Question-1:

Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.

Answer:

Overfitting could be reason for this difference in accuracy between train and test datasets. I believe that Rahul has trained his model so much instead of generalizing the model has memorized all the data from training datasets. This can be resolved with a technique called as Regularization which focuses on building a optimal model which is not so complex and Robust as well.

Question-2:

List at least 4 differences in detail between L1 and L2 regularization in regression.

Answer:

1. L1 regression adds the sum of absolute value magnitude of coefficients to the cost function while the L2 regression adds sum of the squares of coefficients to the cost function.
2. L1 or Lasso regression is used when dealing with a large number of features because it helps in feature elimination (Sparsity). L2 regression does not have this property.
3. L1 regression sometimes provide unstable and multiple solutions. But L2 on the other hand produces stable and one solution because its computationally efficient.
4. L1 regression is more time consuming than the L2 regression.

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 dataset, which one would you prefer and why?

Answer:
I would go with second model L2. Both are 1 degree equations, both would require a floating point representation. But when model internally converts these to binary 1s and 0s to perform computation, L2 model is simpler. Hence, the Occam's Razor comes into play which suggests When in dilemma, choose the Simpler model.

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:
Regularization is a technique which is used to make model robust and generalizable. It is important because the simpler model is general would have a lot of errors(accuracy won't be so high) and complex model would consume more computational time and there is a chance of over fitting(Model's performance could go terribly with unseen data). In both cases, the accuracy of the model is affected. So the regularization can be used to make models simple but not to naïve. A balance between the complexity of model and its learning from it.

Question-5:
As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

Answer:
I would choose Lasso Regression as the number of features is high and It had eliminated almost 50% of the feature while building the model. Lasso regression fits this situation better.