**Churn Prediction in Telecommunication Industry**

**Project Report**

**Abstract**

Telecommunication market is expanding day by day. Companies are facing a severe loss of revenue due to increasing competition hence the loss of customers. They are trying to find the reasons of losing customers by measuring customer loyalty to regain the lost customers. The customers leaving the current company and moving to another telecom company are called churn. Churn rate refers to the proportion of contractual customers who leave a supplier during a given time period. This phenomenon is very common in highly competitive markets such as telecommunications industry. In a statistical setting, churn can be considered as an outcome of some characteristics and past behavior of customers.

The predictive churn model presented in this study is based on the theory of survival analysis. In this report, churn prediction is performed considering a dataset of a telecommunications company. The model used for prediction churn is Logistic Regression. Finally, a prediction model and verification results are presented. Results indicate that some churn determinants influence customer churn, either directly or indirectly through a customer’s status change, or both; therefore, a customer’s status change explains the relationship between churn determinants and the probability of churn. The results are detailed discussed.

**Introduction**

Competition in the wireless telecommunications industry is rampant. To maintain profitability, wireless carriers must control churn, the loss of subscribers who switch from one carrier to another. The focus of telecommunication companies has therefore shifted from building a large customer base into keeping customers in house. For that reason, it is valuable to know which customers are likely to switch to a competitor in the near future. Those customers are so called churned customers. Since acquiring new customers is more expensive than retaining existing customers, churn prevention can be regarded as a popular way of reducing the company’s costs. With this objective, this research was carried out. So, detecting the “going to be churner” customers beforehand is the objective of the telecom companies.

During preprocessing, the available data tables were transformed so that a classification algorithm could be applied. In the resulting data set, each row (that is, each example for classification) corresponded to one customer of the company, and contained many features describing their telecommunication behavior for each of five consecutive months. Whether or not the customer left the company was determined by the binary classification label or target.

The goal of our research is to evaluate the benefits of predicting churn using techniques from statistical machine learning. We designed model that predict the probability of a subscriber churning.

**Variables Description**

To find the solution for the issues using data mining, the first step is to understand the data. We selected the continuous months(July, August, September, October, November) in order to build the model. The following are the three main types of data, which were available for the given months:

1. Data file: This data file contained data related to internet usage( vol2g, vol3g)
2. Payment file: This file was related to customer’s segmentation and the payment details for various type(STD, ISD, Local)
3. Usage file: This file was related to Voice usage of the customers.

Other than the above files we also had Request file, which consisted of the customer’s type of Requests made and the date on which the request was made, we also had disconnection file which included the dates on which the customer’s churned.

**Methodology**

The dataset provided contains all telecom customer information. The data stored in this used for predictive modeling. The data is already monthly aggregated into customer data information which includes internet usage like 2G/3G; Pay information like mobile activation date, customer segmentation, call type and Usage information which includes the incoming, outgoing voice usage of customers.

We worked on R for Data preprocessing and Rhadoop to store the data and build the model. We worked on continuous 5 months data from July to November.

The data contained duplicate values, which we removed and retrieved the unique values in all three files.

For each month we preprocessed the data by first reading the file, we used the package “Data.Table” for this and retrieved the unique values for each of the three files. Then we merged all the files for that particular month by “id”.

We removed the rows, which had more than 27 NA’s that was generated after merging the files.

Next we moved on to detecting and removing the outliers, we use the package “extremevalues” to detect the outliers and used the function “getoutliers” for the same. Which was graphically represented. Once the outliers were detected, we used inter-Quartile Range (I.Q.R) to remove the outliers. We multiplied the IQR with 1.5 is so that a certain proportion of the sample in a normal population will be outside it. Then Added 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier. And Subtracted 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier. These outliers were then replaced with NA’s.

REMOVING NAS

NA’s found in the numeric column were replaced with the mean of the values in that column; as for categorical variables, we replaced the NA’s with the mode.

For the date column, we replaced the NA’s with the first date of that particular month.

(eg-Replace the NA with last date of August month

"01AUG2014")

Since Not all fields of the database are suitable for modeling purposes. Fields with unique values, like customer group, section and repeated columns, such as mobile activation are removed.

DERIVING VARIABLES

We derived 13 new variables using function “mutate” from the package “DPLYR”. Few of the variables were derived, such as, fraction of 2-G and 3-G data, and ratio of incoming to outgoing calls, etc.

Then finally the similar process was repeated for the other 4-months data file and merged to display in a single view. We merged the final view with the disconnection data to derive the target variable “churn-status”. The existing customers till date were the last day of the month i.e. 30th November 2015. Using this assumption we calculated the number of days, which would be an important parameter to decide whether exiting customer would churn or not.

We have also used request file with disconnection date and calculated the No of days between the customers complaining and the disconnection date. Later we derived another variable called Ratio.

We did a check for multi colinearity using function cor(). The dataset was further divided into 70:30 ratio as train and test which was passed into the logistic model. We then developed the model using Logistic Regression on train data. Later we predicted the probabilities and used confusion matrix to get accuracy, which came up to 90.05%.

**Result**

The task was solved using Logistic Regression, which achieved a predictive accuracy of 90.05%. This good result was only possible due to the introduction of relevant derived features for prediction, which were not available in the original data, and due to a representation of the data so that temporal aspects could be included. Thus data preprocessing was a key success factor in this application.

The result of the AUC is 0.7846 and kappa is 0.6469, which are moderately good.

We can state that the extended model gives satisfying results with both sensitivity and specificity.

roc=roc(churn\_status~outputnew,test)

plot(roc)

Call:

roc.formula(formula = churn\_status ~ outputnew, data = test)

Data: outputnew in 327347 controls (churn\_status 0) < 81929 cases (churn\_status 1).

Area under the curve: 0.7846



confusionMatrix(outputnew,test$churn\_status,positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 320119 33492

1 7228 48437

Accuracy : 0.9005

95% CI : (0.8996, 0.9014)

No Information Rate : 0.7998

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

Kappa : 0.6469

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

Sensitivity : 0.5912

Specificity : 0.9779

Pos Pred Value : 0.8702

Neg Pred Value : 0.9053

Prevalence : 0.2002

Detection Rate : 0.1183

Detection Prevalence : 0.1360

Balanced Accuracy : 0.7846

'Positive' Class : 1

**Conclusion**

This report describes and discusses a study in the mobile telephone industry. It demonstrates convincingly the benefits of the application of a knowledge discovery process through data mining for informing decision making on customer defection. Retaining customers is one of the most important issues in customer relationship management (CRM). It is believed that the data mining process undertaken underpins a sound quality service and customer care improvement approach. The model has allowed the company to identify several areas for enhancement and the challenge is now to incorporate the model into marketing strategies and make retention actions on it. Data mining provides a collection of methodology, techniques, and approaches that help to fully integrate the solutions with existing business knowledge and implementation, and as such is likely to become a highly useful customer retention tool in the coming years.

*“Knowledge is information that changes something or somebody either by becoming grounds for actions or by making an individual (or an institution) capable of different or more effective action.”*

The experiment facilitates the company with these basic premises.