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:

For Ridge Regression

As per the analysis (Found in Jupyter Notebook under section)

> **When alpha is double i.e 8**

The optimal alpha that was obtained when the model was build was **4**

I did the calculation Alpha=8

WHEN ALPHA =4

THE values of r2 for train set, r2 for test set, rss for train set

	ALPHA=4	**ALPHA=8**
R2 FOR TRAIN SET	0.9520459366695895	0.9472506303830205
R2 FOR TEST SET	0.9060831317557037	0.9042810859219477
RSS FOR TRAIN SET	5.2118499730671495	5.733024096901077
RSS FOR TEST SET	4.658077256581881	4.747454903754937
RMSE FOR TRAIN SET	0.006340450088889477	0.0069744818697093395
RMSE FOR TEST SET	0.013195686279268783	0.013448880747181124

CONCLUSIONS

1)The library indeed has selected the best alpha from the above table

When alpha is 8 the r2 value is lower indicating lower proportion of variance when compared to alpha=4

RSS is also higher when alpha=8. The lower the RSS , the better the model

RMSE- Rootmean square error, the lower the RMSE, the better the model

Overall higher value of alpha leads to greater RSS and RMSE and lower R2 . All are not good for the model which indicates that the model could be further optimized.

Comparing the Beta coefficients(refer notebook for full list)

```
In [62]: betas = pd.DataFrame(index=X.columns)

In [63]: betas.rows = X.columns

In [64]: betas['Ridge'] = ridge.coef_
         betas['RidgeWhenAlphaDoubled'] = ridgedoubled.coef_

In [65]: pd.set_option('display.max_rows', None)
         betas.head(200)
```

Out[65]:

	Ridge	RidgeWhenAlphaDoubled
Id	-0.000010	-0.000011
LotFrontage	0.000101	0.000095
LotArea	0.000005	0.000005
Street	0.000000	0.000000
Utilities	0.000000	0.000000
YearBuilt	-0.002850	-0.002778
YearRemodAdd	-0.001183	-0.001397
MasVnrArea	-0.000030	-0.000027
BsmtFinSF1	0.000077	0.000081
BsmtFinSF2	0.000073	0.000075
BsmtUnfSF	-0.000027	-0.000027

If the value of alpha is small, then the magnitude of coefficients will be higher.

For Lasso regression

The optimal alpha that was obtained when the model was build was **0.0001**

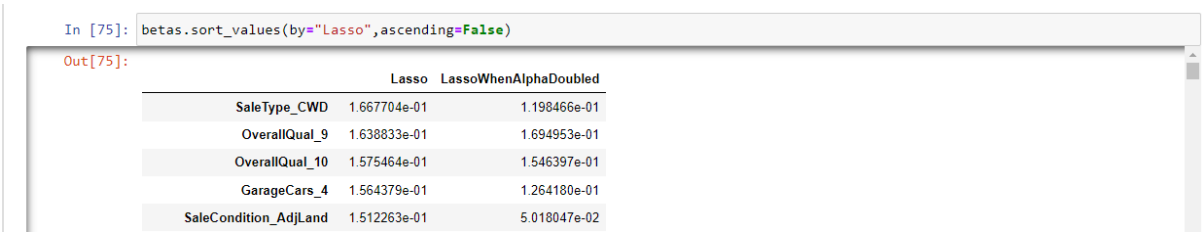
I did the calculation Alpha=0.0002

WHEN ALPHA =0.0001

THE values of r2 for train set, r2 for test set, rss for train set

	ALPHA=0.0001	**ALPHA=0.0002**
R2 FOR TRAIN SET	0.9567950424071987	0.952909420211516
R2 FOR TEST SET	0.8991934854594414	0.903421264963937
RSS FOR TRAIN SET	4.695697119864503	5.118002937754924
RSS FOR TEST SET	4.999789084483058	4.790100197663413
RMSE FOR TRAIN SET	0.005712526909810832	0.0062262809461738735
RMSE FOR TEST SET	0.014163708454626226	0.013569688945222135

TOP 5 PREDICTOR VARIABLES IN LASSO REGRESSION WHEN ALPHA IS 0.0001

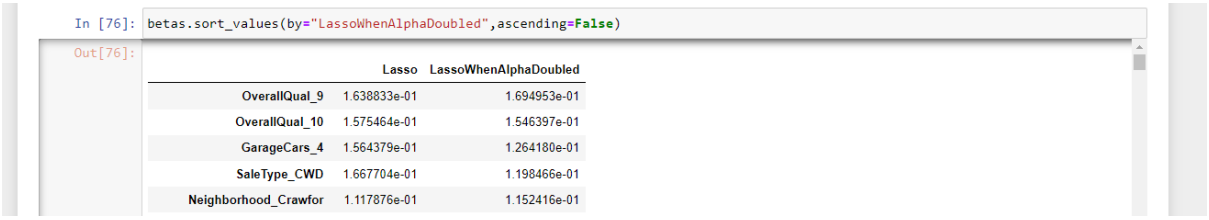


```
In [75]: betas.sort_values(by="Lasso",ascending=False)
```

Out[75]:

	Lasso	LassoWhenAlphaDoubled
SaleType_CWD	1.667704e-01	1.198466e-01
OverallQual_9	1.638833e-01	1.694953e-01
OverallQual_10	1.575464e-01	1.546397e-01
GarageCars_4	1.564379e-01	1.264180e-01
SaleCondition_AdjLand	1.512263e-01	5.018047e-02

TOP 5 PREDICTOR VARIABLES IN LASSO REGRESSION WHEN ALPHA IS 0.0002



```
In [76]: betas.sort_values(by="LassoWhenAlphaDoubled",ascending=False)
```

Out[76]:

	Lasso	LassoWhenAlphaDoubled
OverallQual_9	1.638833e-01	1.694953e-01
OverallQual_10	1.575464e-01	1.546397e-01
GarageCars_4	1.564379e-01	1.264180e-01
SaleType_CWD	1.667704e-01	1.198466e-01
Neighborhood_Crawfor	1.117876e-01	1.152416e-01

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 PREFER LASSO OVER RIDGE AS R2

1)R2 SCORE IS SLIGHTLY BETTER

2)COEFFICIENTS OF ALMOST 100+ PREDICTOR VARIABLES ARE 0 AND SMALLER NUMBER OF SIGNIFICANT PARAMETERS SEEM TO DECIDE THE MODEL

3)Model is more robust in lasso

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?

TOP 5 PREDICTOR VARIABLES IN LASSO REGRESSION WHEN ALPHA IS 0.0001

```
In [75]: betas.sort_values(by="Lasso",ascending=False)
```

Out[75]:

	Lasso	LassoWhenAlphaDoubled
SaleType_CWD	1.667704e-01	1.198466e-01
OverallQual_9	1.638833e-01	1.694953e-01
OverallQual_10	1.575464e-01	1.546397e-01
GarageCars_4	1.564379e-01	1.264180e-01
SaleCondition_AdjLand	1.512263e-01	5.018047e-02

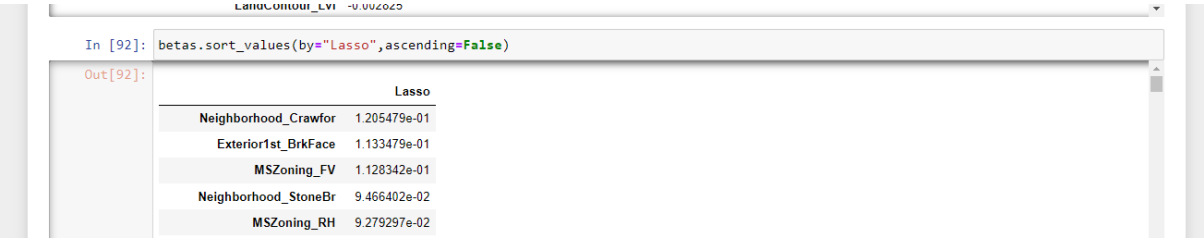
After dropping the top 5 predictor variables

```
In [83]: X_train=X_train.drop(['SaleType_CWD','OverallQual_9','OverallQual_10','GarageCars_4','SaleCondition_AdjLand'], axis=1)

In [84]: X_test=X_test.drop(['SaleType_CWD','OverallQual_9','OverallQual_10','GarageCars_4','SaleCondition_AdjLand'], axis=1)

In [85]: lasso = Lasso()
```

Alpha did not change =0.0001

Hence the rest statistics like r2score, rss, rmse remained the same as question 1

TOP 5 predictor variables now are



```
In [92]: betas.sort_values(by="Lasso",ascending=False)
```

Out[92]:

	Lasso
Neighborhood_Crawfor	1.205479e-01
Exterior1st_BrkFace	1.133479e-01
MSZoning_FV	1.128342e-01
Neighborhood_StoneBr	9.466402e-02
MSZoning_RH	9.279297e-02

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?

Model can be made robust and generalizable by ensuring

1)Outliers are not given lot of weightage

2)Those outliers that are valuable to data set should be made available

3)Delete the other outliers that are unnecessary and which can skew the data

4Tranforming data can also help e.g log transformation

5)By also ensuring that there is a nice trade off between Bias V variance By introducing little bias to reduce variance

6)Also By avoiding over fitting and underfitting

By doing the above 6 points accuracy can be increased and model will be more robust and generalized.