Predicting Customer Churn in Telecommunication

Submitted by **Nabi Sheikh**

**Shridhi Pandya**

**Abstract**

Customer churn in the Telecommunication industry is a continuous problem owing to stiff competition, new technologies, low switching costs, deregulation by governments, among other factors. To address this issue, players in this industry must develop precise and reliable predictive models to identify the possible churners beforehand and then enlist them to intervention programs in a bid to retain as many customers as possible. This paper proposes a new set of features with the aim of improving the recognition rates of possible churners. The features are derived from call details and customer profiles and categorized as contract-related, call pattern description, and call pattern changes description features.

**Introduction/Problem Statement**

Customers become “churners” when they discontinue their subscription and move their business to a competitor. That is, churning is the process of customer turnover. This is a major concern for companies with many customers who can easily switch to other competitors. Examples include credit card issuers, insurance companies and telecommunication companies.

Using a customer data, companies these days are saving a business by identifying the customers which are likely to churn in near future and give a clear view what step should be taken into consideration for retaining their existing customer.

Customer churns are often considered as a Downtime to any leading Organization in any industrial section, here we are talking Customer churn in Telecomm industry. One of the strategies that can be used to achieve this is by developing predictive models that can reliably identify possible churners in the near future. In the recent past, data mining techniques have been used extensively to develop these mode is with satisfactory performance

**OBJECTIVE**

The objectives of this study are in two folds. The first objective is to estimate customer survival function and customer hazard function to gain knowledge of customer churn over the time of customer tenure. The second objective is to demonstrate how survival analysis techniques are used to identify the customers who are at high risk of churn and when they will churn.

**DATASET**

The dataset consists of 10 months of data but there are some usage and payment files missing in some of the months so for the prediction only 5 continuous months of data has been taken. So the months that are been taken are July, August , September , October , and November.

The three files are which consists information of the customers is data usage , payment and voice usage files

The data usage file (mar\_data1.csv) consists of the 2G or 3G data which used by the customers

The payment file (mar\_pay1.csv) consists activation date, the customer group, revenue assurance, recurring charge, Non-recurring charge and other details.

The Voice usage file (mar\_usg1.csv) consists of the local and std charges of the user both outgoing and incoming calls.

The Disconnection file (Discon1.csv) file consists of the user ‘s that are being churned.

**Methodology**

1. The problem is a classification problem so the algorithm that is used is logistic regression, so it will predict if the customer is churned or not after training the model with the training set. For storing the data HDFS is being used. The model is being developed in R. So the steps that are carried out for developing the model are as follows:
2. The first step is to read the three files, i.e.data usage, payment, voice usage of every month in one file and change the column names of the files and then merge all the three files in one dataframe.
3. New variables are being derived from the existing one as in Total outgoing, Total Incoming, Local to STD outgoing, Local to Local outgoing, Local to STD incoming, Local to Local incoming, Roaming outgoing, Roaming incoming.
4. After the new variables which were derived from the existing variables the existing variables were deleted, so the unnecessary columns were removed.
5. There are two columns which defines the same activation date of the user so the other column was deleted. The format of the date was not proper so the format of the date was changed into proper format.
6. Unique id’s are considered from the distinct dataframe and only that data is written in the new csv file and it is done for five months, at the end we have five different csv files with unique id’s.
7. All the five different files then is merged to a single csv file.
8. After this we had Na values and outliers, unnecessary columns are dropped out as it’s not reluctant to the prediction
9. To remove outliers we actually created a outlier function as we were not clear about the particular package as in how would they define the outliers in-build in them as in we didn’t wanted to go with knn algorithm when we use rm.outlier.
10. The Na values in the dataframe is been replace accordingly as in what is the class of the particular column as if the column is numeric it would be replaced by median/mean.
11. There are also NA values in the character column and the character columns are the mobile activation date of the user for every month. So the mode function cannot be used to replace the NA values. To replace it the first date of that month is being used and that date is used to replace the NA values
12. So after the outliers and Na value replacements we actually we created a new file .
13. The new csv is being merged with the “Discon1.csv” and a new column churn is being added which consists 1 and then it is merged with the main csv file and the blank values in the churn column is replaced with NA.
14. Then the merged file is divided into train and test dataset
15. The model is being evaluated by using Confusion matrix and ROC curve.

**Results**

To describe the performance of the model confusion matrix is used on the test data set for which the true values are known.

confusionMatrix(data = test1$final, reference = test$churn\_status)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 331813 31767

1 63 23985

Accuracy : 0.9179

95% CI : (0.917, 0.9187)

No Information Rate : 0.8562

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5633

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.9998

Specificity : 0.4302

Pos Pred Value : 0.9126

Neg Pred Value : 0.9974

Prevalence : 0.8562

Detection Rate : 0.8560

Detection Prevalence : 0.9380

Balanced Accuracy : 0.7150

'Positive' Class : 0

> confusionMatrix(data = test1$final, reference = test$churn\_status, positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 331813 31767

1 63 23985

Accuracy : 0.9179

95% CI : (0.917, 0.9187)

No Information Rate : 0.8562

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5633

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity : 0.43021

Specificity : 0.99981

Pos Pred Value : 0.99738

Neg Pred Value : 0.91263

Prevalence : 0.14383

Detection Rate : 0.06188

Detection Prevalence : 0.06204

Balanced Accuracy : 0.71501

'Positive' Class : 1

