Exploring Customer Churn Prediction

Nikhitha C
Dept of Computer Science
Kalam Technological University
RSET
Email: nikhithanair07@gmail.com

Neenu Jose
Dept of Computer Science
Kalam Technological University
RSET
Email:neenujose20@gmail.com

***Abstract*—Simply put, Customer attrition or customer churn occurs when a customer leaves from a company and becoming the customer of another opposing company. Keeping of existing customers which is more better than procure of new customers. Customer churn obstructs growth of a company. Because of this reason calculating customer churn is an important factor in business analytics. By being aware of and observing churn rate, business are fitted to determine their customer detention success rates and identify some approaches for improvement. Generally churn management system identifies the probability of customers who will be churning in the future. This paper explains the survey done in this area including different papers from late 90s to the current scenario. It describes five different categories and each categories includes different methods for churn prediction.**

***Index Terms*—Customer,Customer Churn,Churn Management System**

I. Introduction

Customer Churn or Customer attrition, in which migrating of a customer from one company to another competing company. In recent years customer churn is increasing because of many reasons. Two types of customer churn are there. First one is voluntary, and next one is involuntary. Voluntary customers who accidentally or knowingly leaves from the company. Next category is involuntary in which because of misuse of services business itself cuts off services from them.[1]

Customer churn rate in a month can be calculated using the following equation

(No of active customers at the start of the month-No of active customers at the end of the month)/ No of active customers at the start of the month

The ideal churn value is five percentage. If the churn rate is high then the business in unsustainable state. Next fact is that acquiring of new customers more expensive, than the customers you already have. Current study argue that if holding of customers increases then it also increases the profit of a company. So There is a great role in retention of existing customers. Holding on to obedient customers rather than taking of new customers can lower marketing costs and increase your profit per customer. Getting a new customer from five to twenty five times will be more costly than reserving an current one. Consider the research done in ”Bain Company” , that shows rising customer retention rates by five percentage, It increases profits by twenty five percentage to ninety five percentage.

Generally Churn management systems identifies the probability of customers who will be churning in the future. The main intention of this paper is to study different churn prediction models that are existing now.

A. *Sources of Customer Churn*

There are different reasons that can lead churning of customers from a company. Poor customer service is the main factor. It includes Something Happened to the Customer, Bad Onboarding, Lack of Ongoing Customer Success, Bad Support, low-quality communications etc.

B. *Customer Churn Disadvantages*

Chances of a growing company reduces if churn rate increases. Customer attrition is an expensive for a business even if the company have good marketing campaigns. Because company spends more than 7 times in acquiring new customers than retaining of existing customers.

C. *Customer Churn Reducing Strategies*

It is absolutely clear that customer churn is not good for companies. Because of this reason organizations should implement some churn management systems for identifying the customers who will be churning in the future. It is required for a company to exactly know their customers, this is the best method for avoiding the customer churn problem.

Customer Relationship Management

Customer Relationship Management is an advancing concept to crystallize with customers and at the same time decreasing the cost and improving productivity and utility. For all business a CRM provides a very good platform to communicate with their customers and meet all their needs. A CRM contains entire knowledge of each individual customer. The main advantage of CRM is it can be used in small as well as large business and it is vast and significant.

TABLE I
CLUSTERING METHODS

Method	Publication Year	Dataset	Advantages	Disadvantages
[21]	2017	Bank Data	High performance	Complexity high
[22]	2017	Patient churn data	K-Means - Simple K-Medoids - Simple Fuzzy-c means - Performance high DBSCAN - used to find arbitrary size and arbitrarily shaped cluster	K-Means - Random choice of initial centroids K-Medoids - If number of clusters increases complexity also increases Fuzzy c means - Complexity is high DBSCAN - neck type of data not supported

II. BACKGROUND STUDY

Some of the works referred during this literature survey has been explained below.

A. *Clustering Methods*

Clustering is one of the method used for churn prediction. The process of distinguishing homogeneous class of data from a data set called clustering. Things in each cluster which have similar properties compared to other cluster.

In bank customer churn prediction is an important one. Customer churn prediction which focuses on banking industry using improved fuzzy C-means clustering algorithm. C-means clustering method is divided into C-clear means clustering method and C-fuzzy means clustering method (FCM).[21] Improved fuzzy C-Means algorithm has high accuracy.

Franciska, [22] suggests different clustering algorithms for customer churn prediction. K-means clustering, k-medoids clustering, Hierarchical clustering, DBSCAN and Fuzzy c means clustering are used to analyze data to predict type of class for customer churn.Pre processing step is required for k-means clustering. K-medoids in which the iteration depends on the number of cluster.Hierarchical clustering takes more time to cluster the data. DBSCAN uses cosine and Euclidean distance functions for distance function.In fuzzy c means we have to define the number of iterations and cluster.

B. *Sampling, Pre-processing, Imbalance Problem based churn prediction*

Sampling is the method in which electing of sample elements from a community. Next is Pre-processing, that contains some initial steps for increasing performance. and last is imbalance data problem which comes when class distribution is dissimilar.

Traditional churn management models focusing on prediction accuracy, but not considering other parameters. Mineh Ghorbani, [1] proposed a framework is CMF, that covers almost demands of a churn prediction model. CMF is a seven step process model that are Churn Causes Analysis,Finding appropriate data,Pre-processing,Feature Selection,Prediction algorithm,Model,Retention Campaign. In which the important blocks are prediction algorithm and evaluation of prediction algorithm. Final Step Company has to do some retention activities, which depends on type of the company and financial status etc.

During customer churn prediction one of the problems is imbalanced data. Peng Li, [8] suggested LRM (logistic regression model) is a solution for this imbalanced data. Logistic regression is a predictive analysis used to describe data, relationship between dependent and independent variable. Stratified sampling in which some examples randomly extract from both positive (not churn) and negative (churn) and merging them for parameter calculation.K-means clustering could be used for this balanced data. This will give result in set of classes with low churn, medium churn, and high churn.

This churn prediction model uses boosting as a method to improve the accuracy. One of the boosting algorithms is AdaBoost algorithm. Good separation of churn data can be achieved by Boosting. [9]Chi-square automatic interaction detection (CHAID) used to finding relevant variables related to customer churn.

Important blocks in the customer churn prediction are feature selection and feature extraction. A customer management system is good if and only if these blocks are good. So feature selection based on pruning technique known as orientation ordering pruning Method (OOPM).[13] Feature extraction based on Random Forest and Transduction is intended to select various features from customer data.These methods give better results when compared to traditional methods.

Customer churn prediction strength is an important criterion in churn prediction. combination of SMOTE and AdaBoost techniques are used for improving strength of churn prediction. Sampling techniques are also used to solve imbalance data problem. AdaBoost is one of the boosting algorithms to improve accuracy.[16] SMOTE is a combined process of oversampling and under sampling technique which solves the problem of imbalanced data.

Customer churn prediction in telecommunication industry can be done by collecting information from telecommunication industries. The framework that consist here based on knowledge discovery data(KDD). This method contains Data Acquisition, Data Preparation, Data Pre-processing, Data Extraction, Decision. WEKA tool is used for implementation. Dahiya, [10] suggested Decision Tree and Logistic Regression can be used for building prediction model. Results indicates that decision tree have high accuracy than logistic regression.

Supervised massive data analysis mainly focuses on three problems. First is imbalanced data problem. Second, the samples in feature space are relatively scattering. Third, the dimension of feature space is high. To pre-process the imbalanced dataset supervised one-side sampling technique is used.[14] For selecting relevant variables random forest method can be used. Finally main classifier is C5.0 decision tree for better performance.

Customer Churn prediction is important in SaaS company. Performance of random forests, Principal Component Analysis (PCA), and Extreme Gradient Boosted (XGBoost) Trees compared with the conventional classification algorithm, logistic regression, to identify the risk of customer churn.[18]XGBoostis built upon the principle of gradient boosting.

In manufacturing industry it produces the products like automobiles, engines etc. These industries in which churn prediction that rarely occurs because of the unavailability of data. Customer churn prediction is very crucial in discrete manufacturing industry.Rama Mohan Rao Dintakurthi, [17] proposed a Decision support system model consist of Data understanding, Data preparation, Data harmonization, Data enrichment. Finding the variables affecting the customer churn is the main step for this model.

C. *Machine Learning and Ensemble Methods*

Ensemble methods comes under machine learning in which combine different methods to increase the accuracy.

Customer Churn prediction have an important role in business analytics. Customer churn is the processes in which customers who have an aim to leave from the company and becoming the customer of another competing company. Accuracy and count of churn customers is small these are the two characteristics of customer churn prediction problem. Yu Zhao, [2] proposed Kernel trick method is used to transform data and finds an optimal bounds between outputs based on these transformations. SVM classifier must be week if the number of churn customers is small. Improved one class SVM which overcomes this problem. Accuracy is also high by using this technique.

TABLE II
SAMPLING, PRE-PROCESSING, IMBALANCE PROBLEM BASED CHURN PREDICTION

Method	**Publication Year**	**Dataset**	**Advantages**	**Disadvantages**
[1]	2009	Telecommunication company	Covers all requirements of churn management system	Data selection phase is poor
[8]	2014	Orange Data	Stable promotion effect.	Imbalance data problem
[9]	2014	Telecommunication Company	Good separation of data	Performance is less
[13]	2017	Telecommunication Company	Removes irrelevant information	Based on distribution information of test samples the application needs are fulfilled
[16]	2016	B2C Ecommerce	Low Cost Accuracy high	Class rarity does not considered
[10]	2015	Telecommunication industry	accurate result	Retention policies not considered
[14]	2016	Telecommunication industry	solves unbalanced and scatter problem	high dimensional problem
[18]	2017	Telecommunication industry	Low Cost	Complexity high
[17]	2016	Discrete manufacturing industry	Interactive tool for better usage	Classification is poor

Acquiring new customers is more complex process when compared with retaining of existing customers. Yaya Xie, [5] suggested IBRF(Improved Balanced Random Forests) which combines both balanced random forests and weighted random forests. Sampling process used in balanced random forests which is more efficient and cost sensitive learning employed in weighted random forests. This IBRF has better accuracy than traditional random forests.

If the class distribution of client's data is imbalanced,then it degrades the performance of an entire system. Customer attrition prediction problem is crucial in business.Uan Wang an [23], suggested TEMID (Transfer Ensemble Model for Imbalanced Data) that integrates both transfer learning and multiple classifiers ensemble. Transfer Learning Method which balances the data , and multiple classifier ensemble method implements the classification. It solves the imbalanced data problem hence this method have high performance compared to other methods.

Churn prediction is used to find the probability of customers who will be churning in the future. Rough customer data decreases performance of the system. To solve this problem Zhao Xin, [4] proposed a prediction model this based on both PCA (Principal Component Analysis) and LS-SVM (Least Square Support Vector Machine). PCA is a statistical tool used for reduction of crude data from different sources hence can extract relevant information. LS-SVM used for customer churn identification. SVM can remove the problems of LS-SVM

Customer attrition in a bank which is high in recent years. Bank data consist of large amount of data and it is also imbalanced. Zhao Jing, [3] suggested a method support vector machine and this method compared to traditional methods. Result indicates that performance is greater in SVM compared to other methods.

Oumya Banerjee, [7] suggested Decision support system, that can be applied to telecommunication data, based on finding the relationship between attributes both the customer and the mobile network of the customer to help make decisions. Relationship between attributes can be computed using entropy computation and probability of occurrence. Entropy can be calculated using Shannons entropy distribution. The relationship between attributes gives the probability of customers who will churn in the future.

Customers churn prediction in which the customer who leaves from the company and becoming the customer of another competing company. Four algorithms suggested by Guo-en Xia, [15] are Bayes, Decision Tree (DT), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) used for churn prediction. Compared to DT, Bayes, ANN, SVM are better methods for prediction.

Customer churn in telecommunication industries have a cost sensitive problem. Partition cost-sensitive CART model solves this problem. It reduces the total mis-classification costs effectively through which it increases the performance of entire model. Huanqi Wang [19] introduces a new technique CPA that can distinguish the mis-classification costs. This model has high performance.

In a fitness industry customer behaviour prediction is an important one. Series of experiments are using here for increasing gym utilization and reducing acquiring new customers. Jas Semrl [20] suggested two platforms such as AzureML and BigML were evaluated here. Data manipulation can be done by AzureML and data generation and filtering can be done by BigML. In comparison BigML is the simple method, AzureML contribute many tools for data manipulation.

TABLE III
MACHINE LEARNING AND ENSEMBLE METHODS

Method	Publication Year	Dataset	Advantages	Disadvantages
[2]	2005	Telecommunication company	Performance high	High dimensionality problem
[27]	2008	Bank data	Scalability	Internal time in-varying variables only used
[23]	2011	UCI California University.	Solves imbalance data problem	High Complexity
[4]	2009	Business bank in China.	Removes latent noise	Complexity
[3]	2008	Bank data	Best accuracy, hit rate, covering rate, lift coefficient	Noise distribution
[7]	2013	Telecommunication company	Proper utilization of resources	Feature suggestion is poor
[15]	2016	Telecommunication company,US	Effective	Classifier is not good
[19]	2017	Indian telecommunication sectors	High accuracy	Not scalable
[20]	2017	Fitness industry	Increases performance	Feature selection poor

D. *Meta-heuristic Methods*

It is a high-level method in which by inducing or finding a heuristic, Meta-heuristic Method gives better result.

Customer Churn is the movement of customers from one company to another competing company. Traditional prediction model consist of two approaches C4.5 and RIPPER. These two techniques used as benchmark to proposed methods [6] AntMiner+ and ALBA. AntMiner+ is a classification technique based on ant colony optimization. ALBA is a rule extraction algorithm. Input data collected from different sources, hence chi-squared based filter could be used for reduction of crude data.ALBA can be combining with either C4.5 or RIPPER for improving accuracy.

Customer migration which is very high in telecommunication area in recent years. Because of large data in telecommunication enterprises, the complexity increases. Another reason for complexity is imbalanced data problem. Because of the large data, T. Sumathi, [24] suggested Meta heuristic based attrition prediction technique is used. Here PSO (Particle Swarm Optimization) algorithm used here for customer churn prediction.

In a business customer retention is an important one when compared to acquiring new customers. Customer churn prediction is important in telecommunication industry.

TABLE IV
META-HEURISTIC METHODS

Method	Publication Year	Dataset	Advantages	Disadvantages
[6]	2010	Telecom data	High performance	Complexity high
[24]	2016	Orange Dataset	High accuracy	High number of false positives
[11]	2015	Telecom data	Performance good	Poor feature selection
[12]	2015	Telecom data	Better dimensionality reduction	Unbalanced data problem

Ramakanta Mohanty [11] suggested four techniques are Counter Propagation Neural Networks (CPNN), Classification and Regression Trees (CART), J48 and fuzzy ARTMAP to predict customer churn and non-churn. CPNN is a hybrid network. It is guaranteed to find the correct weights. CART and J48 are used to construct decision tree. From these four techniques CART provides better performance.

Acquiring new customers is more expensive than retaining of existing customers. Adnan Amin, [12] suggested different algorithms Exhaustive, Genetic, Covering, and LEM2 implemented after that compare the performance of these methods. These four methods used for rule generation. Based on these methods, the decision maker/manager can easily design more suitable strategic plan to retain the churn.

E. Hybrid Churn Prediction Methods

Virtual Word is an environment in which people can interact with each other and they are represented by animated characters. Hsiu-Yu, [25] developed a hybrid classification model for customer attrition prediction in virtual word. Hybrid classification model is the combination of monetary cost, user behavior and social neighbor features. The main advantage of this method is it takes less time for prediction of migrating customers.

Chih-Fong Tsaib et al, [26] suggested combination of back-propagation artificial neural networks (ANN) and self-organizing maps (SOM) which turned into a hybrid neural network model.This method have high performance ,since it have high accuracy.

III. CONCLUSION

Customer churn prediction is a relevant factor in business analytics. Customer churn in which customers leaving from one company and becoming the customer of another competing company.Churn management systems identifies the probability of the customers who will be churning in the future. According to this probability company will do some retention activities.This retention activities depends on the type of company, Financial status of company etc. This Literature survey describes various techniques for churn prediction. In this paper we have made a detailed comparison study on different methods used for customer churn prediction. The analysis done has been tabulated.

TABLE V
HYBRID CHURN PREDICTION METHODS

Method	Publication Year	Dataset	Advantages	Disadvantages
[25]	2015	Roomi dataset	Less time, Accuracy high	Multi-objective problem
[26]	2009	Telecom data, US	Stability high	No dimensionality reduction

REFERENCES

[1] Amineh Ghorbani and Fattaneh Taghiyareh,”CMF: A Framework to Improve the Management of Customer Churn”, *IEEE Asia-Pacific Services Computing Conference (IEEE APSCC)*,2009

[2] Yu Zhao, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren,”Customer Churn Prediction Using Improved One-Class Support Vector Machine”, *Springer-Verlag Berlin Heidelberg 2005*

[3] Zhao Jing and Dang Xing-hua, ”Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Banks VIP Customer Churn as the Example”, *2008 IEEE*

[4] Zhao Xin , Wang Yi and chahongwang,”A New Prediction Model of Customer Churn Based on PCA Analysis”,*2009 IEEE*

[5] Yaya Xie, Xiu Li, E.W.T. Ngai and Weiyun Ying,”Customer churn prediction using improved balanced random forests”, *2008 Elsevier*

[6] Wouter Verbeke, David Martens, Christophe Mues and Bart Baesens ,”Building comprehensible customer churn prediction models with advanced rule induction techniques”, *2010 Elsevier*

[7] Soumya Banerjee ,Nashwa EI-Bendary, Aboul Ella Hassanien and M.F.ToIba,”Decision Support System for Customer Chum Reduction Approach ”, *2013 IEEE*

[8] Peng Li, Siben Li, Tingting Bi, Yang Liu,”Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression”, *2014 International Conference on Software Intelligence Technologies and Applications*

[9] Ning Lu, Hua Lin, Jie Lu, and Guangquan Zhang,”A Customer Churn Prediction Model in Telecom Industry Using Boosting”, *2014 IEEE*

[10] Dahiya and Surbhi Bhatia, ”Customer Churn Analysis in Telecom Industry”, *2015 IEEE*

[11] Ramakanta Mohanty and Jhansi Rani K,”Application of Computational Intelligence to predict churn and non-churn of customers in Indian Telecommunication”, *2015 IEEE*

[12] Adnan Amin, Changez Khan, Sajid Anwar, ”Churn Prediction in Telecommunication Industry Using Rough Set Approach”, *January 2015*

[13] Qiu Yihui and Zhang Chiyu,”Research of Indicator System in Customer Churn Prediction for Telecom Industry”, *The 11th International Conference on Computer Science Education (ICCSE 2016) August 23-25, 2016. Nagoya University, Japan*

[14] Hui LI, Deliang YANG, Lingling YANG, Yao LU and Xiaola LIN,”Supervised Massive Data Analysis for Telecommunication Customer Churn Prediction”, *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)*

[15] Guo-en Xia, Hui Wang, and Yilin Jiang,”Application of Customer Churn Prediction Based on Weighted Selective Ensembles”, *The 2016 3rd International Conference on Systems and Informatics (ICSAI 2016)*

[16] Xiaojun Wu and Sufang Meng,”E-commerce Customer Churn Prediction Based on Improved SMOTE and AdaBoost”, *2016 IEEE*

[17] Rama Mohan Rao Dintakurthi, Balaji Venkatraman, Poornachandran Mahendran, and Sheela Siddappa,”Decision support system for identifying customer churn based on buying patterns in a discrete manufacturing industry”, *2016 IEEE*

[18] Yizhe Ge, Shan He, Jingyue Xiong, and Donald E. Brown,”Customer Churn Analysis for a Software-as-a-service Company”, *2017 IEEE*

[19] Chuanqi Wang, Ruiqi Li, Peng Wang, and Zonghai Chen,”Partition cost-sensitive CART based on customer value for Telecom customer churn prediction” ,*Proceedings of the 36th Chinese Control Conference July 26-28, 2017, Dalian, China*

[20] Jas Semrl and Alexandru Matei, ” Churn Prediction Model for Effective Gym Customer Retention”, *2017 IEEE*

[21] Shaoying Cui and Ning Ding,”Customer Churn Prediction Using Improved FCM Algorithm”, *2017 3rd International Conference on Information Management*

[22] Franciska and Swaminathan, ”Churn Prediction analysis Using Various Clustering Algorithms in KNIME Analytics Platform ”, *2017 IEEE 3 rd International Conference on Sensing, Signal Processing and Security (ICSSS)*

[23] Yuan Wang and Jin Xiao , ”Transfer ensemble model for customer c hurn prediction with imbalanced class distribution”, *2011 IEEE*

[24] T. Sumathi, Churn Prediction on Huge Sparse Telecom Data Using Meta-heuristic, *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol.5, no.7, pp.574-577, 2016*

[25] Hsiu-Yu Liao, Kuan-Yu Chen, Duen-Ren Liu, and Yi- Ling Chiu. ”Customer Churn Prediction in Virtual Worlds.,*” In 2015 IIAI 4th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 115-120.*

[26] Chih-Fong Tsai and Yu-Hsin Lu. ”Customer churn prediction by hybrid neural networks.” Expert Systems with Applications, vol. 36, no. 10, pp. 12547-12553, 2009.