SURVEY ON CUSTOMER CHURN PREDICTION TECHNIQUES

Akshara Santharam[1], Siva Bala Krishnan[2]

[1]Department of Computer Science and Engineering, Student, SRM Institute of Science and Technology, Kancheepuram District, Chennai, Tamil Nadu, India
[2]Department of Computer Science and Engineering, Global Institute of Engineering and Technology, Melvisharam, Vellore, Tamil Nadu, India

---

Abstract—*Customer churn is a critical and challenging problem to various business and industries and in particular affecting the rapidly growing telecommunication sector. With the increase in competitors and innovative business models, the cost of customer acquisition has been increased. It is therefore important for any service providers to perform churn prediction. This paper reviews different techniques used not only in communication sector but also in other sectors where customer participation is highly active.*

***Key Words*: Churn Prediction, Churn Retention, Customer Relation Management, Datasets, Attributes, Churn Prediction Models**

1. INTRODUCTION

Customer churn is one of the mounting issues in any industries. The focus of the industries has been shifted to retaining the customer than acquiring new customer [1]. Studies had revealed that gaining new customers is 5 to 10 times costlier than retaining the existing customer by making them happy and satisfied and an average company loses 10 to 30 percent of customer annually [2].

Customer retention is one of the main objectives of Customer Relation Management and customer churn occurs when the customer terminates from a company or a service. The major cause of churn is due to the dis-satisfaction in the service by the provider. However, there are many other factors which cause customer churn and it also varies for each customer.

Many companies employ many techniques to predict customer churn and develops many strategies to retain customers for a longer tenure. Customer churn prediction has become the number one business goal. In general, many data mining techniques and machine learning techniques are employed to perform customer churn prediction with better accuracy.

N. Kamaraj and A. Malathi presented a review on customer churn prediction using data mining, machine learning and neural networks. To take forward our research this proposed analysis form the base of the upcoming research.

1.1 Types of Churners

Churners are classified into two main categories as Voluntary and Involuntary.

Voluntary churn occurs when the customer initiates termination of the service and are further subdivided into deliberate and incidental churners. Involuntary churn occurs when the company initiates termination of the customer from the subscriber's list [3].

Incidental churners occur due to incident because of some changes in location or change in financial position whereas deliberate churners occur due to customers need to change in technology or price rate [4].

2. ANALYSIS OF CUSTOMER CHURN

PREDICTION TECHNIQUES

As there is a high demand to predict and prevent customer churn in various organization many techniques and methodologies have been developed.

Mumin Yildiz et al., [5] proposed the use of Random Forest Method to achieve better performance in customer churn prediction and the results prove that the Random Forest Method achieves better specificity than AntMiner+ and C4.5 Decision tree. The data set used in this research is the Larose's data set which is obtained from wireless network telecommunication company and the data set consists of 5000 customers information and 21 features for each customer.

Tan Yi Fei et al., [6] proposed a novel method of using K means combined with Naïve Bayes classifier to predict customer churn. K means combined Naïve Bayes classifier method achieved better accuracy and sensitivity than EWD combine Naïve Bayes classifier method. However, it experiences a shortcoming which is the trade-off point in correctly predicting the true positive and true negative output. The dataset is a set of cleaned customer churn data from a telecommunication company. This data set consists of 5,000 customer caller data and 21 attributes. The future work proposed is to improve the model by studying the interval boundaries which might affect the learning rate of the classifier and this technique can be improved by using different machine learning algorithms like Support Vector Machine, Decision Tree and Bayesian Network.

A. Saran Kumar et al., [7] proposed an enhanced method such as SVM with AdaBoost Classification using Feature Discovery based prediction method which combine classifications of SVM, NBTree and SVM AdaBoost to address the limitation of high dimensional classification. The data set used is a bank data set. The proposed method achieves higher classification accuracy which can be used to predict customer churn efficiently.

Sebastiaan Hoppner et al., [8] presented a new churn classification method called ProfTree which develops an evolutionary algorithm to optimize the EMPC in the model construction step of a decision tree. [9] The data sets used are the real-life churn data sets taken from various telecommunication service providers which contains 889 customers and 10 explanatory variables. ProfTree is the overall most profitable model which can balance complexity and profitability compared to other tree-based methods such as EvTree, CART, ctree etc. The future work proposed is that the ProfTree can be combined with Random Forest to form ProfForest which can improve the profit maximizing property by building a large collection of profit induced trees and then aggregating them.

Sepideh Hassankhani Dolatabadi et al., [10] presented a case study of various Data Mining Techniques and Neural Predictor. The data sets used is collected over a period of a year and a half and contains details of every employee and customer. By comparing methods like Decision tree, Naïve Bayes, Support Vector Machine and Neural Network, the paper concludes that Support Vector Machine can be used to build reliable and accurate customer churn prediction models. The future work includes examining of customer and employee features to improve predictive models and increase trust companies to these prediction projects.

Chuanqi Wang et al.,[11] analyses the four groups of comparative experiments, which are common CART model, customer value mean of the training data set for the cost of CARTCS model, customer value mean of the subset A for the cost of CARTCS model and customer value mean of the subset B for the cost of CARTCS model and concludes that partition cost sensitive CART model not only has a good performance but also reduces the total misclassification cost. The data set contains 247,929 samples with 42 attributes. Thus, the main aim of the paper is to increase the profit of the telecommunication companies. The future research work includes generalization of partition cost sensitive model and to explore the impact of the class imbalance problem on the partition-cost sensitive model.

Muhammad Azeem et al., [12] proposes the use of fuzzy classifiers for developing a churn prediction model for prepaid customers. The data set has been taken from telecom company operating in South Asia which contains 600,000 instances with 722 attributes. The proposed model utilizes fuzzy classifiers like FuzzyNN, VQNN, OWANN and FuzzyRoughNN to predict accurately the churners from the large set of customer records. Various techniques like Multilayer Perceptron, Linear regression, C4.5, SVM and Decision Tree have been compared with Fuzzy classifier and it has been concluded that Fuzzy classifier predicts more accurately than other classifier techniques. Future research work includes the use of fuzzy based feature selection methods like fuzzy rough set and a comparative analysis can be done with respect to different classifiers can be done.

Long Zhao et al., [13] projected the use of K-Local Maximum Margin Feature extraction algorithm for churn prediction in telecommunication. The data set used in this research is KDD Cup 09 which contains 230 features and 50,000 samples. It has been concluded that the overall performance of KLMM is much better than other algorithms for KDD Cup 09 data set such as ALH, MIFS, MRMR, CMIM, JMI, DISR, CIFE, ICAP, CONDRED, CMI and RELIFE algorithm.

Ruiyun Yu et al., [14] used particle classification optimization-based BP network for telecommunication customer churn prediction. The data samples used in the research are the real phone call records from the China Mobile Communications Corporations(CMCC). The PBCCP algorithm has more obvious improvement on customer churn prediction accuracy compared with other algorithms like BP and PSOBP (Particle Swam Intelligence) algorithms. Future work will focus more on the elaboration of the range of "nearby" which is decided by the state of particles within the habitat and will focus on classifying the training data into different categories and train the model to predict more precisely and to avoid two-end deviants.

Franciska et al., [15] performed Churn Prediction analysis using various clustering algorithms in the KNIME Analytics

Platform. The KNIME Analytics Platform is used to visualize data flow and analysis. The data set used in the analysis is the patient churn data set. The clustering algorithms which are used are K-Means, K-Medoids, Fuzzy C Means, Hierarchical Clustering and Density Based Spatial Clustering of Applications with Noise(DBSCAN) and their performances have been analyzed successfully.

Muhammad Raza Khan et al., [16] proposed the use of Behavioral Modeling for Churn Prediction. This paper presents a unified analytical platform for predicting churn and assigning a 'Churn Score' to each customer who are likely to churn in the predefined amount of time. The data set used is the several terabytes of data from a South Asian mobile phone operator which contains 10,000 features for approximately 100,000 individuals. The proposed method uses brute force to identify overlapping features from customer transaction logs and then use related technique to identify features and metrics which are most predictive of customer churn. These features are then fed into a series of supervised learning algorithms that can predict customer churn. This approach shows 90 percent accuracy. In the future, the researchers are interested in systematic exploration of the feature space requiring the least involvement from analysts.

Abinash Mishra et al., [17] performs a comparative study of customer churn prediction in Telecom Industry using Ensemble Based Classifiers. The data set used is collected from the web link *http://www.ics.uci.edu~mlearn/MLRepository.html*. The experimental results show that the Random Forest performs better in terms of accuracy, sensitivity, specificity and error rate than other classifiers like Bayes, C4.5, ANN, SVM, LIBSVM, Probability Weighted Integration and Gaussian Weighted Integration. The future research work can include the Reinforcement Learning and Deep Learning to address the Churn Prediction.

Rong Zhang et al., [18] proposes the use of Deep and Shallow Model for churn prediction in Insurance industry. The data set used is the real-life data set from the NEW CHINA LIFE INSURANCE COMPANY LTD. The use of Deep and Shallow Model (DSM) improves the performance and earns better performance than both shallow-only and deep only models. DSM also performs better than CNN, LSTM, Stochastic Gradient Descent, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Gaussian Naive Bayes, AdaBoost, Random Forest, and Gradient Tree Boosting. The future research work can include the possibility of combining other shallow models like Gaussian Naive Bayes, k-Nearest Neighbor and Gaussian Processes; and other deep models, including CNN, LSTM, and Generative Adversarial Nets (GAN).

Qiu Yanfang et al., [19] projects the use EBURM model with logistic regression to predict customer churn in E-Commerce Platform. The data used in the study is the real user data from an E-Commerce platform which contains 6000 data. The experimental results suggest that the model performs better in terms of accuracy, precision, the full rate and the rate of omission and it can predict the customer churn in a confident manner.

Bryan Gregory et al. [20], uses extreme gradient boosting with temporal data to predict customer churn. The data set used in the research came from the WSDM Cup 2018 Challenge and was provided by KKBOX, a music streaming service. In this research, a supervised machine learning ensemble of decision trees is implemented in the modern XGBoost library, to build a highly accurate classification model to predict customer churn. Final accuracy was boosted by using LightGBM library and primary XGBoost model. The proposed model achieved high accuracy and thus outscored other models used submitted for the challenge. The future research includes further optimization of XGBoost and LightGBM model by using StackNet models and further exploration of additional features which are not yet tested.

Li Wang et al., [21] performed time-sensitive customer churn prediction based on PU Learning. The data sets used are the data obtained from Alipay.com, which is the world's leading third-party payment platform. In this research, PU Learning technique has been used and by comparing with other rule-based methods like Recency Rule, Frequency Rule, Logistic Regression, Distributed Factorization Machine, the proposed model achieved better results.

XIA Guo-en et al., [22] projected the use of Support Vector Machine(SVM) to perform Customer Churn prediction. The data sets analyzed in the research has been collected from the machine learning UCI database of University of California and home telecommunication carry. SVM has a good prediction precision, strong generation ability and good fitting precision when compared to BPANN, Decision Tree C4.5, Logistic Regression and Naïve Bayes Classifier.

Adnan Amin et al., [23] uses a rough set approach to predict customer churn in telecommunication sector. The data set used is the publicly available data set with 3333 instances. In this research, four different rules-generation algorithms like Exhaustive Algorithm, Genetic Algorithm, Covering Algorithm and LEM2 Algorithm have been compared and it has been concluded that the rough set theory classification based on Genetic Algorithm outperforms other algorithms in terms of precision, recall, the rate of misclassification, lift, coverage, accuracy, and F-measure. Future research work can investigate the proposed approach as there are certain issues. Firstly, churn data sets exhibit class imbalance problem and the churn class contain less number of

samples which makes it difficult to recognize the minority class for machine learning techniques. Secondly, eliminating and detecting outliers would provide better results and finally the profiles of predicted customer churns were not considered. The future work can improve the above-mentioned flaws.

3. CONCLUSION AND FUTURE WORK

Customer churn prediction is essential as it helps to detect customers who are likely to cancel a product, subscription etc. If a customer cancels their subscriptions or services from the company, the company faces a huge loss and it will cost a huge amount and also takes a lot of time to find new customers. Customer churn is a notorious problem faced by many industries and companies as it leads to a loss of revenue and loss of brand image. It is very important to build reliable models which can predict customer churn so that the company can escape huge loss and it is also difficult to acquire new customers.

The paper presents a review of customer churn prediction in various industries. It projects many attributes and techniques used to predict customer churn. The purpose of the paper is not to introduce a new technique but to review the implementation and understanding of the existing models. In this paper, 18 modeling techniques have been discussed. From the survey conducted it is understood that predicting customer churn is important for customer retention and has become mandatory in many industries to prevent a huge loss. Considering all the studies performed, it is clearly seen that all the studies have been trained and performed on real data provided by different companies like banking etc.

Many companies and industries are focused on the study of customer/user behavior analysis today. These studies will be useful to come up with better plans. Many techniques and models continue to emerge to predict customer churn not only in telecommunication sector but also in many other sectors.

Table 1: CLASSIFICATION OF STUDIES INVESTIGATED

REFERENCE NUMBER	METHODS USED	DATA SET USED
Mumin Yildiz	Random Forest, AntMiner+ and C4.5 Decision tree	5000 customer records with 21 features each customer
Tan Yi Fei	K means combined with Naïve Bayes	5000 customer caller data with 18 attributes
A. Saran Kumar	SVM, NBTree and SVM AdaBoost	Bank Dataset
Sebastiaan Hoppner	ProfTree, EvTree, CART, ctree	889 customer records and 10 explanatory variables.
Sepideh Hassankhani Dolatabadi	Decision tree, Naïve Bayes, Support Vector Machine and Neural Network	9239 customer records with 15 attributes for employee service and 21 attributes for customers
Chuanqi Wang	Partition cost sensitive CART model	247,929 samples with 42 attributes
Muhammad Azeem	Fuzzy classifiers like FuzzyNN, VQNN, OWANN and FuzzyRoughNN	600,000 instances with 722 attributes
Long Zhao	K-Local Maximum Margin Feature extraction algorithm	KDD Cup 09 contains 230 features, 50,000 samples.
Ruiyun Yu	particle classification optimization-based BP network	100,000 customer records with 7 features
Franciska	K-Means, K-Medoids, Fuzzy C Means, Hierarchical Clustering and Density Based Spatial Clustering of Applications with Noise(DBSCAN)	Patient Churn Dataset
Muhammad Raza Khan	Behavioural Modelling	100,000 individual records, 10,000 features

Abinash Mishra	Ensemble Based Classifiers like Random Forest, Bayes, C4.5, ANN, SVM, LIBSVM, Probability Weighted Integration and Gaussian Weighted Integration	3333 records ,15 attributes
Qiu Yanfang	EBURM model with logistic regression	6000 data, 5 features
Bryan Gregory	Extreme gradient boosting with temporal data	subscriber data from 3 distinct sources: user activity logs, transactions, and member data, 208 features
Li Wang	PU Learning technique, Recency Rule, Frequency Rule, Logistic Regression and Distributed Factorization Machine	Dataset from Alipay.com divides into three datasets P(Positive), N(Negative) and U(Unlabelled) each contains about 200 million customers
XIA Guo-en	SVM, BPANN, Decision tree C4.5, Logistic Regression, and Naive Bayesian classifiers	Dataset 1 from University of California and Dataset 2 FROM home telecommunication carry
Adnan Amin	Exhaustive Algorithm, Genetic Algorithm, Covering Algorithm and LEM2 Algorithm	Dataset -3333 instances, Training set-2333 instances, Test set-1000 instances, 12 attributes
Mumin Yildiz	Random Forest, AntMiner+ and C4.5 Decision tree	5000 customer records with 21 features each customer
Tan Yi Fei	K means combined with Naïve Bayes	5000 customer caller data with 18 attributes
A. Saran Kumar	SVM, NBTree and SVM AdaBoost	Bank Dataset
Sebastiaan Hoppner	ProfTree, EvTree, CART, ctree	889 customer records and 10 explanatory variables.
Sepideh Hassankhani Dolatabadi	Decision tree, Naïve Bayes, Support Vector Machine and Neural Network	9239 customer records with 15 attributes for employee service and 21 attributes for customers
Chuanqi Wang	Partition cost sensitive CART model	247,929 samples with 42 attributes
Muhammad Azeem	Fuzzy classifiers like FuzzyNN, VQNN, OWANN and FuzzyRoughNN	600,000 instances with 722 attributes
Long Zhao	K-Local Maximum Margin Feature extraction algorithm	KDD Cup 09 contains 230 features, 50,000 samples.
Ruiyun Yu	Particle classification optimization-based BP network	100,000 customer records with 7 features
Franciska	K-Means, K-Medoids, Fuzzy C Means, Hierarchical Clustering and Density Based Spatial Clustering of Applications with Noise(DBSCAN)	Patient Churn Dataset
Muhammad Raza Khan	Behavioural Modelling	100,000 individual records, 10,000 features
Abinash Mishra	Ensemble Based Classifiers like Random Forest, Bayes, C4.5, ANN, SVM, LIBSVM, Probability Weighted Integration and Gaussian Weighted Integration	3333 records ,15 attributes
Qiu Yanfang	EBURM model with logistic regression	6000 data, 5 features
Bryan Gregory	Extreme gradient boosting with temporal data	subscriber data from 3 distinct sources: user activity

		logs, transactions, and member data, 208 features
Li Wang	PU Learning technique, Recency Rule, Frequency Rule, Logistic Regression and Distributed Factorization Machine	Dataset from Alipay.com divides into three datasets P(Positive), N(Negative) and U(Unlabelled) each contains about 200 million customers
XIA Guo-en	SVM, BPANN, Decision tree C4.5, Logistic Regression, and Naive Bayesian classifiers	Dataset 1 from University of California and Dataset 2 FROM home telecommunication carry
Adnan Amin	Exhaustive Algorithm, Genetic Algorithm, Covering Algorithm and LEM2 Algorithm	Dataset -3333 instances, Training set-2333 instances, Test set-1000 instances, 12 attributes

4. REFERENCES

[1] J. Hadden, A. Tiwari, R. Roy, D. Ruta "Computer assisted customer churn management: state-of the-art and future trends"

Comput. Oper. Res. 34 (10) (2007) 2902–2917.

[2] Kotler, P., Keller, K. L. 2009. Marketing Management. Pearson Prentice Hall.

[3] N. Kamalraj, A. Malathi "A Survey on Churn Prediction Techniques in Communication Sector" International Journal of Computer Applications (0975 – 8887) Volume 64– No.5, February 2013; 39-42

[4] Sindhu M E1 and Vijaya M S2 "Predicting Churners in Telecommunication Using Variants of Support Vector Machine" American Journal of Engineering Research Volume-4, Issue-3,pp-11-18

[5] Mumin Yıldız, Songul Albayrak "Customer Churn Prediction in Telecommunication with Rotation Forest Method" DBKDA 2017: The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications; 26-29

[6] Tan Yi Fei1, Lam Hai Shuan1, Lai Jie Yan1 Guo Xiaoning1, Soo Wooi King. "Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier" Int. J. Advance Soft Compu. Appl, Vol.9, No.3,Nov 2017

[7] A. Saran Kumar, Dr. D. Chandrakala "An Optimal Churn Prediction Model using Support Vector Machine with Adaboost "International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT. Volume 2. Issue 1

[8] Sebastiaan Hoppnera, Eugen Striplingb, Bart Baesensbc, Seppe vanden Brouckeb, Tim Verdonck "Profit Driven Decision Trees for Churn Prediction" arXiv:1712.08101v1 ;1-30

[9] Verbraken, T., Verbeke, W., Baesens, B., " A novel profit maximizing metric for measuring classification performance of customer churn prediction models. "IEEE Transactions on Knowledge and Data Engineering; 2013, 25 (5), 961-973

[10] Sepideh Hassankhani Dolatabadi, Farshid Keynia "Designing of Customer and Employee Churn Prediction Model Based on Data Mining Method and Neural Predictor" The 2nd International Conference on Computer and Communication Systems 23 October 2017;74-77

[11] Chuanqi Wang, Ruiqi Li, Peng Wang, 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 ;5680-5684

[12] Muhammad Azeem1, Muhammad Usman, Fong "A churn prediction model for prepaid customers in telecom using Fuzzy classifiers" Telecommunication Systems-Modelling, Analysis, Design and Management December 2017, Volume 66, Issue 4, pp 603–614

[13] Long Zhao, Qian Gao, XiangJun Dong, Aimei Dong, Xue Dong "K local maximum margin feature extraction algorithm for churn prediction in telecom" Cluster Computing June 2017, Volume 20, Issue 2, pp 1401–1409

[14] Ruiyun Yu, Xuanmiao An, Bo Jin, Jia Shi, Oguti Ann Move, Yonghe Liu "Particle classification optimization-based BP network for telecommunication customer churn prediction" Neural Computing and Applications February 2018, Volume 29, Issue 3, pp 707–720

[15] Franciska, Swaminathan "Churn Prediction Analysis Using Various Clustering Algorithms in KNIME Analytics Platform" May 2017; 166-170

[16] Muhammad Raza Khan, Joshua Manoj , Anikate Singh, Joshua Blumenstock "Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty" December 2015 ;1-4

[17] Abinash Mishra, U. Srinivasulu Reddy "A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers" Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) November 2017; 721-725

[18] Rong Zhang, Weiping Li, Tong Mo, Wei Tan "Deep and Shallow Model for Insurance Churn Prediction Service" 2017 IEEE 14th International Conference on Services Computing 25-30 June 2017; 346-353

[19] Qiu Yanfang, Li Chen "Research on E-commerce User Churn Prediction Based on Logistic Regression" 15-17 Dec. 2017 ;87-91

[20] Bryan Gregory "Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data" arXiv.org; 9 Feb 2018;1-4

[21] Li Wang, Chaochao Chen, Jun Zhou, Xiaolong Li "Time-sensitive Customer Churn Prediction based on PU Learning" arXiv.org;27 Feb 2018;1-4

[22] XIA Guo-en, JINWei-dong "Model of Customer Churn Prediction on Support Vector Machine" Systems Engineering — Theory & Practice Volume 28, Issue 1, January 2008 28(1): 71–77

[23] Adnan Amina, Sajid Anwara, Awais Adnana, Muhammad Nawaza, Khalid Alawfib, Amir Hussainc, Kaizhu Huang "Customer churn prediction in the telecommunication sector using a rough set approach" Neurocomputing 237 (2017) ;242–254