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 1

In our analysis, the optimal value of alpha (lambda) was found to be 10 for Ridge regression and

0.001 for Lasso regression.

If we double the value of alpha for Ridge regression, the penalty for larger coefficients will become more severe,
potentially causing the coefficients of less important variables to shrink towards zero, but not reaching zero. This may
slightly decrease the model's accuracy, particularly if some variables that were important at the original alpha level
become less influential at the new alpha level.

In contrast, doubling the value of alpha for Lasso regression (from 0.001 to 0.002) will still maintain the feature selection property of the Lasso model. However, more coefficients may shrink to zero, possibly losing some important
variables and decreasing the model's accuracy.

The most important predictor variables after these changes would need to be reevaluated by running the models with the
new alpha values and examining the coefficients. However, in general, we might expect the most robust predictors from
the original model to remain important.

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?

Answer 2

The choice between Ridge and Lasso depends on the specific situation and goals of the analysis. In your case, the Lasso
regression model was suggested because it performed comparably to the Ridge model and had the added benefit of feature
selection. Lasso regression could simplify the model by reducing the number of features, potentially making it easier to
interpret and less prone to overfitting.

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?

Answer 3

If the top five predictors in the original Lasso model are not available, you would need to rebuild the model excluding

these variables. The new set of most important predictor variables would be identified by fitting the Lasso regression
model on this updated dataset and examining the coefficients. Since this is a hypothetical scenario, I can't provide the
actual variables without running the model on the updated dataset.

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?

Answer 4

Ensuring a model is robust and generalizable involves several strategies:

Generalisation is important and test accuracy needs to be higher than the training score. However, the difference should not be exceedingly high. The model should generalise during training, but if the score is very high in training, and lower in testing, it means that the model has memorised the data, meaning that it is overfitting.

Overall, there should not be large differences between the results. Low test scores could result from splitting the data set too early in the preprocessing step, so that some steps may be missed on the test data. Robustness of a model is generally not solely based on high test scores, but also depends on the assumption that the train scores are higher than the test scores. Both scores have to be high enough to be acceptable for the specific business case and expectations of the model. It is also important to consider the values obtained for train and test, so that the model will perform well on unseen data.

This means that the data should retain some outliers to help with predictions. As demonstrated in the assignment, accuracy of the model will vary, depending on the way data is processed and how features are selected. There may be no perfect model, but different steps are available to ensure that the model developed is fit for purpose for the specific context and the uniqueness of the business case. This is in line with Occam’s razor, that is, the model to be chosen should not be more complex than it needs to be.