membran

RoofMatl_WdShake

RoofMatl_Roll

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 –

To make model robust and generalisable, we need below 3 things to perform well for us.

I. Model accuracy should be > 70-75% (our case its coming 80%(Train) and 81%(Test) which is correct)
II. P-value of all the features is < 0.05
III. Lastly we need VIF of all the features are < 5 Thus we are sure that model is robust and generalisable.