Advance Regression Assignment – Problem Statement II

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.

This can be called as Overfitting. The model has not been generalized and has catered to the structures of the training dataset only. This could be because of the limited test data or not good split of train and test split. The model has been influenced a lot by the training data so when the model is fitted onto the test data the accuracy degrades. This can be solved by testing the model in the unseen data to eliminate overfitting. Also the train and test split if has not worked with usual 70-30 train-test split, the k-fold split of train and test split can be adopted.

Question-2:

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

L1 regularization creates Sparse models while L2 creates dense models

Feature selection works well with L1 regularisation.

L2 tends to result in small non-zero regression coeffecients whereas L1 penalty results in exactly zero regression coefficients

L2 is computationally efficient while L1 is computationally inefficient

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?

Typically model with higher r2-score can be picked up. Now here since both L1 and L2 perform equally well, I think we can check the p-values of the model and decide. Other way could, if we substitute x with equal value in both the equations, the Output of L2 would be higher, so I think L2 can be preferred.

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?

Regularization needs to be applied to make sure that the model is robust and generalisable. As this creates an optimally complex model which requires few training samples.

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?

I would choose Lasso as it eliminates lot of feature variables and creates many zero coefficients as that helps in determining the predictive features that influences the target variable.