# Dropout Prediction: A Systematic Literature Review

CAPSI2021

Pedro Sobreiro, Domingos Martinho, Javier Berrocal, & José Garcia Alonso

University of Extremadura *pdealexa@alumnos.unex.es*

21º Conferência da Associação Portuguesa de Sistemas de Informação, 15-10-2021

# Summary

1. Introduction
2. Methodology
3. Results
4. Conclusion

# Some background. . .

- Customer analysis is fundamental to develop business and marketing intelligence (Sheth, Mittal, & Newman, 1998), which is also know as datamining (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);
- Machine learning allow the extraction of knowledge to understand dropout and effective retention strategies (Verbeke, Martens, Mues, & Baesens, 2011);

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

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

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.

# Some background. . .

- The costs of retaining customers are lower than attracting new ones (Edward & Sahadev, 2011) and reducing it 5% (e.g., from 15% to 10% per year) could could represent an increase in profits up to the double (Reichheld, 1996);
- A customer that dropout represents a loss of money, if an organization can predict the dropout is possible to develop counter measures to avoid it;
- This study analyses state of the art and identifies Machine Learning studies to predict customer dropout using a systematic literature review as advocated by Kitchman & Charters (Kitchenham & Charters, 2007);

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

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

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

# Research questions

- What are the trends in machine learning algorithms to predict dropout? This question aims at identifying the ML techniques that have been used to predict the customer's dropout
- When the dropout occurs? This intends to understand if the timing related to the customer dropout is considered.
- What are the more relevant features related to predicting customer dropout?
- What is the accuracy of the machine learning algorithms to predict dropout?

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

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

# Research methodology

- Was developed a Systematic Literature Review (SLR) which according to Fink (Fink, 2010) is a reproducible method for identifying, evaluating and synthesize existing body of knowledge developed by researchers;
- SLR followed the recomendations of Kitchenham & Charters (Kitchenham & Charters, 2007) in three stages: Planning, Implementation and Results;
- The search strategy was based in the Population, Intervention, Comparison, Outcomes and Context (PICOC) as suggested Kitchenham and Charters (Kitchenham & Charters, 2007);

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

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

# Research methodology

- The search criteria was ((“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 was Books, Non-English articles, patents, and thesis;
- A total of 218 studies were found in the first step and the selection process 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;

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

# Research methodology

- After we started the reviewing process labelling the subsequent papers as irrelevant or relevant until ASReview started suggesting only irrelevant papers

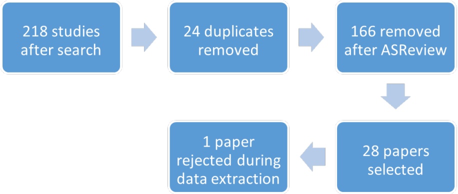

218 studies after search
24 duplicates removed
166 removed after ASReview
28 papers selected
1 paper rejected during data extraction

Figure: Filtering process

# Research methodology

- 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 in each research question;
- The selected papers were reviewed to answer the quality questions;

# Research results

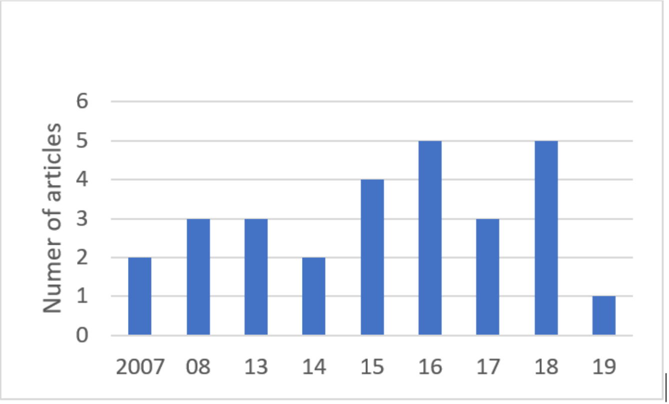


Figure: Articles per year after quality assessment

# Research results - Machine Learning Algorithms

- Addressing the research question ”trends in machine learning algorithms to predict dropout” were identified common algorithms used to address the dropout problem in different business contexts;

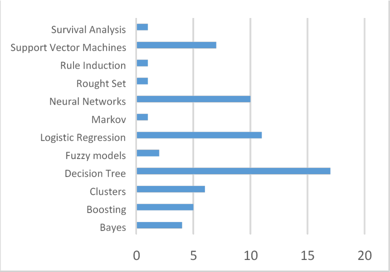


Figure: Main algorithms used in the research papers

# Research results - Machine Learning Algorithms

- 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 Google Scholar
- Runge et al. (2014) predict churn for high value players of casual social games and Bi et al. (2016) propose a new clustering algorithm exploring a case study of China Telecom;
- Phadke et al. (2013) developed a churn prediction using social analysis quantifying the strength of social ties between users;

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

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

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

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

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

# Research results - When dropout occurs?

- Only three studies addressed the timings related to the dropout;
- Were used average retention time, times series and survival ensemble;

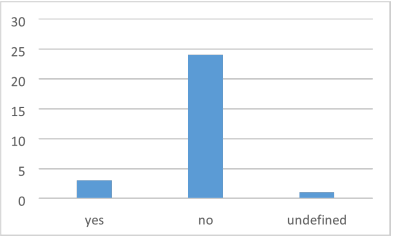
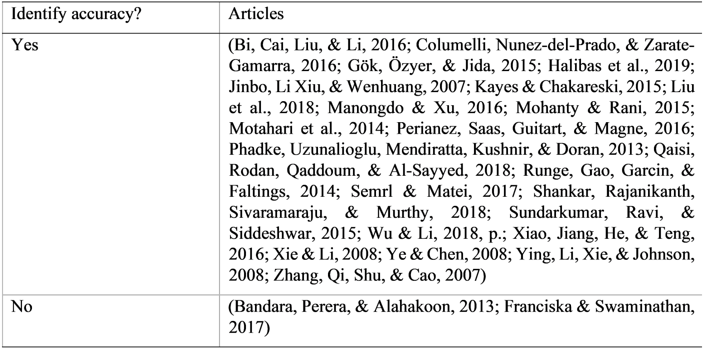


Figure: studies addressing the dropout timings

# Research results - Accuracy in the prediction?

- 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



| 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) |

Figure: studies addressing the prediction accuracy

# Research results - Studies by business area

- The sector of telecommunications is the main researched area, followed by financial institutions

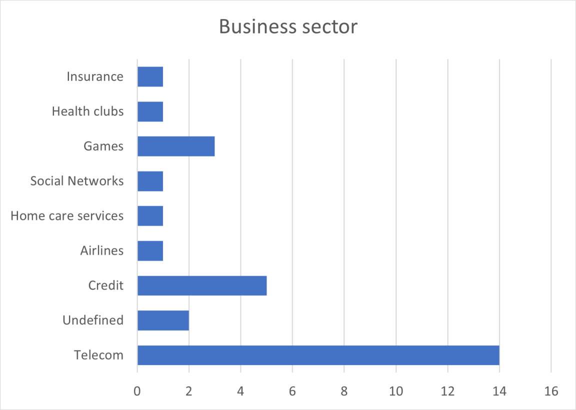


Figure: Number of studies by business area

# Conclusion

- The adoption machine learning techniques to predict customer dropout is using manly ensemble methods integrating different approaches;
- The implementation of algorithms to predict dropout using survival analysis approaches is underresearched, only three research papers, but if is considered the number of citations there are interest in the researchers;
- 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;

# Thanks!

Start where you are. Use what you have. Do what you can. **Arthur Ashe**