Question1:

Rahul built a logistic regression model with a training accuracy of 97% and a test accuracy of 48%. What could be the reason for the gap between the test and train accuracies, and how can this problem be solved?

Answer 1 **(a)**:

The reason for gap between train and test Accuracy is **Overfitting.**

The problem can be solved by **regularizing** the model with ridge or lasso Regularization Technique.

Question 2:

List at least four differences in detail between L1 and L2 regularisation in regression.

Answer 2 :

1. In ridge(L2) regression, the penalty is the sum of the squares of the coefficients and for the Lasso(L1), it's the sum of the absolute values of the coefficients.
2. Lasso trims down the coefficients of redundant variables to zero and, thus, indirectly performs feature selection as well. Ridge, on the other hand, reduces the coefficients to arbitrarily low values, though not zero.
3. Lasso regression is computationally more intensive than Ridge regression.
4. When number of features(variables) are low, Ridge (L2) regression have better prediction power than LASSO. Though Ridge won't help in feature selection and model interpretability is low.

Question 3:

Consider two linear models:

L1: y = 39.76x + 32.648628

And

L2: y = 43.2x + 19.8

Given the fact that both the models perform equally well on the test data set, which one would you prefer and why?

Answer3 :

According to **Occam's razor** Principles, simpler model should be preffered over complex model

Hence second model **L2: y = 43.2x + 19.8** will be preferred over the first model as it simpler then L1.

The number of bits required by the computer(or any other machine) to save the L1 will be higher.

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:

The model can be made robust and generalizable by applying Regularization on the model.

Accuracy for Train Data:-

With Increase in value of hyper parameter lambda of Regularization, the error term increase constantly resulting into decrease of constantly decrease in accuracy

Accuracy for Test Data:-

With Increase in value of hyper parameter lambda of Regularization, the error term decrease initially and then increase constantly. Hence Accuracy will increase initially and then decrease constantly (as the hyper parameter increases).

Accuracy for Overall Model:-

With Increase in value of hyper parameter lambda of Regularization, the error term increase constantly. Hence Accuracy will decrease constantly (as the hyper parameter increases).

Question 5:

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 5:

Lasso Regularization - On applying the optimal value of lambda , the coefficients of redundant variables becomes zero . Thus Lasso indirectly performs feature selection on the dataset.

Ridge Regularization - On applying the optimal value of lambda , the coefficients of redundant variables becomes close to zero but does not exactly zero. Thus Ridge does not perform feature selection on the dataset.

As the number of variables in the dataset provided in assignment is very high, Lasso Regression is more preferable as it performs feature selection.