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 – After performing Ridge and lasso regression my optimal values are –

- Ridge – 41
- Lasso – 0.005

If we double the value of alpha for ridge and lasso a more heavier penalty will be imposed on the cost term which might lead to underfitting of the data , in case of lasso more features' coefficient will become zero and for ridge the coefficient will tend to be zero

Most Important predictors are –

- Age of the house(how old is property) - inverse relation
- total baseemnt surface
- Overrall condition
- overrall quality
- grliving area

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?

Answer –

For ridge regression I obtained the rsme – 0.352174 and for lasso I obtained a value of 0.3478 which are already comparable but lasso provides for feature selection as well which would help in reducing significant amount of work of manual feature elimination or by using rfe hence will suggest lasso regression

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?

Answer – the top features after lasso are –

1. 'GrLivArea'
2. , 'OverallQual',
3. 'OverallCond',
4. 'TotalBsmtSF',
5. 'Age of House'

If we have to eliminate these features then we can go on to these features –

	Features	rfe_support	rfe_ranking	Coefficient
2	2ndFlrSF	True	1	0.343444
1	1stFlrSF	True	1	0.294310
0	BsmtQual	True	1	0.088736
4	GarageCars	True	1	0.084560
3	KitchenQual	True	1	0.083817

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?

Answer –

while building a model , one thing should be kept in mind that model should be kept as simple as possible adding , adding too much complexity (by adding too many unnecessary features which leads to a situation of curse of dimensionality or using a more complex model architecture) in the model will cause a high variance causing the model perform poorly in actual real world scenario. Model tend to perform poorly in case outliers also considered while training process hence outliers should be properly identified according to the business use case and then treated.

Proper regularization should be used in case the data starts to overfit the samples so that when tested in real world scenario it should not perform poorly

The data included in the training should be a proper subset of the population that the model will be used in the production or test environment, if not done so then we will have a reduction in accuracy against the accuracy stated when model was handed over