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?

In Ridge and Lasso regression, doubling the \(\alpha\) values (initially set at 0.5) intensifies regularization. Ridge restricts coefficient magnitudes more strongly, promoting a constrained model without driving coefficients to zero. In Lasso, the higher \(\alpha\) increases the likelihood of exact zeroing of some coefficients, resulting in a sparser model. Ridge aims for balanced shrinkage, while Lasso, with its feature selection property, may exclude certain predictors entirely. The optimal \(\alpha\) values depend on the dataset characteristics and the trade-off between regularization and model interpretability.

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 between Ridge and Lasso regression depends on the specific characteristics of the dataset and the objectives of the modeling task. The decision is often influenced by the trade-off between regularization strength and the desire for feature selection.

If the dataset has many correlated features and there is no strong prior belief that only a subset of features is relevant, Ridge regression is generally preferred. Ridge regression tends to shrink coefficients towards zero, but it rarely drives them exactly to zero. This can be beneficial when retaining all features is important, and avoiding complete elimination based on correlations is desired.

On the other hand, if there is a suspicion that only a subset of features is truly important and others can be excluded, Lasso regression may be more appropriate. Lasso has a stronger feature selection property, driving some coefficients exactly to zero. This can result in a simpler and more interpretable model when dealing with high-dimensional datasets.

In summary, if a balanced approach is needed with all features retained, Ridge regression is suitable. If feature selection and sparsity are crucial, Lasso regression may be the preferred choice. The optimal choice often involves experimentation and cross-validation to assess model performance on specific datasets and tasks.

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?

In Lasso regression, the importance of predictor variables is determined by the magnitude of their coefficients. If the five most important predictor variables from the initial Lasso model are not available in the incoming data, the next set of most important predictors can be identified by examining the coefficients in the updated model excluding the original five.

After excluding the unavailable variables, the five most important predictor variables in the new Lasso model are those with the largest absolute coefficients. These variables contribute significantly to the model's predictions, even after the exclusion of the initially identified important predictors. Inspecting the absolute values of the coefficients in the updated model will reveal the new set of most important predictors.

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?

Ensuring a model's robustness and generalizability is vital for its real-world effectiveness. Employing techniques like k-fold cross-validation, train-test splitting, and the use of validation sets allows for a comprehensive evaluation of a model's performance on diverse data subsets. Feature engineering, regularization, outlier handling, and prevention of data leakage contribute to model stability. The implications for model accuracy are profound. A well-generalized model demonstrates consistent performance across different datasets, safeguarding against overfitting and ensuring reliability in real-world scenarios. Achieving the right balance between complexity and generalization is critical for accurate and trustworthy predictions on unseen data.