Assignment Advanced Regression Subjective Questions

Wednesday, November 22, 2023 8:04 PM

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?

Lasso : 20

Ridge : 0.3

→ original values

```
R2 Score of Lasso Training - 0.9049
R2 Score of Lasso Testing - 0.9
Index(['1stFlrSF', 'OverallQual', '2ndFlrSF', 'LotArea', 'OverallCond',
       'GarageArea', 'BsmtFinSF1', '60_MSSubClass', 'No Basement_BsmtFinType1',
       'Crawfor_Neighborhood'],
      dtype='object')

1stFlrSF                  88077.395260
OverallQual               86650.256810
2ndFlrSF                  57657.207599
LotArea                   32783.915251
OverallCond               24882.076161
GarageArea                24068.571639
BsmtFinSF1                23394.243135
60_MSSubClass             22098.850542
No Basement_BsmtFinType1  17848.691188
Crawfor_Neighborhood      17758.376903
dtype: float64
```

Result for Lasso Original optimal alpha

```
R2 Score of Ridge Training - 0.91
R2 Score of Lasso Testing - 0.9
Index(['1stFlrSF', 'OverallQual', '2ndFlrSF', 'LotArea', 'OverallCond',
       'BsmtFinSF1', 'GarageArea', '60_MSSubClass', 'No Basement_BsmtFinType1',
       'NWAmes_Neighborhood'],
      dtype='object')

1stFlrSF                  84566.153377
OverallQual               83832.213125
2ndFlrSF                  58654.231171
LotArea                   34500.214408
OverallCond               25685.434476
BsmtFinSF1                25334.953133
GarageArea                24253.592286
60_MSSubClass             22654.744455
No Basement_BsmtFinType1  20577.967067
NWAmes_Neighborhood       18597.088906
dtype: float64
```

Doubled alpha = 40 Ridge

Even after change most important Predictor for Lasso is 1st FlrSF

```
R2 Score of Lasso Training - 0.9039
R2 Score of Lasso Testing - 0.9
Index(['1stFlrSF', 'OverallQual', '2ndFlrSF', 'LotArea', 'GarageArea',
       'OverallCond', '60_MSSubClass', 'BsmtFinSF1', 'Crawfor_Neighborhood',
       'NWAmes_Neighborhood'],
      dtype='object')

1stFlrSF                90230.994607
OverallQual             86778.341950
2ndFlrSF                55414.518474
LotArea                 30445.577510
GarageArea              24423.517292
OverallCond             23153.409791
60_MSSubClass           21534.445627
BsmtFinSF1              21467.585436
Crawfor_Neighborhood    18353.049865
NWAmes_Neighborhood     14958.341562
dtype: float64
```

Original Ridge

For Ridge two the most Important feature before & after change is 1st FlrSF

```
R2 Score of Ridge Training - 0.91
R2 Score of Lasso Testing - 0.9
Index(['1stFlrSF', 'OverallQual', '2ndFlrSF', 'LotArea', 'BsmtFinSF1',
       'OverallCond', 'GarageArea', '60_MSSubClass',
       'No Basement_BsmtFinType1', 'NWAmes_Neighborhood'],
      dtype='object')

1stFlrSF                  83102.032463
OverallQual               81417.002249
2ndFlrSF                  57612.030653
LotArea                   33924.073525
BsmtFinSF1                25547.451168
OverallCond               25030.103629
GarageArea                24728.576938
60_MSSubClass             22606.041182
No Basement_BsmtFinType1  20187.751263
NWAmes_Neighborhood       17943.833023
dtype: float64
```

Alpha = 0.6 Ridge α doubled

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?

I would apply Lasso Regression as the Scores for ridge & lasso are extremely close in the

"Q ... Regression as the scores for ridge & lasso are extremely close, in this case choosing Lasso which applies feature selection & reduces model complexity is best.

Unf_BsmtFinType1	1693.396770
HeatingQC	1012.833631
Stone_MasVnrType	902.864089
BrkFace_MasVnrType	0.000000
No Masonry_MasVnrType	0.000000

dtype: float64

Lasso has eliminated 2 features in the model

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?

```
R2 Score of Lasso Training - 0.8414
R2 Score of Lasso Testing - 0.88
Index(['GarageArea', 'BsmtFinSF1', 'BsmtUnfSF', 'No Basement_BsmtFinType1',
       'StoneBr_Neighborhood'],
      dtype='object')

GarageArea                51040.427852
BsmtFinSF1                50685.485774
BsmtUnfSF                 44426.127861
No Basement_BsmtFinType1  43278.467027
StoneBr_Neighborhood      38614.199810
dtype: float64
```

these are the 5 most important features after dropping original 5 most imp features

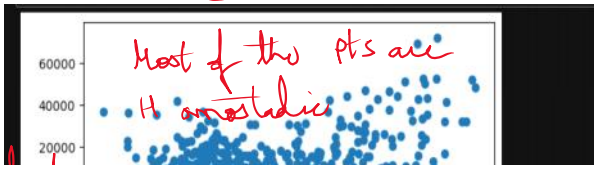
1stFlrSF	88077.395260
OverallQual	86650.256810
2ndFlrSF	57657.207599
LotArea	32783.915251
OverallCond	24882.076161

→ these are original 5 most important features

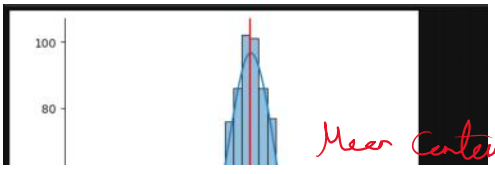
Question 4
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

To make sure the model is robust & generalizable I applied hyperparameter tuning on the train set by doing a K-Fold Grid Search. Thus the test data was completely unseen.

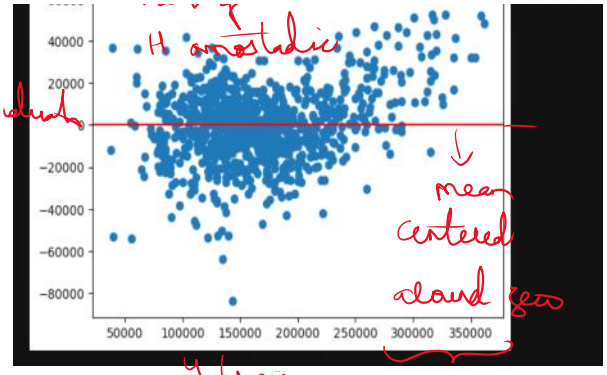
Also By checking the residual distribution & making sure the residuals are normally distributed by being centered around zero.



Most of the pts are in a straight line



Mean Centered



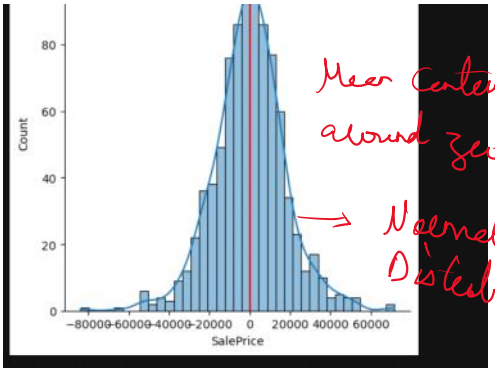
residuals

Homoscedicity

mean centered around zero

y_train

↳ pts in this region have a different variance, but this is within acceptable large given data



Mean Centered around zero

→ Normal Distribution

```
R2 Score of Lasso Training - 0.9049
R2 Score of Lasso Testing - 0.9
```

↳ Similar scores on train & test data

↳ unseen