Problem Statement - Part II

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: Following reasons can be the cause of this discrepancy:

1. The created model has learned everything (overfitting model) and that is why it is unable to produce the same results when it is exposed to unknown data.
2. He might have done a biased split in the dataset.
3. Wrong model selection

This problem can be resolved by:

1. Introduce regularization so that the model does not depend too much on feature and information is gathered from all sources. Check the degree of the polynomial
2. He should at least split his data into training and testing datasets (70:30). But for better results split them into 3 datasets i.e. training, cross-validation, and testing (70:10:20) and can also use k-fold cross-validation.
3. Various models should be checked and which can be thought as useful.

Question 2: List at least four differences in detail between L1 and L2 regularisation in regression.

Answer 2: Differences are:

1. L1 simply selects the features from the set of features and hence is also used where feature selection is needed. L2 does not choose any features and hence strikes to maintain the balance between all variable so important information is lost.
2. Since L1 involves iterative process, it is computationally intensive. Whereas in L2, it can be solved using matrix operations on the invertible matrix(not possible in L1). L1 uses absolute value but L2 uses squares of the term
3. L1 yields sparse model whereas L2 does not as former makes some coefficients zero.
4. L2 gives a unique solution whereas L1 does not. For example, in a matrix, if we have to reach diagonally opposite point then L2 will give the shortest direct path but the solution from L1 can vary every time.
5. L1 generates model that are simple and interpretable but cannot learn complex patterns whereas L2 can.
6. L2 gives better prediction when output variable is a function of all input features.
7. L1 is robust to outliers whereas L2 is not.

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 3: Both are a Linear equation, but preference will be given to L2 as it requires fewer bits of memory as compared to L1. Additionally, L2 will have a corporately lower bias.

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: Model should not have a high bias(underfit) and at the same time should not high variance (overfit condition). They both should be kept at an optimum value. Another thing that we need to keep in mind that data should be divided properly and model should not see a test before actual testing. Regularization should be performed to each feature its required weight.

Accuracy should be checked from validation data and not from training data. Accuracy should not be tried to increase unnecessary as it can also lead to overfitting condition. Residuals represent the portion of the target that the model is unable to predict. It should be zero-cantered bell shape. Adding more variables to the model might help the model capture the pattern that is not captured by the current 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: Optimal value of lambda in Ridge is 6 and in Ridge its 0.001. After checking RMSE, it would be advisable to with Lasso regression. Additionally, it will select top variables that are actually needed to gain insights for the company with fewer errors.