Question 1

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: Overfitting is the answer. This means that the model learned specifically only to train set and it cannot generalize for test set as well. In order to avoid overfitting, we can cross validate. We can split the training data into multiple train-test sets and then test it on them.

____

Question 2

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

Answer:

1. The main difference between as an error function between L1 and L2 is that L1-normalization function (Least absolute Deviation) is minimizing the absolute differences between target and estimated values. L2 normalization function (Least Square Error) is minimizing the sum of the square of differences between target and estimated value.
2. The Lasso regression uses L1 normalization and Ridge regression uses L2 normalization.
3. L1 function is robust and L2 is not robust, also there is a possibility of multiple solutions for L1, whereas for L2 there is always only one solution.
4. L1 loss function has unstable solution and L2 has stable solution.

____

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?

Answer: Preference will be given for L1 as it seems more accurate

____

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: We can fit a series of binary logistic models for different cutoffs, plot them against the regression coefficients, and check if it is constant. The model accuracy will not be accurate.

____

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: Ridge regression will be chosen as it has only one solution and it is stable