Problem Statement - Part II

Assignment Part-II

Name: Arunkumar

Assignment-based Subjective Questions

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?

Optimal Alpha value for Ridge: {'alpha': 1.6}

Optimal Alpha value for Lasso: {'alpha': 0.0001}

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.536295e-01	0.947998	0.943149
1	R2 Score (Test)	8.834611e-01	0.898155	0.905573
2	RSS (Train)	2.709449e+11	0.585965	0.640602
3	RSS (Test)	2.696234e+11	0.454402	0.421307
4	RMSE (Train)	1.661058e+04	0.024428	0.025541
5	RMSE (Test)	2.527683e+04	0.032814	0.031597

Double the Optimal Alpha value for Ridge: {'alpha': 3.2}

Double the Optimal Alpha value for Lasso: {'alpha': 0.0002}

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.536295e-01	0.943438	0.934957
1	R2 Score (Test)	8.834611e-01	0.898562	0.904065
2	RSS (Train)	2.709449e+11	0.637352	0.732914
3	RSS (Test)	2.696234e+11	0.452586	0.428036
4	RMSE (Train)	1.661058e+04	0.025476	0.027319
5	RMSE (Test)	2.527683e+04	0.032749	0.031848

By theoretically, After doubling the optimum alpha, model bias will increase. Here In out, R2_train score has decreased

Assignment-based Subjective Questions

1. What will be the most important predictor variables after the change is implemented?

Optimal Alpha value

```
beta_coef['Ridge'].sort_values(ascending=False)
```

1stFlrSF	0.136124
BsmtFinSF1	0.094172
2ndFlrSF	0.089157
OverallQual_V_Excel	0.083377
RoofMatl_WdShngl	0.067372
OverallQual_Excellent	0.056303
TotRmsAbvGrd	0.047201
BsmtUnfSF	0.047002
MasVnrArea	0.043498
Neighborhood_StoneBr	0.040209
LotArea	0.040076
SaleType_New	0.036778
BsmtFinSF2	0.036281
GarageArea	0.033832
Neighborhood_NoRidge	0.031802
Exterior1st_BrkFace	0.029798
OverallCond_Excellent	0.029722
OverallQual_V_good	0.028819
ScreenPorch	0.028533
FullBath	0.027615

```
beta_coef['Lasso'].sort_values(ascending=False)
```

1stFlrSF	0.228027
2ndFlrSF	0.124537
OverallQual_V_Excel	0.122056
BsmtFinSF1	0.084217
OverallQual_Excellent	0.079979
RoofMatl_WdShngl	0.062750
SaleType_New	0.041203
MasVnrArea	0.040651
Neighborhood_StoneBr	0.035421
OverallQual_V_good	0.033610
GarageArea	0.031208
LotArea	0.028993
Neighborhood_NridgHt	0.028634
Neighborhood_NoRidge	0.028437
OverallCond_Excellent	0.027292
Exterior1st_BrkFace	0.026959
BsmtUnfSF	0.025601
BsmtExposure_Gd	0.023706
BsmtFinSF2	0.023312
Neighborhood_Crawfor	0.022966

Double Optimal Alpha value

```
beta_coef2['Ridge'].sort_values(ascending=False)
```

1stFlrSF	0.110734
BsmtFinSF1	0.081349
OverallQual_V_Excel	0.074142
2ndFlrSF	0.068898
RoofMatl_WdShngl	0.056522
OverallQual_Excellent	0.052298
TotRmsAbvGrd	0.049884
MasVnrArea	0.043679
BsmtUnfSF	0.039525
LotArea	0.038788
Neighborhood_StoneBr	0.037957
GarageArea	0.034623
Neighborhood_NoRidge	0.034594
FullBath	0.031834
OverallQual_V_good	0.028395
SaleType_New	0.028138
Exterior1st_BrkFace	0.027078
BsmtExposure_Gd	0.026258
OverallCond_Excellent	0.026179

```
beta_coef2['Lasso'].sort_values(ascending=False)
```

1stFlrSF	0.234709
OverallQual_V_Excel	0.132964
2ndFlrSF	0.114980
OverallQual_Excellent	0.087676
BsmtFinSF1	0.072872
SaleType_New	0.044355
RoofMatl_WdShngl	0.041820
MasVnrArea	0.038328
OverallQual_V_good	0.037150
GarageArea	0.034854
LotArea	0.029803
Neighborhood_StoneBr	0.026977
Neighborhood_NridgHt	0.025543
Neighborhood_NoRidge	0.023295
BsmtExposure_Gd	0.023043
Exterior1st_BrkFace	0.022476
Functional_Typ	0.022453
Neighborhood_Crawfor	0.021970
GarageCars	0.018101

The Important predictor variables after change is implemented are

```
'1stFlrSF',
'OverallQual_V_Excel',
'2ndFlrSF',
'OverallQual_Excellent',
'BsmtFinSF1',
'SaleType_New',
'RoofMatl_WdShngl',
'MasVnrArea',
'OverallQual_V_good',
'GarageArea',
'LotArea',
'Neighborhood_StoneBr',
'Neighborhood_NridgHt',
'Neighborhood_NoRidge',
'BsmtExposure_Gd',
'Exterior1st_BrkFace',
'Functional_Typ',
'Neighborhood_Crawfor',
'GarageCars'
```

Assignment-based Subjective Questions

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?

From ridge & lasso regression, I'll choose Lasso Regression model because of the high R2_Score(Test) - 0.90 and it is close to its R2_Score(Train) - 0.94

Lasso Regression model uses very less number of predictor variables when compared to the Ridge regression model

RMSE value is also very less in Lasso

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.536295e-01	0.947998	0.943149
1	R2 Score (Test)	8.834611e-01	0.898155	0.905573
2	RSS (Train)	2.709449e+11	0.585965	0.640602
3	RSS (Test)	2.696234e+11	0.454402	0.421307
4	RMSE (Train)	1.661058e+04	0.024428	0.025541
5	RMSE (Test)	2.527683e+04	0.032814	0.031597

Assignment-based Subjective Questions

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?

After removing the top 5 significant variable, the next 5 significant variable will be

```
'SaleType_New',
'RoofMatl_WdShngl',
'MasVnrArea',
'OverallQual_V_good',
'GarageArea',
'LotArea',
```

Assignment-based Subjective Questions

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?

   Model which learns the generic pattern of the data from seen data & able to predict the output of target variable in unseen data is robust & generalisable

   r2_score is the one of measure of linear regression model along with RMSE & MSE.

   r2_score represents the amount of variability explained by the model

   For ex:

   a. **If the r2_score of test data is too low when compare to it's train data, it means the model has memorised the noise along with pattern and it is a clear sign of overfitting**

   b. **If the r2_score of train data itself is too low, then the model was not able to find any pattern and it is under-fitting**

Thank you

Arunkumar