A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics

V. Umayaparvathi[1], K. Iyakutti[2]

[1] Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India
[2] Professor-Emeritus, Department of Physics and Nanotechnology, SRM University, Chennai, Tamilnadu, India

---***---

Abstract - *In this competitive world, business is becoming highly saturated. Especially, the field of telecommunication faces complex challenges due to a number of vibrant competitive service providers. Therefore, it has become very difficult for them to retain existing customers. Since the cost of acquiring new customers is much higher than the cost of retaining the existing customers, it is the time for the telecom industries to take necessary steps to retain the customers to stabilize their market value. In the past decade, several data mining techniques have been proposed in the literature for predicting the churners using heterogeneous customer records. This paper reviews the different categories of customer data available in open datasets, predictive models and performance metrics used in the literature for churn prediction in telecom industry.*

Key Words: Customer relationship management (CRM), Data mining, Customer churn prediction, Predictive models, and Performance metrics.

1. INTRODUCTION

Today is the competitive world of communication technologies. Customer Churn is the major issue that almost all the Telecommunication Industries in the world faces now. In telecommunication paradigm, Churn is defined to be the activity of customers leaving the company and discarding the services offered by it due to dissatisfaction of the services and/or due to better offering from other network providers within the affordable price tag of the customer. This leads to a potential loss of revenue/profit to the company. Also, it has become a challenging task to retain the customers. Therefore, companies are going behind introducing new state of the art applications and technologies to offer their customers as much better services as possible so as to retain them intact. Before doing so, it is necessary to identify those customers who are likely to leave the company in the near future in advance because losing them would results in significant loss of profit for the company. This process is called Churn Prediction.

Data mining techniques are found to be more effective in predicting customer churn from the researches carried out during the past few years. The construction of effective churn prediction model is a significant task which involves lots of research right from the identification of optimal predictor variables (features) from the large volume of available customer data to the selection of effective predictive data mining technique that is suitable for the feature set. Telecom Industries collect a voluminous amount of data regarding customers such as Customer Profiling, Calling pattern, Democratic data in addition to the network data that are generated by them. Based on the history of the customers calling pattern and the behavior, there is a possibility to identify their mindset of either they will leave or not. Data mining techniques are found to be more effective in churn prediction from the researches carried out for the past one decade. Especially Predictive modeling techniques are often found to be more accurate in churn prediction.

In this paper, we review the existing works on churn prediction in three different perspectives: datasets, methods, and metrics. Firstly, we present the details about the availability of public datasets and what kinds of customer details are available in each dataset for predicting customer churn. Secondly, we compare and contrast the various predictive modelling methods that have been used in the literature for predicting the churners using different categories of customer records, and then quantitatively compare their performances. Finally, we summarize what kinds of performance metrics have been used to evaluate the existing churn prediction methods. Analyzing all these three perspectives is very crucial for developing a more efficient churn prediction system for telecom industries.

While there are other churn prediction surveys available in the literature [1][2][3], they primarily focused on different modelling techniques. To the best of our knowledge, none of those surveys reviewed the datasets and metrics for evaluating the churn prediction models. Hence, we believe that this survey can provide a roadmap for both researchers and customer relationship managers to better understand the domain and challenges in detail.

The rest of the paper is organized as follows. In Section 2, we present the details about a model churn prediction system, In Section 3, we explain different categories of customer details (data collected by telecom industries) and also summarizes the properties of publicly available datasets for customer churn prediction. In Section 4, we review the existing predictive models proposed in the literature, followed by in Section 5, we explore the various performance metrics used in evaluating the existing churn prediction methods. Finally in Section 6, we conclude the paper.

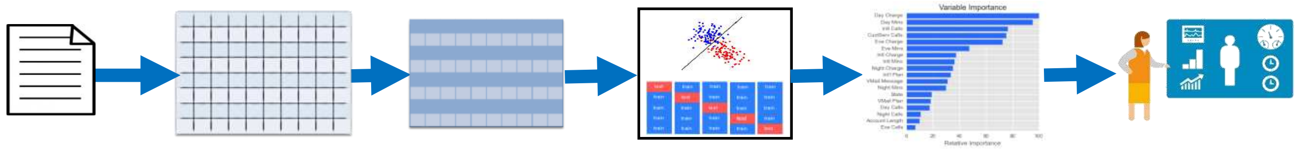

Figure-1: Different phases of a model churn prediction system proposed by us in our earlier research work. It consists of five phases: 1) Preprocessing the input customer records, 2) Extracting the required features for developing churn models, 3) Construction models using different classifiers and cross validate the models, 4) Calculation of prediction accuracy and variable importance report, and 5) Providing customer retention polices to CRM executives.

2. CHURN PREDICTION PROBLEM

In a business environment, the term, customer attrition simply refers to the customers leaving one business service to another. Customer churn or subscriber churn is also similar to attrition, which is the process of customers switching from one service provider to another anonymously. From a machine learning perspective, churn prediction is a supervised (i.e. labeled) problem defined as follows: Given a predefined forecast horizon, the goal is to predict the future churners over that horizon, given the data associated with each subscriber in the network.

Churn Prediction is a phenomenon which is used to identify the possible churners in advance before they leave the network. This helps the CRM department to prevent subscribers who are likely to churn in future by taking the required retention policies to attract the likely churners and to retain them. Thereby, the potential loss of the company could be avoided.

The input for this problem includes the data on past calls for each mobile subscriber, together with all personal and business information that is maintained by the service provider. In addition, for the training phase, labels are provided in the form of a list of churners. After the model is trained with highest accuracy, the model must be able to predict the list of churners from the real dataset which does not include any churn label. In the perspective of knowledge discovery process, this problem is categorized as predictive mining or predictive modeling.

A model churn prediction system is illustrated in Figure-1, which is proposed in [4][5]. It consists of five phases: 1) Preprocessing the input customer records, 2) Extracting the required features for developing churn models, 3) Construction of models using different classifiers and cross validate the models, 4) Calculation of prediction accuracy and identification of variable importance report, and 5) Providing customer retention polices to CRM executives.

3. CUSTOMER ATTRIBUTES AND DATASETS

In this section, we present the details about the list of existing public datasets and summarize under what categories, the customer related attributes are grouped and how these groups are used for churn prediction.

3.1 Category of Customer information

Telecom companies collect enormous amount of customer details every day and whenever a customer subscribes to service. These customer details can be grouped into following six categories.

1) Customer care service details.
2) Customer demography and personal details.
3) Customer credit score.
4) Bill and payment details.
5) Customer usage pattern.
6) Customer value added services.

The following datasets contain most of the above categories of customer details. For each dataset, we present their characteristics, number of customers, training and testing set size, number of churners and so on.

3.1.1. PAKDD 2006 Data Mining Competition Dataset

This dataset was provided as part of the PAKDD 2006 data mining competition[1]. The dataset was released by an Asian telecom company who has successfully launched a 3G mobile telecommunications network. The company wanted to identify which customers are likely to switch to using their 3G network, by making use of existing customer usage and demographic data. It consists of 24,000 customers. Each customer is described by 250 attributes and a class label (2G/3G). The training set consists of 18000 records, whereas the test set consists of 6000 customer records. Among this, 3150 customers were churners and the remaining 20850 were non-churners, in both training and test set. It consists of all the categories of customer attributes explained in Section 3.1. The complete details of all attributes and their descriptions are given in Appendix A.

[1] http://www3.ntu.edu.sg/SCE/pakdd2006/competition/overview.htm

Dataset	Total no. of consumers	Total no. of variables	No. of non-churners	No. of churners	% of non-churners	% of churners	Country
PAKDD 2006	24,000	250	20,850	3,150	83	13	Asian
KDD Cup 2009 small	50,000	15,000	46,328	3,672	92.5	7.5	Europe
Cell2Cell	70,831	75	50,326	20,505	71	29	USA
CrowdAnalytix	3,333	20	2,850	483	86	14	NA

Table-1: List of various publicly available churn prediction datasets and their properties

3.1.2 ACM KDD Cup 2009 Orange Labs Dataset

This dataset was provided as part of the KDD Cup 2009 data mining competition[2]. The dataset was released by Orange Labs, an European telecom company. The company has developed its own prediction models as part of its CRM system for identifying the churners. But the challenge of this competition is to beat the in-house system developed by Orange Labs. Two versions of datasets were released for the competition: Small and Large. The small version consists of 50,000 records in total, including training and test sets. Both small and large datasets have numerical and categorical variables. But, due to privacy reasons, the actual variables names are not revealed in this dataset. For the large dataset, the first 14,740 variables are numerical and the last 260 are categorical. For the small dataset, the first 190 variables are numerical and the last 40 are categorical. In the small dataset, 3672 customers are churners and the remaining 46328 are non-churners.

3.1.3 Cell2Cell Dataset

This dataset was provided by Teradata Center for Customer Relationship Management of Duke University[3]. This real dataset was collected from customers of Cell2Cell Telecom Company. Cell2Cell is one of the largest wireless companies in the USA with more than 10 million customers and its average monthly churn rate is 4%. For churn prediction, Cell2Cell collected several data about its customers including 1) customer care service details, 2) customer demography and personal details, 3) customer credit score, 4) bill and payment details, 5) customer usage pattern, and 6) customer value added services, totaling to 75 variables from 71, 047 customers. The dataset is divided into calibration and validation set. The calibration set (training data) contains 50% churners, 20,000 among 40,000 customers. Whereas, the validation set contains approximately 2% churners, 609 among 31,047 customers.

[2] http://www.kdd.org/kdd-cup/view/kdd-cup-2009
[3] http://www.fuqua.duke.edu/centers/ccrm/index.html

3.1.4 CrowdAnalytix Community

This public dataset is provided by the CrowdAnalytix community[4] as part of their churn prediction competition. The real name of the telecom company is anonymized. It contains 20 predictor variables mostly about customer usage patterns. There are 3333 records in this dataset, out of which 483 customers are churners and the remaining 2850 are non-churners. Thus, the ratio of churners in this dataset is 14%.

Table 1 summarizes the list of publicly available churn prediction datasets and their characteristics.

4. CHURN PREDICTION MODELS

There are a lot of researches being carried out in the area of customer churn prediction modeling. In this section, we survey some of the researches carried out in this area in the past few years.

Authors of [1] depict phases of a general churn prediction model such as data collection, preparation, classification and prediction. It also describes that identifying the right grouping of variables has significant influence in improving the percentage of true predictions (TP). A churn prediction model was proposed by [1], which works in 5 steps: i) problem identification; ii) dataset selection; iii) investigation of data set; iv) classification; v) clustering, and vi) using the knowledge. It seems to be a complete model. Classification techniques are used for distinguishing Churners. Clustering is used for model evaluation. For classification, Decision tree, Support vector machine and Neural Network are used. And for clustering Simple K-Means was used. It concluded that SVM was the best among the three methods in distinguishing churners from non-churners.

Building an effective customer churn prediction model using various techniques has become a significant topic for business and academics in recent years [2]. The identification of why customers give up their relationships has been focus of marketing research for the past few years [3]. Due to the enormous growth of customer related data and call detail

[4] https://www.crowdanalytix.com/contests/why-customer-churn

Author	Dataset	Features	Methods	Metrics
Adem Karahoca [13]	GSM operator, Turkey 24,900 customers 22 attributes	Demography, Usage pattern, Value added services	x-Means clustering, Adaptive Neuro Fuzzy Inference System	Precision and Recall
Clement Kirui [14]	European operator 106,405 customers 112 attributes	Contract, usage pattern patterns, and calls pattern	Naïve Bayes, Decision Tree	Confusion matrix, accuracy, precision, recall
Ballings, Michel [15]	Unknown 129,892 customers 113 attributes	Demographic, Value added, usage pattern	Logistic regression, Bagging, Decision Tree	AUC
Ismail, Mohammad [16]	Unknown, 169 customers 10 attributes	Demographic, Billing data, usage pattern, customer relationship	Neural network, Regression	Confusion matrix, accuracy, precision, recall
H Lee [17]	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral information, Customer care and demographics	Stepwise variable selection partial least squares	Proportion of hit records
Anuj Sharma [18]	ML Dataset at UCI 2,427 customers 20 attributes	Demographics, Usage pattern, Value added services	Artificial Neural Network	Confusion matrix
Abbas Keramati [19]	Iranian telco operator 3150 customers 15 attributes	Demographic, call usage pattern, customer care service	Binomial logistic regression model	Statistical hypothesis test
Kristof Coussement [20]	Belgian 134, 120 customers 27 attributes	Demographic Usage patter, bill and payment	generalized additive models (GAM)	AUC top-decile lift
Marcin Owczarczuk [21]	Polish mobile operator 122098 customers 1381 attributes	Demographic, call data records, customer care services	Logistic regression Decision tree	Lift curves
Umayaparvathi [5]	Cell2Cell Dataset 100,000 customers 171 attributes	Behavioral information, Customer care and demographics	Gradient Boosting, Decision Tree, Support Vector Machine, Random Forest, K-NN, Ridge Regression and Logistic Regression	Confusion matrix, accuracy, precision, recall, F1-score

Table-2: Literature survey of datasets, features, models and metrics used by various churn prediction systems

data collected and maintained by the companies in the recent years, more sophisticated metrics have evolved to describe customer behaviour and better understand how behavioural characteristics can be linked to customer retention and firm performance [4].

Authors in [5] proposed a decision tree based Random Forest method for feature extraction. From the original data with Q features, exactly N samples with q < Q are randomly selected for each tree to form a forest. The number of trees depends on the number of features randomly combined for each decision tree from the whole Q features. Development of a predictive model based on data-centric approach for detecting the early warning signs of churn, and for developing a Churn Score to identify subscribers who are likely to end their relationship with the company was discussed in [6], where the customer's calling pattern played the major role in predicting churn.

An overview of socially grouped users and their behavioural pattern are elaborately identified in [7]. It also explores the impact of centrality features among customers. It concludes that when a customer in a group leaves the network, there is a high probability of others in that group to leave the network. It classifies the feature variables in two types: i) dependent variables (call duration of in-degree and out-degree), ii) independent variables (social forum like services and the customer's involvement in the forum). The level of customer's active participation was used as the measure of probability of churn. Customers who are members of multiple community forums are at high risk than those who are members in less number of forums. A review of existing work on churn prediction datasets, features, methods and metrics are listed in Table 2.

5. PERFORMANCE METRICS

There are several standard performance metrics proposed in the literature to compare the effectiveness of the different classifiers for churn prediction. These metrics are suitable for analyzing the performance of any model which is built using both balanced and unbalanced dataset. The metrics are described below.

1. ***Confusion matrix***: It is a table with two rows and two columns that reports the number of false positives

(FP), false negatives (FN), true positives (TP), and true negatives (TN). It provides the required information for analyzing the churn prediction accuracy in terms of false.

2. ***Accuracy***: Accuracy of the given prediction model is defined as below

   ACC = (TP + TN) / (TP + TN + FP + FN)

3. ***Precision and Recall***: It is defined as below

   Precision = TP / (TP + FP)

   Recall = TP / (TP + FN)

4. ***F1-score***: It is defined as below

   F1 = 2 TP / (2TP + FP + FN)

5. ***AUC curve****:* In contrast with other metrics AUC is not influenced by any threshold value as it takes into account all possible thresholds on the predicted probabilities.

6. ***Decile lift or Lift curve****:* The top-decile lift focuses exclusively on the most critical group of customers and their churn risk.

A review of existing works which use different evaluation metrics are given in Table-2.

6. CONCLUSION

In this paper, initially, we introduced the churn prediction problem and the significance of using predictive modeling methods to overcome the problem of customer churn in telecom industry. We surveyed the existing churn prediction methods in detail and summarized them. Unlike other surveys, which primarily focused only on the prediction models and the accuracy of churn prediction, in this survey we presented the characteristics of the existing publicly available churn prediction datasets. Further, we focused on different customer related variables that are used for churn prediction and categorized them. Finally, we surveyed the list of the commonly used metrics proposed in the literature for evaluating the performance of various churn prediction methods

ACKNOWLEDGEMENT

Authors would like to thank the following research communities for making the churn prediction datasets publicly available for research purpose. 1) Teradata Center for Customer Relationship Management (CCRM) of Duke University, 2) CrowdAnalytix community, and 3) PAKDD 2006 and ACM KDD cup Data mining competition organizing committee.

REFERENCES

[1] Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." Expert Systems with Applications 39, no. 12 (2012): 11303-11311.

[2] Kamalraj, N., and A. Malathi. "A survey on churn prediction techniques in communication sector." International Journal of Computer Applications 64, no. 5 (2013).

[3] N.Hashmi, N.ButtandM.Iqbal. Customer Churn Prediction in Telecommunication A Decade Review and Classification. International Journal of Com-puter Science Vol.10(5),2013

[4] V. Umayaparvathi, K. Iyakutti, " Applications of Data Mining Techniques in Telecom Churn Prediction", International Journal of Computer Applications, Vol. 42, No.20, 2012

[5] V. Umayaparvathi, K. Iyakutti,, "Attribute Selection and Customer Churn Prediction in Telecom Industry", Proceedings of the IEEE International Conference On Data Mining and Advanced Computing, 2016 (to be appeared).

[6] Huang, Bingquan, Mohand Tahar Kechadi, and Brian Buckley. "Customer churn prediction in telecommunications." Expert Systems with Applications 39, no. 1 (2012): 1414-1425

[7] Shaaban, Essam, Yehia Helmy, Ayman Khedr, and Mona Nasr. "A proposed churn prediction model." IJERA 2 (2012): 693-697.

[8] Jain, Dipak, and Siddhartha S. Singh. "Customer lifetime value research in marketing: A review and future directions." Journal of interactive marketing 16, no. 2 (2002): 34-46.

[9] Gupta, Sunil, and Valarie Zeithaml. "Customer metrics and their impact on financial performance." Marketing Science 25, no. 6 (2006): 718-739.

[10] Yihui, Qiu, and Mi Hong. "Application of Feature Extraction method in customer churn prediction based on Random Forest and Transduction." Journal of Convergence Information Technology 5, no. 3 (2010): 73-78.

[11] Khan, Muhammad Raza, Joshua Manoj, Anikate Singh, and Joshua Blumenstock. "Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty." In Big Data (BigData Congress), 2015 IEEE International Congress on, pp. 677-680. IEEE, 2015.

[12] Karnstedt, Marcel, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. "The effect of user features on churn in social networks." In Proceedings of the 3rd International Web Science Conference, p. 23. ACM, 2011.

[13] Adem Karahoca, Dilek Karahoca,"GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system", Expert Systems with Applications 38 (2011) 1814–1822.

[14] Kirui, Clement, Li Hong, Wilson Cheruiyot, and Hillary Kirui. "Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining." IJCSI International Journal of Computer Science Issues 10, no. 2 (2013): 1694-0784.

[15] Ballings, Michel, and Dirk Van den Poel. "Customer event history for churn prediction: How long is long enough?." Expert Systems with Applications 39, no. 18 (2012): 13517-13522.

[16] Ismail, Mohammad Ridwan, Mohd Khalid Awang, M. Nordin A. Rahman, and Mokhairi Makhtar. "A Multi-Layer Perceptron Approach for Customer Churn Prediction." International Journal of Multimedia and Ubiquitous Engineering 10, no. 7 (2015): 213-222.

[17] H Lee, Y Lee, H Cho, K Im, YS Kim, "Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model", Decision Support Systems, Volume 52, Issue 1, 2011, Pages 207–216.

[18] Anuj Sharma, Dr.Prabin Kumar Panigrahi, "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services", International Journal of Computer Applications, Volume 27– No.11, 2011, pp. 0975 – 8887.

[19] Abbas Keramati, Seyed M.S.Ardabili, "Churn analysis for an Iranian mobile operator", Telecommunications Policy, 35 , 2011, pp. 344–356.

[20] Kristof Coussement, Dries F. Benoit, Dirk Van denPoel, "Improved marketing decision making in a customer churn prediction context using generalized additive models", Expert Systems with Applications, Volume 37, Issue 3, 2010, Pages 2132–214.

[21] Marcin Owczarczuk, "Churn models for prepaid customers in the cellular telecommunication industry using large data marts", Expert Systems with Applications, 37, 2010,pp. 4710–4712.

APPENDIX A

List of attributes from the Cell2Cell Dataset which are grouped into six different categories, as show in the below table.

Attribute	Description
	Customer care Service
blckvce	Mean number of blocked voice calls
custcare	Mean number of customer care calls
dropblk	Mean number of dropped or blocked calls
mailres	Responds to mail offers
mailflag	Has chosen not to be solicited by mail
retcalls	Number of calls previously made to retention team
incmiss	Income data is missing
income	Income (0=>missing)
retcall	Customer has made call to retention team
	Customer Demography
months	Months in Service
uniqsubs	Number of Unique Subs
actvsubs	Number of Active Subs
csa	Communications Service Area
phones	# Handsets Issued
models	# Models Issued
eqpdays	Number of days of the current equipment
age1	Age of first HH member
age2	Age of second HH member
children	Presence of children in HH
prizmrur	Prizm code is rural
prizmub	Prizm code is suburban
prizmtwn	Prizm code is town
refurb	Handset is refurbished
webcap	Handset is web capable
truck	Subscriber owns a truck
rv	Subscriber owns a recreational vehicle
occprof	Occupation - professional
occcler	Occupation - clerical
occcrft	Occupation - crafts
occstud	Occupation - student
occhmkr	Occupation - homemaker
occret	Occupation - retired
occself	Occupation - self-employed
ownrent	Home ownership is missing
marryun	Marital status unknown
marryyes	Married
marryno	Not Married
mailord	Buys via mail order
travel	Has traveled to non-US country
pcown	Owns a personal computer
creditcd	Possesses a credit card
newcelly	Known to be a new cell phone user
newcelln	Known not to be a new cell phone user
refer	Number of referrals made by subscriber
mcycle	Owns a motorcycle
setprcm	Missing data on handset price
setprc	Handset price (0=>missing)
	Customer Credit Score
credita	Highest credit rating
creditaa	High credit rating
creditb	Good credit rating
creditc	Medium credit rating
creditde	Low credit rating
creditgy	Very low credit rating
creditz	Lowest credit rating
creditad	Number of adjustments made to customer credit rating (up or down)
	Bill & Payment Analysis
revenue	Mean monthly revenue
recchrge	Mean total recurring charge
changer	% Change in revenues
	Customer Usage Pattern (Behaviour pattern)
mou	Mean monthly minutes of use
overage	Mean overage minutes of use
roam	Mean number of roaming calls
changem	% Change in minutes of use
dropvce	Mean number of dropped voice calls
unansvce	Mean number of unanswered voice calls
outcalls	Mean number of outbound voice calls
incallss	Mean number of inbound voice calls
peakvce	Mean number of in and out peak voice calls
opeakvce	Mean number of in and out off-peak voice calls
	Value added services
directas	Mean number of director assisted calls
threeway	Mean number of three-way calls
mourec	Mean unrounded mou received voice calls
callfwdv	Mean number of call forwarding calls
callwait	Mean number of call waiting calls
retcalls	Number of calls previously made to retention team
retaccpt	Number of previous retention offers accepted