Dropout Prediction: A Systematic Literature Review

Preliminar results

Pedro Sobreiro, Javier Berrocal (Thesis Supervisor), José Garcia Alonso (Co-supervisor),

Webinar, 26 June 2020

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

Summary

Research goals

Research goals

What is the current state of machine learning research studies to predict dropout in contractual settings?

Introduction

Data mining

- 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)
- According to Han, Kamber, and Pei (2012), these tasks present many similarities between data mining and machine learning

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.

Data mining

- Machine learning could be used to extract knowledge to understand dropout with the development of effective retention strategies (Verbeke, Martens, Mues, & Baesens, 2011)
- Machine learning algorithms have been used to predict customer dropout (Bandara, Perera, & Alahakoon, 2013), without however to consider the timings of the dropout
- Machine learning can be used to develop of customer retention strategies based on existing data (Verbeke et al., 2011), extracting patterns from data (Kelleher et al., 2015), that support the development of counteractions before an event occurs

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

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. doi: 10/ggtgjg

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

Contractual settings

- The identification of the dropout can be developed in different contexts: customers that buy in contractual settings and non-contractual settings where a firm have to infer if the customer is still active (Gupta et al. ,2006)
- The main characteristic of a contractual setting is a contact of the customer cancelling a subscription (Fader & Hardie, 2007);
- This research analyses state of the art and identifies Machine Learning studies to predict customer dropout.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., ... Sriram, S. (2006). Modeling Customer Lifetime Value. Journal of Service Research, 9(2), 139–155. doi: 10.1177/1094670506293810

Fader, P. S., & Hardie, B. G. S. (2007). How to project customer retention. Journal of Interactive Marketing, 21(1), 76–90. doi: 10.1002/dir.20074

Methodology

Methodology

- Was developed a Systematic Literature Review (SLR) in three stages (Kitchenham & Charters, 2007): Plan, Conduct and Report;
- Plan: definition of the research need, identification of the research questions and the development of the review protocol;
- Conduct: research identification, study selections, quality assessment, data extraction, finishing with the data synthesis;
- Report: stage that develops the activity report review.

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.

Systematic Literature Review Phases

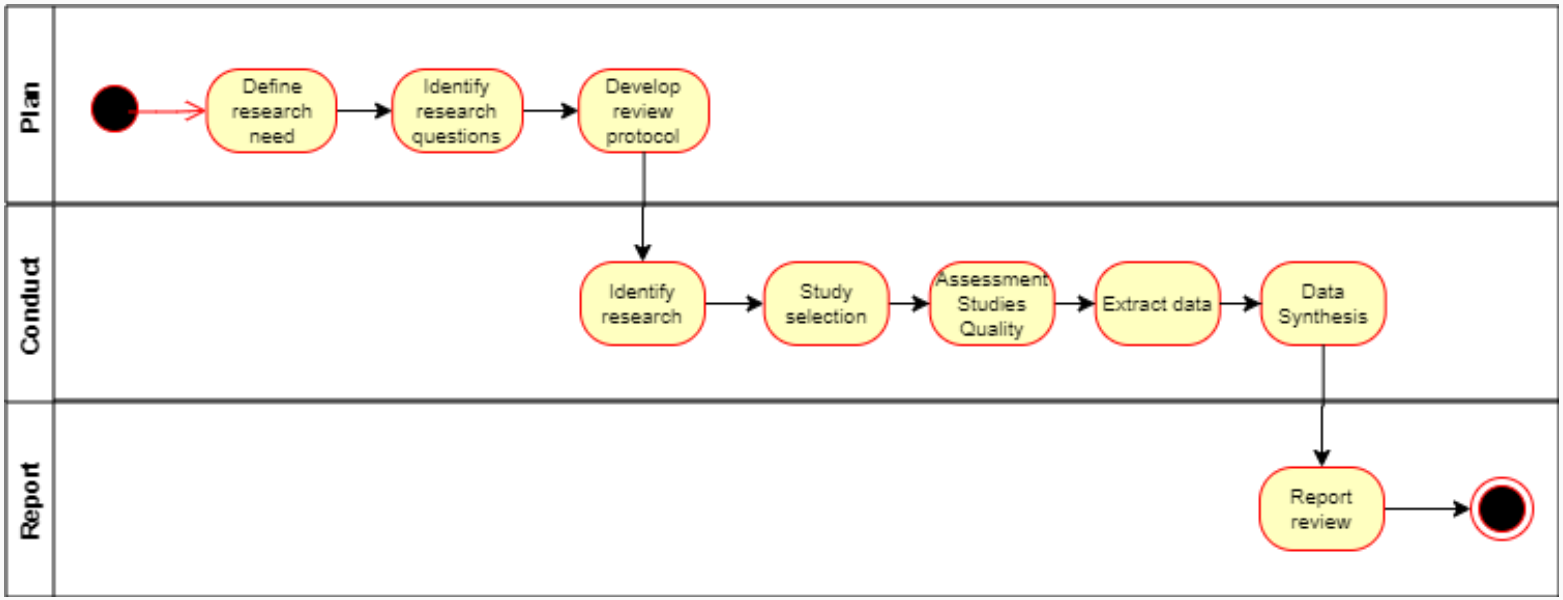

Figure 1: SLR phases based on Kitchenham and Charter (2007)

Research questions

What is the current state of machine learning research studies to predict dropout in contractual settings? Based in this question were identified the following questions:

- RQ1: What studies have been published?;
- RQ2: Which algorithms have been used to predict the dropout?
- RQ3: What are the more relevant features related to predicting customer dropout?
- RQ4: When the dropout occurs?
- RQ5: What is the accuracy of the machine learning algorithms to predict dropout?

Population, Intervention, Comparison, Outcomes and Context

Table 1: PICOC criteria

PICOC	Description
Population	Research papers about dropout with contractual settings
Intervention	Machine learning algorithms to predict dropout
Comparison	Studies addressing machine learning algorithms to predict dropout
Outcome	Synthesis identifying research questions, gaps in the research domain and also best practices identified
Context	Academia and industry

Note: Context (PICOC) as suggested Kitchenham and Charters (2007) and proposed by Petticrew and Roberts (Petticrew & Roberts, 2006) to support the development of the search string.

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.

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

Search

- Search string: ((“customer dropout”) OR (“customer churn”) AND “machine learning” AND (“contractual” OR “membership”));
- Applied to the title, abstract, and keywords in the search period between January 2000 and June 2020
- The exclusion criteria were Books, Non-English articles, patents, and thesis
- 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 (pp. 1–26) [Joint technical report]. Australia: Keele Univ., and Empirical Software Eng., Nat I ICT.

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

Results

Results IEEE

- 218 studies were found in the first step of the conduct (Identify research);
- 24 duplicates were removed
- 166 were removed after ASReview
- 1 paper rejected during data extraction
- 28 papers selected

RQ1. What studies have been published?

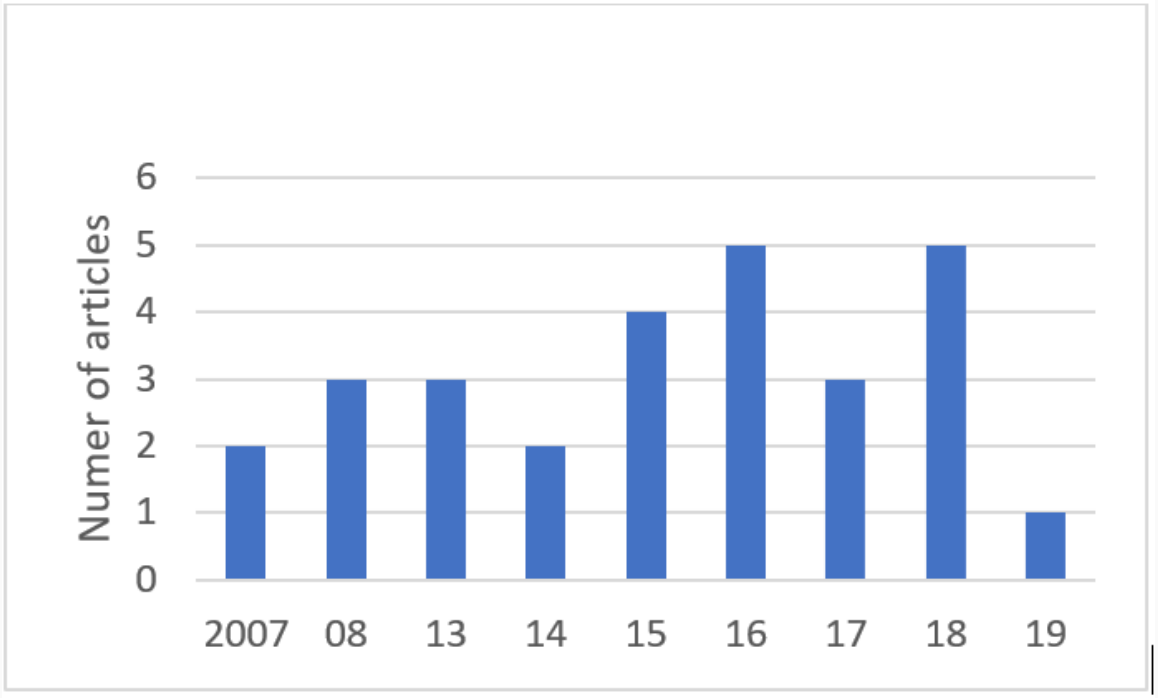

Figure 2: Articles per year after quality assessment

RQ1. What studies have been published?

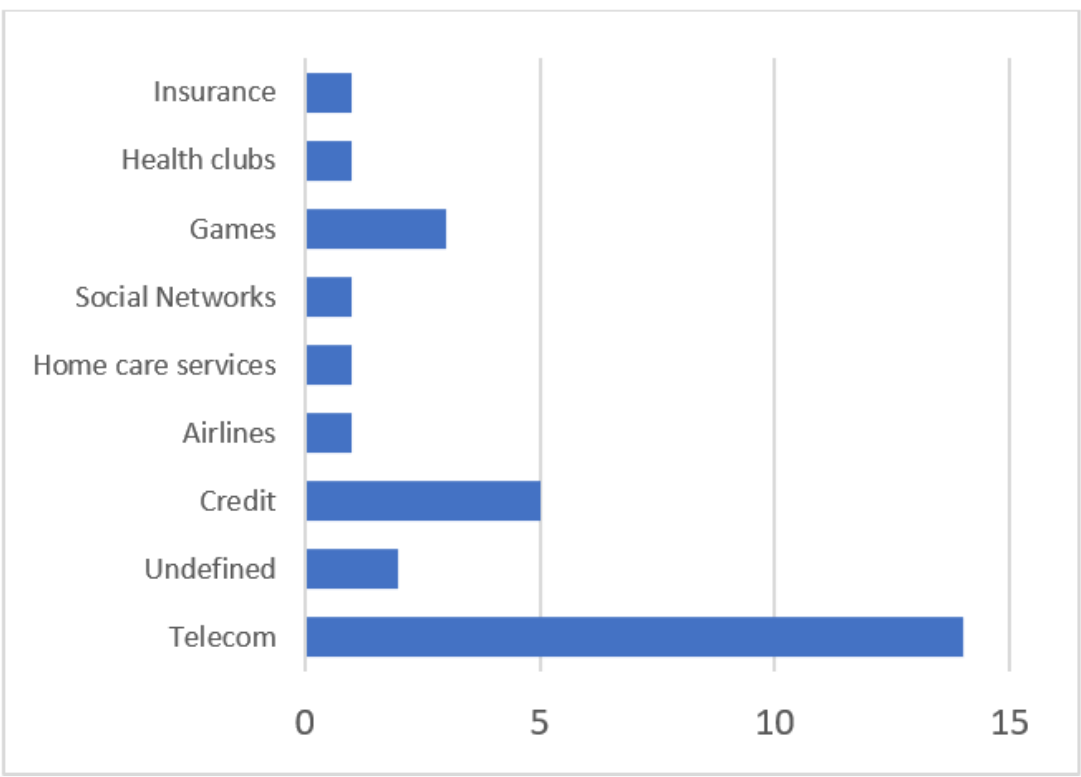

Figure 3: The number of studies per business context

RQ2: Which algorithms have been used to predict the dropout?

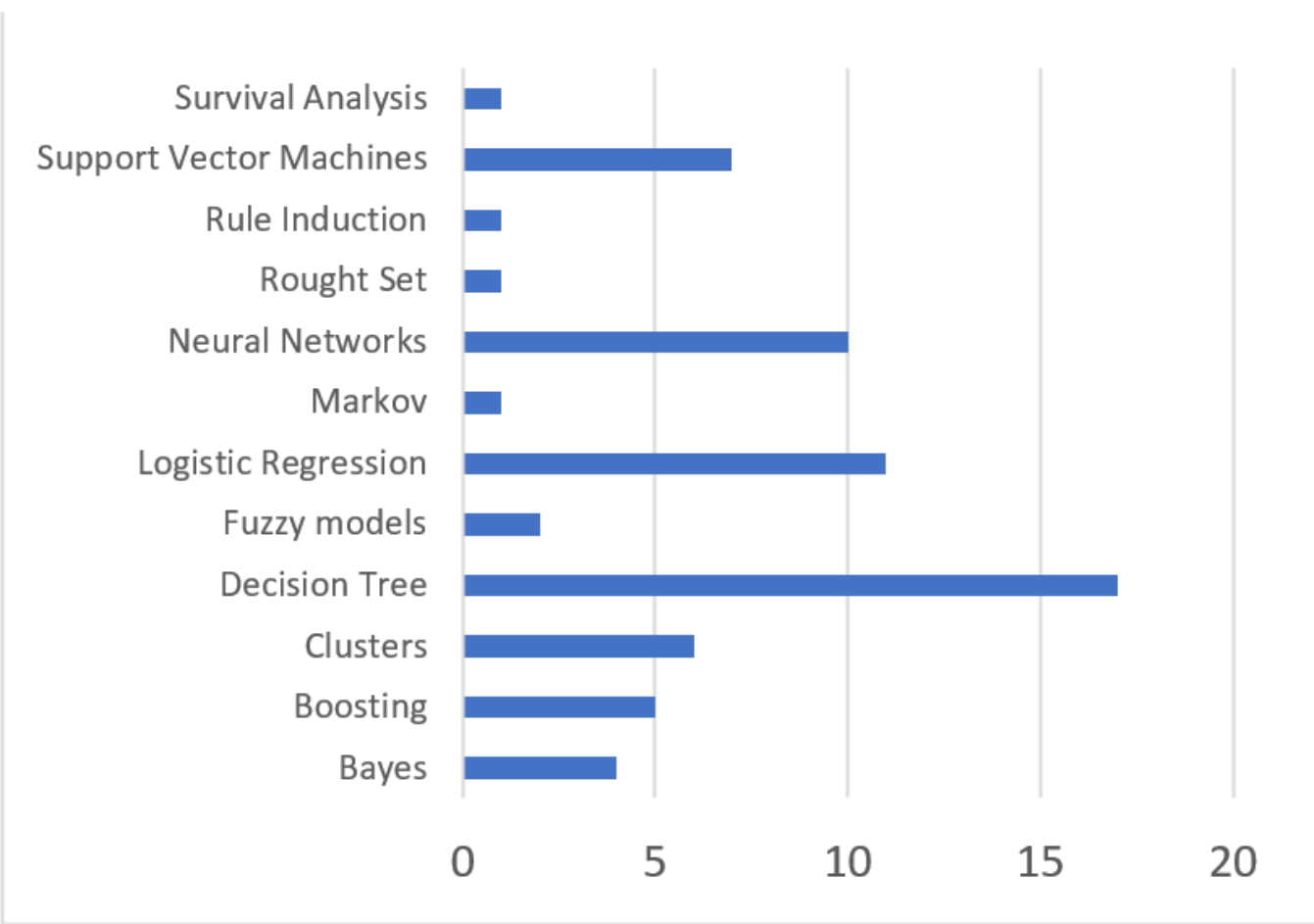

Figure 4: Main algorithms used in the analysed papers

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

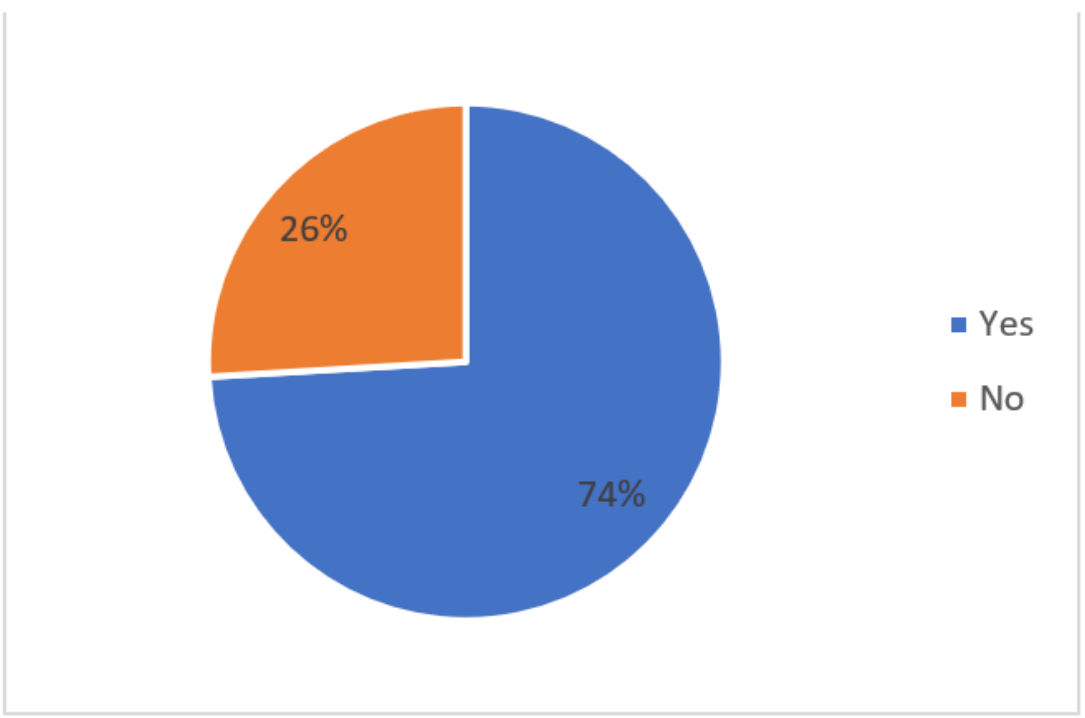

Figure 5: Percentage of studies identifying the relevant features

RQ4: When the dropout occurs?

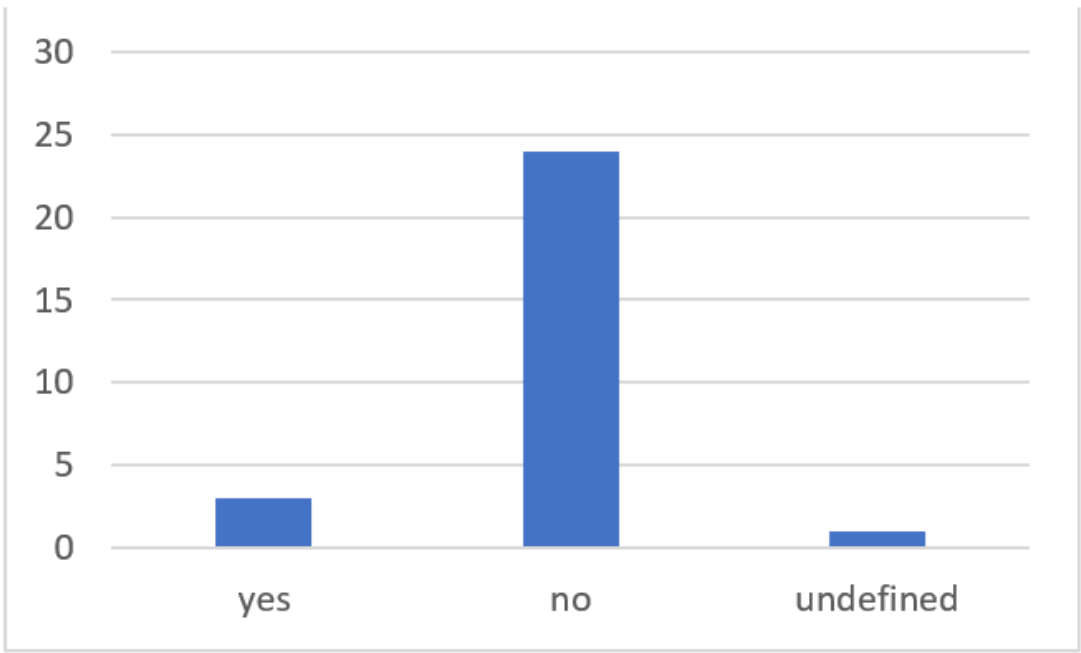

Figure 6: The number of studies addressing the dropout timings

RQ5: What is the accuracy of the machine learning algorithms to predict dropout?

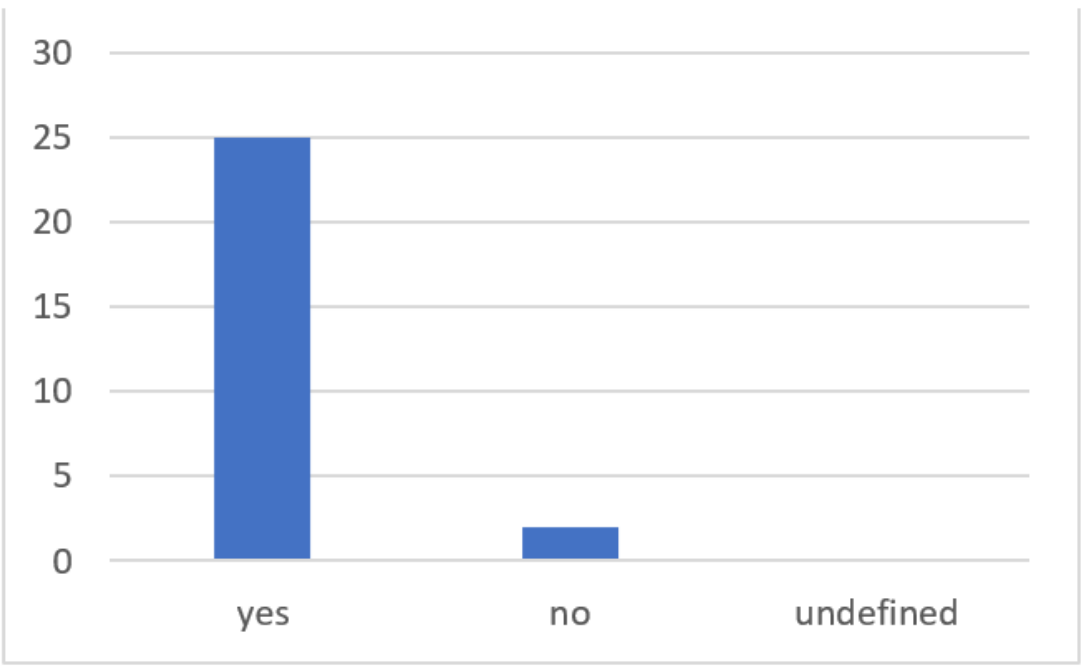

Figure 7: Number of studies identifying the prediction accuracy

Conclusion

Conclusion

- The telecommunications sector is the area where are being developing most of the studies, which identifies some research areas gaps that need to be addressed;
- Algorithms to predict dropout using also survival analysis approaches is an area under researched, only three research papers, however considering the number of citations these approaches getting the attention (Perianez et al., 2016);
- The use of algorithms to explore the timings when the dropout will occur is an approach that could complement the dropout prediction, supporting the development of actions considering both the probability and when should be developed countermeasures to avoid the customer dropout

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. doi: 10/ggtgjh

Thanks!

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