Fall 10-16-2021

Dropout Prediction: A Systematic Literature Review

Pedro Sobreiro
ESDRM-IPSantarem, sobreiro@esdrm.ipsantarem.pt

Domingos Martinho
ISLA Santarém, domingos.martinho@islasantarem.pt

Javier Berrocal
Quercus Software Engineering Group, University of Extremadura, jberolm@unex.es

José Garcia Alonso
Quercus Software Engineering Group, University of Extremadura, jgaralo@unex.es

Dropout Prediction: A Systematic Literature Review

Pedro Sobreiro, Quality of Life Research Centre, Polytechnic Institute of Santarém, Portugal, sobreiro@esdrm.ipsantarem.pt

Domingos Martinho, ISLA Santarém, Portugal, domingos.martinho@islasantarem.pt

Javier Berrocal, Quercus Software Engineering Group, University of Extremadura, Spain, jberolm@unex.es

José Garcia Alonso, Quercus Software Engineering Group, University of Extremadura, Spain, jgaralo@unex.es

Abstract

Dropout predicting is challenging analysis process which requires appropriate approaches to address the dropout. Existing approaches are applied in different areas such as education, telecommunications, retail, social networks, and banking services. The goal is to identify customers in the risk of dropout to support retention strategies. This research developed a systematic literature review to evaluate the development of existing studies to predict dropout using machine learning, following the guidelines recommended by Kitchenham and Peterson. The systematic review followed three phases planning, conducting, and reporting. The selection of the most relevant articles was based on the use of Active Systematic Review tool using artificial intelligence algorithms. The criteria identified 28 articles and several research lines where identified. Dropout is a transversal problem for several sectors of economic activity, where it can be taken countermeasures before it happens if detected early.

Keywords: dropout prediction; customers; machine learning

1. INTRODUCTION

Customer analysis is fundamental to develop business and marketing intelligence (Sheth, Mittal, & Newman, 1998), supporting the understanding of historical data identifying trends and patterns (Berry & Linoff, 2004). This process is also known as data mining, the extraction of knowledge from data (Han & Kamber, 2006). Data mining encompasses techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, and high-performance computing (Han, Kamber, & Pei, 2012). Data mining explores and analyses data to discover relevant patterns using task such as classification; regressions; clusters analysis (Han & Kamber, 2006). According to Han, Kamber, and Pei (Han et al., 2012), these tasks present many similarities between data mining and machine learning. Machine learning is understood as an automated process to extract patterns from the data (Kelleher, Namee, & D'Arcy, 2015), generalizing from the examples in the training set (Domingos, 2012). Machine learning could be used to extract knowledge to understand dropout with the development of effective retention strategies (Verbeke, Martens, Mues, & Baesens, 2011). The use of machine learning allows the discovery of patterns supporting the identification of hypothesis addressing existing problems.

The dropout is a problem that needs to be addressed. The costs of retaining customers are lower when compared to the costs of attracting new ones (Edward & Sahadev, 2011). Reichheld (Reichheld, 1996) evidenced that reducing dropout rates by 5% (e.g., from 15% to 10% per year) could represent an increase in profits up to the double. A customer that dropout represents a loss of money, if an organization can predict the dropout is possible to develop counter measures to avoid the desertion. Machine learning algorithms have been used to predict customer dropout (Bandara, Perera, & Alahakoon, 2013), without however to consider the timings of the dropout. Survival analysis, or more generally, time-to-event analysis, refers to a set of methods to describe the probability of surviving past a specified time point, or more generally, the probability that the event of interest has not yet occurred by this time point (Schober & Vetter, 2018). To our knowledge there is a lack of systematic literature review addressing the dropout using machine learning techniques.

This research analyses state of the art and identifies Machine Learning studies to predict customer dropout. It is developed under a methodology of systematic literature review applied by Kitchman & Charters (Kitchenham & Charters, 2007) to perform a systematic literature review.

2. RESEARCH METHODOLOGY

According to Fink (Fink, 2010), systematic literature review (SLR) is a systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners.

The importance to understand customer dropout and the diversity of employed algorithms requires an understanding of trends and existing problems to create a ground base of knowledge. For the development of the systematic literature review was adopted, the methodology applied by Kitchenham & Charters (Kitchenham & Charters, 2007) developed in three stages: Planning, Implementation and Results.

There were identified four research questions to determine the main aspects related to the customer dropout with contractual settings.

RQ1.What are the trends in machine learning algorithms to predict dropout? RQ1 aims at identifying the ML techniques that have been used to predict the customer's dropout.

RQ2. When the dropout occurs? RQ2 intends to understand if the timing related to the customer dropout is considered.

RQ3. What are the more relevant features related to predicting customer dropout?

RQ4. What is the accuracy of the machine learning algorithms to predict dropout?
This phase requires the identification of the search strategy. The authors adopted the Population,

Intervention, Comparison, Outcomes and Context (PICOC) as suggested Kitchenham and Charters (Kitchenham & Charters, 2007) and proposed by Petticrew and Roberts (Petticrew & Roberts, 2006).

The adopted search criteria were ((“customer dropout”) OR (“customer churn”) AND “machine learning” AND (“contractual” OR “membership”)), which was applied to the title, abstract, and keywords in the search period between January 2000 and December 2019 using the IEEE Digital Library database.

The exclusion criteria were Books, Non-English articles, patents, and thesis.

A total of 218 studies were found in the first step. The selection process of the identified articles was developed using ASReview (ASReview Core Development Team, 2019) creating a dataset of the identified articles, providing five relevant papers and five irrelevant papers to train Machine Learning model Naïve Bayes. After we started the reviewing process labelling the subsequent papers as irrelevant or relevant until start suggesting only irrelevant papers. The results were exported and analyzed only 28 relevant papers. During the data extraction one paper was excluded, remaining 27 relevant papers; this process is represented in Figure 1.

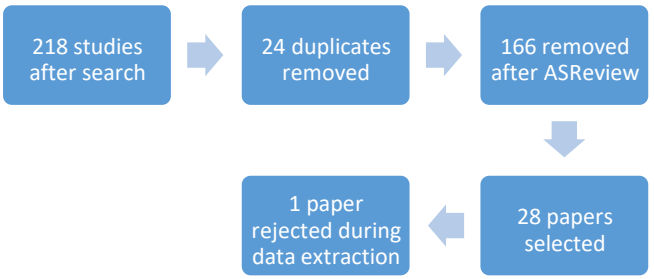

Figure 1 – Filtering process to identify the final studies for review

The quality assessment criteria was developed for the four research questions based on the score schema on Kitchenham et al. (2010). Was adopted three-level scale Yes = 1.0, Undefined = 0.5 and No = 0. The selected papers were reviewed to answer the quality questions.

The data extraction process was developed while the papers were reviewed identifying the dropout domain, the type of organization (e.g., airline company, insurance company or telecommunication) and the dropout prediction techniques (e.g., Decision trees, logistic regression, or support vector machine).

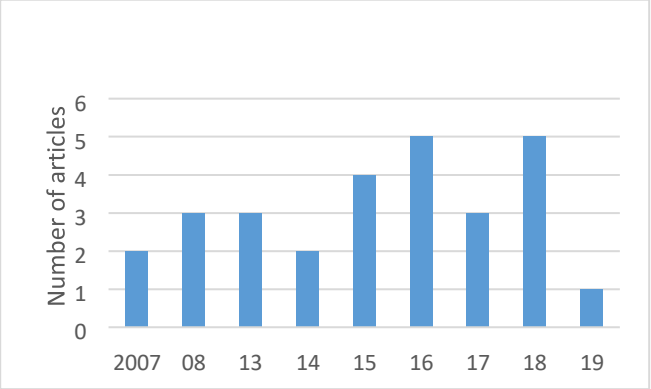

Figure 2 – Articles per year after quality assessment

3. RESULTS

During the development of the systematic literature review, the authors considered the following to be the most important research organized according to the research questions.

RQ1.What are the trends in machine learning algorithms to predict dropout? RQ1 aims at identifying the ML algorithms that have been used to predict the customer's dropout.

Figure 3 presents the most common algorithms used to address the dropout problem in different business contexts. The algorithms based in decision trees are present in almost 17 articles, followed by logistic regression and neural networks.

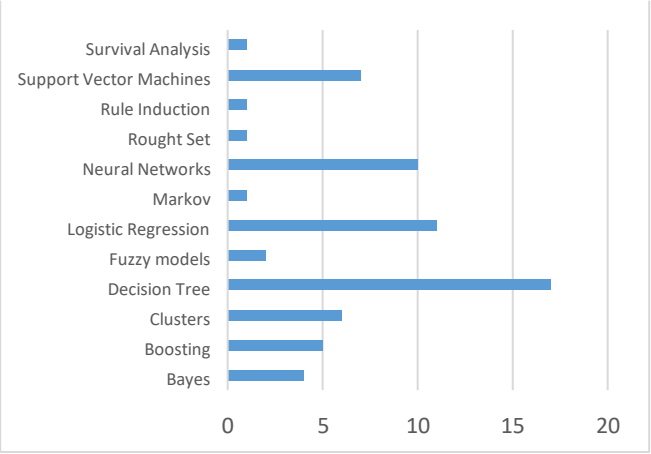

Figure 3 – Main algorithms used in the analyzed papers.

Focusing further on the citations, the five most cited studies, are the one of Runge, Gao, Garcin, & Faltings (2014) with 119 citations in total, Bi, Cai, Liu, & Li (2016) with 73 citations, Phadke, Uzunalioglu, Mendiratta, Kushnir, & Doran (2013) with 62, Perianez, Saas, Guitart, & Magne (2016) with 42 and Jinbo, Xiu, & Wenhuang (2007) with 25, as per May of 2020 according to Google Scholar.

The most cited article, by Runge et al. (2014) predict churn for high value players of casual social games and attempts to assess the business impact that can be derived from a predictive churn model—indicating that contacting players shortly before the predicted churn event improves the effectiveness of communication with players.

Bi et al. (2016) propose a new clustering algorithm exploring a case study of China Telecom. The study address also management suggestions to develop marketing strategies to ensure profit maximization.

Phadke et al. (2013) employ churn prediction algorithms based on service usage metrics, network performance indicators, and traditional demographic information. However, they develop the churn prediction based also on a social analysis of the call graph to quantify the strength of social ties between users. The Study was developed in the telecom sector.

Perianez et al. (2016) for the first time in the social games' domain, develop a survival ensemble model which provides a comprehensive analysis together with an accurate prediction of churn. Their approach predicts the probability of churning as function of time providing more accurate and more stable prediction results than traditional approaches.

Jinbo et al. (2007) employ the use of AdaBoost which is a main branch of boosting algorithms to predict the customer churn in a bank using a credit debt customer database.

The most cited articles are manly in the business sector games (two), telecommunications (two) and credit (one). Only study address the use of the survival

Weiyun Ying et al. (2008) investigate the effectiveness of the random forests approach in predicting customer churn in the banking industry.

Motahari et al. (2014) investigate various churn models profiling the customers and assigning churn probabilities to them and showing that churn prediction methods do not adequately model subscriber churn. Their research suggests that other subscribers' churn in their social network can influence the dropout. Their approach considers the influencers and their level of importance to increase the performance of churn prediction models.

Sundarkumar, Ravi, & Siddeshwar (2015) adopted the models' Decision Tree, Support Vector Machine, Logistic Regression, Probabilistic Neural Network (PNN) and Group Method of Data Handling (GMDH) in an Automobile Insurance fraud dataset and Credit card customer churn dataset is taken from literature. Creating rules using the obtained decision trees to identify groups to develop pre-emptive actions.

Halibas et al. (2019) implemented exploratory data analysis and feature engineering in a public domain Telecoms dataset. They applied seven classification techniques, namely, Naïve Bayes,

Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees. The best classifier identified was the Gradient Boosted Trees.

RQ2. When the dropout occurs? RQ2 intends to understand if the timing related to the customer dropout is considered.

To address this research question in the selected studies we analyzed the articles identifying which ones address the timings.

Only three studies addressed the timings related to the dropout (Figure 4). Liu et al. (Liu et al., 2018) determine the dropout and the average mobile game retention as a function of time, showing that 95% if the users end the relationship after 40 days. Employing the algorithms logistic regression, decision tree-based and support vector machines.

Runge et al. (2014) compare the prediction performance of four different classification algorithms and attempt to explore the temporal dynamics of time series data using a hidden markov model. Comparing its prediction performance against neural networks, logistic regression, decision tree and support vector machine. The test results indicated that contacting players shortly before the predicted churn event substantially improves the effectiveness of communication with players.

Perianez et al. (2016) model based on survival ensembles outputs accurate predictions of when players churn and provides information about the risk factors that affect the exit of players as well. This approach allowed to extract the median survival time and use as a life expectancy threshold. Using this allowed to label players as being at risk of churning, to act beforehand to retain valuable players, and ultimately improve game development to enhance player satisfaction.

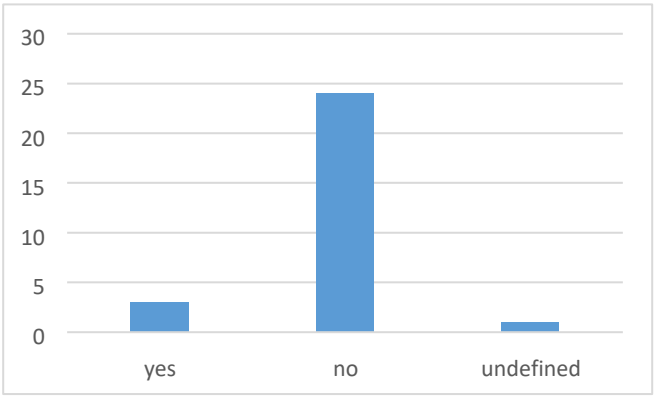

Figure 4 – The number of studies addressing the dropout timings. RQ3.

Regarding to RQ3, what are the more relevant features related to predicting customer dropout?

To answer this research, question the selected papers where reviewed looking if the articles identified relevant features to predict dropout. The number of studies that identified the relevant features is approximately 71% (Figure 5). This represents a large proportion of the studies suggesting important features that should be considering predicting dropout, using manly demographic (e.g., age or gender and behavioral data (e.g., type of service, usage, payments) related to the use of the service being purchased manly telecommunications and financial sector.

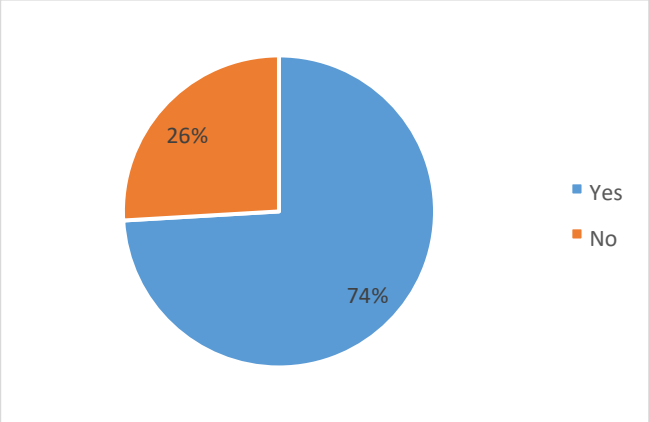

Figure 5 – Percentage of studies identifying the relevant features.

RQ4. What is the accuracy of the machine learning algorithms to predict dropout?

In order to answer this research question, the selected papers where reviewed looking if the articles identified the accuracy in the prediction of the dropout, 25 studies (Table 1). Two studies didn't address the dropout accuracy, Franciska & Swaminathan (2017) identify several clustering algorithms to predict dropout, Bandara et al. (Bandara et al., 2013) described the efforts to build successful churn prediction models highlighting their characteristics developing a survey in the telecommunication sector. The other 25 studies identify de accuracy in the prediction of the dropout representing a large majority of the studies using manly the confusion matrix.

Identify accuracy?	Articles
Yes	(Bi, Cai, Liu, & Li, 2016; Columelli, Nunez-del-Prado, & Zarate-Gamarra, 2016; Gök, Özyer, & Jida, 2015; Halibas et al., 2019; Jinbo, Li Xiu, & Wenhuang, 2007; Kayes & Chakareski, 2015; Liu et al., 2018; Manongdo & Xu, 2016; Mohanty & Rani, 2015; Motahari et al., 2014; Perianez, Saas, Guitart, & Magne, 2016; Phadke, Uzunalioglu, Mendiratta, Kushnir, & Doran, 2013; Qaisi, Rodan, Qaddoum, & Al-Sayyed, 2018; Runge, Gao, Garcin, & Faltings, 2014; Semrl & Matei, 2017; Shankar, Rajanikanth, Sivaramaraju, & Murthy, 2018; Sundarkumar, Ravi, & Siddeshwar, 2015; Wu & Li, 2018, p.; Xiao, Jiang, He, & Teng, 2016; Xie & Li, 2008; Ye & Chen, 2008; Ying, Li, Xie, & Johnson, 2008; Zhang, Qi, Shu, & Cao, 2007)
No	(Bandara, Perera, & Alahakoon, 2013; Franciska & Swaminathan, 2017)

Table 1 – Studies identifying the prediction accuracy.

The sector of telecommunications is the main researched area, followed by financial institutions (Figure 7)

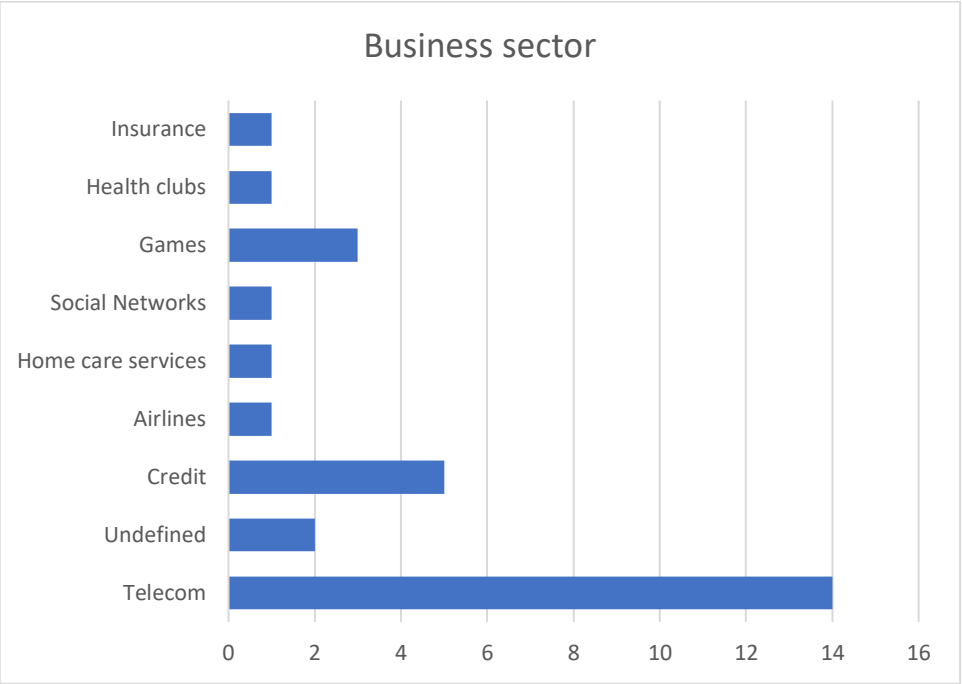

Figure 7 – The number of studies per business context.

4. CONCLUSION

The development in the adoption of machine learning techniques to predict customer dropout is using more ensemble methods integrating different approaches. The telecommunications sector is the area were being developing most of the studies, which identifies some business areas that need to be addressed.

The implementation of algorithms to predict dropout using survival analysis approaches is under-researched, only three research papers, but if we considering the number of citations for those articles this could stand for an interest using survival analysis to predict dropout (Perianez et al., 2016).

The use of algorithms to explore the timings when the dropout will occur is an approach that allow to complement the dropout prediction, giving more information to support the development of actions considering both the probability and when should be developed countermeasures to avoid the customer dropout.

REFERENCES

ASReview Core Development Team. (2019). ASReview: Active learning for systematic reviews. Utrecht, The Netherlands: Utrecht University. doi: 10.5281/zenodo.3345592

Bandara, W. M. C., Perera, A. S., & Alahakoon, D. (2013). Churn prediction methodologies in the telecommunications sector: A survey. 2013 International Conference on Advances in ICT for Emerging Regions (ICTer), 172–176. Colombo, Sri Lanka: IEEE. doi: 10.1109/icter.2013.6761174

Berry, M. J. A., & Linoff, G. (2004). Data mining techniques: For marketing, sales, and customer relationship management (2nd ed). Indianapolis, Ind: Wiley Pub.

Bi, W., Cai, M., Liu, M., & Li, G. (2016). A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn. IEEE Transactions on Industrial Informatics, 12(3), 1270–1281. doi: 10/f8swxp

Columelli, L., Nunez-del-Prado, M., & Zarate-Gamarra, L. (2016). Measuring churner influence on pre-paid subscribers using fuzzy logic. 2016 XLII Latin American Computing Conference (CLEI), 1–10. Valparaíso, Chile: IEEE. doi: 10/ggtgjr

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. Commun. ACM, 55(10), 78–87. doi: 10.1145/2347736.2347755

Edward, M., & Sahadev, S. (2011). Role of switching costs in the service quality, perceived value, customer satisfaction and customer retention linkage. Asia Pacific Journal of Marketing and Logistics, 23(3), 327–345. doi: 10.1108/13555851111143240

Fink, A. (2010). Conducting Research Literature Reviews: From the Internet to Paper. SAGE.

Franciska, I., & Swaminathan, B. (2017). Churn prediction analysis using various clustering algorithms in KNIME analytics platform. 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), 166–170. Chennai, India: IEEE. doi: 10/ggtgjp

Gök, M., Özyer, T., & Jida, J. (2015). A Case Study for the Churn Prediction in Turksat Internet Service Subscription. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15, 1220–1224. Paris, France: ACM Press. doi: 10/ggtgh9

Halibas, A. S., Cherian Matthew, A., Pillai, I. G., Harold Reazol, J., Delvo, E. G., & Bonachita Reazol, L. (2019). Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling. 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), 1–7. doi: 10/ggtgbw

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed). Amsterdam; Boston: San Francisco, CA: Elsevier; Morgan Kaufmann.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3. ed). Amsterdam: Elsevier; Morgan Kaufmann.

Jinbo, S., Xiu, L., & Wenhuang, L. (2007). The Application of AdaBoost in Customer Churn Prediction. 2007 International Conference on Service Systems and Service Management, 1–6. Changdu, China: IEEE. doi: 10/fn2m26

Kayes, I., & Chakareski, J. (2015). Retention in Online Blogging: A Case Study of the Blogster Community. IEEE Transactions on Computational Social Systems, 2(1), 1–14. doi: 10/ggtgjt

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (1 edition). Cambridge, Massachusetts: The MIT Press.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing structural literature reviews in software engineering (pp. 1–26) [Joint technical report]. Australia: Keele Univ., and Empirical Software Eng., Nat'l ICT.

Kitchenham, B., Pretorius, R., Budgen, D., Pearl Brereton, O., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering – A tertiary study. Information and Software Technology, 52(8), 792–805. doi: 10.1016/j.infsof.2010.03.006

Liu, X., Xie, M., Wen, X., Chen, R., Ge, Y., Duffield, N., & Wang, N. (2018). A Semi-Supervised and Inductive Embedding Model for Churn Prediction of Large-Scale Mobile Games. 2018 IEEE International Conference on Data Mining (ICDM), 277–286. Singapore: IEEE. doi: 10/ggtgh8

Manongdo, R., & Xu, G. (2016). Applying client churn prediction modeling on home-based care services industry. 2016 International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), 1–6. Durham, NC, USA: IEEE. doi: 10/ggtgjj

Mohanty, R., & Rani, K. J. (2015). Application of Computational Intelligence to Predict Churn and Non-Churn of Customers in Indian Telecommunication. 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 598–603. doi: 10/ggtgb3

Motahari, S., Jung, T., Zang, H., Janakiraman, K., Li, X.-Y., & Hoo, K. S. (2014). Predicting the influencers on wireless subscriber churn. 2014 IEEE Wireless Communications and Networking Conference (WCNC), 3402–3407. Istanbul, Turkey: IEEE. doi: 10/ggtgjf

Perianez, A., Saas, A., Guitart, A., & Magne, C. (2016). Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 564–573. Montreal, QC, Canada: IEEE. doi: 10/ggtgjh

Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Malden, MA; Oxford: Blackwell Pub.

Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D., & Doran, D. (2013). Prediction of Subscriber Churn Using Social Network Analysis. Bell Labs Technical Journal, 17(4), 63–75. doi: 10/ggtgjq

Qaisi, L. M., Rodan, A., Qaddoum, K., & Al-Sayyed, R. (2018). Customer churn prediction using data mining approach. 2018 Fifth HCT Information Technology Trends (ITT), 348–352. doi: 10/ggtgb4

Reichheld, F. F. (1996, March 1). Learning from Customer Defections. Harvard Business Review, (March–April 1996). Retrieved from https://hbr.org/1996/03/learning-from-customer-defections

Runge, J., Gao, P., Garcin, F., & Faltings, B. (2014). Churn prediction for high-value players in casual social games. 2014 IEEE Conference on Computational Intelligence and Games, 1–8. Dortmund, Germany: IEEE. doi: 10/ggtgjk

Semrl, J., & Matei, A. (2017). Churn prediction model for effective gym customer retention. 2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC), 1–3. doi: 10.1109/BESC.2017.8256379

Shankar, R. S., Rajanikanth, J., Sivaramaraju, V. V., & Murthy, K. V. S. S. R. (2018). PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING. 2018 IEEE International Conference on System, Computation, Automation and Networking (Icscan), 1–8. Pondicherry: IEEE. doi: 10/ggtgjs

Schober, P., & Vetter, T. R. (2018). Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesthesia and Analgesia, 127(3), 792–798. doi: 10.1213/ANE.0000000000003653

Sheth, J. N., Mittal, B., & Newman, B. (1998). Customer Behavior: Consumer Behavior and Beyond (1 edition). Fort Worth, TX: South-Western College Pub.

Sundarkumar, G. G., Ravi, V., & Siddeshwar, V. (2015). One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 1–7. Madurai, India: IEEE. doi: 10/ggtgjb

Wu, L., & Li, M. (2018). Applying the CG-logistic Regression Method to Predict the Customer Churn Problem. 2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS), 1–5. doi: 10/ggtgb5

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(3), 2354–2364. doi: 10.1016/j.eswa.2010.08.023

Xie, Y., & Li, X. (2008). Churn prediction with Linear Discriminant Boosting algorithm. 2008 International Conference on Machine Learning and Cybernetics, 228–233. Kunming, China: IEEE. doi: 10/ftr6fz

Ye, D., & Chen, Z. (2008). A rough set based minority class oriented learning algorithm for highly unbalanced data sets. 2008 IEEE International Conference on Granular Computing, 736–739. Hangzhou: IEEE. doi: 10/d4qj3t

Ying, W., Li, X., Xie, Y., & Johnson, E. (2008). Preventing customer churn by using random forests modeling. 2008 IEEE International Conference on Information Reuse and Integration, 429–434. Las Vegas, NV, USA: IEEE. doi: 10/csjz8s

Zhang, Y., Qi, J., Shu, H., & Cao, J. (2007). A hybrid KNN-LR classifier and its application in customer churn prediction. 2007 IEEE International Conference on Systems, Man and Cybernetics, 3265–3269. Montreal, QC, Canada: IEEE. doi: 10/fqhd5m