Subjective Questions

Question 1:

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

Solution:

1) Optimal Value of Alpha for:
    a. Ridge Regression: 4
    b. Lasso Regression: 0.0001
2) Observation when Alpha was doubled for both Ridge & Lass [Details in Jupyter Notebook]
    a. **Ridge**:

<table>
<tr><th>Parameters</th><th>Observation</th><th>Comments</th></tr>
<tr><td>Model Metrics</td><td>
<table>
<tr><th></th><th>Metric</th><th>Doubled Alpha</th><th>Optimal Alpha</th></tr>
<tr><td>0</td><td>R2 Score (Train)</td><td>0.900</td><td>0.906</td></tr>
<tr><td>1</td><td>R2 Score (Test)</td><td>0.869</td><td>0.876</td></tr>
<tr><td>2</td><td>RSS (Train)</td><td>1.085</td><td>1.013</td></tr>
<tr><td>3</td><td>RSS (Test)</td><td>1.044</td><td>0.992</td></tr>
<tr><td>4</td><td>MSE (Train)</td><td>0.034</td><td>0.033</td></tr>
<tr><td>5</td><td>MSE (Test)</td><td>0.051</td><td>0.049</td></tr>
</table>
</td><td>Both train and test data the R2 score value has dropped and MSE and RSS has increased</td></tr>
<tr><td>Model Coefficients</td><td>Alpha: 8
<table>
<tr><th></th><th>Feature</th><th>Coef</th></tr>
<tr><td>0</td><td>MSSubClass</td><td>0.053</td></tr>
<tr><td>4</td><td>OverallCond</td><td>0.052</td></tr>
<tr><td>14</td><td>BsmtFullBath</td><td>0.046</td></tr>
<tr><td>70</td><td>Neighborhood_NridgHt</td><td>0.044</td></tr>
<tr><td>11</td><td>2ndFlrSF</td><td>0.042</td></tr>
<tr><td>7</td><td>BsmtFinSF2</td><td>0.042</td></tr>
<tr><td>12</td><td>LowQualFinSF</td><td>0.040</td></tr>
<tr><td>5</td><td>MasVnrArea</td><td>0.036</td></tr>
<tr><td>77</td><td>Neighborhood_Timber</td><td>0.032</td></tr>
<tr><td>10</td><td>1stFlrSF</td><td>0.032</td></tr>
</table>
Alpha: 4</td><td>On doubling the alpha, the model coefficients has reduced.</td></tr>
</table>

<table>
<tr><td></td><td>
<table>
<tr><th></th><th>Feature</th><th>Coef</th></tr>
<tr><td>0</td><td>MSSubClass</td><td>0.092</td></tr>
<tr><td>14</td><td>BsmtFullBath</td><td>0.060</td></tr>
<tr><td>4</td><td>OverallCond</td><td>0.060</td></tr>
<tr><td>12</td><td>LowQualFinSF</td><td>0.054</td></tr>
<tr><td>11</td><td>2ndFlrSF</td><td>0.053</td></tr>
<tr><td>70</td><td>Neighborhood_NridgHt</td><td>0.048</td></tr>
<tr><td>7</td><td>BsmtFinSF2</td><td>0.047</td></tr>
<tr><td>5</td><td>MasVnrArea</td><td>0.043</td></tr>
<tr><td>77</td><td>Neighborhood_Timber</td><td>0.039</td></tr>
<tr><td>10</td><td>1stFlrSF</td><td>0.037</td></tr>
</table>
</td><td></td></tr>
</table>

a. **Lasso Regression:**

<table>
<tr><th>Parameter</th><th>Observation</th><th>Comments</th></tr>
<tr><td>Model Metrics</td><td>
<table>
<tr><th></th><th>Metric</th><th>Doubled Alpha</th><th>Optimal Alpha</th></tr>
<tr><td>0</td><td>R2 Score (Train)</td><td>0.897</td><td>0.905</td></tr>
<tr><td>1</td><td>R2 Score (Test)</td><td>0.881</td><td>0.883</td></tr>
<tr><td>2</td><td>RSS (Train)</td><td>1.118</td><td>1.030</td></tr>
<tr><td>3</td><td>RSS (Test)</td><td>0.949</td><td>0.934</td></tr>
<tr><td>4</td><td>MSE (Train)</td><td>0.034</td><td>0.033</td></tr>
<tr><td>5</td><td>MSE (Test)</td><td>0.048</td><td>0.048</td></tr>
</table>
</td><td>On doubling alpha the for both train and test data the R2 score has reduced slightly and the MSE has increased</td></tr>
<tr><td>Model Coefficients</td><td>Alpha: 0.0002
<table>
<tr><th></th><th>Feature</th><th>Coef</th></tr>
<tr><td>14</td><td>BsmtFullBath</td><td>0.218</td></tr>
<tr><td>4</td><td>OverallCond</td><td>0.089</td></tr>
<tr><td>70</td><td>Neighborhood_NridgHt</td><td>0.051</td></tr>
<tr><td>7</td><td>BsmtFinSF2</td><td>0.051</td></tr>
<tr><td>5</td><td>MasVnrArea</td><td>0.049</td></tr>
<tr><td>142</td><td>ExterQual_Fa</td><td>0.045</td></tr>
<tr><td>77</td><td>Neighborhood_Timber</td><td>0.041</td></tr>
<tr><td>10</td><td>1stFlrSF</td><td>0.034</td></tr>
<tr><td>197</td><td>KitchenQual_Fa</td><td>0.032</td></tr>
<tr><td>71</td><td>Neighborhood_OldTown</td><td>0.029</td></tr>
</table>
</td><td>On doubling alpha the model coefficients has increased</td></tr>
</table>

Alpha: 0.0001

	Feature	Coef
14	BsmtFullBath	0.225
4	OverallCond	0.082
70	Neighborhood_NridgHt	0.053
5	MasVnrArea	0.052
77	Neighborhood_Timber	0.047
7	BsmtFinSF2	0.047
142	ExterQual_Fa	0.046
3	OverallQual	0.043
10	1stFlrSF	0.041
71	Neighborhood_OldTown	0.031

3) Most Important Variable after change is implemented [alpha is doubled]:

   a. **Ridge Regression:**

      i.

	Feature	Coef
0	MSSubClass	0.053
4	OverallCond	0.052
14	BsmtFullBath	0.046
70	Neighborhood_NridgHt	0.044
11	2ndFlrSF	0.042
7	BsmtFinSF2	0.042
12	LowQualFinSF	0.040
5	MasVnrArea	0.036
77	Neighborhood_Timber	0.032
10	1stFlrSF	0.032

   b. **Lasso Regression:**

	Feature	Coef
14	BsmtFullBath	0.218
4	OverallCond	0.089
70	Neighborhood_NridgHt	0.051
7	BsmtFinSF2	0.051
5	MasVnrArea	0.049
142	ExterQual_Fa	0.045
77	Neighborhood_Timber	0.041
10	1stFlrSF	0.034
197	KitchenQual_Fa	0.032
71	Neighborhood_OldTown	0.029

i.

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?

Solution:

Comparing R2 Score(test) for Lasso vs Ridge, [0.883 vs 0.876] and other metrics, Lasso has an edge over Ridge. Therefore we will work with Lasso as it gives the option of feature selection along with regularization. It removes unwanted features from the model without affecting the model accuracy. In Lasso, some of the coefficients become 0, thus resulting in feature selection and, hence, easier interpretation, particularly when the number of coefficients is very large.

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?

Solution:

Top 5 Variables identified by Lasso are:

	Feature	Coef
14	BsmtFullBath	0.225
4	OverallCond	0.082
70	Neighborhood_NridgHt	0.053
5	MasVnrArea	0.052
77	Neighborhood_Timber	0.047

If we exclude these variables, the new five most important predictor variables are[refer Jupyter Notebook]:

	Feature	Coef
12	BsmtHalfBath	0.211
4	BsmtFinSF1	0.096
5	BsmtFinSF2	0.053
137	ExterQual_Fa	0.051
67	Neighborhood_OldTown	0.050

Metrics:
R2 train : 0.8989486557032434
R2 test : 0.8760921599704945
RSS train : 1.0933267728297622
RSS test : 0.9907016254405436
MSS train : 0.0011496601186432832
MSS test : 0.002428190258432705

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?

Solution:

A robust and generalisable model is one which has low training error and low testing error. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable.

To make such model following things are essential:

1. Avoiding overfitting by doing Regularization while model making. In overfitting, a model fits the training data but fails to generalize and hence, cannot be used as the model to predict on new data or out-of-sample data. Regularization helps to avoid overfitting as well underfitting, keeping bias & variance trade off at its best. We use regularization because we want our models to work well with unseen data, without missing out on identifying underlying patterns in the data

Implications on model accuracy

By making robust and generalized model i.e. by introducing regularization we compromise accuracy to some extent as we allow a little bias for a significant reduction in variance. A robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Reason for this implication

This happens because Regularization introduces a penalty, which grows in relation to the size of the coefficients and reduces its impact, thus making the model less sensitive to small changes in the variables. More extreme model coefficients values gives better accuracy but lead to a large variance. Regularization prevents this by shrinking the coefficients towards 0.