Problem Statement - Part II

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?

Ans:
Optimal value of alpha for Ridge regression - 50
Optimal value of alpha for Lasso regression - 500

Important predictors before the change

Top 5	Ridge	Lasso
1	GrLivArea(18783.02)	GrLivArea(24549.1)
2	TotalBsmtSF(16682.67)	TotalBsmtSF(15243.85)
3	OverallQual_10(9970.464)	OverallQual_9(11738)
4	OverallQual_9(9768.693)	YearBuilt(11396.95)
5	YearBuilt(9051.595)	OverallQual_10(10710.95)

r2 score	Ridge	Lasso
Train	0.922627453	0.920546798
Test	0.894278801	0.894314996

After doubling,
Optimal value of alpha for Ridge regression - 100
Optimal value of alpha for Lasso regression - 1000

Top 5 features after alpha is doubled

Top 5	Ridge	Lasso
1	GrLivArea(16319.77)	GrLivArea(24464.59)
2	TotalBsmtSF(15029.6)	TotalBsmtSF(14134.25)
3	OverallQual_10(9970.464)	OverallQual_9(11589.36)
4	OverallQual_9(9592.473)	YearBuilt(10909.79)
5	YearBuilt(7562.401)	OverallQual_10(10413.65)

r2 score after alpha is doubled

r2 score	Ridge	Lasso
Train	0.918272138	0.913953619
Test	0.890103858	0.887514245

We can clearly see that when we increase the alpha (double) the coefficients get decreased but the features remain the same and also the r2score of the train is getting decreased since when we keep on increasing the alpha it will lead to underfitting and getting high bias in the model.

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?

Ans:
I am choosing Lasso's lambda. Since we are getting better r2 score in lasso when comparing the ridge and also since the lasso makes the feature selection as well. I am choosing Lasso.

Optimal value of alpha for Lasso regression - 500

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?

Ans:
Five most important predictor variables before exclusion are,

- GrLivArea
- TotalBsmtSF
- OverallQual_9
- YearBuilt
- OverallQual_10

After excluding these five variables, building a lasso model again and getting the below five variables as the next five important predictors.

- OverallQual_5
- OverallQual_6
- TotRmsAbvGrd_5
- TotRmsAbvGrd_6
- TotRmsAbvGrd_7

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?

Ans:
The model will be robust and generalisable when we have low bias and low variance.
I.e,With the help of regularization by tuning the lambda, we can reduce the coefficients of the variables which eventually reduces the model's bias.
Ridge will reduce the coefficients and make it go towards zero.
Lasso will reduce and will make it zero which helps in feature selection as well.

So, when the difference in r2 score is less in between train and test(variance) and also we have good amount of accuracy in them(bias), our model will be robust and generalisable.

Note: Given the code for the questions in the jupyter notebook itself.