Q1. 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?

A1. The default value of alpha for lasso regression is 1. as the value of alpha increases, the model complexity reduces and thus model starts underfitting. But if we keep on decreasing the values of alpha the coefficients become too large and every parameter becomes important hence causing overfitting. Thus, alpha should be chosen wisely. A widely accept technique is cross-validation, i.e. the value of alpha is iterated, range of values being provided and the one giving higher cross-validation score is chosen.

For lasso regression I have chosen the alpha 0.01. Similarly, for ridge regression I have chosen the alpha 2.

When we double the value of alpha, the penalty the model has applied on the curve will increase, increasing the generalization of the model. The model can hence get oversimplified because more coefficients get reduced to zero. Also the cost function decreases.

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?

A2. Lasso seems to perform well if there are a small number of significant parameters and the others are close to zero

Ridge seems effective if there are many large parameters of about the same value

Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables.

It is always advisable to use simple yet robust model.

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?

A3. The five most important predictor variables excluded according to my model are:

1. GrLivArea
2. OverallQual
3. OverallCond

4. TotalBsmtSF
5. BsmtFinSF1

The next five predictor variables should be:

1. GarageArea
2. Fireplaces
3. LotArea
4. LotFrontage
5. BsmtFullBath

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?

A4. The model should be an optimum balance between accuracy and robustness and generalisability. Too much accuracy might lead to an unstable model while too stable a model might lead to underfitting.

This can be better understood using bias variance trade-off.

Bias
The bias is known as the difference between the prediction of the values by the ML model and the correct value. High bias gives a large error in training and testing data. An algorithm should always be low in bias to circumvent the problem of underfitting.

Variance
Variance of a model tells us spread of our data. The model with high variance has a very complex fit to the training data. Thus it is not able to fit accurately on the unseen data. As a result, such models perform very well on training data but has high error on test data.
When a model is high on variance, it is said as Overfitting.

Bias Variance Trade-off

If the algorithm is too simple then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.