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:

Looks like Rahul's logistic model is "Over fitting" the test set, hence the accuracy there is quite high as it has learnt quite a lot(seems like a complex model), it has memorized it(the noise and outliers as well). Due to this when tested on the test set, the accuracy goes downhill.

This classic problem can be solved by simplifying the model by putting regularization during model building as doing so the number of parameters and features are used as required, reducing the over fitting problem.

Question 2:

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

Answer:

L1 Regularization:

- A regression model which uses L1 regularization technique is called Lasso regression.
- L1 regularization is computationally inefficient.
- L1 regularization produces sparse models.
- It shrinks or eliminates coefficients to zero, reducing the features, which helps in feature selection.
- We add absolute value of magnitude of coefficient as penalty term to the loss function.

L2 regularization:

- A regression model which uses L2 regularization technique is called Ridge regression.
- L2 regularization is computationally efficient.
- L2 regularization produces dense models.
- There is no feature selection in this regularization.
- We add squared magnitude of coefficient as penalty term to the loss function.

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:

L2 model (***L2: y = 43.2x + 19.8)*** is the simpler model with only 3 bits to accommodate, so as this is a simpler model, this is less prone to over fit the train data and will be a generalized and will be less penalized.

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:

To make sure that the model is robust and generalized we can use regularization, so that our model is as simple as possible, it should not use redundant features, required coefficients to be used.

We should keep our eyes on the AIC, BIC, R squared and advanced r squared values. First 2 should be minimized and later 2 to be maximum(near to 1). This will make sure our model is simple i.e. more generalized. This is avoid our model to be over fitted and they will be robust as well.

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:

Well looking at the question where we could see the advantages and disadvantages of lasso and ridge (L1 and L2 regularization), we should definitely choose the value of lambda for lasso regression as it has the capability of reducing the variability, by feature selection and improves the accuracy of linear regression models.