**PROJECT REPORT**

**CHURN PREDITCTION FOR TELECOM INDUSTRY**

P.Chandrashekhar

Snehal Salaskar

**ABSTRACT**

Customer value analysis is critical for a good marketing and a customer relationship management strategy. An important component of this strategy is the customer retention rate. Customer retention rate has a strong impact on the customer lifetime value, and understanding the true value of a possible customer churn will help the company in its customer relationship management. Conventional statistical methods are very successful in predicting a customer churn. The goal of this study is to apply logistic regression techniques to predict a customer churn and analyze the churning and no-churning customers by using data from a personal retail banking company.

**INTRODUCTION**

In the telecommunication industry, customers are able to choose among multiple service providers and actively exercise their rights of switching from one service provider to another. In this fiercely competitive market, customers demand tailored products and better services at less prices, while service providers constantly focus on acquisitions as their business goals. Given the fact that the telecommunications industry experiences an average of 30-35 percent annual churn rate and it costs 5-10 times more to recruit a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition. For many incumbent operators, retaining high profitable customers is the number one business pain. Many telecommunications companies deploy retention strategies in synchronizing programs and processes to keep customers longer by providing them with tailored products and services. With retention strategies in place, many companies start to include churn reduction as one of their business goals. In order to support telecommunications companies manage churn reduction, not only do we need to predict which customers are at high risk of churn, but also we need to know how soon these high-risk customers will churn. Therefore the telecommunications companies can optimize their marketing intervention resources to prevent as many customers as possible from churning. In other words, if the telecommunications companies know which customers are at high risk of churn and when they will churn, they are able to design customized customer communication and treatment programs in a timely efficient manner. Conventional statistical methods (e.g. logistics regression, decision tree, and etc.) are very successful in predicting customer churn. These methods could hardly predict when customers will churn, or how long the customers will stay with. However, survival analysis was, at the very beginning, designed to handle survival data, and therefore is an efficient and powerful tool to predict customer churn.

**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**

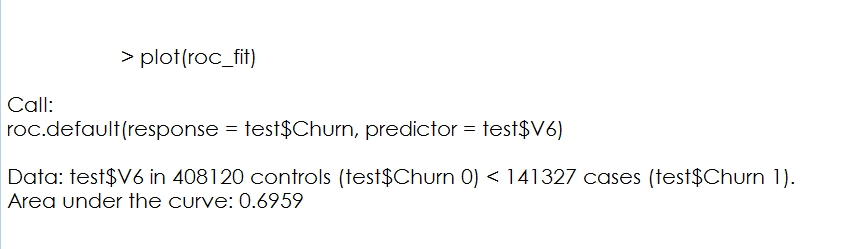
1. The problem is a classification problem so the algorithm that is used is logistic regression, so it will predict if the customer is churned or not after training the model with the training set. For storing the data HDFS is being used. The model is being developed in R. So the steps that are carried 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. The new variables are local and std outgoing call charge and then the total outgoing , local and std incoming call charges and the total incoming charges, roaming charges and then the whole incoming and outgoing and roaming charges of the particular user.
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. The unique id’s are taken from the dataframe and only that data is written in one single csv file. So this is done for all the five months, in the end there are five different csv files with unique id’s.
7. All the five months files is now being merged into one single csv file.
8. Now the data only contains outliers and NA values. So the single csv file is being read and the unnecessary columns such as “cust\_value\_seg”, ”cust\_group”, ”section” is deleted because this columns are of no use for prediction.
9. For the outliers the outlier function is being written and this function will detect and remove the outliers.
10. The NA values is being replaced by median for the numeric columns.
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 removing the outliers and replacing the NA Values new csv file is being created namely “new\_months\_without\_na.csv”
13. The “new\_months\_without\_na.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 train set consists of 1282042 rows and 93 variables. The test set consists of 549447 rows and 93 variables.
16. The model is being evaluated by using Confusion matrix and ROC curve

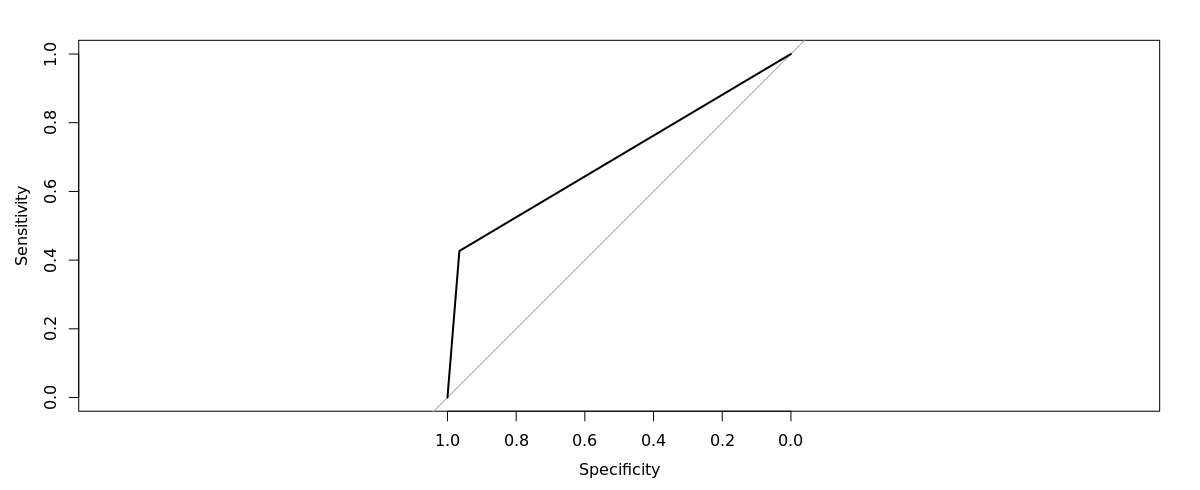
The result is explained in the below point.

**RESULT**

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

****

The ROC curve:

****

**CONCLUSION**

In the context of big customer behavior data, a customer churn prediction based on customer segmentation and misclassification cost is developed. The proposed model conducts customer segmentation first, which is of benefit for model to enhance its ability of churn customer recognition as well as facilitating companies to develop customer maintain strategy. Then, we make customer churn prediction with classification cost based on different customer groups.The results suggest that the proposed model performances better than those models without customer segmentation and misclassification cost in terms of the accuracy and coverage of model. Furthermore, the proposed technique is easy to handle and the outputs of decision tree method are simple and clear, which are of advantage for business purpose. Moreover, good performance in customer churn prediction of our model is able to strengthen the ability of enterprise for customer retention. For another, with new telecom products and services coming up continuously, how to increase and distinguish potential consumers has been an urgent requirement for telecom enterprises. The proposed technique would make contributions to realize precise marketing by identifying different potential product consumers according to their consumer behaviors.