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?

Solution:

The model built by Rahul have very high train score and a very poor test score. This may be is because Rahul has used a very complex model for training, which has memorized the training data because of its high complexity. Therefore, when the same model is run over the test data the results are poor. This is a classic case of overfitting and this can be avoided by using a simpler model, simple enough to capture the pattern but not memorizing the training data.

Question 2:
List at least four differences in detail between L1 and L2 regularisation in regression.

Solution:

1. Penalty term:
   L1 also known as Lasso regression has penalty term as sum of absolute values of coefficients where as L2 also known as Ridge regression has penalty term as Sum of squared of coefficients
2. Type of solution:
   L2 can be solved using linear algebra techniques, whereas solution for L1 is not direct but rather should be done stepwise. L1 has a more stable solution where as L2 has a less stable solution
3. Computational time:
   Since L2 can be solved by simpler techniques, it takes far less time when compared to L1 regression
4. Lasso regression can be used as feature elimination technique as it makes coefficients of non-important features 0.

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?

Solution:

We should choose model2 as it is less complex when compared to model1. Model1 requires more bits for representation and also takes longer time for computation because of higher precision. Therefore, model2 should be chosen when compared to model1.

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?
Solution:
Always try to choose a simple enough model that explains the given dataset well. Simple models are robust and flexible to changes in data when compared to complex models therefore changes in data will not affect the performance of the model. According to bias variance trade off choosing simpler model resulting in having a model with higher bias, therefore there will a dip in accuracy when compared to other complex models like neural networks or tree methods. This dip is accuracy is because the model being simple is not able to capture all the relationships present in the data. But advantage we get is that simple models give insights into relation between output and input models which is not given by the complex models like neural nets.

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?
Solution:
I have decided to choose lasso regression, because the test and train scores of both the methods were similar, but the model given by the lasso is much simpler when compared to the one given by the ridge. The lasso model has only 22 variables when compared to the ridge model with 35 variables, which implies that I have almost same accuracy model but with 10 less variables. Therefore, I have decided to use lasso model for this problem.