Question 1

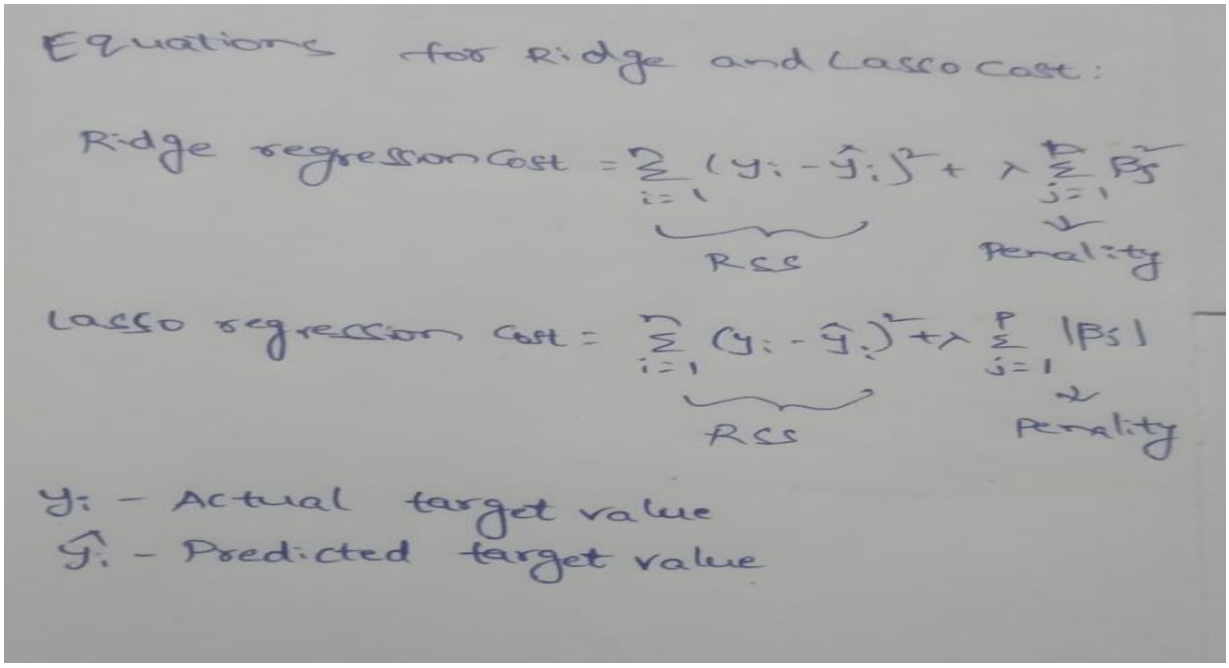
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Optimal Value of alpha:

The optimal value of Ridge:2

The optimal value of Lasso:0.002

If we double the optimal value of alpha:



From the above equations also we can observe if we increase the value of alpha the penalty also be high. So that Ridge and Lasso models will leads the coefficients near to zero. If we increase the value of alpha in the ridge model coefficients are near to zero and in lasso model coefficients is zero.

Top 10 features after doubling the alpha value in Ridge regression:

	Params	**Coef**
4	OverallQual	0.083
13	GrLivArea	0.082
9	TotalBsmtSF	0.069
7	BsmtFinSF1	0.064

10	1stFlrSF	0.061
5	OverallCond	0.057
63	Neighborhood_StoneBr	0.049
22	GarageArea	0.048
56	Neighborhood_NoRidge	0.046

Top 10 features after doubling the alpha value in Lasso regression:

	Params	**Coef**
13	GrLivArea	0.148
135	BsmtFinType1_GLQ	0.030
7	BsmtFinSF1	0.021
9	TotalBsmtSF	0.021
4	OverallQual	0.020
122	Foundation_PConc	0.019
3	LotArea	0.016
169	SaleCondition_Partial	0.014
152	GarageType_Attchd	0.013

After doubling the alpha the most important predictor variables are:

Ridge: OverallQual

Lasso: GrLivArea

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?

The optimal values of Ridge and Lasso regression is:

- Ridge-2
- Lasso-0.002

The results of test data for ridge and Lasso:

Ridge:

- R2 score: 0.9168089469736156
- RMSE: 0.04035242711079925

Lasso:

- R2 score: 0.8318682696079184
- RMSE: 0.05736621447361267

Model which is very complex is not best model. We can observe that Lasso regression model has eliminated more features than ridge model. In Lasso it was tending the coefficients to be almost zero or near to zero. As simpler model was better I will choose Lasso 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?

First top 5 features in Lasso model are as below:

GrLivArea 0.169

OverallQual 0.115

BsmtFinSF1 0.060

TotalBsmtSF 0.044

Neighborhood_NridgHt 0.032

GarageArea 0.030

After excluding these variables now top 5 important variables was:

SaleCondition_Partial 0.029

LotArea 0.027

BsmtFinType1_GLQ 0.021

BsmtExposure_Gd 0.019

Foundation_PConc 0.014

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?

The model which is simple to be preferable. The model should be complex too but not very complex.

Occam's Razor also suggests that simpler model is be preferred over a more complex model.

If we choose complex model there will be chance of overfitting. Normally overfitting will occur when the variance was high. If the bias is very high the underfitting will occur which can lead incorrect predections.

If model is too simple the variance is low and bias is high the cause underfitting. If model is too complex high variance and low bias then it will go under overfitting.

A robust model should have low variance and low bias so that it will not go under overfitting and underfitting. To make model robust we have to follow bias variance trade-off.