Review on Churn Rate Prediction Using Data Mining Techniques

Pallavi Ghanshani
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
pallavighanshani@gmail.com

Mr. Vaibhav Chandrakar
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
vaibhavchandrakar@gmail.com

Abstract

In this competitive world, business is winding up profoundly saturated. Particularly, the field of telecommunication faces complex difficulties because of various energetic focused service providers. Along these lines, it has turned out to be exceptionally troublesome for them to hold existing customers. Since the cost of securing new customers is substantially higher than the cost of holding the current customers, it is the time for the telecom ventures to find a way to hold the customers to balance out their fairly estimated worth. In the previous decade, a few information mining systems have been proposed in the writing for anticipating the churners utilizing heterogeneous customer records. This paper surveys the distinctive classes of customer information accessible in open datasets, predictive models and performance metrics utilized as a part of the writing for churn prediction in the telecom business.

Keywords— *Data mining, Customer churn prediction, Predictive models, and Performance metrics.*

I. INTRODUCTION

Today is the focused universe of communication advances. Customer Churn is the real issue that all the Telecommunication Industries on the planet faces now. In telecommunication worldview, Churn is characterized to be the movement of customers leaving the organization and disposing of the administrations offered by it because of the disappointment of the administrations as well as because of the better offering from other network suppliers inside the reasonable sticker price of the customer. This prompts a potential loss of income/profit to the organization. Additionally, it has turned into a testing undertaking to hold the customers. Along these lines, companies are going behind presenting another best in class applications and innovations to offer their customers however much better administrations as could reasonably be expected in order to hold them in place. Before doing as such, it is important to distinguish those customers who are probably going to leave the organization sooner rather than later ahead of time on the grounds that losing them would bring about noteworthy loss of profit for the organization. This procedure is called Churn Prediction.

Information mining strategies are observed to be more viable in anticipating customer churn from the explores completed amid a previous couple of years. The development of compelling churn prediction display is a huge assignment which includes bunches of research ideal from the ID of ideal indicator factors (features) from the huge volume of accessible customer information to the choice of successful prescient information mining method that is reasonable for the list of capabilities. Telecom Industries gather a voluminous measure of information seeing customers, for example, Customer Profiling, Calling design, Democratic information notwithstanding the network information that is produced by them. In light of the historical backdrop of the customers calling design and the conduct, there is a probability to recognize their outlook of possibly they will leave or not. Information mining procedures are observed to be more powerful in churn prediction from the specialists did for as long as one decade. Particularly Predictive demonstrating methods are often observed to be more precise in churn prediction.

II. DATA MINING

Information mining goes under the procedure of KDD (Knowledge Discovery in Database). It is utilized to extricate helpful learning as examples from the distinctive web sources, for example, databases, records and so forth. These days information mining instruments are be utilized to answer questions business that was prior to time-consuming and hard to reply. The systems of information examination and the devices that assistance in the extraction of intriguing shrouded designs assume an indispensable part of the decision

making process. The model has six stages which is appeared in the figure 1.

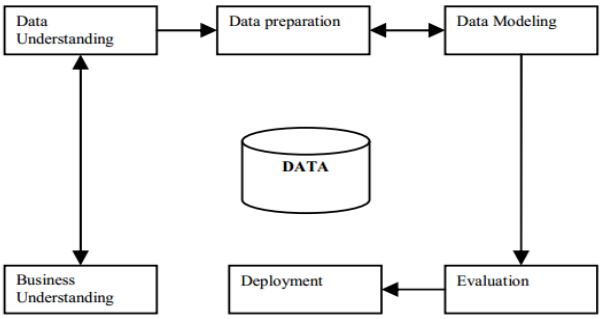

Fig. 1. Data Mining Model

In this paper, we survey the current deals with churn prediction in three alternate points of view: datasets, methods, and metrics. Initially, we display the insights about the accessibility of open datasets and what sorts of customer subtle elements are accessible in each dataset for anticipating customer churn. Also, we thoroughly analyze the different prescient demonstrating techniques that have been utilized as a part of the writing for foreseeing the churners utilizing distinctive classifications of customer records, and afterward quantitatively look at their exhibitions. At last, we condense what sorts of execution measurements have been utilized to assess the current churn prediction techniques. Analyzing all these three points of view are extremely critical for building up a more effective churn prediction system for telecom businesses.

While there are other churn prediction reviews accessible in the writing, they fundamentally centered on various demonstrating procedures. To the best of our insight, none of those overviews assessed the datasets and measurements for assessing the churn prediction models. Thus, we trust that this study can give a guide to the two researcher and customer relationship supervisors to better comprehend the area and difficulties in detail.

III. HADOOP

Hadoop is a free open source stage, which helps in putting away data and parallel handling in a distributed domain. Hadoop parts the expansive database into pieces of data and disperses over the clusters in the distributed condition. To process the data, MapReduce is utilized for parallel processing on the clusters, in this way lessening the execution time.

The Hadoop Distributed File System (HDFS) is fundamentally a distributed document framework which is intended to keep running on item equipment. It is numerous like the current distributed record frameworks. Be that as it may, there are a few contrasts amongst HDFS and other distributed document frameworks which makes it noteworthy. HDFS is exceedingly fault tolerant and is planned such that it can be conveyed on minimal effort equipment. HDFS likewise gives high throughput access to application data and is exceptionally appropriate for applications that have substantial data sets.

Figure 2 demonstrates the HDFS master/slave architecture. A HDFS cluster comprises of two sections viz., a single NameNode and more than one DataNode. NameNode is a master server that directs access to records by clients and deals with the record framework namespace. There are various DataNodes in HDFS, generally one for every node in the cluster. The DataNode deals with the capacity which is connected to the nodes that they are running on. HDFS uncovered a record framework namespace and enables the client data to be stored in documents. Inside in a HDFS, a document is part into at least one pieces and these blocks are then put away in an arrangement of DataNodes.

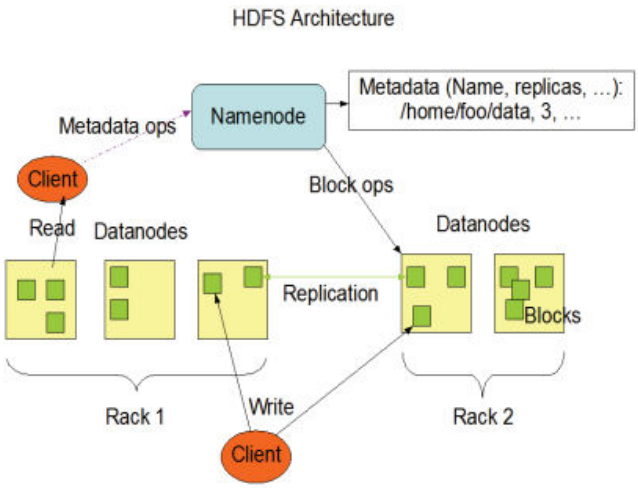

Fig. 2. Architecture of HDFS

The NameNode is likewise used to execute document framework namespace activities which incorporate opening a document, closing a record and renaming records and catalogs in the HDFS. It additionally plays out the mapping of pieces of data to the DataNodes. On the client's side, the DataNodes are in charge of serving the read and compose demands from the HDFS. The DataNodes likewise perform tasks, for example, block

creation, deletion, and replication upon the direction gave from the NameNode.

IV. LITERATURE SURVEY

Adem Karahoca et al. [1], Churn management is essential and basic issue for Global Services of Mobile Communications (GSM) administrators to create procedures and strategies to keep its endorsers of pass other GSM administrators. In the first place period of churn management begins with profile creation for the endorsers. Profiling process assesses call detail information, money related data, calls to customer benefit, contract points of interest, showcase subtle elements and geographic and populace information of a given state. In this examination, input features are clustered by x-means and fuzzy c-means clustering algorithms to put the supporters into various discrete classes. Adaptive Neuro Fuzzy Inference System (ANFIS) is executed to build up a delicate prediction demonstrate for churn management by utilizing these classes.

Clement Kirui et al. [2] Customer churn in the mobile communication industry is a persistent issue attributable to firm rivalry, new innovations, low switching costs, deregulation by governments, among different components. To address this issue, players in this industry must create exact and solid prescient models to distinguish the conceivable churners in advance and after that enroll them to intercession programs in an offer to hold whatever number customers as could be expected under the circumstances. This paper proposes another arrangement of features with the point of enhancing the recognition rates of conceivable churners. The features are gotten from call points of interest and customer profiles and sorted as contract-related, call design depiction, and call design changes portrayal features. The features are assessed utilizing two probabilistic information mining algorithms Naïve Bayes and Bayesian Network, and their outcomes contrasted with those acquired from utilizing C4.5 decision tree, a broadly utilized algorithm in numerous arrangement and prediction errands. Test comes about show enhanced prediction rates for every one of the models utilized.

Ballings, Michel et al. [3], the key inquiry of this examination is: How long should customer occasion history be for customer churn prediction? While most investigations in prescient churn displaying mean to enhance models by information increase or algorithm change, this examination centers around another measurement: time window improvement as for prescient execution. This paper first displays a formalization of the time window choice technique, alongside a writing survey. Next, utilizing logistic regression, characterization trees and packing in blend with arrangement trees, this investigation breaks down the change in churn demonstrate execution by broadening customer occasion history from one to sixteen years. The outcomes demonstrate that, after the fifth extra year, prescient execution is just hardly expanded, implying that the organization in this investigation can dispose of 69% of its information with no abatement in prescient execution.

Ismail et al. [4], Nowadays, the telecommunication ventures are confronting generous rivalry among the suppliers keeping in mind the end goal to catch new customers. Numerous suppliers have confronted lost profitability because of the current customers relocating to different suppliers. Customer maintenance program is one of the primary procedures received so as to keep customers faithful to their supplier. In any case, it requires a high cost and in this way the best technique that companies could rehearse is to center around distinguishing the customers that can possibly churn at a beginning period. The restricted measure of research on examining customer churn utilizing machine learning procedures has lead this exploration to investigate the capability of a counterfeit neural network to enhance customer churn prediction. The exploration proposes Multilayer Perceptron (MLP) neural network way to deal with foresee customer churn in one of the main Malaysian's telecommunication companies. The outcomes are thought about against the most famous churn prediction procedures, for example, Multiple Regression Analysis and Logistic Regression Analysis. The outcome has demonstrated the amazingness of neural network (91.28% of prediction exactness) over the measurable models in prediction undertakings.

H Lee et al. [5], in an exceptionally focused mobile telecommunication business condition, showcasing supervisors require a business knowledge display that enables them to keep up an ideal (no less than a close ideal) level of churners adequately and effectively while limiting the expenses all through their advertising programs. As an initial move toward ideal churn management program for showcasing administrators, this paper centers around building an exact and compact

prescient model with the end goal of churn prediction using a Partial Least Square (PLS)- construct system in light of exceptionally corresponded informational indexes among factors. A preparatory trial exhibits that the displayed show gives more precise execution than customary prediction models and recognizes key factors to better comprehend churning practices.

Anuj Sharma et al. [6], Marketing writing states that it is more exorbitant to draw in another customer than to hold a current steadfast customer. Churn prediction models are created by scholastics and specialists to adequately oversee and control customer churn keeping in mind the end goal to hold existing customers. As churn management is an imperative action for companies to hold faithful customers, the capacity to accurately anticipate customer churn is fundamental. As the cell network administrations advertise winding up more focused, customer churn management has turned into a vital errand for mobile communication administrators. This paper proposes a neural network (NN) based way to deal with foresee customer churn in membership of cell remote administrations. The aftereffects of tests show that neural network based approach can foresee customer churn with precision over 92%.

Abbas Keramati et al. [7], To get by in the testing condition of a global market, associations must perceive and examine customer states of mind. To be aggressive, associations must perceive and gauge customer inclinations and practices to expand customer maintenance before their adversaries do as such. This exploration distinguishes factors that influence customer churn, the absolute most significant of an association's advantages. One year's information from call log documents identifying with 3150 customers were chosen haphazardly from an Iranian mobile administrator call-focus database. Binomial Logistic Regression was the strategy for examination utilized as a part of this exploration. The consequences of this exploration show that a customer's disappointment, their measure of administration use and certain statistic qualities have the most impact on their decision to remain or churn. The outcomes likewise infer that customer status (dynamic or latent status) intercedes the connection amongst churn and the reason for churn. The Iranian government's present intend to privatize the telecommunications business without deregulation prompts a non-square rivalry condition. Deregulation for designating more experts of customer mind is fundamental with a specific end goal to build up a square private rivalry condition in the Iranian mobile telecommunications industry.

Kristof Coussement et al. [8], Nowadays, companies are putting resources into an all-around thought about CRM procedure. One of the foundations in CRM is customer churn prediction, where one tries to foresee regardless of whether a customer will leave the organization. This examination centers around how to better help promoting decision producers in distinguishing hazardous customers by utilizing Generalized Additive Models (GAM). Contrasted with Logistic Regression, GAM unwinds the linearity requirement which takes into consideration complex non-direct fits to the information.

Marcin Owczarczuk [9], In this article, we test the value of the mainstream information mining models to anticipate churn of the customers of the Polish cell telecommunication organization. When contrasting with past investigations on this subject, our exploration is novel in the accompanying regions: (1) we manage prepaid customers (past examinations managed postpaid customers) who are significantly more prone to churn, are less steady and considerably less is thought about them (no application, demographical or individual information), (2) we have 1381 potential factors got from the customers' use (past investigations managed information with no less than many factors) and (3) we test the strength of models crosswise over time for every one of the percentiles of the lift bend – our test is gathered a half year after the estimation of the model.

Umayaparvathi et al. [10], In this competitive world, business is winding up exceptionally immersed. Particularly, the field of telecommunication faces complex difficulties because of various dynamic focused specialist organizations. In this way, it has turned out to be exceptionally troublesome for them to hold existing customers. Since the cost of getting new customers is substantially higher than the cost of holding the current customers, it is the time for the telecom enterprises to find a way to hold the customers to balance out their fairly estimated worth. This paper investigates the use of information mining methods in foreseeing the possible churners and quality determination on distinguishing the churn. It likewise looks at the proficiency of a few classifiers and records their exhibitions for two genuine telecom datasets.

Table I. Shows comparisons of existing methods and its features

Author	Dataset	Features	Methods	Metrics
Adem Karahoca [1]	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 [2]	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 [3]	Unknown 129,892 customers 113 attributes	Demographic, Value added, usage pattern	Logistic regression, Bagging, Decision Tree	AUC
Ismail, Mohammad [4]	Unknown, 169 customers 10 attributes	Demographic, Billing data, usage pattern, customer relationship	Neural network, Regression	Confusion matrix, accuracy, precision, recall
H Lee [5]	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 [6]	ML Dataset at UCI 2,427 customers 20 attributes	Demographics, Usage pattern, Value added services	Artificial Neural Network	Confusion matrix
Abbas Keramati [7]	Iranian telco operator 3150 customers 15 attributes	Demographic, call usage pattern, customer care service	Binomial logistic regression model	Statistical hypothesis test
Kristof Coussement [8]	Belgian 134, 120 customers 27 attributes	Demographic Usage patter, bill and payment	generalized additive models (GAM)	AUC top-decile lift
Marcin Owczarczuk [9]	Polish mobile operator 122098 customers 1381 attributes	Demographic, call data records, customer care services	Logistic regression Decision tree	Lift curves
Umayaparvathi [10]	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

V. TOOLS USED

There are many tools available for processing data and extracting features from telecom dataset. Some of them are presented below.

a. Hadoop Mapper Tool
b. Hadoop Reducer Tool
c. Hadoop Distributed File System
d. Hadoop Clustering Tool

VI. CONCLUSION

Telecommunication industry has experienced high churn rates and gigantic churning misfortune. In spite of the fact that the business misfortune is unavoidable, yet at the same time churn can be overseen and kept at a satisfactory level. Great techniques should be produced and existing strategies must be improved to keep the telecommunication business to face challenges. This paper reviewed the diverse classifications of customer information accessible in open datasets, predictive models and performance metrics utilized as a part of the writing for churn prediction in the telecom business.

REFERENCES

[1] 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.

[2] 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.

[3] 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

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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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).