Prediction customer dropout using machine learning

Current situation

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

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

Webinar, 11 September 2020

Summary

Research goals

This thesis focuses in applying machine learning techniques aiming to find answers to the following questions:

- Q1: Which are the patterns related to dropout?
- Q2: There are temporal patterns related to the dropout?
- Q3: Is possible to increase customer lifetime value employing machine learning techniques?

These questions are supported in a Systematic Literature Review, establishing the state of the art in this research area. Our aim is to identify information that can be employed to increase their Customer Lifetime Value, which allows us to develop insights to support retention systems to reduce dropout in customers in a contractual setting.

Research plan

1. Developing the systematic literature review supporting the thesis: "Dropout Prediction: A Systematic Literature Review" (in progress)
2. Exploring existing machine learning techniques to predict dropout (supported also in other developments already executed)
3. Creation of an ensemble method to improve dropout prediction accuracy
4. Developing article addressing an ensemble method using a case study
5. Finishing and thesis delivery

Research plan

Name
Developing the systematic literature review supporting the thesis
Exploring existing machine learning approaches to predict dropout
Creation of an ensemble method to increase dropout prediction accuracy
Developing article addressing an ensemble method
Finishing and thesis delivery

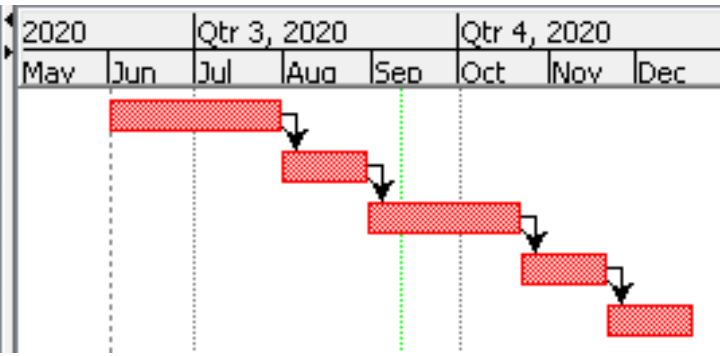

Figure: Activities timeline

Introduction

Why this research?

- 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.

Introduction

Why this research?

- Machine learning could be used to extract knowledge to understand dropout for 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

Introduction

Why this research?

- 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 canceling a subscription (Fader & Hardie, 2007);
- Customer dropout prediction should consider the context, where there is a contractual or non-contractual setting.

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

Introduction

Why this research?

- It is also know, that the costs of retaining customers are lower when compared to the costs of attracting new ones (Edward & Sahadev, 2011), reinforced by that the reduction of the dropout rates could represent an increase in the profits (Reichheld, 1996);
- Machine learning algorithms have been used to predict customer dropout (Bandara, Perera, & Alahakoon, 2013);
- But, to our knowledge, there is a lack of an overview of research related to the use of machine learning techniques to target customer dropout with contractual settings considering also the timings of the dropout.

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

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

Systematic Literature Review

- 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

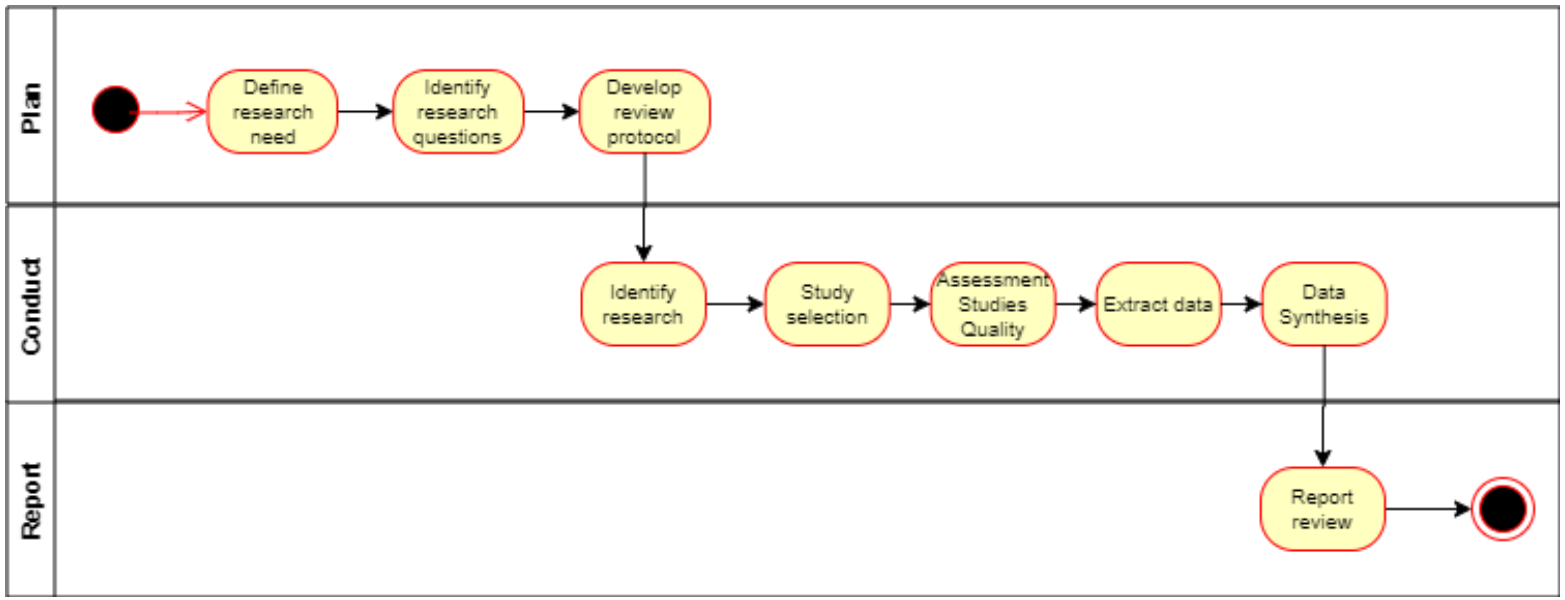

Figure: SLR phases based on Kitchenham and Charter (2007)

Research questions

What is the current state of machine learning research in existing 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? This question will also address algorithms used by business area (as suggested in activity 6)
- RQ3: What are the more relevant features to predict 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: 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;
- Sources SpringerLink, Scopus, Science@Direct, ISI Web of Science, IEEE Digital Library, and ACM Digital Library;
- The selection process using the abstract was developed with ASReview (van de Schoot et al., 2020), creating a dataset of the identified articles based in the identification of at least five relevant papers and five irrelevant papers to train a Machine Learning algorithm;

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., ... Oberski, D. (2020). ASReview: Open Source Software for Efficient and Transparent Active Learning for Systematic Reviews. arXiv:2006.12166 [cs]. Retrieved from http://arxiv.org/abs/2006.12166

Search results

Table: Articles identified in the initial dataset

Source	Articles
Scopus	210
IEEE	20
SpringerLink	79
Science Direct	126
ISI Web of Knowledge	6
ACM	8
Total	449

Note: After the initial dataset where removed 20 incomplete entries, 16 duplicates and 335 papers that don't addressed the research topic after abstract analysis. The next steps related to the quality assessment of the 78 studies remaining are being developed.
The steps developed and the data used for the SLR are available in github here and the script in R to process the initial dataset here.

Exploring existing machine learning techniques

- Was tested class train/test stratification in the target variable, imbalance datasets, grid search optimization targeting AUC for classification (dropout,non-dropout);
- The class imbalance approaches to adjust the weights inversely proportional to class frequencies in the input data using the library scikit-learn (Pedregosa et al., 2011);
- Hyper-parameters optimization developed using grid search targeting AUC as the optimization goal considering the discriminatory power (Emeterio et al., 2016);
- Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and Gradient Boosting Classifier (GBC) best performances in GBC (accuracy, sensitivity, precision, F1 Score) and RFC in AUC;
- Code book, dataset and Jupyter notebook available in github here.

Exploring existing machine learning techniques

1732 *Journal of Reviews on Global Economics*, 2019, *8*, 1732-1740

Predicting High-Value Customers in a Portuguese Wine Company

Pedro Sobreiro[1], Domingos Martinho[2,*], António Pratas[2], Jose Garcia-Alonso[3] and Javier Berrocal[3]

[1]*Escola Superior de Desporto de Rio Maior, Instituto Politécnico de Santarém, Portugal*

[2]*ISLA Santarém, Portugal*

[3]*Quercus Software Engineering Group, University of Extremadura, Spain*

Abstract: Wine companies operate in a very competitive environment in which they must provide better-customised services and products to survive and gain advantage. The high customer turnover rate is a problem for these companies. This work aims to provide wine companies with new knowledge about customers that help to retain the existing ones. The study applies a collected dataset from a transaction database in a medium-sized Portuguese wine company to determinate: (1) customer lifetime value; (2) cluster customer value as output (customer loyalty). The measurement of the customer lifetime value (CLV) was analysed using the Pareto/NBD model and gamma-gamma model. Clustering techniques are employed to segment customers according to Recency, Frequency, and Monetary (RFM) values. Study findings show that exists three clusters with different interest to the marketing strategies, identifying the high-value customers, to target using marketing to increase their lifetime value effectively. The implications for the marketing strategy decisions is that using techniques based on the RFM model can make the most from data of customers and transactions databases and thus create sustainable advantages.

Keywords: Customer lifetime value, clustering, wine marketing, RFM model.

Figure: Publication using an technique for research goal (Q3) "Is possible to increase customer lifetime value employing machine learning techniques?"

Case study

- Dataset for the case study will use data from Portuguese Software company that provides solution for sport clubs using a business model of a software as a service (SaS);
- They want to improve their software to support the customer retention of organizations using their solution;
- This case study support the achievement of our thesis goals Q1 (Patterns related to dropout), Q2 (Temporal patterns related to dropout) and Q3 (Increase the lifetime value).

Research outcomes

- SLR is in progress and is expected to produce an article submitted to a indexed Journal addressing existing research to predict dropout of customer with contractual settings until the end of this year;
- Publication of an article addressing a case study of an ensemble method to create an approach for dropout prediction, selecting the most relevant features, to support lines of action that managers and marketing researchers can employ, and allowing the identifications of the timings related to the dropout;
- Thesis wrapping SLR, existing approaches to predict dropout and a case study in the context of the research being developed;
- The potential research direction is to investigate how the dropout prediction can increase the Customer Lifetime Value using existing information about customers and historical data.

Thanks!

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