**PROBLEM UNDERSTANDING:**

The problem is about predicting the telecom customers who are likely to exit from contract, to identify the underlying reasons and derive patters out of the given data. It is more important for the company to identify customers who are likely to leave than those who are staying, this way the company can target customers and provide them additional offers, services or discounts to make them stay. Therefore, ‘Recall’ is our desired metric here.

**PRE-PROCESSING:**

The data has been provided in four different datasets therefore, they had to be merged before performing analysis. The missing values have been labelled differently in different datasets. All the missing values have been changed to ‘NA’ before processing. Instances with ‘NA’ values have been omitted as they are very minimum as compared to the total number instances in the dataset. The categorical variables have been made into factors.

After pre-processing on train and test data is done, a level mismatch has been observed in the summary of the train and test data on the ‘PaymentMethod’ attribute. The levels have been matched and after removing unnecessary variables such as CustomerID, country, state and date of data collection, the data is now ready for model building.

**MODEL BUILDING:**

Models like c5.0, rpart, KNN , Logistic Regression, SVM and Boosting. Even though models like svm and boosting gave good accuracies, the recall value was very low. Logistic regression model has features like ROC curves which provide visualisations that aid in deciding the threshold for improving the recall.