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: Ridge and Lasso regression's ideal alpha value is usually found by evaluating the model's performance over a range of alpha values, a procedure similar to cross-validation. By balancing variance and bias, this ideal alpha creates a model that performs well when applied to fresh data.

The regularization strength improves when the alpha value is doubled. This will reduce all coefficients in Ridge Regression toward zero, but not precisely to zero. A larger alpha in Lasso regression may lead to more coefficients being lowered to zero, hence carrying out feature selection. This could result in a more straightforward model with fewer variables, which could improve interpretability at the expense of increased bias.

The greatest absolute coefficients in the model indicate which predictor variables are most significant after doubling alpha. These variables are thought to be more significant in predicting the target variable since regularization penalizes them less. The features of the response variable and the dataset will determine the precise variables.

Recall that your data's unique context and the objectives of your study should be taken into account when selecting an alpha and interpreting predictor importance.

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?

Selecting between Lasso and Ridge regression relies on your modelling goals and the particular features of your data:

If you think that the majority of the variables in your dataset are significant or if there is a high correlation between the predictors in your dataset, then Ridge Regression is usually the better option. Since Ridge doesn't set coefficients to zero, multicollinearity has less of an influence while still incorporating all features.
If you believe that only a portion of the predictors actually have an impact on the target variable, Lasso Regression may be more appropriate. By setting certain variables' coefficients to zero, Lasso can perform automatic feature selection and totally delete those variables. When you have a lot of features and want to pick the most significant ones, this can be helpful.

In the end, the choice is based on:

The quantity of observations and features in your dataset.
The level of feature multicollinearity.
If choosing features is a top concern.
The significance of model interpretability for your analysis.
Lasso could be more beneficial for your model if it gains from being simple and recognizing important factors. Ridge might be a preferable option if managing multicollinearity and keeping all variables is your main concern.

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?

After removing the top five predictor factors from the Lasso model, you would need to retrain the Lasso regression model without these variables in order to determine the next five most significant predictor variables. This procedure includes:

Eliminating the Leading Predictors: Initially, take out of your dataset the five leading predictor variables that were previously found.
Retraining the Model: Using this adjusted dataset, train the model using the same Lasso regression setup (with the identified optimal alpha value).
Assessment of Coefficients: Analyze the model's coefficients after training. The variables with the biggest absolute coefficients in Lasso regression are deemed to be the most significant.
Finding Fresh Top Predictors: From this retrained model, identify the five variables with the highest coefficients. These are going to be your new best guesses.

This method is predicated on the idea that the other variables still offer enough data to accurately forecast the target variable. Keep in mind that when the model is retrained using various sets of predictors, the significance of the variables may change.

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?

Cross-validation: The dataset is divided into segments, with the model being trained on some and validated on others. It aids in evaluating the model's effectiveness over several data subsets to make sure it doesn't overfit to one.

Regularization: Techniques such as Ridge and Lasso penalize the model's complexity, promoting the use of more basic models that are more likely to be well-suited to fresh data and less likely to overfit.
The performance of a model can be enhanced by carefully choosing and engineering its features. This covers choosing pertinent variables, scaling features, and managing outliers.

Model Complexity: Select a model complexity that makes sense given the available information. While extremely basic models may underfit, unduly complex models may overfit.
Various Information To make sure the model learns general patterns rather than just particular peculiarities of the training data, train it on a variety of data points.

Metrics of Performance: Utilize suitable metrics for assessing the model. A number of performance metrics for the model, including area under the ROC curve, recall, accuracy, precision, and F1-score, can provide light on various areas.

It is assumed that a resilient and generalizable model will continue to be accurate when used on both new and unseen data, not simply the data it was trained on. But model complexity and generalizability frequently trade off, and striking the correct balance is essential for useful applications.