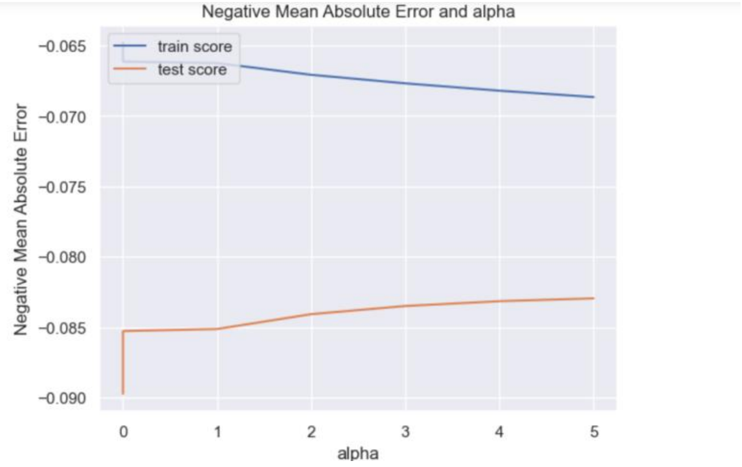
Problem Statement - Part II

Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans:

Ridge Regression:

As I shown below **Negative Mean Absolute Error vs Alpha value**, we can observe that error term is decreasing when alpha value is increasing from zero. I decided to choose the alpha value is 2 because error term is very minimum; same thing is observed in the graph and table which are showing the train, test and RMSE value.



Lasso Regression: As per different alpha value model is generating different train and test values. I compared the train, test and RMSE and decided to chose the alpha=0.01 where test and train values closest each other and RMSE is also low.

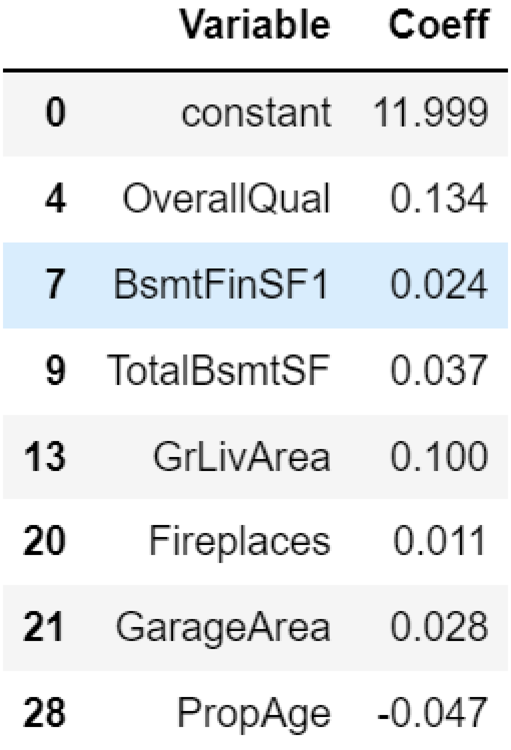
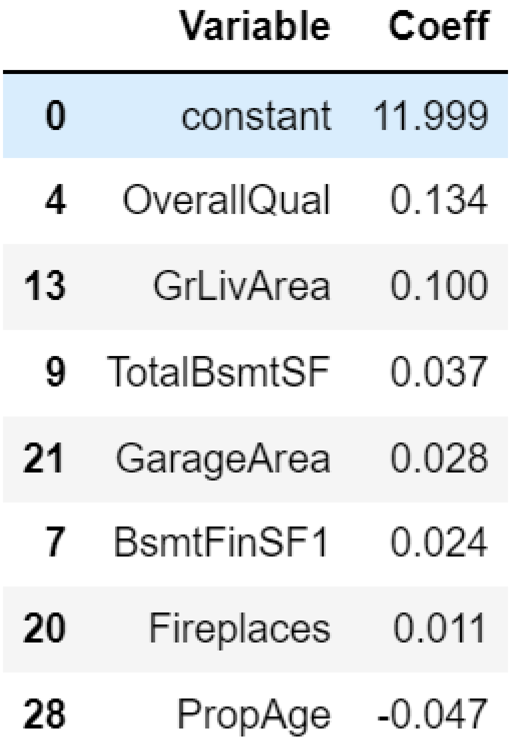
Sl.No	Lasso Regression	Ridge Regression
1	alpha = 0.001 0.9206709153542387 0.9067784241823187 RMSE : 0.11546718823765789	alpha = 1 0.9379382207144428 0.9068442884415822 RMSE : 0.11542639025415641
2	**alpha = 0.01** **0.8854624158407248** **0.8894603158029368** RMSE : 0.12573595366706636	**alpha = 2** **0.9364594823911135** **0.9077597079466583** RMSE : **0.11485785595060986**
3	alpha = 0.05 0.810576713549257 0.8291485394300617 RMSE : 0.15631826218847628	alpha = 3 0.935338638540601 0.9082889363061554 RMSE : 0.1145278836886629
4	alpha = 0.1 0.7012250007915783 0.7190041619079308 RMSE : 0.20047042640027601	alpha = 4 0.9344289827010624 0.9086565586717956 RMSE : 0.11429811158196207

Double the value of alpha for both Ridge and Lasso:

If I double the alpha value for the ridge regression then alpha value equal to 4, the model will apply more penalty on the curve. We can see table section, not much difference in train, test and RMSE values except the increase penalty. We are trying to make the model more generalized that is making model simpler.

Meanwhile, when I doubled value of alpha **i.e.** 0.05 for Lasso Regression then increase the error term in both train and test as well as penalize more when we increase the value of alpha.

If changes are implemented then the most important predictor variables are as shown below table:

<table>
<tr><th>Lasso Regression</th><th>Ridge Regression</th></tr>
<tr>
<td>
<table>
<tr><th></th><th>Variable</th><th>Coeff</th></tr>
<tr><td>0</td><td>constant</td><td>11.999</td></tr>
<tr><td>4</td><td>OverallQual</td><td>0.134</td></tr>
<tr><td>7</td><td>BsmtFinSF1</td><td>0.024</td></tr>
<tr><td>9</td><td>TotalBsmtSF</td><td>0.037</td></tr>
<tr><td>13</td><td>GrLivArea</td><td>0.100</td></tr>
<tr><td>20</td><td>Fireplaces</td><td>0.011</td></tr>
<tr><td>21</td><td>GarageArea</td><td>0.028</td></tr>
<tr><td>28</td><td>PropAge</td><td>-0.047</td></tr>
</table>
</td>
<td>
<table>
<tr><th></th><th>Variable</th><th>Coeff</th></tr>
<tr><td>0</td><td>constant</td><td>11.999</td></tr>
<tr><td>4</td><td>OverallQual</td><td>0.134</td></tr>
<tr><td>13</td><td>GrLivArea</td><td>0.100</td></tr>
<tr><td>9</td><td>TotalBsmtSF</td><td>0.037</td></tr>
<tr><td>21</td><td>GarageArea</td><td>0.028</td></tr>
<tr><td>7</td><td>BsmtFinSF1</td><td>0.024</td></tr>
<tr><td>20</td><td>Fireplaces</td><td>0.011</td></tr>
<tr><td>28</td><td>PropAge</td><td>-0.047</td></tr>
</table>
</td>
</tr>
<tr>
<td>OverallQual
BsmtFinSF1
TotalBsmtSF
GrLivArea
Fireplaces
GarageArea
PropAge</td>
<td>OverallQual
GrLivArea
TotalBsmtSF
GarageArea
BsmtFinSF1
Fireplaces
PropAge</td>
</tr>
</table>

Q2. 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:

As per my analysis, I have chosen the alpha/delta: **0.01** for Lasso and alpha/delta: **2** for Ridge regressions. Please find respective details like R2 train and test, RMSE values:

Lasso Regression:

alpha = 0.01,

R2 Train: 0.8854624158407248

R2 Test: 0.8894603158029368

RMSE : 0.12573595366706636

Ridge Regression:

alpha = 2

R2 Train: 0.9364594823911135

R2 Train : 0.9077597079466583

RMSE : 0.11485785595060986

I have chosen the **Lasso regression** even though, the model performance by Ridge Regression was better in terms of R2 values of Train and Test. Because it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables easily. It is become a simple model.

Log(Y) = C + 0.125(x1) + 0.112(x2) + 0.050(x3) + 0.042(x4) + 0.035(x5) + 0.034(x6) + 0.024(x7) + 0.015(x8) + 0.014(x9) + 0.010(x10) + 0.010(x11) + 0.005(x12) - 0.007(x13) - 0.007(x14) - 0.008(x15) - 0.095(x16) + Error term(RSS + alpha * (sum of absolute value of coefficients)

Regularization helps with managing model complexity by essentially shrinking the model coefficient estimates towards zero. This discourages the model from becoming too complex, avoiding the risk of overfitting.

Q3. 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?

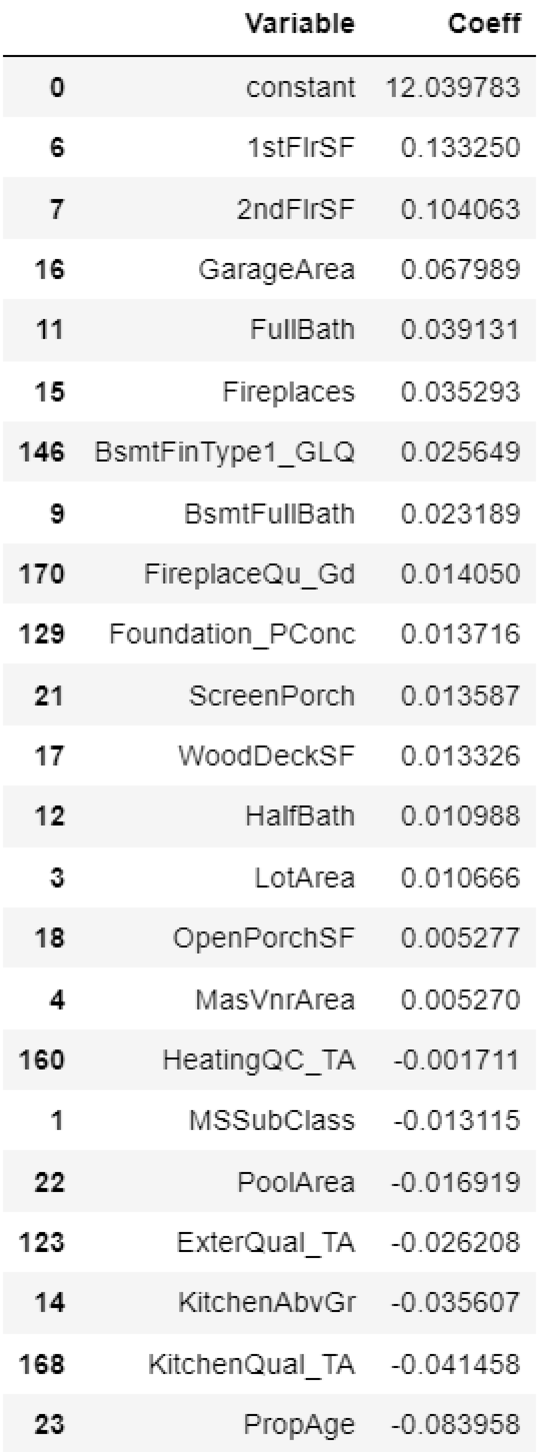
Ans:

New five most important predictor variables are:

1. 1stFlrSF

2. 2ndFlrSF
3. GarageArea
4. FullBath
5. Fireplaces

Now new model gave 23 predictor variables



	Variable	Coeff
0	constant	12.039783
6	1stFlrSF	0.133250
7	2ndFlrSF	0.104063
16	GarageArea	0.067989
11	FullBath	0.039131
15	Fireplaces	0.035293
146	BsmtFinType1_GLQ	0.025649
9	BsmtFullBath	0.023189
170	FireplaceQu_Gd	0.014050
129	Foundation_PConc	0.013716
21	ScreenPorch	0.013587
17	WoodDeckSF	0.013326
12	HalfBath	0.010988
3	LotArea	0.010666
18	OpenPorchSF	0.005277
4	MasVnrArea	0.005270
160	HeatingQC_TA	-0.001711
1	MSSubClass	-0.013115
22	PoolArea	-0.016919
123	ExterQual_TA	-0.026208
14	KitchenAbvGr	-0.035607
168	KitchenQual_TA	-0.041458
23	PropAge	-0.083958

Q4. 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?

Ans:

A model needs to be made robust and generalizable so that it's not impacted by outliers in the training data. The model should be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. More weightages should not give to the outliers so that the accuracy predicted by the model is high. The simpler the model the more the bias but less variance and more generalizable.

Bias is error in model, when the model is weak to learn from the data. High bias means that model is unable to learn from the data. Model performs poor on training and testing data.

Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data. But performs very poor on testing data as it's unseen data for the model.

It’s important that we have to balance in Bias and Variance to avoid overfitting and under-fitting of data.