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?

We use Ridge Regression when there is multicollinearity in the data. Least square estimates will be unbiased, but their variances are large(i.e, they may be for from the true value) when you have multicollinearity in the data. Ridge regression is good when OLS is overfitting

Lasso regression shrinks the coefficients towards the zero. Lasso procedure encourages simple model which mean model with fewer parameters. Primary difference between Lasso and Ridge is their penalty term. Lasso performs variable selection and can be exactly zero unlike Ridge

Below Formula has written in my notebook and took the screen print. Hope this is fine.

FORMULA – Ridge → $\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

LASSO → $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

For Ridge regression optimal value is 20 where in for lasso 1 is the optimal value.

If we double the alpha value for lasso and ridge regression, the complexity of the model will have greater contribution to cost.

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 major difference between Ridge and Lasso is, ridge regression cannot zero out the coefficients whereas lasso can be zero and does parameter shrinkage and variable selection. For a given problem, Ridge produces only one solution whereas Lasso can produce multiple solutions to the problem

For a given lambda value, You use Ridge regression when there is multicollinearity in the data and use lasso when you want to perform variable selection

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 of all you should have subject area knowledge to find out the most important predictor variables. Statical measure's can show the qualified importance of the different forecaster variables, however those measure's cannot determine these variables are important in a practical sense.

How you collect and measure your sample can bias the seeming importance of the variables in your sample related to their true importance over population

If you randomly sample the observations, the variability of the predictor values in your sample likely reflects that unpredictability of the population

Consider the accuracy and precision of the measurements for the predictors because this can affect their apparent importance. Ex – Lower quality measurements can cause a variable to appear less projecting than it really is

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 designed model should be robust and generalizable, so that they won't get impacted by the outliers and the model should be generalizable to test the accuracy is not less than the training score. Model should be accurate for the datasets that was not trained on during the training. No need of giving, too much weight to the outliers, so that the accuracy projected by the model is high. Before creating the model, drop the outliers data that doesn't make sense. Closer to 100% accuracy indicated all the predictions are correct. One thing, we should keep in mind is there should be some balance between accuracy and complexity.