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: The regression model can have a high training accuracy at the same time it can also have a really low test accuracy as shown in the above model. The reason for such a difference between the two accuracies is simply a condition that is known as **Overfitting.**

Overfitting occurs in a scenario where the model memorizes the training data completely and can appear 100% of the variance in the data. In such a case the model hasn't captured the trends in the data rather it simply has memorized every data point in the training set. This problem mainly arises due to the complexity of the model. As we increase the number of independent variables we do get a more accurate model on the training data however this model will not perform well on test data. In regression analysis, overfitting can lead to misleading R-squared values, regression coefficients and p-values.

To avoid overfitting, we should draw a large random sample size to handle all of the terms that one can expect to include in the model. The goal here is to find which are the relevant variables and terms that need to be used to provide an optimal model.

In some cases, we don't have enough sample size. In such cases we can use what is called as **Cross-validation**.

Cross Validation is a technique to efficiently use the data regardless of how big our dataset is. The idea is to use your initial training data to generate multiple mini train-test splits. These multiple splits can be used to tune the model. In a standard k-fold cross-validation, we divide the data into k subsets, also known as folds. Then, we iteratively train the algorithm on k-1 folds while using the remaining folds as the test set,at each iteration the model is trained on all but one paratition and this model is used to predict the remaining partition

Another technique we can use is **Regularization**. It is the process of penalizing the model when it includes more independent variables. This is done by adding hyperparameter to the regression formula. Adjusted R^2 ,AIC,BIC are used to validate the accuracy of model. two common techniques of regularization are Lasso and ridge.

Question 2

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

Answer: The two regularisation techniques work on the same principle i.e. they add a regularisation term at the end.

1) L1 or Lasso adds absolute sum of coefficient values whereas L2 or Ridge adds the sum of the square of coefficient values.

L1 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

L2 regularization on least squares:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left(t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} w_i^2$$

2) Lasso or L1 regularization is computationally inefficient on non-sparse cases(cases in which coefficients are not going to become 0) while Ridge or L2 regression is computationally more efficient in such cases.

3) Both the techniques L1 and L2 regularise the coefficients by reducing the magnitude to as minimum as possible almost close to 0 or 0 . They cause shrinkage of the coefficients but differently. L1 or Lasso shrinks some of the coefficients to zero, thus performing *Feature selection*.

4) L2 or Ridge always has a matrix representation of the solution, whereas L1 or Lasso requires a few iterations to get to the final solution.

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: At first glance we don't really find much difference In the two solutions. Both L1 and L2 are comprise of one on slope (m) and an intercept (c). So the only thing separating the two solutions is the number of decimal places on the slope(m) and intercept(c).
Consider L1:

Y= 39.76x + 32.648628

Taking It into consideration as the standard format y=mx+c

m=39.76

c=32.648628

m has 2 values after decimal points and c has 6 values after decimal point.

Consider L2:

y = 43.2x + 19.8

Taking into consideration format y=mx+c

M=43.2x

C= 19.8

m has 1 value after decimal point and c has 1 value after decimal point

Therefore it is evident that the model L2 is optimal to be chosen as it requires lesser number of bits to represent its slope(m) and intercept(c)

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 a model is robust and generalisable the easiest solution is to make the model simple. Simple model requires lesser amount of training data and is easily generalizable. However, if the model is too simple it can lead to underfitting. Thus, one needs to find the right amount of complexity to be achieved in a model to maintain optimal accuracy. This is represented by what is known as bias-variance trade-off.

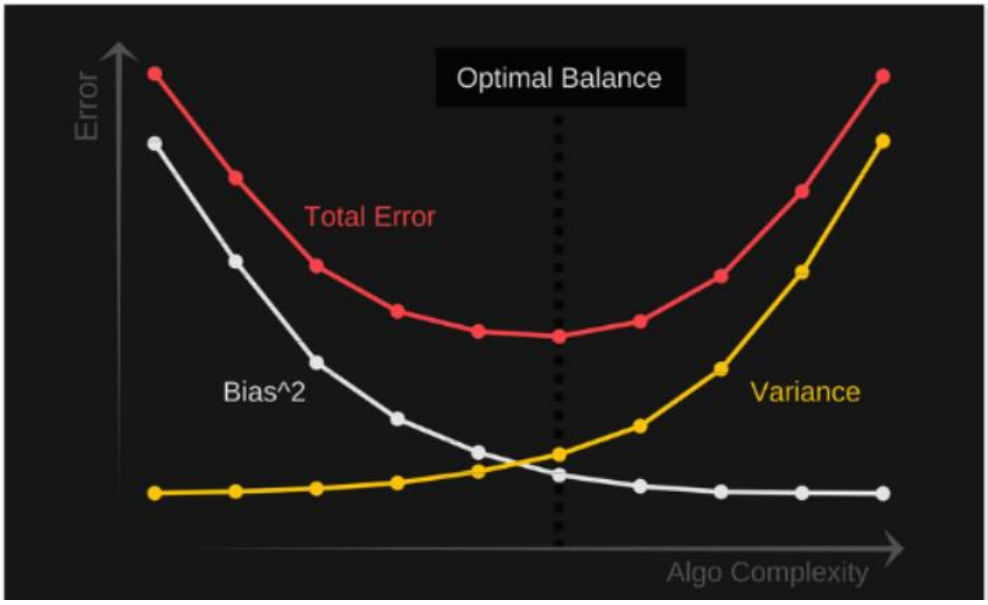
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

An overly simple model will have bias and low variance and will lead to high error on training and test data. But a very complex model will have high variance and low bias and will perform very well on training data but will have high error rates on test data.

To build a good model, we need to find the right balance between bias and variance such that it minimizes the total error.

Total Error = Bias^2 + Variance + Irreducible Error



Thus, An optimal balance of bias and variance would never overfit or underfit 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: After performing the ridge and lasso regression from the dataset we found the optimal value of ridge regression to be at 2 and the optimal value for lasso regression to be at 100. On fitting the model on our dataset. We can see in Ridge Regression 389 non zero values and Lasso regression has 109 non zero values . So essentially Lasso regression has performed feature selection and thus we will apply lasso regression on the dataset as it is computationally efficient as it only adds the absolute sum of value of coefficients also as it has performed feature selection which even RFE wasn't able to successfully get.