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 1:

The fact that the training accuracy is so high but the test accuracy returned to be as low as 48% indicates a clear case of the model **overfitting**. The reason this happens is because the model may be too complex, such that it learns the datapoints on the training set far too well, and as a result, the model is not a generalized one.

This is where the concept of **Variance-Bias Trade-off** comes, whereby we need to find a common point between how much variance is being explained by the model (complexity) and how much bias (simplicity) is there in the model.

In order to solve this, we use **regularization techniques** – this is a process of creating optimally complex models i.e. a model that is as simple as possible while it still performs well on training data. Basically, it tries to strike a balance between keeping the model as simple as possible and yet explain enough variance in the data.

There are 2 common techniques of regularization – **Ridge** and **Lasso**. Both these techniques introduce a bias component to the regression model that reduces its complexity. The difference between the 2 is that Ridge introduces the bias using the sum of squares of the coefficients, where as Lasso's regularization term is the absolute sum of the coefficients.

Question 2:

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

Answer 2:

The L1 regularization technique is used in Lasso regression and L2 regularization technique is used in Ridge regression. To list the differences between the 2:

1. Lasso Regression technique uses the **sum of absolute values of the coefficients** of the model as the regularization term whereas the Ridge regression technique uses the **sum of squares of the coefficients** of the model as the regularization term
2. Lasso is **computationally more intensive** compared to Ridge because Ridge almost always has a matrix representation for the solution, whereas Lasso requires several iterations to get to its final solution
3. Lasso technique of regularization can be used for feature selection whereas Ridge cannot because Lasso regression is capable of making the lesser important variables' coefficients to become exactly 0 – hence Lasso regression can produce **sparse** outputs
4. The solution produced from the L1 regularization technique is a lot **more robust** than the L2 regression technique. This is because the L2 regression technique squares the error, which means it is more susceptible to large errors due to existence of outliers in the data compared to L1 regression technique

Question 3:

Consider two linear models:

L1: y=39.76x + 32.648628

L2: y=43.2x + 19.8

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

Answer 3:

The model L2 would be the preferred model in this case because it is **less complex** than L1. If we look at the **number of digits** in the constant term, we see that L1 term is more complex than the L2 model and considering the performance of both the models on the test set is equal, we would prefer to choose the simpler model. The L2 model would be computationally less intensive, and is **more generalized** than the L1 model.

Question 4:

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Answer 4:

A model can be made more robust and generalizable by using **generalization techniques such as Lasso and Ridge regression**. Both these techniques introduce bias into the model to ensure the model is not too complex. These techniques also use a hyperparameter, the regularization coefficient, that ensures that the model is a lot more robust and stable. The regularization coefficient, lambda, ensures that it does not allow the model coefficients to become too large as it **clamps down any such swings** in the coefficients.

In the process of making the model more robust and generalizable, we usually tend to **lose out on accuracy**. This is because we're trying to ensure that the model doesn't memorize the training data as we want to be generalized to so as to be applicable on data it hasn't seen yet. As a result, the bias in the model would have to be increased, which will compromise on the variance explained by the data, which would result in lower accuracy of the model

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:

The **Lasso regression** model would be the most optimal model that we would apply in this case. That is because Lasso has the capability of performing feature selection by making the coefficient of insignificant variables as 0. This was the case in our model too, whereby the Lasso regression produced a model that had fewer coefficients than the Ridge regression technique, despite ensuring that the performance (Error and R-squared) is not compromised much. This means that the model produced by the Lasso regression technique is more generalized than the Ridge regression model and hence preferable in our case.

The optimal values of lambda in our analysis for:

- Ridge regression was 2.5
- Lasso regression was 100