A Critical Examination of Different Models for Customer Churn Prediction using Data Mining

Seema, Gaurav Gupta

Abstract: ***Due to competition between online retailers, the need for providing improved customer service has grown rapidly. In addition to reduction in sales due to loss of customers, more investments are needed to be done to attract new customers. Companies now are working continuously to improve their perceived quality by way of giving timely and quality service to their customers. Customer churn has become one of the primary challenges that many firms are facing nowadays. Several churn prediction models and techniques are proposed previously in literature to predict customer churn in areas such as finance, telecom, banking etc. Researchers are also working on customer churn prediction in e-commerce using data mining and machine learning techniques. In this paper, a comprehensive review of various models to predict customer churn in e-commerce data mining and machine learning techniques has been presented. A critical review of recent research papers in the field of customer churn prediction in e-commerce using data mining has been done. Thereafter, important inferences and research gaps after studying the literature are presented. Finally, the research significance and concluding remarks are described in the end.***

Keywords: CRM, e-commerce, dataset, pre-processing, data mining, customer churn prediction, model building, machine learning.

I. INTRODUCTION

1.1 CRM in e-commerce

CRM can be defined as the strategies, technologies and practices, companies used in order to manage customer interactions and data to improve relationships with valued customers by collecting their information through websites, email, telephone, news and social media etc. [1]. The CRM systems analyze customer data and use it for decision making regarding customers behavior about the company and its products [1]. CRM solution helps to retain existing customers, add new customers and keep them satisfied by organizing customer information. Development of the www service has led to the growth of e-commerce.

E-Commerce uses ICT in customer service and business operations for creating, converting and redefining relationships between buyers, sellers [2]. This made it possible for businesses to become close to potential and existing consumers and to develop the more loyal relationship between them. E-commerce along with CRM is the tool that has been designed to communicate with customers, to find out the needs, possibilities, and preferences of the customers. CRM helps in the growth of a business by using key strategies to impress people to buy online [3].

1.2 Data Mining and CRM

Data mining can be defined as the technique of storing, extracting, deleting and editing data in large databases. The main aim of data mining technique is to extract the needed information from dataset and convert it into an understandable form so as to use it further. Figure 1 shows the interdependency of data mining and CRM. The CRM interact with data warehouse through ETL system that extracts valuable data, transforms it into a required format, and loads it into the database and update data for the purpose of analytics, data mining, and reporting.

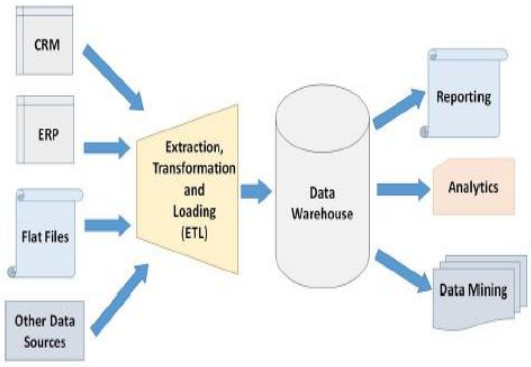

Figure 1: Data Mining and CRM [4]

Data mining is used to extract data patterns such as the data record groups, unusual data records and data interdependencies addressed using cluster analysis, anomaly detection and association rule mining respectively. Data mining normally consist of six tasks [2] namely (i) anomaly detection – to identify unusual data records (ii) association rule mining – to search for relationships between variables (iii) clustering – to discover structures in data (iv) classification – to generalize to apply to new data (v) regression – to estimate relationships among datasets and model the data with least error (vi) summarization – to represent the data in compact form.

A primary reason for using data mining is to draw inferences and predictions from a collections of observations of customer behavior. With the growth of technology, data mining has emerged as an important tool for CRM based applications [3]. A

Revised Manuscript Received on July 22, 2019.
***** Correspondence Author
Seema*, Research Scholar, Computer Science and Engineering Department, Punjabi University Patiala, Punjab.
Gaurav Gupta#, Assistant Professor, Computer Science and Engineering Department, Punjabi University Patiala, Punjab.
#Corresponding Author Email: seemabaghla@pbi.ac.in

company should concentrate to offer some gifts and discounts those have chances to reflect to an offer rather than contacting the customer directly through their customer support office or email. A number of predictive methods and models are available to predict the offer to which a customer is likely to respond and also the customers to churn in spite of the offer [4].

1.3 Customer Churn

Customer churn can be defined as the behavior of a customer that leaves his service with a company by ending the relationship, cancellation of a subscription, ending of a membership, closure of an account, not renewing a service agreement and deciding not to buy [4]. Customer churn can be caused by unsatisfied customer, bad customer support, lack of satisfaction for current customers, or natural causes [4].

All businesses are working for increase in customer loyalty and reduce customer complaints, and consequently reducing customer churn. CRM assists firms to customize their services based on information of customer interactions and other data [5]. Hence customer churn prediction plays an important role in the success of a company.

II. LITERATURE REVIEW

Potential research work carried out on various techniques for churn prediction in different areas such as e-commerce, telecom and banking etc. has been discussed in the following paragraphs. GA, PSO, NB, SVM, DT, RF, LR, NN signifies genetic algorithm, particle swarm optimization naïve Bayes, support vector machine, decision tree, random forest, linear regression and neural network respectively in this paper.

Azeem et al. [2] developed a churn prediction model for prepaid customers in telecom using fuzzy classifiers, NN, LR, SVM, AdaBoost & RF techniques were compared with a fuzzy nearest-neighbor classifier to predict an accurate set of churners on a real-time dataset of prepaid telecom customers from south Asia. Parameters such as TP rate & AUC were considered and enhanced using the model. *Authors suggested that prediction accuracy can be improved by using large data volume. It was also suggested that churn prediction model in other application such as e-commerce can also be developed.*

Bahari and Eloyidom [3] proposed a CRM framework using neural network and data mining for the prediction of customer behavior in banking. The UCI dataset containing direct bank marketing campaigns of Portuguese bank was taken. It was concluded that NN was better than NB algorithm for accuracy & specificity while NB was better than NN algorithm for sensitivity, TPR, FPR, and ROC area, neural network classified 4007/514 & Naive Bayes classified 3977/544 instances correctly/incorrectly. Authors suggested that time for model building is very high and needed to be reduced. Further, improved algorithm can be designed for banking and e-commerce for model building.

Baumann et al. [5] developed an algorithm based on clickstream data of a website to extract information and tested the predictive power of the model based on data such as number of clicks, repeated visits, repetitive purchases, etc. Correlation analysis of graphical information was done to find the surfing structure of a surfer on a website. It was concluded that graph metrics and regression analysis resulted into increase in predictive power of the models. *Authors suggested that weighted graph can be used to improve the accuracy of prediction and surfing pattern of the customer on the websites.*

Vijaya and Sivasankar [7] proposed a system for customer churn prediction in telecom. The feature selection model was developed using and simulated annealing for a UCI telecom dataset. PSO model was used for churn prediction and results were compared with DT, NB, NN & SVM techniques. Parameters such as accuracy, TP rate, TN rate, FP rate, precision were compared for churn prediction. It was concluded that PSO perform well on huge imbalanced data. *Authors suggested that algorithm in enhanced way can be used for customer churn prediction in other sectors such as e-commerce.*

Lee et al. [8] proposed a customer churn prediction model in mobile industry using data mining based on influence of words in online news from Korea. Churn prediction from the effect of texts within advertisements in the online news was done using NN, DT, and LR. Authors concluded that the prediction is on the bases of web information. They added that the surveys can also be included for better prediction. Other markets, such as online market, have a greater scope and they must be explored. *Authors considered macro perspective so ignores churn factors of individual tendencies, which can further be studied.*

Jie et al. [9] analyzed purchase data of customers based on OWA and k-mean clustering in order to improve customer services. The method was also used to identify potential losing customers based on customer lifetime value. Authors suggested that the improved questionnaire can be used to accurately predict potential churners. Customer service can be improved to retain customers. *More factors can be considered influencing customer satisfaction*.

Fridrich [10] proposed an ANN based optimization model to predict customer churn using GA in e-commerce. The prediction model is developed to identify customers at risk of defection. The proposed model lead to improved customer churn prediction ability on the basis of parameters such as TP rate, FP rate, and accuracy. *Authors suggested that techniques such as multiple objective optimizations, improved LR, DT& fuzzy logic can be used. Additionally, more parameters addition in e-commerce for better prediction accuracy was advised.*

Machado and Ruiz [11] developed a churn prediction model based on mobile application usage. The accuracy and precision of the proposed model was found to be better than previous studies for a UCI dataset about Portugal taxi service. *It was advised that improved algorithm can be used for another application such as e-commerce in future. It was also suggested that data mining based mobile and web apps can also be developed for prediction of customer churn.*

Gordini and Veglio [12] proposed a churn prediction SVM model for Italian online e-commerce customers. The parameters such as recency, frequency, length, product category, failure, monetary, age, profession, gender, request status etc. were taken for performance comparison. The prediction power of the proposed method was found to be better as compared to LR, NN & SVM especially for noisy, imbalance & nonlinear data. *It was suggested by the authors that staying power of the model is not predicted as it is important to check the time period of the*

customer to stay with the company and will certainly help to take more preventive measures to avoid churn. Authors also advocated for selection of SVM kernel function more accurately & inclusion of more prediction variables.

Li et al. [13] developed a BG/NB model to analyze the purchase behavior records of existing e-commerce customers to identify pre-losing customers. Customers' purchase behavior and personal preference were analyzed to classify customers by using data mining. It was concluded that the success rate of retaining existing churners is 25% and success rate of acquiring new customers is 6.25%. *Authors were only able to forecast the number of consumers to buy, while the amount of consumption and the types of items purchased are not accounted for, which can also be considered. Additionally, the data related to potential customers who have visited the web pages and didn't make a purchase can also be included in the dataset.*

Berger and Kompan [14] developed a prediction model based on the intent of user churn based on web based interaction of the user. The performance of the model was predicted by prediction of churn of real data from an online retailers. They concluded that the prediction using the proposed model outperforms the churn prediction based on baseline models. *Authors suggested that the developed methodology can be extended in e-commerce applications.*

Zhao [15] developed a GA, SVM and NN based customer churn prediction model in e-commerce. It was concluded that the accuracy and precision of the developed model is better than previous models. *It was suggested that in future, an improved model may be developed with more attributes and parameters.*

Yu et al. [16] developed an e-SVM technique to predict customer churn in e-commerce. The developed method worked better than NN, SVM & DT based on coverage rate, hit rate, accuracy, and lift coefficient. *It was suggested that customer churn can be predicted for multiple e-commerce sites with imbalance data. Further, data mining methods can be improved by considering actions such as discount, cashback, return policy etc.*

Hosseini et al. [17] proposed a method of clustering by joining WRFM-method to the K-mean algorithm in data mining to classify customer loyalty based on recency, frequency & monitoring for an Iranian company product dataset. Authors concluded that the results of statistical tests for model validation are acceptable. Authors designed customer loyalty assessment function to find customer loyalty value (CLV) of a customer. *Authors suggested that the proposed method can be used to evaluate customer churn areas such as telecom, e-commerce etc.*

Poel and Buckinx [18] proposed a Logit modeling based model to predict the purchasing behavior at an online store. Four variables namely customer demographics, clickstream information, historical purchase behavior and clickstream data were used in the prediction. It was concluded that clickstream behavior is important in order to predict the buying tendency. *In future, the enhanced algorithm may be extended to detailed data sets for many websites for customer churn prediction.*

Kumar and Ravi [19] proposed a credit card customer churn prediction model to predict customer churn in banks using data mining. The data was balanced using oversampling, synthetic minority oversampling technique and parameters such as accuracy, sensitivity, and specificity were used for performance evaluation. Techniques such as multilayer perceptron, DT, RF, SVM were used for churn prediction. It was concluded that the proposed system performs well with unbalanced original data. *It was suggested by the authors that a warning system can be designed based on a set of rules for churn prediction. It was also found that macro perspective was taken into consideration so individual tendencies can also be included. Further, data dictionary can be used for prediction.*

Comparison of key research papers

Table 1 outlines the summary of key research papers discussed in previous section. The comparison has been done on the basis of area of prediction, type of dataset, parameter studied, techniques used, the results derived from each research work done by various authors. The future scope of the work has also been identified.

Table 1: Summary of key research papers

Citation	Area of prediction and Data Type	Parameter	Techniques Used	Results	Future Scope
Vijaya and Sivasankar (2017) [7] Springer	Telecom, UCI dataset	TP rate, TN rate, FP rate, precision, Accuracy, F-measure,	PSO based SA, NB, DT, SVM, NN	PSO better than other methods	• Enhanced algorithm can be used for other application. • The latest tool can be used.
Azeem et al. (2017) [2] Springer	Telecom, Real-Time dataset (South Asia)	TP rate, Accuracy, AUC	NN, LR, SVM, RF, Fuzzy based NN	Fuzzy based NN is better	• Churn prediction in other areas can be done. • Prediction accuracy can be improved with large data.

Lee et al. (2017) [8] Emrald Insight	Mobile, Online news dataset (Korean market)	External (e.g. death) and Internal (e.g. change of service)	NN, LR, DT	NN is better	• Questionnaires can be included for better prediction. • Other markets have a greater scope. • Current studies can be added design database.
Bahari and Eloyidom [3] (2015) Elsevier	Banking, UCI dataset	Accuracy, TPR, FPR, ROC, accuracy, sensitivity, specificity.	NB, NN, WEKA Tool	NN better (accuracy & specificity), NB better (TP Rate, FP Rate, ROC & sensitivity)	• Improved algorithm can be used for banking and e-commerce. • Time for model building needs to be reduced.
Jie et al. (2015) [9] Springer	e-commerce, Questionn-aires	CLV, RFM Value	OWA method	CLV & RFM are good parameters	• Improved questionnaires can be used. • More factors are needed to be taken for prediction.
Fridrich (2017) [10]	e-commerce, Analytical database of Alza.cz	Average revenue, claims, invoices issued, TP Rate, FP Rate, accuracy	NN, GA	NN better	• Improved LR, decision tree, fuzzy methods can be used. • E-commerce datasets and more parameters can be used for better prediction accuracy.
Machado and Ruiz (2017) [11] IEEE	Mobile, UCI dataset	Accuracy, Precision, FPR	Data Stream Clustering	The proposed method is better	• Improved algorithm and more features can be taken. • The system can be used for other applications.
Li et al. (2017) [13] Wuhan Conference	e-commerce, Historical transaction records of a website	Purchase behavior, personal preferences	Prediction BG/NBD Model	Retaining better than adding a new customer	• The number of customers to buy and types of items purchased needed to be included. • Customers who have visited the web pages and didn't make a purchase may be considered.
Berger and Kompan (2019) [14] IEEE	Website data of online retailers	Accuracy and precision	Proposed NN Based method	Proposed method works better	• Authors suggested that the developed methodology can be extended in e-commerce applications.
Zhao (2014) [15] Scopus	e-commerce, Questionnaire	Accuracy, precision	GA, SVM, NN	Hybrid SVM & NN is better than SVM & GA	• The improved model needs to be developed. • More data attributes and parameters can be used in e-commerce better prediction.
Yu et al. (2011) [16] Elsevier	e-commerce, Website	Hit Rate, Lift Coefficient, Coverage Rate, Accuracy	SVM, NN	SVM better	• Imbalance data was not accounted for. • Multiple e-commerce websites can be taken. • More actions can be taken to prevent churning.
Kumar and Ravi (2008) [19] Inder-science	Banking, Credit Card Data American Bank	Accuracy, sensitivity, and specificity	MLP, LR, SVM	MLP is better	• A warning system for churn prediction can be designed. • Improved tools and data mining can be used for accurate churn prediction in other applications.

III. RESEARCH GAPS

After studying the relevant published literature related to the prediction of customer churn in various application areas such as banking, telecom, e-commerce etc. using various methods, the following research gaps have been identified.

- Enhanced algorithms are needed to be developed in order to predict customer churn in e-commerce [2, 3, 7, 9, 14, 15, 18].
- Highly accurate model is needed to be developed [12, 13].

- Latest available data was not included previously in designing the databases that limited the scope of work [14, 17].
- The time for the model building was quite high and is required to be reduced using improved tools and techniques [3, 5, 7, 8, 10].
- Comprehensive questionnaires were not used and important parameters were not considered [11, 12, 16].
- The customer buying behavior and type of items purchased were not included in designing the database [14, 15, 16].
- The prediction models are not available on commercial platforms [6, 7, 13].

IV. RESEARCH SIGNIFICANCE AND CONCLUDING REMARKS

The need to improve customer service has grown steeply. Companies are tirelessly engaged in providing accurate and in time processing of customer orders for better quality and service. The companies usually spend a lot of money to study the purchase behavior of their customers in order to minimize their churn. As per current practice, company representatives contact the customers for their feedback about service quality and other issues. Company persons usually try to convince them to stay with their company and solve their problems by assuring better service along with offering some monetary or discount benefits. Company also give some special benefits and services to retain the valued customers. Losing customers may not only lead to reduced sales but increased cost of adding new customers.

In this paper, a comprehensive review of various churn prediction models in order to predict customer churn in e-commerce using various techniques such as machine learning and data mining etc. has been performed. An attempt has been made through this paper to understand and explore the existing work done by various researchers in the field of customer churn prediction models development using data mining for various applications. Quality research papers has been studied and reviewed critically in the field of customer churn prediction in e-commerce using data mining. Important inferences and research gaps has been worked out. These research gaps form the basis of future work by the authors.

Authors are working further to develop a model for customer churn prediction in e-commerce using data mining. With this model, the businesses will be able to predict the churning customers that are going to leave them in advance. This will give them a chance to plan their strategies to retain them. The model to be developed by authors will not only help to predict the customer churn in e-commerce but explain the reasons for the churn also. Use of the model will definitely be helpful to reduce the expenditure of a company in activities related to retention of valued customers and churn prediction. The developed model will surely be accurate and transparent based on the real e-commerce customer data.

BIBLIOGRAPHY

1. Au, W. H.; Chan, K. C. C.; Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, 7, pp.532–545.
2. Azeem, M.; Usman, M.; Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Springer Telecommunication Systems*, 66 (4), pp.603-614.
3. Bahari, T. F.; Elayidom, M. S. (2015). An efficient CRM-Data mining framework for the prediction of customer behavior. *Elsevier Procedia Computer Science*, 46, pp.725 – 731.
4. https://www.tutorialspoint.com/customer_relationship_management/crm_quick_guide.htm (assessed at 5.10pm on 22.03.2019)
5. Baumann, A.; Haupt, J.; Gebert, F.; Lessmann, S. (2018). Changing perspectives: Using graph metrics to predict purchase probabilities. *Expert Systems with Applications*, 94, pp.137-148.
6. Chen, C. P.; Weng, J. Y.; Yang, C. S.; Tseng, F. M. (2018). Employing a data mining approach for identification of mobile opinion leaders and their content usage patterns in large telecommunications datasets. *Technological Forecasting & Social Change*, 130, pp.88-98.
7. Vijaya, J.; Sivasankar, E. (2017). An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing, *Springer Cluster Computing*, pp.1-12.
8. Lee, E. B.; Kim, J.; Lee, S. G. (2017). Predicting customer churn in the mobile industry using data mining technology. Industrial Management and Data Systems, 117 (1), pp.90-109.
9. Jie, C.; Xiaobing Y.; Zhifei Z. (2015). Integrating OWA and data mining for analyzing customers churn in e-commerce. *Springer Journal of Systems Science and Complexity*, 28, pp.381–392.
10. Fridrich, M. (2017). Hyperparameter optimization of artificial neural network in customer churn prediction using Genetic Algorithm. *Trends in Economics and Management*, 28(1), pp.9–21.
11. Machado, N. L. R.; Ruiz, D. D. A. (2017). Customer: A novel customer churn prediction method based on mobile application usage. IEEE Wireless Communications and Mobile Computing Conference, pp.2146-2151.
12. Gordini, N; Veglio, V. (2016). Customer churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter selection technique in B2B e-commerce industry. *Industrial Marketing Management.* 8, pp. 1-8.
13. Li, H.; Guan, Z.; Cui, Y. (2017). Customer churn prediction based on BG / NBD model. *Wuhan International Conference on E-Business - Emerging Issues in E-Business*, pp.386-393.
14. Berger, P.; Kompan, M. (2019). User modelling for churn prediction in E-commerce. IEEE Conference on Intelligent Systems, pp.1-6.
15. Zhao, X. (2014). Research on E-commerce customer churning modeling and prediction. *The Open Cybernetics & Systemics Journal*, 8, pp.800-804.
16. Yu, X.; Guo S.; Guo J.; Huang X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, 38, pp.1425–1430.
17. Hosseini, S. M. S.; Maleki, A.; Gholamian M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37, pp. 5259–5264.
18. Poel, D. V. D.; Buckinx, W. (2005). Predicting online-purchasing behavior. *European Journal of Operational Research*, 166, pp.557–575.
19. Kumar, D. A.; Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *Data Analysis Techniques and Strategies*, 1(1), pp.4-28.