PART II: SUBJECTIVE QUESTIONS AND ANSWESRS

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.

The reason for such a huge difference between the training and testing accuracy is because of 'overfitting'.

This can be overcome by performing any of the below suitable methods:

1. Cross validation
2. Regularization
3. By removing the unimportant features.

Question 2

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

1. L1 model give sparse, while the L2 are non-sparse. I.e., many of the features have zero coefficient value, when other features have non zero coefficient value.
2. The one of the major difference is the feature selection, where in L1 method does feature selection based on the importance of them, but the L2 doesn't perform feature selection.
3. L1 model type is robust to outliers and won't get affected by outliers, while the L2 model type in not robust and does get affected by outliers.
4. L1 model is good when we have high dimensional, i.e. data set which is very huge and vast, where in feature selection helps to handle it efficiently, but L2 is not suitable for high dimensional data.
5. L1 model type uses sum of absolute value of coefficients as regularization term, while L2 model type uses sum of squares of the coefficients as regularization term.

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?

Looking at the equations of the two models, even though it is said that both perform equally well, it appears to be that the L2 model is simpler than the L1 model.

Hence the main reason to go with L2 model is that it is simpler and it would take less number of bits, compared to the L1.

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?

Usually a model is said to be robust and generalizable when it is a simple model.

If a model us not robust, then it would impact the model and would lead to over fit, which will affect the overall performance of the model.

Implications of the model:

A simpler model is more generic and robust than the complex model.

Total number of data needed for simpler model is less

A simple model has more tendency of making more errors

To overcome these concerns we can do a bias- variance trade-off to work on it, to make the model with good performance.

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?

The evaluation metrics obtained for Ridge and Lasso regression are as follows:

From the analysis we ca observe that the optimal value of lambda for Lasso and ridge regression is 50 and 6 respectively

And the R2 score for train and test in ridge and lasso regressions are 0.93 and test 0.91

However the comparing the AIC, BIC and adjusted R2 score, we can conclude that the Lasso regression is better than the Ridge regression.