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?

Value of alpha depends on the dataset. We use iterative method to find the optimal value of alpha. As value of alpha doubles the variance in the model starts decreases and model becomes more generalized.

In ridge regression if you double the alpha model will be less likely to overfit and move toward underfit.The lesser important coefficients of predictor variables will reduce further towards zero(but not zero)

In Lasso regression if we double the alpha it will increase the cutoff for predictor variables coefficients and less number of predictor variable will make it to the model. This will simplify the model further but there also chance of underfit.

In ridge regression all the predictor variables will remain but the impact of the predictor variables will be reduced further.
In lasso regression more predictor variables will have zero coefficients and will lead to pruning of variables leaving only the most important ones.

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 choice of ridge and lasso will depend on dataset, business needs and domain expertise.

Ridge regression will be applied if we think or business tells us that all the parameters have meaningful impact on the predicted variables and yet you want to reduce impact of multi collinearity. Ridge regression will add penalty such that coefficients impact is reduced without eliminating it totally.
Lasso regression will be applied if we think or business tells us that not all parameters have meaningful impact on the predicted variables. Lasso is mostly used for feature selection.Some coefficients will be pushed to zero based on value of alpha.This will reduce model complexity.

Selection of Lasso and Ridge
Select Ridge Regression if we want to retain most features and reduce multi collinearity.

Select Lasso Regression if we want feature selection. Model will simplified with most important features.

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?

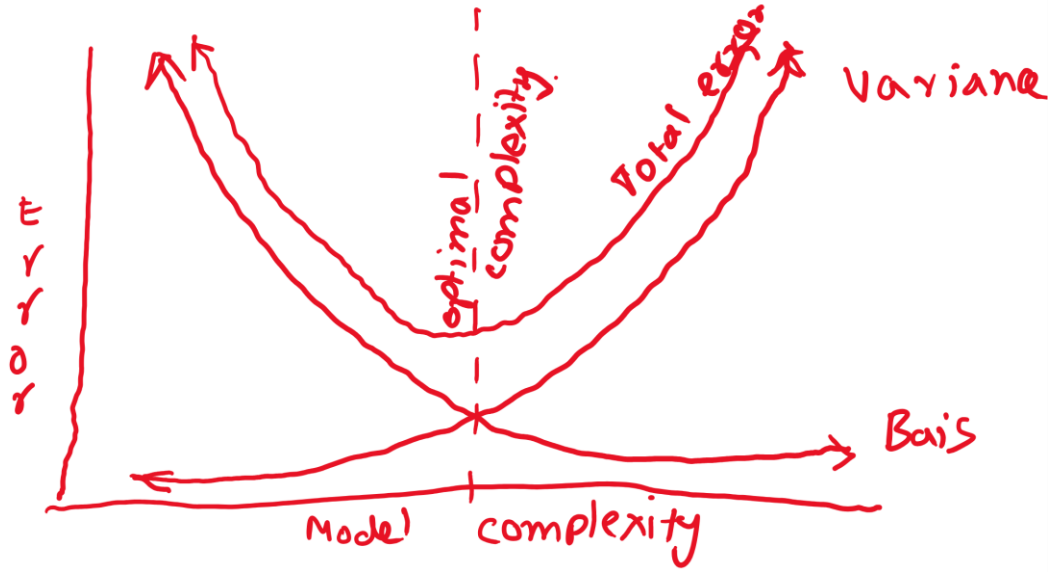
If five most important predictor variables are not available for training the model, then we need to retrain the model using the remaining predictor variables. The predictor variables with the next set of five highest absolute coefficients post training the model will determine the next most important set of predictor variables in the updated model.

The variable selection is based on dataset. Most relevant variables will make the cutoff to the top in Lasso based on feature significance.

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?

Model should not be over complex(overfit) and also should not be over generalized(underfit).It should have proper balance between variance and bias such as to minimize the model errors.



There are multiple steps that can help achieve the above balance between over fit and under fit problem. The steps are mentioned on next page….

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?

- Data used for model building should have all scenarios covered so model will pickup relevant patterns while making prediction.
- Do test train split and ensure model never uses any part for test data in training phase of model building.
- K fold validation to evaluate model performance
- Do proper feature selection and avoid multicollinearity this can be done by using Ridge and Lasso based on dataset
- Hyper parameter tuning by tuning alpha.
- Continuous retraining of model based on real life scenario.
- Check of residual plot if nonlinear relations are present do log,square or exponential transformation.