Customer Churn Prediction Model for Telecom

by

Vijay Singh and Sudarsan R

A project report
submitted to the Aegis School of Business, Data Science and Telecommunication, Mumbai
in fulfilment of the requirements for the degree of
PGP in Business Analytics and Big Data

May, 2016

ABSTRACT
Given the dynamic nature of pre-paid mobile phone subscribers and the ease with which they can stop using their phone services without giving any notice, combined with the increasing influence of their group of close friends/family/peers, the task of managing churn has become of prime importance to telecom service providers. In this study we used R programming and Hadoop ecosystem to predict customer churn. We used continuous 5-month data to model customer churn and model takes as input data of customer data usage, tariff plan, payment details. logistic regression model used for churn prediction model, and it yields 85.15% accuracy.

INTRODUCTION
Mobile phones are now fast becoming a commodity. Most cellular circles in growing countries like India now have as many as 5-7 offerings such as GSM and CDMA being provided by service providers. In such a competitive market, it becomes very easy to switch your cellular service provider merely at the drop of a hat. Rapid advancement in next-generation services has drawn the attention of teenagers and rural subscribers by features like free SMS. It has also enabled the service providers with tools to attract working professional by offering them free STD minutes, bundling and value added services. Tumbling average revenue per user (ARPU) figures, an ever increasing MOU (Minutes of Usage), together with the fact that ARPU's rate of decline has outpaced the MOU's rate of increase, there is an increasing pressure on service providers to maintain their margins. (ARPU is declining because more and more subscribers are signing up and a constant decline in rate plans and an upsurge in attractive tariff plans.) Soon, such a commoditization will cause the market price of mobile telephone services to fall to the marginal cost of lowest-cost volume producer.

The problem of churn originated from European countries where the matured markets gave an incentive to the operators to try to attract customers from competitors. From a global perspective, churn of mobile operators led them to loss of ~$100 billon USD per year (Berson et al., 2000). It has assumed alarming proportions in growth markets like India as well. According to Gartner research, India's churn rate is a high 3.5 – 6% a month, aggregating to ~40- 50% every year. This fact combined with the high installation and marketing costs, makes it 5-10 times more expensive to acquire a new customer than to retain an existing one (Ruta et al., 2006). All this shows that the churn of an existing customer, especially in pre-paid category, could hit the bottom-line of these operators. Also, today, most of Chinese and South African service providers are selling handsets along with the tariff plans, which means that handset in itself is giving the customers an incentive to churn.

These factors make the task of predicting churn as one of high priority for service providers. To stay competitive in this market, they must be able to correctly predict risky subscribers on whom the subsequent retention efforts should be focused. On top of that, MNP (Mobile Number Portability) is looming as a threat to service providers. MNP can provide flexibility to subscriber by letting him change the service provider at will. It can also fuel a hard-nosed battle between providers. The only flip side to MNP for the customer is that the operators will charge maintenance and monthly fees from them, and the time taken to port the number from one provider to another will cause some inconvenience to the customer. So, MNP will cause churn. Unlike post-paid subscribers, pre-paid subscribers can annul their service without giving any prior indication because they are not bound by any contract. They might churn if their current needs charge or if they get influenced by their social network of family and friends. This dynamic situation makes the task of predicting the likelihood (and timing) of churn very important in the context of pre-paid segment. So, in order to survive competition,

telecommunications service providers must detect the main reasons for both the expected churn and the churn that happens after the event has taken place in pre-paid category because this information can help them to customize their offers. It can be a tool to effectively anticipate the demands of their key customers who have the highest churn propensity, fully knowing that retention can have a huge impact on life time value.

An enterprise can increase its profits by 25-95% by reducing its churn by just 5% (Reichheld et al., 1990), which shows the impact of doing analytics. In order to reduce the losses caused by churn, operators have to find the most valuable customers who are inclined to churn, and then carry out retention policies for them.

Here is a formula to correlate churn problem with the ultimate goal of achieving loyalty: Higher the churn, lower the chances of being loyal. So, Churn% ~= 100% – Loyalty%.

It is very important to know the business goal along with the data mining goal. For e.g., the business goal could be to reduce churn rate by 15% in next 6 months. Whereas, the data mining goal could be to achieve ~95% accuracy in prediction with a Lift of > 2 being captured in top 20 percentile users (where accuracy is defined as the ratio of predicted churn to actual churn). Also, it is the proportion of correct churn predictions, not the number of absolute correct predictions, which should be more important for the business to analyze.

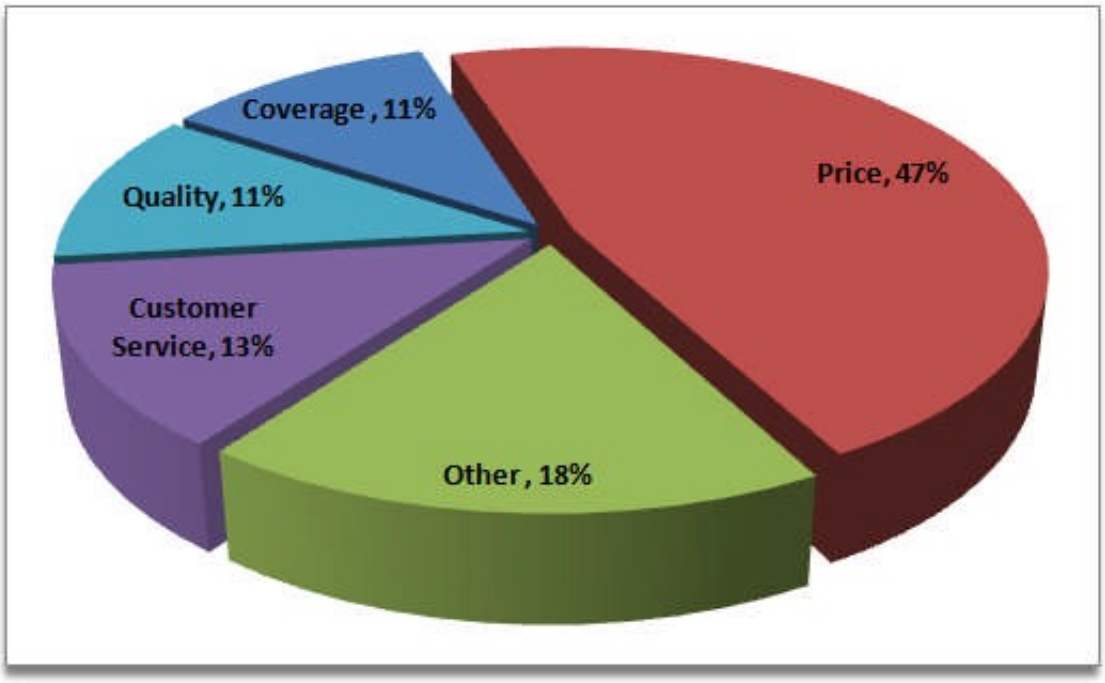

Figure 1: Main Churn Contributors for The Telecommunications Industry (Chu et al., 2007)

The main contributors to churn as illustrated in Figure 1 have a significant implication to researchers attempting to develop a model to capture it. According to these contributors, the main reason for customer churn is pricing issues. Pricing is responsible for 47% of defection. It is anticipated that pricing is the major contributor of what could be regarded as spontaneous churn. The customer has moved competitor through no real fault of the service provider but because he/she has found a similar service at a lower price. This means that any strategy will be at best 53% effective if a methodology targets the churn of the customers who defect due to reasons of dissatisfaction.

According to Burez et al. (2009), Lift is defined as ratio of precision to overall churn rate. According to http://www.siam.org/proceedings/datamining/2010/dm10_064_richtery.pdf, only a small fraction of the subscriber base can be contacted at any given time, and the subscribers with the highest churn scores are assigned top priority. So, a churn prediction system should be measured by its ability to identify churners within its top predictions. Performance is measured using lift. For any given fraction $0 < T < 1$, lift is the ratio of the number of churners among the fraction of T subscribers that are ranked highest by the proposed system, to the expected number of churners in a random sample from the general subscriber's pool of equal size. For e.g., Lift of 5 at a fraction $T = 0.01$ means that if we contact the 1% of subscribers ranked highest by the proposed system, we expect to see five times more people who planned to churn in this population than in a 0.01 fraction random sample of the population.

DATA AND VARIABLES

A Call Detail Record (CDR) is the computer record produced by a telephone exchange containing details of a call that passed through it (http://en.wikipedia.org/wiki/Call_detail_record). It is the automated equivalent of the paper toll tickets that were written and timed by operators for long distance calls in a manual telephone exchange.

There were total 5-month's (July to November) data and each month containing 3 files of data sets which containing Uses data, Pay data and Data files. All month data combined together and having 1700646 records of customers and 136 variables, out of which the Partition node in R programming did the split (a 70:30 split ratio was chosen), thereby causing 1190452 records in training data set and 510194 records in validation data set.

METHODOLOGY

Data Mining can either be supervised or unsupervised in nature. In our case, since we already know that we have to do churn forecasting for the next month or so, hence it is supervised technique. Association mining and clustering are some examples of unsupervised techniques.

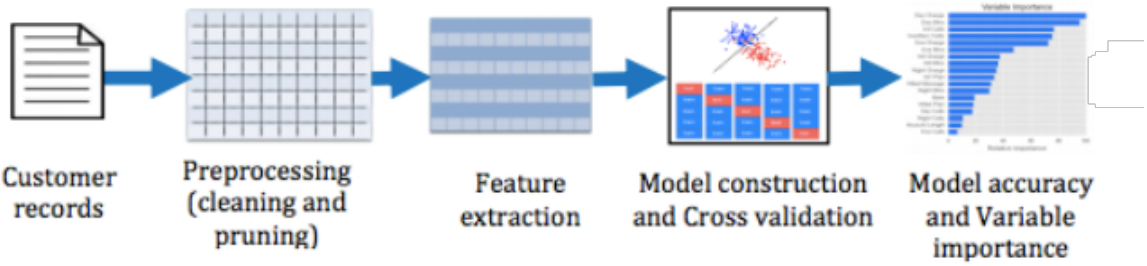

Figure-2: Different phases of a model churn prediction system used by us

Reading Data Set from CSV file: All 5-month datasets imported in R.

Data Merging and Single window creation: After successfully importing the month wise datasets in R, all are merged together and created a final single file of all datasets.

Data Cleansing: There are various type of data redundancy and can result in inconsistent data if not cleaned properly. So, first, we cleaned the data given to us. Similarly, there could be some outlier values which we replaced with the 5% and 95% of the all columns of data.

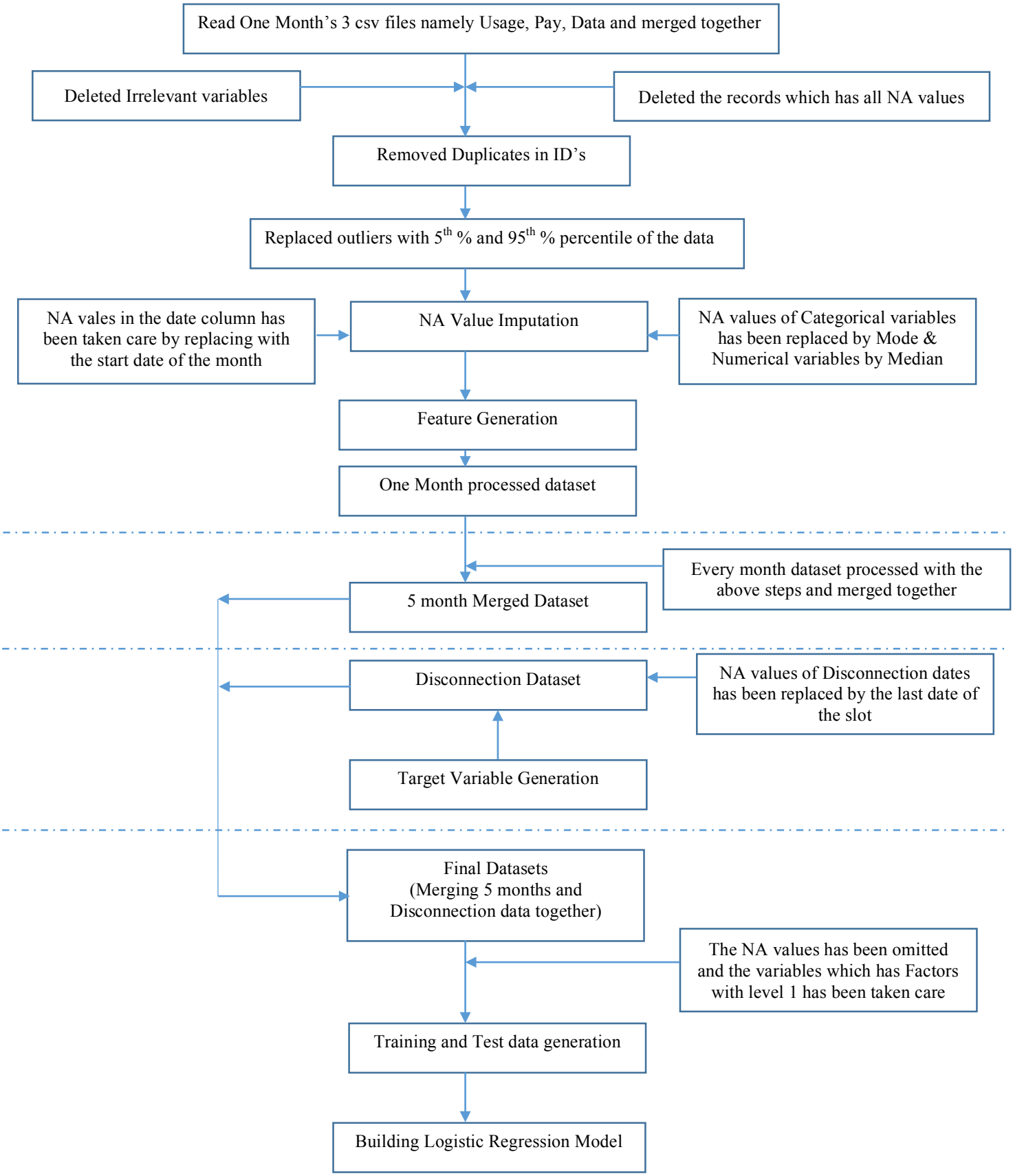

Figure-3 Methodological Flow Chart

Missing Value Analysis & Missing Value Imputation

There are certain records which are missing in the dataset and some of the NA are generated during merging process of datasets. The Na values of numerical columns has been replaced by the median of that particular column and categorical columns has been replaced by Mode of the respective column. All this exercise is required only for logistic regression. It is not required when we use decision trees.

The NA values of data columns of every months has been replaced by the starting date of the particular month. After merging all five-month data together, there were 5 mobile activation dates (July to November) whereas we have to keep one activation (July activation date) and one disconnection date in order to calculate the age on network of each customer, so in date column during the merging NA's are generated in July activation date considering there may be new customers joined in the further months, so the July activation date's NA values has been replaced by the respective further month's dates.

Variable selection

We can use Information Value/Weight of Evidence, Feature Selection in Logistic Regression to select the most significant variables. Remove the highly correlational variables using some basic statistical techniques. Also, we make sure that we end up with only one dependent variable for our study.

R Libraries Used:

- dplyr - dplyr is a package which provides a set of tools for efficiently manipulating datasets in R. dplyr is a powerful R-package to transform and summarize tabular data with rows and columns.
- data.table – It is used for fast aggregation of large data (e.g. 100GB), fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns and a fast file reader (fread)
- rhdfs – It provides basic connectivity to the Hadoop Distributed File System. R programmers can browse, read and write files stored in HDFS from R
- lubridate – Used for date formatting
- caret – The caret package (short for classification and regression training) contains functions to streamline the model training process for complex regression and classification problems. Calculates a cross-tabulation of observed and predicted classes with associated statistics.
- pROC – Tools for visualizing, smoothing and comparing receiver operating characteristic (ROC curves). (Partial) area under the curve (AUC) can be compared with statistical tests based on U-statistics or bootstrap.

RESULTS AND OBSERVATIONS

Model Generation:

```
Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = train_new)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2945  -0.6882  -0.3342  -0.0315   4.9629

Coefficients: (43 not defined because of singularities)
                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)       1.222e+00  8.608e-03  141.938  < 2e-16 ***
VAR2.x1           1.391e-01  1.417e-02    9.817  < 2e-16 ***
VAR3.x1           9.148e-02  1.056e-02    8.667  < 2e-16 ***
vol2g.x           2.024e-04  4.054e-05    4.993 5.95e-07 ***
```

Confusion Matrix and Statistics:

```
          Reference
Prediction      0      1
         0 389128  65640
         1  10115  45311

               Accuracy : 0.8515
                 95% CI : (0.8505, 0.8525)
    No Information Rate : 0.7825
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4675
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.40839
            Specificity : 0.97466
         Pos Pred Value : 0.81750
         Neg Pred Value : 0.85566
             Prevalence : 0.21747
         Detection Rate : 0.08881
   Detection Prevalence : 0.10864
      Balanced Accuracy : 0.69153

       'Positive' Class : 1

Call:
roc.formula(formula = as.numeric(test[, 136]) ~ as.numeric(test$final),     da
ta = test)

Data: as.numeric(test$final) in 399243 controls (as.numeric(test[, 136]) 0) <
110951 cases (as.numeric(test[, 136]) 1).
Area under the curve: 0.6915
```

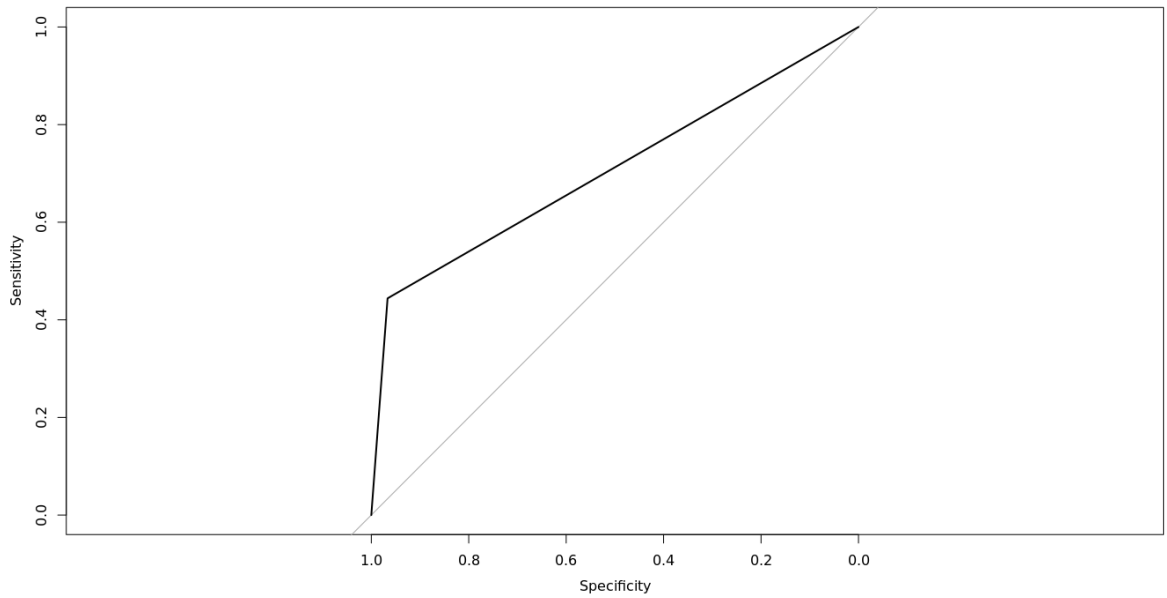

Figure 4: ROC Curve

The following are the list of rates that are computed from a confusion matrix:

- **Accuracy**: Overall, how often is the classifier correct
  - Accuracy = (True Positive + True Negative)/Total
- **Misclassification Rate**: Overall, how often is it wrong?
  - Misclassification rate = (False Positive + False Negative)/Total
  - equivalent to 1 minus Accuracy
  - also known as "Error Rate"
- **True Positive Rate**: When it's actually yes, how often does it predict yes
  - True Positive rate = True Positive /Actual Yes
  - also known as "Sensitivity" or "Recall"
- **False Positive Rate:** When it's actually no, how often does it predict yes
  - False Positive rate = False Positive / Actual No
- **Specificity**: When it's actually no, how often does it predict no?
  - Specificity = True Negative / Actual No
  - equivalent to 1 minus False Positive Rate
- **Precision**: When it predicts yes, how often is it correct?
  - Precision = True Positive / Predicted Yes
- **Prevalence**: How often does the yes condition actually occur in our sample?
  - Prevalence = Actual Yes / Total
- **Positive Predictive Value**: This is very similar to precision, except that it takes prevalence into account. In the case where the classes are perfectly balanced (meaning the prevalence is 50%), the positive predictive value (PPV) is equivalent to precision.
- **Cohen's Kappa:** This is essentially a measure of how well the classifier performed as compared to how well it would have performed simply by chance. In other words, a model will have a high Kappa score if there is a big difference between the accuracy and the null error rate.
- **F Score:** This is a weighted average of the true positive rate (recall) and precision.

- **ROC Curve:** This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.

In relation to Bayesian statistics, the sensitivity and specificity are the conditional probabilities, the prevalence is the prior, and the positive/negative predicted values are the posterior probabilities.

CONCLUSIONS AND FUTURE WORK

This project has focused to identify the factors that influence customer churn in telecommunication. The analysis focused on churn prediction based on logistic regression and R programming. The logistic regression model predicted the actual churners with 85.15% accuracy, which is quite good. The findings of this study indicate that the user should update the logistic regression model to be able to produce predictions with more accuracy. The effect of derived variables on the accuracy of the model was also studied.

However, there is one caveat while using any modeling tool such as R programming. If the customer has been showing an increasing trend of his phone's usage, then our churn prediction model would tend to suggest that the CSP should focus its efforts on sending him a higher rate plan that can eventually generate potentially higher revenues. However, he may have increased his recent usage only because of a battery problem that resulted in frequent email synchronization. So, at times, the churn prediction itself could be misleading. If we can detect such patterns in data that can give meaningful predictions, then it could help enhance the value of next offer for him. This is akin to the false alert, and needs to be studied carefully.

Here are the additional variables required to increase our model's predictive power:

- **Migration data** (say 'Active' status to 'Grace period', 'Grace' to 'Active', 'Grace' to 'Churn', etc. Hypothesis: If there is a pattern of such migration, then it can be a predictor of future behavior).
- **Number of days between inactivity and recharge**. (Hypothesis: If there is a pattern of say, 10 days of inactivity before which he generally recharges, and if he has been inactive recently for more than 10 days, then we can use it as a churn indicator).
- **Remaining balance**: (Hypothesis: Lower the remaining account balance, higher his churn likelihood. Also, it could be important to analyze the ratio of balance to top-up. In other words, how much of his balance goes unutilized when his validity period gets expired. Another pattern worth exploring could be to explore how many customers opt for recharge just one or two days within the expiration of their validity period).
- **Location data** (Hypothesis: Voluntary churn could be higher in Kolkata than in Bangalore because of higher social connectivity of customers in Kolkata).
- **Watching churn events more closely** (Hypothesis: If a negative event is followed by another negative event (e.g., filing a complaint twice), it can lead to churn).

ATTACHMENTS: Project R programming code file attached herewith

REFERENCE

Berson, S. Smith, and K. Thearling, "*Building data mining applications for CRM," New York: McGraw-Hill* (2000). F. F. Reichheld, W. E. Jr. Sasser, "*Zero defections: Quality comes to services," Harvard Business Review, Vol.68, 1990, pp. 105-111* (1990)

Ruta, D., Nauck, D., Azvine, B.: K nearest sequence method and its application to churn prediction. *In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 207–215. Springer, Heidelberg* (2006)

Rosset S., Abe N.: Data Analytics for Marketing Decision Support, *IBM T.J. Watson Research Center* (2006)

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., & Nanavati, A. A.: Social Ties and their Relevance to Churn in Mobile Telecom Networks. *Proceedings of the 11th international conference on Extending database technology pp. 668—677* (2008)

S. Y. Hung, D. C. Yen and H. Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications, Vol.31, pp. 515–524* (2006)

Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction, Expert Systems with Applications 36 (2009) 4626–4636
http://www.tmforum.org/BestPracticesStandards/1669/home.html
http://www.indepay.com/is_telecom.htm
http://www.mobilephone-news.com/2010/11/mnp-to-cost-rs-19/
http://www.outlookindia.com/article.aspx?264134
http://www.mshare.net/why/customer-loyalty.html
http://strategy-redefined.blogspot.com/2010/09/customer-churn-management-in-telecom.html
http://retailbusinessnewsletter.com/page/3/
http://www.tmcnet.com/usubmit/2008/01/29/3237095.htm
http://www.norusis.com/pdf/SPC_v13.pdf
http://userwww.sfsu.edu/~efc/classes/biol710/logistic/logisticreg.htm