Service Provider Churn Prediction for Telecoms Company using Data Analytics

Freddie Mathews Kau, Hlaudi Daniel Masethe and Craven Klaas Lepota, Member, *IAENG*

Abstract— Company initiated churn has a direct impact on the revenue of the company since most of the time companies are not able to recover the money subscribers are supposed to pay. Churn prediction aims to identify customer that are more likely to be churned by service providers due to non-payment or fraud. It is very expensive for companies to retain fraudulent or non-paying subscribers, meaning that the company is losing double, that is expected revenue and revenue loss due to write-offs. Predictive models can help organization to identify customers that are more likely to churn, this will help the organizations to plan beforehand and put in preventative measures. The research applies logistic regression and decision trees using R package for data analytics to predict the churn. The data set consist of 7000 instances and 17 attributes, but only 10 attributes are important for the study.

***Index Terms*— Churn prediction, data mining, logistic regression, decision trees**

I. INTRODUCTION

In the telecom industry, the broad definition of churn is the action that a customer's telecom service is cancelled or hung. This includes both service-provider initiated churn and customer initiated churn [1]. There are two types to customers namely prepaid and post-paid. Prepaid customers are the customers that pay for the services before they use them.

Churn Prediction is an important problem studied across several areas like banking, insurance, retailing, telecommunications, etc. A wide variety of techniques have been applied to predict churn in the diverse applications. A decision tree based approach has been most widely used in the churn prediction [2]. In South Africa, mobile telecommunication industry has reached a saturated point where the only way for companies to grow is by stealing customers from other networks [3]. Customer attrition refers to customers leaving a business service to another similar business service owned by different industry; churn is also comparable to attrition, which is a practice of customers changing from one service provider to another [4].

The cost associated with customer acquisition is much greater than the cost of customer retention, churn prediction has emerged as crucial business intelligence (BI) application for modern telecommunication operators. Telecommunication industry has a severely competitive market, customer demand products tailored to their needs and good services at affordable prices; their business goal is to acquire and recruit new customers as it costs 10 times more, to retain high profitable existing customers [5]. Most companies focus on predicting customer initiated churn and gives less focus on company initiated churn. Operators believe big data will play a critical role in helping them meet business objectives, promote growth, drive efficiencies and profitability across the entire telecom value chain [6].

Company initiated churn is hard to predict for telecommunication industries and often this result in revenue loss for the company. The majorities of company initiated churn are only detected after the successful activation of a subscriber by this time the fraudster or client is already using the device and incurred a huge usage.

Today most companies invest money in acquiring the customer and often spend less in protecting revenue. Fraud and arrears are the direct enemies to revenue. The new credit act protect the customers to only pay for what they afford not what they have used. The companies are no longer allowed to lend irresponsible.

The problem statement is defined as follows:

Telecommunication companies are not able to predict company initiated churn.

The following question will be addressed in this project.

How can telecommunication company use data mining techniques for churn prediction?

The aim of the research is to predict service provider initiated churn. Customer retention addresses the subject of customer churn, whereby churn pronounces turnover of customers, and supervision of churn designates efforts a business makes to detect and control the customer churn problem [7].

Manuscript received July 23, 2016; revised August 30, 2016. This work was supported by the Tshwane University of Technology. The authors, F. M. Kau is a Software Development MTech IT student at the Tshwane University of Technology, Soshanguve Campus, Pretoria, South Africa,(phone: Tel +27 83 2090042; e-mail: lovekau@gmail.com.)

H.D. Masethe is with the Department of Computer Science at Tshwane University of Technology, Soshanguve Campus, Pretoria 0001, South Africa (phone: +27 382 9714; fax: +27 866-214-011; e-mail: masethehd@tut.ac.za)

C.K. Lepota is with the Department of Computer Science at Tshwane University of Technology, Soshanguve Campus, Pretoria 0001, South Africa (phone: +27 382 9014; fax: +27 866-214-011; e-mail: lepotack@tut.ac.za)

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. RELATED WORK

Telecommunication companies in South Africa invest considerable amount of money in customer acquisition than in customer retention, the biggest enemy for telecom is customer churn, and the main driver for churn prevention in many companies is cost savings. In addition to the cost-saving benefit in churn prevention, there is the realization of a long-term continuous stream of revenue which would have otherwise been lost by increasing the customer lifetime value [8]. Logistic regression and decision tree are well known conventional statistical methods effective in predicting customer churn [5].

Managing customer churn is of great concern to global telecommunications service companies and it is becoming a more serious problem as the market matures[9]. Having the capability to accurately predict subscribers at risk of churn, with a high degree of certainty is valuable to telecom companies [8]. The churn rate is increasing dramatically at 30%, especially in the telecom sector, in order to resolve this problem, predictive models are important to be implemented to categorise customers who are at risk of churning [10].

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 [11]. Customer churn arises as a critical issue for management and customer retention in the business, thus churn prediction is relevant and treasured to retain customers and reduce the loss [12].

Researchers [9] believes that customer churn adversely affects telecommunication companies because they stand to lose a great deal of price premium, decreasing profit levels and a possible loss of referrals from continuing service customers.

There are two basic categories of churn, voluntary and involuntary; involuntary churners are the subscribers that the telecommunication company decides to remove for reasons such as fraud and non-payment; On the other hand, voluntary churn can be described as the termination of service by the subscriber; and the latter is difficult to determine, and is classified as incidental and deliberate churn [8][13].

III. RESEARCH METHODOLOGY

The research uses the following research methodologies:

A. *Literature Review*

Scoping Review is chosen to synthesize evidence from published papers, and explore literature relevant to telecommunication Churn in academic journals, books and conference proceedings [14][15]. Scoping reviews "aim to map rapidly the key concepts underpinning a research area and the main sources and types of evidence available, and can be undertaken as stand-alone projects in their own right, especially where an area is complex or has not been reviewed comprehensively before [16].

B. *Data Mining*

The research makes use of two data mining techniques named logistic regression and decision trees to build a model for predictions using R package as a tool.

C. *Decision Tree*

C4.5 algorithm is a well-known tree based classifier, whereby a tree is created; the algorithm search for an attribute with the highest information gain, and partition data into classes based on the attribute's value, further recursive partition continues on each sub-tree until a leaf node is reached [12].

D. *Logistic Regression*

Logistic regression (LR) is a regression analysis widely applied in probabilistic classification applications estimated through the formula [12]:

$$P(y = 1|x_1, \ldots, x_k) = \frac{e^{b_0 + b_1 x_1 + \cdots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + \cdots + b_k x_k}}$$

E. *Experimental Set-up*

In order to investigate service provider churn comprehensively, the dataset was divided into test data and training data, so as to conduct the experiment. The experiments were conducted using R package tool, the data set that was used had seventeen (17) attributes as indicated in table 1 below and only ten (10) attributes were used in the prediction model. The two classifiers being decision tree and logistic regression are used to determine which subscriber will churn.

Table 1: Data Set

Field Name	**Field Description**
CustomerID	Unique identity
Gender	Male
SeniorCitizen	Yes = 1, No = 0
Partner	Yes, No
Dependents	Yes , No
Tenure	Duration of contract
PhoneService	Data or Voice line
MultipleLines	Subscriber with Multiple lines
InternetService	Is it Data only or not
DeviceProtection	Device insured or not
Contract	Duration of contract
PaperlessBilling	Is the address valid
PaymentMethod	Debit or Cash
MonthlyCharges	Subscription
TotalCharges	Total Usage
AboveCreditLimit	Yes = 1, No = 0
Churn	Churn or not?

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. RESEARCH RESULTS

A. Literature Review

Table 2 Review of Data Mining techniques and benefits

Author	Technique	Benefit
[5]	Logistic Regression and Decision Tree	Predicting Customer Churn
[4]	Decision Tree and Neural Network	Churn prediction in mobile telecom system
[7]	Genetic Programming	Intelligent churn prediction
[13]	J48 Data mining algorithm	Churn prediction in telecom
[17]	Naïve Bayes, Bayesian Network, C4.5 decision tree	Predicting customer churn
[18]	Decision tree, Support Vector Machine and Neural Network	Churn prediction
[10]	Support Vector Machine	Predictive model
[12]	Novel Hybrid model-based learning system	Accurate predictive results

Table 2 above shows a summary of data mining techniques used for churn prediction through the literature.

B. Decision Tree

The diagram in figure 1 below illustrates the results from a decision tree model. The model used four main attributes Churn, InternetService, AboveCreditLimit and Contract. The results shows that 'AboveCreditLimit' is the most influential attribute because every customer that churned was above credit limit but there are customers who didn't churn but were above credit limit, the second attribute followed by the Internet service where majority of people that churn were using 2G, 3G or had no internet service to the tree, the last attribute which contributed was the Contract the majority of the churners were on month-to-month which contributed to 90% of churners.

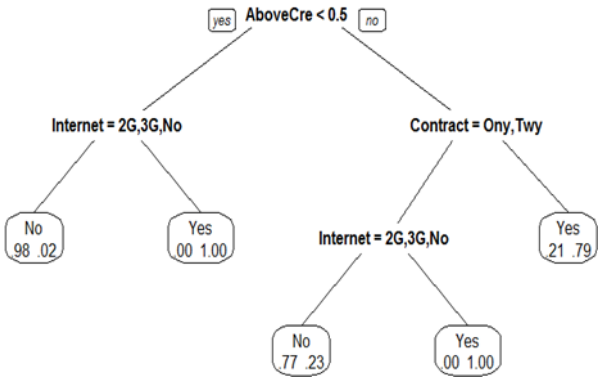

Fig 1: Decision tree model

Table 3 below shows results on how the model performed as compared to the real data. Our data was divided into two data set 'test data' and 'training data.' In the table 'Yes' indicate that the customer churned and 'No' indicate that the customer didn't churn, the model was executed using the 'Test Data' and below are the results. The model correctly predicted that 3335 would not churn and incorrectly predicted that 164 customers have churned. The model also correctly predicted that 1148 have churned and incorrectly predicted that 283 customers have not churned.

Table 3 Prediction using test data

	Actual Data	
Model Prediction	Non-Churned	Churned
Non-Churned	3335	164
Churned	283	1148

The misclassification errors for the model using test data are nine percent (9%).

The model was executed using the 'training data and table 4 below shows the results. The model correctly predicted that 1444 would not churn and incorrectly predicted that 73 customers have churned. The model also correctly predicted that 484 have churned and incorrectly predicted that 112 customers have not churned.

Table 4 Prediction using training data

	Actual Data	
Model Prediction	Non-Churned	Churned
Non-Churned	1444	73
Churned	112	484

The misclassification errors for the model using testing data are eight percent (8.7%).

C. Logistic Regression

Table 5 below shows the results from a Logistic Regression model. The model used four main attributes Churn, InternetService, AboveCreditLimit and Contract.

Our data was divided into two data set 'test data' and 'training data. 'Yes' or '1' indicate that the customer churned and 'No' or '0' indicate that the customer did not churn.

The model was executed using the 'training data'. The model correctly predicted that 1467 would not churn and incorrectly predicted that 96 customer have churned. The model also correctly predicted that 461 have churned and incorrectly predicted that 89 customers have not churned.

Table 5 Prediction using training data

	Actual Data	
Model Prediction	Non-Churned	Churned
Non-Churned	1444	73
Churned	112	484

The misclassification error for the model using test data is eight percent (8.7%).

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Models Comparison

Table 6 below shows the empirical analysis of the models. The model that performed better is Logistic Regression; the accuracy of the model is 91.3% with less the same execution time as decision tree.

Table 6 Empirical Analysis of the Models

Data Mining Technique	Accuracy	Mean Errors	Model Construction Time
Decision Tree	91%	9%	0.09 seconds
Logistic Regression	91.3%	8.7%	0.09 Seconds

V. Conclusion

Churning in the Telecommunication industry is unavoidable but can be managed, the results in this paper shows that non-payment is the major reason why service providers churn customers. The main contributing factor to non-payment is because the customers are allowed to go over their credit limit. Telecommunications companies must regularly adjust credit limits by checking if customers have not increased their debts. This paper used data mining analytics to develop models to predict customers that would churn. Decision tree and Logistic regression were used to predict churn and both of them did so successfully with logistic regression being more by 0.3%. Data mining techniques are relevant for the churn prediction.

References

[1] T. Li, P. Lu, Z. He, and Q. Wang, “A Customer Retention System Based on the Customer Intelligence for A Telecom Company,” *Proc. 9th Jt. Conf. Inf. Sci.*, 2006.

[2] J. Kawale, A. Pal, and J. Srivastava, “Churn prediction in MMORPGs: A social influence based approach,” in *Proceedings - 12th IEEE International Conference on Computational Science and Engineering, CSE 2009*, 2009.

[3] Y. Richter and N. Slonim, “Predicting customer churn in mobile networks through analysis of social groups,” pp. 732–741, 2010.

[4] D. M. Balasubramanian and M. Selvarani, “Churn Prediction in Mobile Telecom System Using Data Mining Techniques,” *Int. J. Sci. Res. Publ.*, vol. 4, no. 1, pp. 2250–3153, 2014.

[5] J. Lu, “Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS,” *Data Min. Tech.*, vol. 114–27, pp. 114–27, 2002.

[6] I. Brief, “Big Data Use Cases for Telcos Key Big Data Use Cases for Telcos,” pp. 1–4, 2014.

[7] I. Khan, I. Usman, T. Usman, G. U. Rehman, and and Ateeq Ur Rehman, “Intelligent Churn prediction for Telecommunication Industry,” *Int. J. Innov. Appl. Stud.*, vol. 4, no. 1, pp. 165–170, 2013.

[8] M. Constantinou, “Prepaid Churn Prediction,” 2014.

[9] J. Ahn, S. Han, and Y. Lee, “Customer churn analysis : Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry,” vol. 30, pp. 552–568, 2006.

[10] I. Brandusoiu and G. Toderean, “Churn prediction in the telecommunications sector using support vector machines,” *Ann. Oradea Univ.*, no. 1, pp. 19–22, 2013.

[11] L. Junxiang, “Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Modeling Using SASâ,” 2001.

[12] Y. Huang and T. Kechadi, “An effective hybrid learning system for telecommunication churn prediction,” *Expert Syst. Appl.*, vol. 40, no. 14, pp. 5635–5647, 2013.

[13] M. Kaur and P. Mahajan, “Churn Prediction in Telecom Industry Using R,” *Int. J. Eng. Tech. Res.*, vol. 3, no. 5, pp. 46–53, 2015.

[14] S. C. Wangberg and C. Psychol, “Personalized technology for supporting health behaviors,” in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, 2013, pp. 339–344.

[15] M. Fanti and W. Ukovich, “Discrete event systems models and methods for different problems in healthcare management,” in *IEEE Emerging Technology and Factory Automation (ETFA)*, 2014, pp. 1–8.

[16] a Boyd and M. Bastian, “What Is a Scoping Study?,” vol. 4, no. October, pp. 1–5, 2011.

[17] C. Kirui, L. Hong, W. Cheruiyot, H. Kirui, C. Engineering, and I. Technology, “Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining,” *IJCSI Int. J. Comput. Sci. Issues*, vol. 10, no. 2, pp. 165–172, 2013.

[18] E. Shaaban, Y. Helmy, A. Khedr, and M. Nasr, “A Proposed Churn Prediction Model,” *Int. J. Eng. Res. Appl.*, vol. 2, no. 4, pp. 693–697, 2012.

ISBN: 978-988-14047-5-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)