Received: 29 November 2019 | Revised: 3 April 2020 | Accepted: 23 June 2020

DOI: 10.1111/exsy.12617

SPECIAL ISSUE PAPER

Expert Systems WILEY

A bi-objective procedure to deliver actionable knowledge in sport services

Paulo Pinheiro[1] | Luís Cavique[2,3]

[1]CEDIS, Lisbon, Portugal

[2]MAS-BioISI, FCUL, Lisbon, Portugal

[3]Universidade Aberta, Lisbon, Portugal

Correspondence
Luís Cavique, MAS-BioISI, FCUL, Lisbon, Portugal.
Email: luis.cavique@uab.pt

Abstract

The increase in retention of customers in gyms and health clubs is nowadays a challenge that requires concrete and personalized actions. Traditional data mining studies focused essentially on predictive analytics, neglecting the business domain. This work presents an actionable knowledge discovery system that uses the following pipeline (data collection, predictive model and retention interventions). In the first step, it extracts and transforms existing real data from databases of the sports facilities. In the second step, predictive models are applied to identify user profiles more susceptible to dropout, where actionable withdrawal rules are based on actionable attributes. Finally, in the third step, based on the previous actionable knowledge, some of the values of the actionable attributes should be changed in order to increase retention. Simulation of scenarios is carried out, with test and control groups, where business utility and associated cost are measured. This document presents a bi-objective study in order to choose the more efficient scenarios.

KEYWORDS

actionable knowledge, business utility, retention intervention, sport services

1 | INTRODUCTION

The promotion of physical activity as a means to prevent increasing rates of obesity and maintenance of well-being has provoked a proliferation of gyms and health clubs that compete with public sports facilities, and consequently there has been an increasing pressure on the providers to maintain competitive advantage through services that provide higher levels of customer satisfaction (Howat & Assaker, 2016). In the context of great supply, the sports services sector is characterized by a high dropout rate (Avourdiadou & Theodorakis, 2014), and the largest dropout is in health clubs and gyms where the most promoted activity is fitness, than in sports facilities that have other facilities, namely swimming pools (Frota, 2011).

Some gyms promote activities in which users subscribe continuously, becoming withdrawn when they fail to pay their monthly subscription or by the customer initiative. Other facilities work during sports seasons, and they need to encourage their users at the end of each season to renew their inscriptions for the following season. In both cases, with monthly subscription or seasonal subscription, the increase in retention is nowadays a challenge that requires concrete and personalized actions, in addition to generic actions to improve the quality of services and facilities.

The widespread use of ERP and access control systems by these sports facilities nowadays enables a collection of quality data that allow us to find and measure user preferences and behaviour from admission to abandonment. We intend to use the existing data from those systems to get actionable knowledge that allows the identification of users of regular sports services at risk of moving out and those who keep being loyal to the service. The use of machine learning techniques, namely with decision trees, is able to generate rules that allow to predict the change of behaviour of the dropout group.

In this context, the concept of actionable knowledge can be successfully applied, since the goal is not focused only on the predictive algorithms but in solving business problems (Cao, 2010).

This work extends the previous works of Pinheiro and Cavique (2015, 2018, 2019a, 2019b) and complements the survival analysis studies proposed by Sobreiro (Sobreiro, Pinheiro, & Santos, 2018), (Sobreiro et al., 2019). These works fill a gap in the study of retention through the use of machine learning techniques in regular sports services. The novelty of this work includes not only the creation of knowledge rules, but also the bi-objective study of scenarios in order to estimate the incremental effect of a retention campaign.

In this study, we use the words "action," "treatment," "test," "intervention" and "experiment" as synonyms. Moreover, "incremental effect", "causal effect" and "impact" have similar meanings.

This document has the following structure. In Section 2, the related work in sports service retention, actionable knowledge and causality is presented. In Section 3, we present the proposed method in three steps: the data collection, the predictive model and finally the application of retention interventions. In Section 4, we study different retention scenarios regarding the optimization of the business utility measure. Finally, in Section 5, we draw the conclusions of this work.

2 | RELATED WORK

In this document, we develop a bi-criteria procedure to deliver actionable knowledge in sports services to avoid dropout. In this section, first, we present the related work in sports services retention. Then, we introduce actionable knowledge, highlighting the difference between data-driven and domain-driven approaches. Causality concepts are mentioned in order to better understand the way how knowledge rules can be operationalized. Bi-criteria decision-making concepts are presented to close the partial views of the problem. Finally, the subject of this document, the actionable knowledge in sports services, is reported.

2.1 | Sport service retention

According to different authors and societies like the International Health, Racquet and Sports club Association (IHRSA), almost 60% of people who join a club continue to be a member after more than 1 year (Bedford, 2009; IHRSA, 2012; San-Emeterio, Iglesias-Soler, Gallardo, Rodriguez-Cañamero, & García-Unanue, 2016). In other words, the annual dropout is around 40%, which reveals the importance of customer retention strategies.

Frota (2011) studied the retention management based on information from associations of mature markets like the IHRSA from USA and Fitness Industry Association (FIA) from UK. He argues that 25% of withdrawals are motivated by club-related issues, of which 45% are recoverable; 22% are related to money, of which 31% are recoverable; 29% are related to situational problems, of which 44% are recoverable; and finally, 24% leave for personal reasons, and the recovery of these is extremely difficult. Doing the calculation, we can conclude that about 30% of the annual dropouts are recoverable. These recoveries can be made through the implementation of loyalty actions that are preferably carried out before the withdrawal takes place. The solution to the problem is to answer the following questions: who will be the target of these actions and what actions should be developed to avoid the dropout.

As already used by some authors (Bedford, 2009; Sobreiro et al., 2019), survival analysis seeks to relate the duration of enrollment with the likelihood of dropping out. Instead, we seek to answer the first question, using databases from CRM applications, to build a decision tree from a set of examples (customers) described by a rich set of attributes including customer personal information (such as name, gender, birthday), financial information (such as yearly income), family information (such as lifestyle, number of children), and so on. Because decision trees can be converted to rules for explicit representation of the classification, one can easily obtain the characteristics of customers belonging to a certain class (such as loyal customer or churner) (Yang, Yin, Ling, & Pan, 2007).

Once one tries to define the profiles of behaviours that lead with churn, it is necessary to find the characteristics or attributes that somehow allow tracing those profiles. Work related to retention in sports services (Avourdiadou & Theodorakis, 2014; Frota, 2011; Gonçalves, 2012; Howat & Assaker, 2016; Surujlal & Dhurup, 2012) allows systematizing and identifying attributes necessary to characterize users and their behaviour, both those who continue to use the services and those who leave.

For the second question, Gorgoglione (2011) identifies five possible approaches for the creation of personalized actions: the computational approach, the similarity approach, the bottom-up approach, the top-down approach and the personalized approach. The computational approach generates actions based on the profiles of clients, with no human intervention. The similarity-based approach is used by recommendation systems and web content personalization methods; this type of approach assumes that actions are related to customer preferences, can be inferred through customer profiles, and it is assumed that similar customers behave similarly and similar actions cause similar reactions. The bottom-up approach includes the methods of knowledge discovery and is implemented in two separate steps: by creating customer profiles and deciding what actions are appropriate. The top-down approach consists of the same two steps of the bottom-up approach, but the decision on which actions to implement is taken before defining customer profiles. Finally, the personalized approach offers customers a number of different options, being at their discretion to choose which they prefer.

2.2 | Actionable knowledge

Traditional data mining studies concentrated primarily on predictive mining, where the cause and effect scenarios are described. But this information alone is not sufficient as it does not provide much benefit to the final user. What becomes more interesting and critical to organizations is to mine patterns in order to create knowledge actionable (Cao, 2010).

As Yang et al. (2007) say, a common problem in current applications of data mining, particularly in intelligent CRM, is that people tend to focus on, and be satisfied with, building up the models and interpreting them, but not to use them to get profit explicitly. More specifically, most data mining algorithms (predictive or supervised learning algorithms) only aim at the construction of customer profiles, which predicts the characteristics of customers of certain classes. This knowledge is useful but it does not directly benefit the enterprise. To improve customer relationship, the enterprise must know what actions to take to change customers from an undesired status (such as churner) to a desired one (such as loyal customers).

Knowledge is considered actionable if users can take direct actions based on such knowledge to their advantage. Actionability should be a criterion that can measure the utility of the mined patterns. Among the most important and distinctive actionable knowledge are actionable behavioural rules that can directly and explicitly suggest specific actions to take to influence (retain and encourage) the behaviour of customers (Su, Zhu, & Zeng, 2014).

To face the increasingly complex challenges of data mining in real-life world problems, Cao (2007, 2010) presents a new approach, which opposes data-driven to domain-driven. Data-driven corresponds to the traditional data mining, while domain-driven is related to the business domain, or business area. The domain-driven data mining (D^3M) close the gap between researchers and practitioners, by generating actionable knowledge for real user needs.

The D^3M evaluates a pattern (p) using the utility measure $U(p)$ from both technical and business perspectives. $U(p)$ is measured in terms of technical significance (technical$U(p)$) and business utility (business$U(p)$), that is, $U(p) = f(\text{technical}U(p), \text{business}U(p))$. An example of technical utility can be given by: accuracy = 87.90% and precision = 91.54%. While an example of business utility is, for instance, the average frequency in days and the average customer value in euros.

Table 1 presents eight different aspects of data-driven and domain-driven, extracted from Cao (2007, 2010). In domain-driven data mining, the object mined is not only the data but the business domain, where the goal is to develop effective problem-solving and discover actionable knowledge to satisfy real users, using real data and information related. D^3M is a multiple-step, iterative and interactive process, where the human cooperates, in a customizable environment, to provide actionable knowledge, which is evaluated in a trade-off between technical significance and business utility.

2.3 | Causality

Causality is the key in implementation of actionable knowledge. The outcome, or the impact, of each actionable rule should be measured in order to be evaluated. Therefore, the way to operationalize knowledge is by combining the two fields of machine learning and causal inference.

Pearl and Mackenzie (Pearl & Mackenzie, 2018) summarize the theme of causality with the three ascending rungs of what they call the "ladder of causation".

The lowest rung deals with the observation that can be exemplified by a conditional probability $P(Y \mid X)$ and obtained using machine learning algorithms.

TABLE 1 Data-driven versus domain-driven

Aspects	Data-driven	Domain-driven
Object	Data tells the story	Data and business domain tell the story
Objective	Effective algorithms, discover knowledge of research interest	Effective problem-solving, discover actionable knowledge to satisfy real users
Data	Abstract, synthetic data	Real-life data and information related
Process	One step	Multiple-step, iterative and interactive
Mechanism	Automated	Human mining cooperation
Usability	Predefined models and process	Customizable models and process
Deliverable	Patterns	Actionable knowledge
Evaluation	Technical metrics	Trade-off between technical significance and business utility

The second rung of the ladder of causation climbs from seeing to doing. Pearl and Glymour introduce the operator, 'do,' which represent a possible intervention (Pearl & Glymour, 2016). The conditional probability, $P(Y \mid \text{do } X = 1)$, expresses the outcome of Y, given an explicit intervention, $X = 1$. The experimental design was introduced by Fisher (1966), which uses randomized samples, an experiment sample and a control sample, also known as A/B test. This method is a standard in science. However, in social science and healthcare, A/B test is not considered adequate, not only because of the high number of particular cases but also because is considered unfeasible or unethical.

The top rung of the ladder of causation deals with a new approach, introduced by Rubin (2005). Since each individual has different characteristics and the potential outcome cannot be measured twice, with treatment and no treatment, the concept of "counterfactual" will be used. The counterfactual is not observed, so the challenge is to create a model that fills in the missing data. The Rubin causal model does not use randomized samples but instead similar groups. Counterfactual impact evaluation compares the outcomes of those that have been benefitted from an intervention (the treated group) with those of a similar group (the control group) that have not been exposed to the hypothetic intervention. The conditional probability is expressed by $P(Y \mid C = \text{c}, \text{do } X = 1)$, where C represents the individual characteristics. The individual intervention effect, IIE_i, is given by $\text{IIE}_i = P(Y_i \mid C_i = \text{c}, \text{do } X_i = 1) - P(Y_i \mid C_i = \text{c}, \text{do } X_i = 0)$. The causal effect of the treatment is given by the difference of the outcomes, and the average causal effect is given by the mean of the causal effect of the units.

Considering the outcomes, $\text{Y0} = P(Y \mid C = \text{c}, \text{do } X = 0)$ and $\text{Y1} = P(Y \mid C = \text{c}, \text{do } X = 1)$, where X is the hypothetic intervention (or treatment), C is the characteristic and the data in Table 2, our goal is to calculate the individual intervention effect or individual treatment effect. In this numeric example, Y can represent the blood pressure in mmHg of the patients, C1 the age, C2 the gender and C3 the weight in Kg.

The common procedure is to create pairs of individuals (or units) with similar characteristics. The matching returns pairs, (1,4) and (2,3). In Table 3, the individual treatment effect can be calculated for each unit and also the average treatment effect (ATE).

Finally, we would like to highlight that Pearl and Mackenzie see data mining in the first step of causality rather than the finial step (Pearl & Mackenzie, 2018). This vision is aligned with the operationalization of actionable knowledge where the fields of machine learning and causal inference should be combined.

2.4 | Bi-criteria

Multi-objective optimization or multi-criteria optimization deals with the optimization of two or more conflicting objectives, which are subject to a set of constraints.

Given the vector $x = (x_1, \ldots, x_m)$ of decision variables, M is the number of objectives and J the number of constrains, multi-objective optimization can be stated as follows (Collette & Siarry, 2011).

Minimize/Maximize $f_m(x)$, $m = 2, \ldots, M$ objetives

Subject to $c_j(x)$ $\{\leq, =, \geq\}$ 0 $j = 1, 2, \ldots, J$ constrains

and $x_i \geq 0$.

In multi-objective optimization, more than one optimal solution can be obtained, whereas classic optimization has only one objective. Variable S represents the set of feasible solutions associated with equality and inequality constraints. $F(x) = (f_1(x), f_2(x), \ldots, f_m(x))$ is the vector of objectives to be optimized.

In multi-objective optimization, the dominance concept is central. In the maximization problem, the objective vector $u = (u_1, \ldots, u_m)$ dominates $v = (v_1, \ldots, v_m)$, denoted as $u > v$, if and only if, $u_i \geq v_i$: $\forall$i, and at least one component of v is smaller, $u_i > v_i$: $\exists$i.

A solution is non-dominated, or Pareto solution, if and only if there is no solution that dominates it. In other words, a solution $x^* \in$ S is Pareto optimal if, for every $x \in$ S, $F(x)$ does not dominate $F(x^*)$.

TABLE 2 Example of missing data

Unit(i)	C1(i)	C2(i)	C3(i)	X(i)	Y0(i)	Y1(i)	Y1(i)-Y0(i)
1	72	1	60	0	183	?	?
2	37	2	78	0	161	?	?
3	38	2	88	1	?	137	?
4	73	1	63	1	?	164	?

TABLE 3 Matching and ATE

Pair(i)	Y0(i)	Y1(i)	Y1(i)-Y0(i)
1,4	183	164	−19
2,3	161	137	−24
Average treatment effect			−22

The Pareto optimal set, P^*, includes all the Pareto solutions, that is, the set of all solutions whose associated vectors are non-dominated.

When plotted in space, non-dominated vectors are collectively known as the Pareto front, PF^*. The procedure to generate Pareto front is to compute as many points as possible and then build a surface that includes those points. The surface created by the Pareto front can be linear, convex or concave. The Pareto front is defined as $PF^* = \{F(x), x \in P^*\}$.

The ideal vector contains the best solutions considering the *m* objectives separately at the same point. A point $y^* = (y^*_1, y^*_2,...,y^*_m)$ is an ideal vector if it optimizes each objective function f_i in $F(x)$.

Figure 1, adapted from (Talbi, 2009), shows a bi-objective optimization. The black circles of the Pareto solution dominate the solutions represented by triangles. The efficient front is given by the curve that includes all the Pareto solutions. In this case, we want to minimize f_1 and maximize f_2.

The ideal solution is obtained by the combination of the solution that minimizes f_1 with a second solution that maximizes f_2.

2.5 | Actionable knowledge in sport services

Predictive analysis in the field of regular sports services is a research area that few authors take into account. One of these authors is Sobreiro, who used machine learning algorithms to identify dropout levels (Sobreiro & Santos, 2017) (Sobreiro et al., 2018), and subsequently used methods of survival analysis to determine the duration of the frequency of users in gyms and health clubs (Sobreiro et al., 2019).

Pinheiro and Cavique (2018) used the decision tree algorithm to identify abandonment profiles and, based on these profiles, create loyalty actions that seek to increase the retention of regular service customers in sports facilities.

The same authors (Pinheiro & Cavique, 2019a) added actionable knowledge discovery with experimental planning, using test and control groups, in order to find and measure the effectiveness of actions aimed at customer retention. Subsequently, they added the business utility measure of a sequence of actions to prevent the withdrawal of users of regular sports services (Pinheiro & Cavique, 2019b).

In this document, we add the concepts of causality and bi-criteria decision making to the work previously developed, and create three scenarios to analyse their impact.

3 | PROPOSED MODEL

In the proposed model, we reuse some concepts of Database Marketing (Cavique, 2006) and IDIC model, <Identify, Differentiate, Interact, Customize>, (Peppers & Rogers, 2004) to obtain customer retention in services. Our model has three steps, aggregating Differentiate and Customize, of IDIC model, in a single step, called knowledge discovery.

The proposed model, which includes data preparation, knowledge discovery using predictive models and retention interventions with evaluation, can be presented in the following data pipeline:

Data Collection → Knowledge Discovery → Retention Intervention.

The proposed framework, presented in Cavique (2006), Pinheiro and Cavique (2018) and Pinheiro and Cavique (2019a, 2019b), works in a multiple-step pipeline, as shown in Table 4, where the columns, phase, outcome, evaluation and groups, are presented. The column, 'groups,' is presented in order to clarify and differentiate groups that have similar names but different meanings.

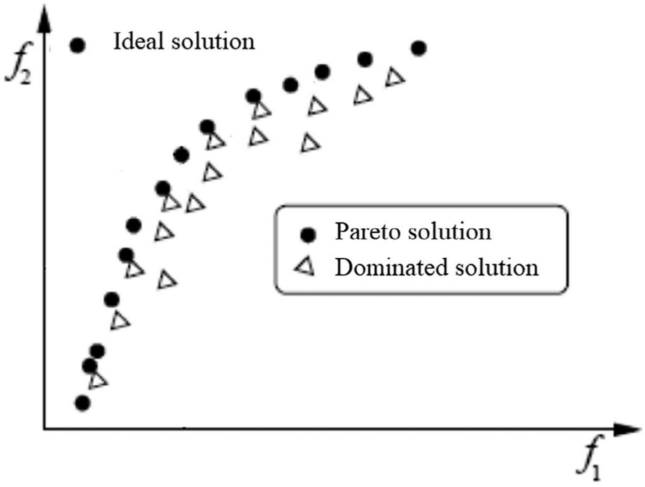


FIGURE 1 Pareto set to minimize f_1 (number of features) and maximize f_2 (accuracy)

In the data collection phase, the outcome is a clean data table that will be used in the following steps. In the predictive model phase, actionable rules are generated; the evaluation is given by technical metrics as accuracy, and the performance of the model is measured using training and test groups. Finally, in the retention intervention, the business utility is the outcome, measured in euros, and the significance of the actions is measured using treated and control groups.

This method corresponds to the domain-driven approach presented by Cao (2007, 2010), where the predictive model delivers business-friendly decision-support actions materialized in actionable knowledge rules, and the evaluation includes the trade-off between technical significance and business utility.

3.1 | Data collection

Following the first step of the pipeline, we need to create a support table with a set of suitable attributes to support the predictive model.

Taking into account that the data refer to individuals who use sports services, we want to find attributes in the three categories pointed out by Brito & Lencastre (2000) as important from a relational marketing point of view: attributes that define each customer's profile, the transactions that each customer performed with the installation, and the communication actions that each customer was exposed to, as well as how they responded to them.

Besides, the selection of attributes must consider the existence of factors with a greater or lesser impact on the retention of sports facilities and the possibility of being able to be extracted from the data registered in the databases of the ERP systems used by the sports facilities.

In order to develop the experiment and test the results, we used data from a Lisbon sports facility that uses a market application (e@sport), considering the entire history of users who started their enrollment in aquatic or fitness activities between June 1, 2014, and October 31, 2017. For this period, the original database contained 21,755 users, 122,805 subscriptions and 3,344,947 card passes in the access control, to which we applied the Extraction, Transformation and Loading (ETL) processes as detailed in previous works (Pinheiro & Cavique, 2018, 2019a).

The application of ETL resulted in the construction of a table with 8.381 users/records that have been enrolled in aquatic and/or fitness activities as shown in Table 5, and 51 relevant attributes, although only 45 present valid data, as some relevant attributes belonging to transactional and communication groups were not filled in due to lack of data. The considered attributes are presented in Table 6.

The 51st attribute, “withdrawal”, is what characterizes the state of the user to date and, therefore, the predicted attribute (the target attribute). For practical reasons, it was decided to define the attribute as a binary value, corresponding to a value of 1 for a quitter user, and a value of 0 to an active member.

Since the performance of some machine learning techniques is limited to the manipulation of values of a particular type or the performance itself is influenced by the range of values (Gama, Carvalho, Faceli, Lorean, & Oliveira, 2017), in addition to the attributes directly mapped from the source database, some attributes have been transformed, discretized through numeric–symbolic conversions. Moreover, new attributes have been created that derive from classifications and transformations made to the original data or attributes.

Some attributes have variations that correspond to derivations that aim to discretize the value of the original attribute, and sometimes more than one method was used.

In order to discretize the attributes, Number of months of enrollment, Days without attendance and Average frequency, the records were sorted in ascending order and five classes were created with an equal number of examples with labels from 1 to 5, as specified by the Hughes (Hughes, 2005) method.

In other cases, the interval classes were created according to the values indicated in the literature mentioned in the related work, in order to allow actionable profiling in terms of loyalty actions. In the case of the Age attribute, the ages were grouped into the following ranges: “<20”, “<35”, “<49”, “<65” e “> = 68”. For the Days without frequency attribute, the following day intervals were created: [00–07],]07–15],]15–30],]30–60],]60 – +∞[.

Pipeline phase	Outcome	Evaluation	Groups
1. Data collection	Clean data table	—	—
2. Knowledge discovery	Actionable knowledge	Accuracy	Training and test
3. Retention intervention	Business utility	Euros	Intervention and control

TABLE 4 Pipeline phases, outcomes, evaluation and groups

Users	In aquatic activities	In fitness activities	Total
Active	1,226	803	1,927
Dropouts	1,697	4,926	6,454
Total	2,923	5,729	8,381

TABLE 5 Number of users in data table

TABLE 6 Considered attributes grouped by Brito and Lencastre (2000) categories

Category	Attributes
Customer profile	Age (2 attributes) Gender References (2 attributes) Distance to the facility
Transactions	Number of months of enrollment (3 attributes) Turnover (2 attributes) Free use Attended activities (10 attributes) Number of activities attended Contracted frequency (2 attributes) Number of renewals Days without attendance (3 attributes) Average frequency (3 attributes) Total number of frequencies (2 attributes) Number of classes (2 attributes) Average frequency of classes (2 attributes) Ratio (real frequency/contracted frequency) (2 attributes) Training duration (2 attributes) Last response net promoter score, NPS Indications of dissatisfaction (3 attributes) Number of manifestations of dissatisfaction
Communications	Number of contacts established Number of assessments of physical condition Number of prescriptions

In addition to the operations referred to the attributes, missing values were also corrected through the strategy of removing the respective records.

3.2 | Predictive model and actionable rules

We intend to use predictive analysis to find profiles that characterize pre-dropout users in order to act on them before dropout occurs. In other words, we intend to formulate a model or hypothesis capable of relating the values of the attributes in the table mentioned in the previous section with the value of the target attribute, the “withdrawal” attribute, which is a nominal attribute. In practice, we intend to construct a predictive model that is able to identify the sets of characteristics that allow to rank a user against the level of their pre-dropout status.

Being a classification problem, the decision tree algorithm is a viable alternative to the problem posed. To obtain the decision trees, we use Microsoft Decision Trees algorithm available in Microsoft SQL Server Analysis Services Designer ver. 13.0.1701.8. This version uses a hybrid algorithm that incorporates different methods to create a tree. The algorithm offers three formulas for scoring information gain: Shannon's entropy, Bayesian network with K2 prior and Bayesian network with a uniform Dirichlet distribution of priors (Microsoft, 2017).

3.2.1 | Decision tree

Since, in this case, the attribute to be predicted is discrete (it can assume the values 1 or 0, for dropout and not dropout, respectively), the algorithm classifies based on the relationships between the input attributes of the data set. Uses the values of these attributes, known as states or classes, to predict the states of the attribute to predict. The model is created by the algorithm that adds nodes to the tree whenever it determines that an input attribute is significantly correlated with the attribute to predict.

Some attributes result from different forms of classification or discretization of the same characteristic, as explained in Section 3.1. and shown in Table 6, and if used simultaneously, the proposed models end up using redundant information. Gama et al. (2017) states that since the process of constructing a tree selects the attributes to use, they result in models that tend to be quite robust in relation to the addition of irrelevant and redundant attributes. Thus, it is desirable to remove the redundant attributes in order to obtain intelligible actionable profiles, even if it is necessary to slightly decrease the precision measures.

After removing the redundant attributes, like Days without frequency attribute with values obtained from the database and the attribute transformed with the Hughes classification, some changes were made to the optimization parameters of the algorithm in order to try to improve the accuracy of the model.

The Microsoft Decision Trees algorithm, available in Microsoft Analysis Services, allows changing the 7 parameters shown in Figure 2, although the Force_Regressor parameter does not apply in this case, as it is intended only for regression trees.

The Complexity_Penalty parameter inhibits the growth of the decision tree. The value of this parameter varies between 0, where the possibility of a split is very high, and 1.0, where the possibility of having a split is very low. In this case, the parameter was left with the default value of 0.9.

The Maximum_Input_Attributes and Maximum_Output_Attributes parameters are intended to limit the attributes that the algorithm must consider before invoking the feature selection. By setting the value 0 to these parameters, the feature selection is turned off. In these two cases, the maximum allowed value for these attributes was considered to be 255.

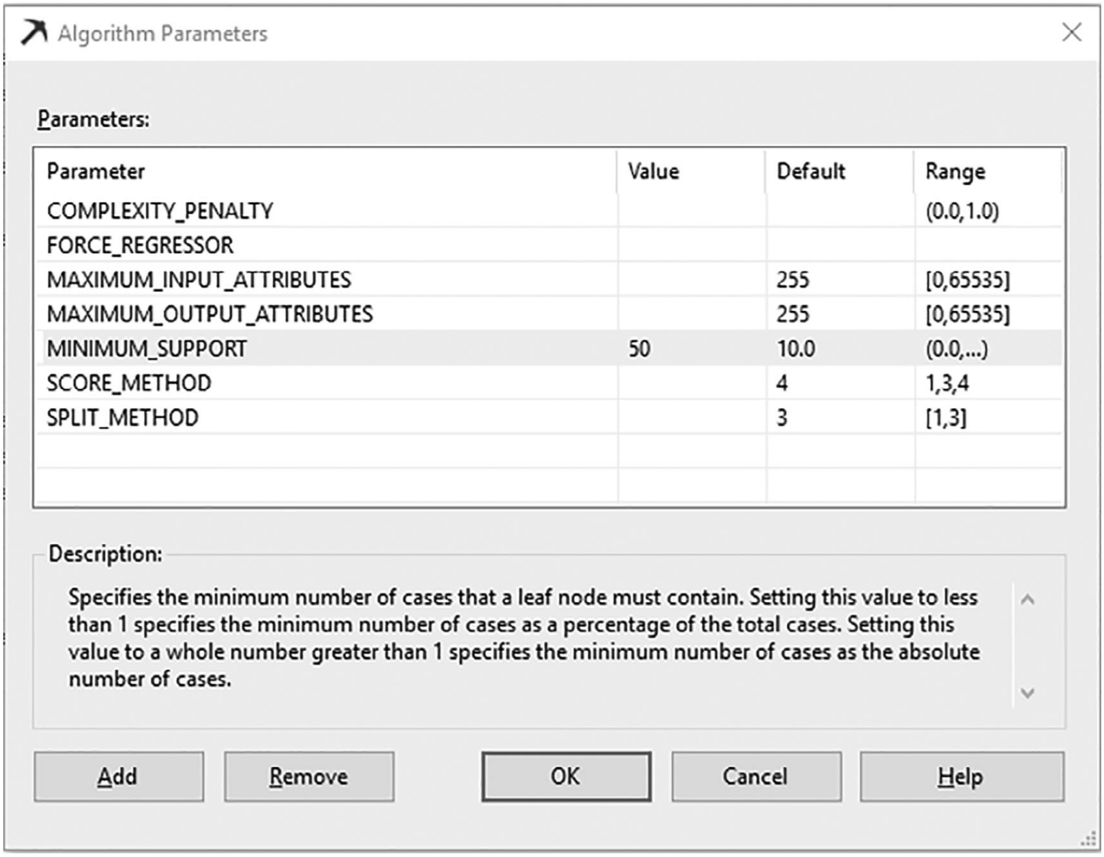
The Minimum_Support parameter indicates the minimum number of examples that each leaf of the tree must have. To avoid overfitting, this parameter was set to 50.

The Score_Method parameter allows you to choose the split score method. There are three methods available: Entropy (a); Bayesian with K2 Prior (b); and finally the Bayesian Dirichlet Equivalent with Uniform prior (c), the default method, which presented the best results.

Finally, the Split_Method parameter specifies the split method, which can be Binary (a), which indicates that regardless of the actual number of values for the attribute, the tree should be split into two branches, Complete (b), which indicates that the tree can create as many splits as there are attribute values or Both (c), which specifies that Analysis Services can determine whether a binary or complete split should be used to produce the best results. The latter was the value used for the parameter.

The resulting model corresponds to a shallow tree with leaves that always have a number of examples greater than 50. The evaluation metrics, obtained with the Holdout method, considering 70% of data for training and 30% for testing of the model are presented in Table 7.

In the tree built by the algorithm, shown in Figure 2, each leaf is defined by a set of rules that characterize it, defined by the path from leaf to root. In each leaf of the tree, there are examples that correspond to users quitting and examples that correspond to active users. The relationship between these quantities defines, on each leaf, a probability's threshold of withdrawal for the set of rules that define it. It is thus possible to draw dropout profiles from the rules on each leaf that shows a dropping rate above a considered threshold.



Parameters:

Parameter	Value	Default	Range
COMPLEXITY_PENALTY			(0.0,1.0)
FORCE_REGRESSOR			
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_OUTPUT_ATTRIBUTES		255	[0,65535]
MINIMUM_SUPPORT	50	10.0	(0.0,...)
SCORE_METHOD		4	1,3,4
SPLIT_METHOD		3	[1,3]

Description:

Specifies the minimum number of cases that a leaf node must contain. Setting this value to less than 1 specifies the minimum number of cases as a percentage of the total cases. Setting this value to a whole number greater than 1 specifies the minimum number of cases as the absolute number of cases.

FIGURE 2 Microsoft analysis services parameters for decision trees algorithm

3.2.2 | Actionable attributes

In order to create loyalty actions, specific attributes from the decision tree should be chosen to generate actionable knowledge. Some attributes cannot influence or be changed, such as the attribute, "age" or "gender", denominated by "non-actionable attributes". On the other hand, customer retention strategies can change the content of some attributes that reflect user behaviour. These attributes, which allow operational changes, are called 'actionable attributes'. An example of an actionable attribute is the number of "days without frequency", since a strategy can be implemented that causes, at least to some users, to return to the sports facility, after some time without attending (Figure 3).

Table 8 presents the six actionable attributes selected by the decision tree. Using dropout conditional probabilities, P(dropout | Xi), we can find the ideal and anti-ideal values, or interval, for each attribute. The histograms in Figure 4 show the ideal and anti-ideal values for each attribute.

Using the data from the histogram it is possible to calculate the conditional probabilities of dropout from each actionable attribute. For instance, for attribute X1, given the contingency table shown in Table 9, it is possible to calculate the conditional probabilities of *P*(dropout | X1 ∈ [0..7]) and *P*(dropout | X1 = otherwise).

Attribute X3 shows that dropout increases after the second month and reduces after the 12 months, so the conditional probability is *P*(dropout | X3 ∈ [0..2] ^]12..inf]). Attribute X5 shows that dropout enlarges for people with a frequency contract of once a week and free entrance, that is, 7 days a week, so the conditional probability is *P*(dropout | X5 ∈ [2..6]). The other attribute is easier to understand. We would like to notice that the ideal and anti-ideal values of the six attributes are consensual with the sports services bibliography.

Considering the outcome of the dropout, the impact of an intervention by changing from non-ideal to ideal is given by Impact = *P*(dropout | *Xi* = 1) − *P*(dropout | *Xi* = 0).

TABLE 7 Evaluation metrics of predictive model with holdout method

No. of nodes	Depth	Accuracy	Sensitivity	Specificity	Precision	*F* score
30	6	87.90%	92.69%	72.53%	91.54%	92.11%

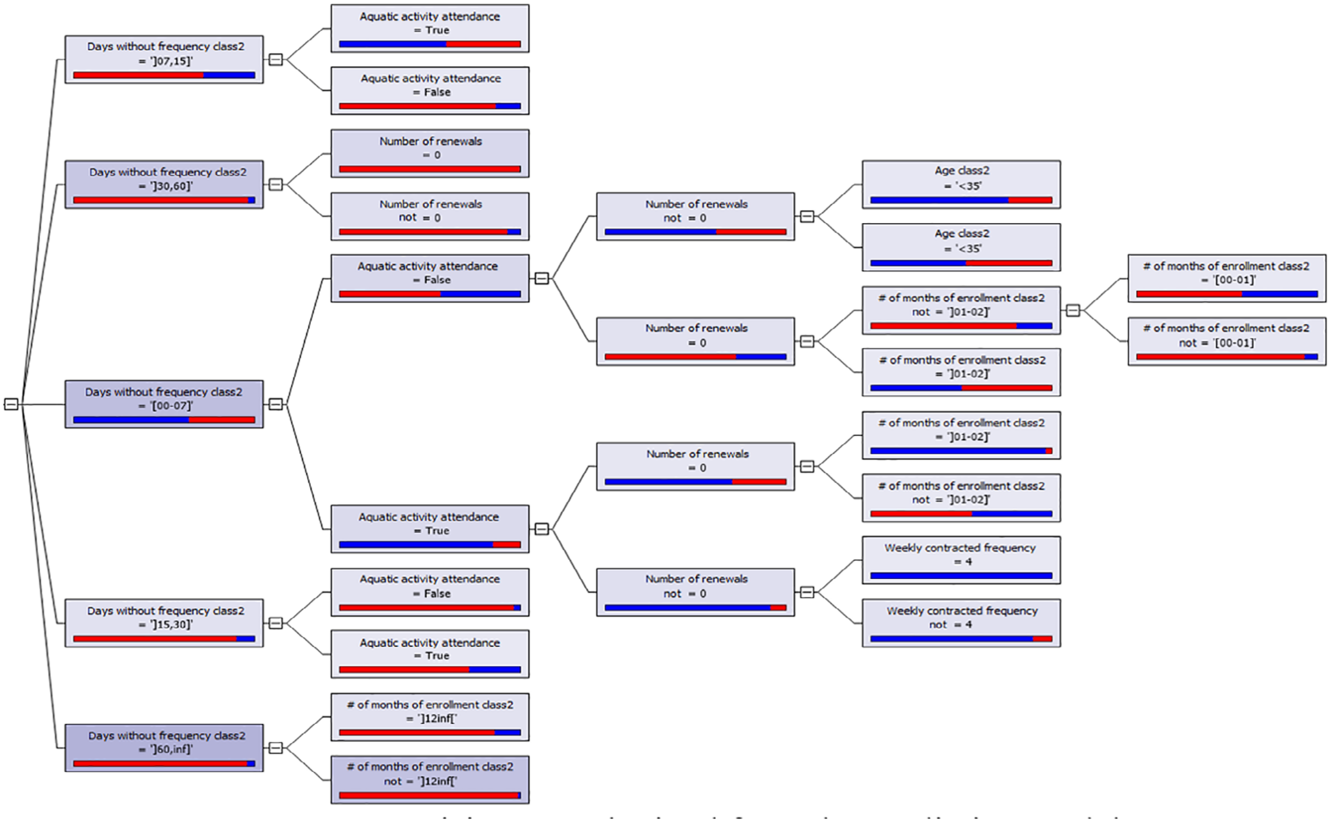

Days without frequency class2
= ']07,15]'
Aquatic activity attendance
= True
Aquatic activity attendance
= False
Days without frequency class2
= ']30,60]'
Number of renewals
= 0
Number of renewals
not = 0
Aquatic activity attendance
= False
Days without frequency class2
= '[00-07]'
Aquatic activity attendance
= True
Days without frequency class2
= ']15,30]'
Aquatic activity attendance
= False
Aquatic activity attendance
= True
Days without frequency class2
= ']60,inf]'
# of months of enrollment class2
= ']12inf['
# of months of enrollment class2
not = ']12inf['
Number of renewals
not = 0
Number of renewals
= 0
Number of renewals
= 0
Number of renewals
not = 0
Age class2
= '<35'
Age class2
= '<35'
# of months of enrollment class2
not = ']01-02]'
# of months of enrollment class2
= ']01-02]'
# of months of enrollment class2
= ']01-02]'
# of months of enrollment class2
not = ']01-02]'
Weekly contracted frequency
= 4
Weekly contracted frequency
not = 4
# of months of enrollment class2
= '[00-01]'
# of months of enrollment class2
not = '[00-01]'

FIGURE 3 Decision tree obtained from the predictive model

TABLE 8 Actionable attributes

Attribute	Description	Ideal	Anti-ideal
X1	Days without frequency	[0..7]	>60
X2	Aquatic activity attendance	True	False
X3	Number of months of enrollment	[0..2]^]12..Inf]	Otherwise
X4	Number of renewals	> 4	0
X5	Weekly contracted frequency	[2..6]	[1,7]
X6	Fitness activity attendance	True	False

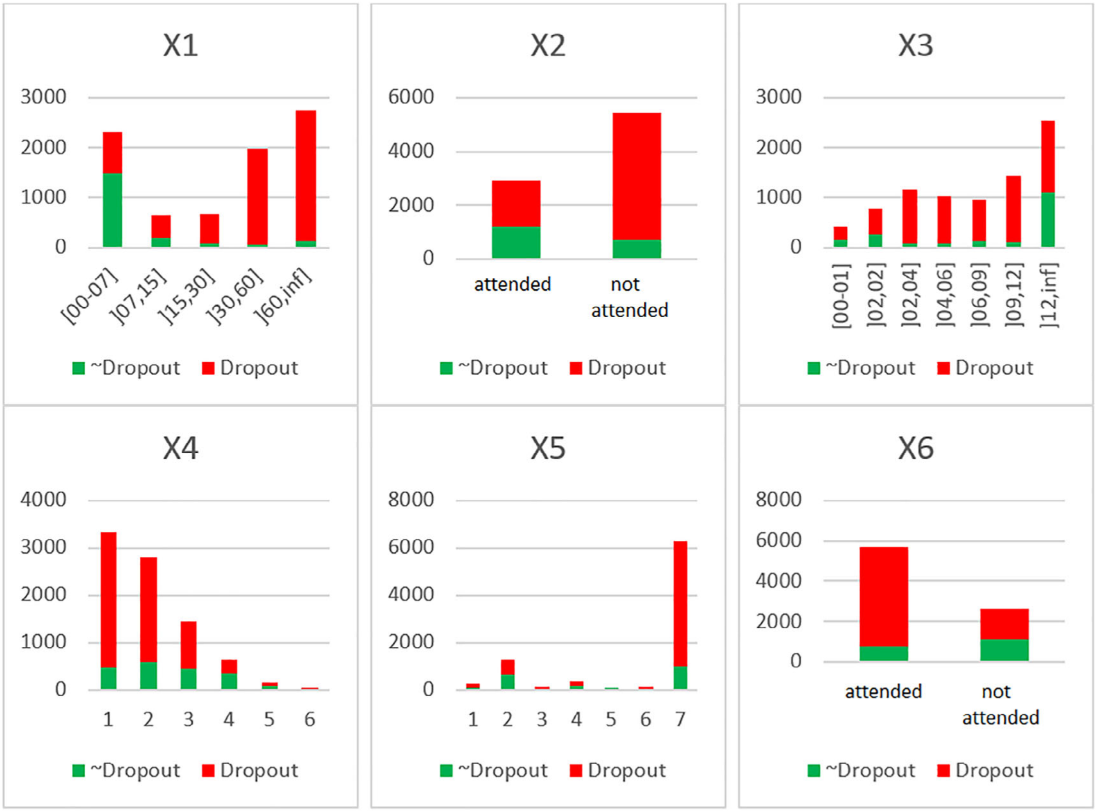


FIGURE 4 Histograms from actionable attributes

TABLE 9 X1 contingency table

	~dropout	Dropout	Total
Recent [0..7]	1,478	849	2,327
~recent	452	5,596	6,048
Total	1,930	6,445	8,375

Table 10 shows the impact of each attribute. The attribute X1 represents the largest attribute, which is also consensual to database marketing bibliography, where recency is the first item to be mentioned in RFM (recency, frequency, monetary) analysis (Hughes, 2005). All the attributes present impact except X6, which shows that dropout is very large in the fitness activity.

3.2.3 | Actionable knowledge

As already mentioned, each branch of the decision tree forms a splitting rule, where each node includes an attribute used by the algorithm. In each leaf of the tree, there are examples that correspond to users quitting and examples that correspond to active users. Thus, it is possible to draw

TABLE 10 Impact of the actionable attributes

<table>
<tr><th>Actionable attribute</th><th>P(dropout | xi = 1)-P(dropout | xi≠1)</th><th>Impact</th></tr>
<tr><td>X1</td><td>P(dropout | X1∈[0..7])-P(dropout | X1 = otherwise)</td><td>–0.570</td></tr>
<tr><td>X2</td><td>P(dropout | X2 = true)-P(dropout | X2 = false)</td><td>–0.293</td></tr>
<tr><td>X3</td><td>P(dropout | X3∈ [0..2]^]12..Inf])-P(dropout | X3 = otherwise)</td><td>–0.316</td></tr>
<tr><td>X4</td><td>P(dropout | X4 > 0)-P(dropout | X4 = otherwise)</td><td>–0.140</td></tr>
<tr><td>X5</td><td>P(dropout | X5∈ [2..6])-P(dropout | X5∈ [1,7])</td><td>–0.341</td></tr>
<tr><td>X6</td><td>P(dropout | X6 = true)-P(dropout | X6 = false)</td><td>0.284</td></tr>
</table>

TABLE 11 Actionable knowledge based on actionable rules

Actionable rules	
Actionable rule A: If X_1 (days without frequency) ∈]30, 60] And X_4 (number of renewals) = 0 Then dropout = 99.29%	Actionable rule B: If X_1 (days without frequency) ∈]60, ∞[And X_3 (number months of enrollment) ∉]12, ∞[Then dropout = 98.15%
Actionable rule C: If X_1 (days without frequency) ∈]15, 30] And X_2 (aquatic activity attendance) = false Then dropout = 96.47%	Actionable rule D: If X_1 (days without frequency) ∈ [0, 7] And X_2 (aquatic activity attendance) = false And X_3 (number of months of enrollment) > 2 And X_4 (number of renewals) = 0 Then dropout = 93.03%
Actionable rule E: If X_1 (days without frequency)∈]30, 60] And X_4 (number of renewals) > 0 Then dropout = 92.84%	

dropout profiles from the splitting rules on each leaf that shows a dropping rate above a considered threshold. These profiles allow us to segment users according to the criteria mentioned by Kotler and Keller (2012), which indicates that segmentation is only useful if the segments meet five criteria: they are measurable, substantial, accessible, differentiable and actionable.

Actionable knowledge is supported by splitting the tree into actionable rules, which include the actionable attributes. Table 11 presents several actionable rules above the 90% threshold that should be considered. Each actionable rule contains actionable attributes with values that should be avoided.

Actionable rule A shows a user who does not visit the facilities between 31 and 60 days and never renewed. Actionable rule B shows a user profile who does not visit the facilities for more than 60 days and whose enrolment is inferior to 12 months. Actionable rule C shows a user who does not visit the facilities between 16 and 30 days and does not attend aquatic activities. Actionable rule D shows a user who does not visit the facilities in the last 7 days, does not attend aquatic activities, whose enrolment has more than 2 months and has never renewed subscription. Finally, actionable rule E shows a user who does not visit the facilities between 31 and 60 days and has already renewed at least once.

3.3 | Retention interventions and evaluation

After defining the retention strategy, it is necessary to evaluate its effectiveness (Ascarza, 2018). The experiments are constructed through the implementation of A/B tests and evaluated through the chi-square method, which will allow computing a statistical conclusion. A/B tests manipulate a causal variable in order to determine the impact of the variable in two different groups of individuals, the test group and the control group. The groups are created by splitting randomly the users into two groups with the same number of elements. In the test group, the loyalty actions are applied, and in the control group, the loyalty actions will not be applied.

The users who received the communication and return to attend the facilities are called “recovered”. The recovered users allow evaluating the loyalty campaign impact in euros. This last estimation is used to measure the business utility we were looking for.

Similar to ideal and anti-ideal actionable attributes, shown in Table 8, there are also ideal and anti-ideal actionable rules, presented in Table 12. Actionable rule A has a large dropout, while actionable rule L, from loyal users, has a dropout equal to zero. Both actionable rules are extracted from the decision tree presented in Section 3.2.1, and the fundamental question is how can we migrate customers from rule A to rule L.

TABLE 12 Anti-ideal and ideal profile

Actionable rule A:	Actionable rule L:
If X_1 (days without frequency) $\in]30, 60]$	If X1 (days without frequency) $\in [0, 7]$
And X_4 (number of renewals) = 0	And X2 (aquatic activity attendance) = true
Then dropout = 99.29%	And X4 (number of renewals) > 0
	And X5 (weekly contracted frequency) = 4
	Then dropout = 0.00%

The actionable attributes of rule L can help us to find retention actions. In sports services, it is consensual to every actionable attribute value of rule L: the recency of X1, the practice of aquatic activities of X2, the large number of renewals of X4 and the intermediate value of X5.

The anti-ideal rule A can be rewritten as follows: $P(\text{dropout} \mid X_1 \in]30,60], X_4 = 0) = 0.9929$; and the equivalent ideal rule $P(\text{dropout} \mid X_1 \in [0,7], X_4 = 0)$. The impact is given by the difference of the rules.

In a more summarized form:

Y0—outcome without intervention or control.

Y1—outcome with intervention.

Y1-Y0—impact of the intervention or causal effect.

So, using the probabilities of dropout:

$Y0 = P(\text{dropout} \mid X_1 \in [30,60], X_4 = 0) = 0.9929$.

$Y1 = P(\text{dropout} \mid X_1 \in [0,7], X_4 = 0)$.

$Y1\text{-}Y0 = P(\text{dropout} \mid X_1 \in [0,7], X_4 = 0) - P(\text{dropout} \mid X_1 \in [30,60], X_4 = 0)$.

Operational marketing involves performing a set of retention experiments with subsequent impact evaluation, given by Y1-Y0.

4 | RETENTION INTERVENTION SCE (OK)

In causality studies, given the outcome without treatment, Y0, and the outcome with treatment, Y1, it is possible to calculate the intervention effect, Y1-Y0.

In this work, we study the conditional probability of dropout $P(\text{dropout} \mid X)$. The procedure to calculate the intervention effect can be summarized as follows:

1. given the controlY0 = $P(\text{dropout} \mid X = 0)$
2. given the interventionY1 = $P(\text{dropout} \mid X = 1)$
3. calculate the impactY1-Y0 = $P(\text{dropout} \mid X = 1) - P(\text{dropout} \mid X = 0)$

This procedure can be applied in randomized samples (A/B test) and also in counterfactual models.

In our study, we are dealing with real data and actionable rules, which are given by the dropout model. However, no companies are at the moment available to apply retention strategies, given the new constrains imposed by the General Data Protection Regulation.

As a way to continue our work, in order to estimate the incremental effect of the campaign, we are using the percentages given by Frota (2011) and IHRSA (2012). The intervention effect, Y1-Y0, can be estimated and scenarios can be built. To exemplify, given the expected annual impact given by Frota (Frota, 2011) is 30% and a hypothetical actionable rule has the conditional probability of 97% of dropout, the estimative for the dropout after the retention intervention is 67%. The procedure of the scenario to induce the outcome Y1 can be summarized and exemplified as follows:

1. given Y1-Y0 = $P(\text{dropout} \mid X = 1) - P(\text{dropout} \mid X = 0) = -0.30$
2. given Y0 = $P(\text{dropout} \mid X = 0) = 0.97$
3. induce Y1 = $P(\text{dropout} \mid X = 1) = 0.67$

The conditional probability of 67% is an estimate for the campaign and should be confirmed after the actual retention intervention.

Counterfactual data are not observed but they are different from scenarios. On the one hand, counterfactual data are obtained by matching real data from different sources. On the other hand, scenarios use also real data but in a different succession of steps. The steps needed to generate an experimental study and to build a scenario are the same, but with a different sequence.

In this section, retention intervention, we develop an approach based on scenarios, in order to estimate the incremental effect of the campaign.

4.1 | Interface of the scenario-simulator

To simulate the realization of retention interventions, the possible effect that they will have on the behaviour of the users, the respective consequences in the retention measures and the decision trees resulting from the application of the actions, a scenario generator has been developed. The interface of the scenario generator (or scenario-simulator) is shown in Figure 4.

This scenario generator selects a segmented subset of users that represents a possible real case in which a leaf of the tree, which represents an actionable rule, may contain some users who, although yet active, presents the profile of dropout, and should, therefore, be the target of the loyalty actions proposed for this profile.

At the end of each scenario-simulation, the generator changes the properties of an attribute in a given number of users, as if they had responded positively to a campaign directed at them. It is expected that the number of users will no longer be in the pre-dropout state at the end of the campaign, so the generator changes their dropout state to false.

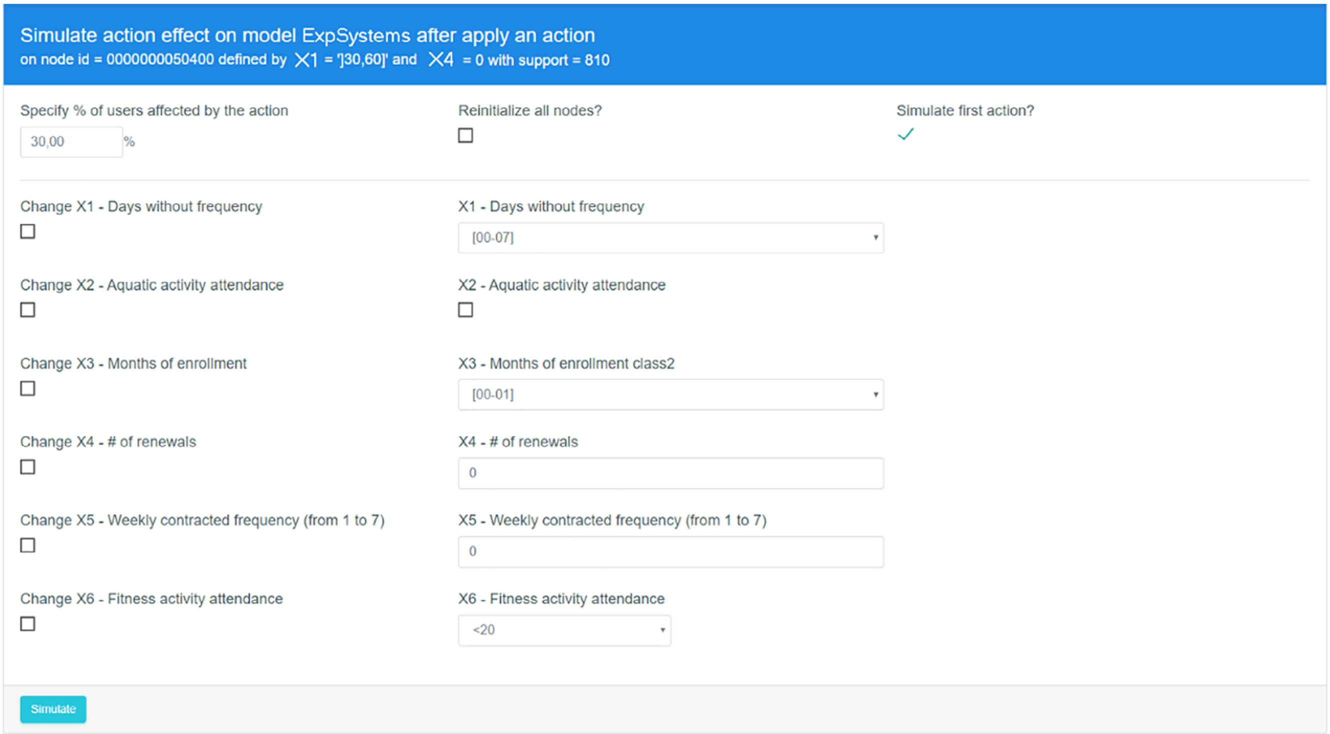
The most important fields of the form in Figure 5 are:

1. The field "Specify % of users affected by the action" allows to specify Y1-Y0, that is, the percentage of users that presents a certain profile defined by the selected decision tree leaf and that we intend to consider recovered after the retention actions;
2. The following two fields, "Reinitialize all nodes" and "Simulate first action," allow you to restart the scenario-simulator after a scenario has been previously created;
3. Finally, there are six sets of two fields. Each set allows us to indicate to the scenario-simulator that we want to change, the Xi attribute of the records of recovered users;

To be able to apply A/B testing, the scenario generator divides users on each tree leaf into two similar groups, alternating each user's location by the testing group, which groups users to whom retention actions will be directed, and control, which groups users who will not be the target of retention actions.

The number of users who responded positively to the campaign is determined by applying a percentage to the total number of users in the test group in the tree leaf defined by the actionable rule. This percentage should be a value appropriate to the profile obtained by the actionable rule according to the previously referenced authors (Frota, 2011; IHRSA, 2012).

For each scenario, we intend to evaluate its statistical relevance and relate the cost of the campaign to the business utility.



Simulate action effect on model ExpSystems after apply an action
on node id = 0000000050400 defined by X1 = ']30,60]' and X4 = 0 with support = 810

Specify % of users affected by the action 30,00 %	Reinitialize all nodes? □	Simulate first action? ✓
Change X1 - Days without frequency □	X1 - Days without frequency [00-07]	
Change X2 - Aquatic activity attendance □	X2 - Aquatic activity attendance □	
Change X3 - Months of enrollment □	X3 - Months of enrollment class2 [00-01]	
Change X4 - # of renewals □	X4 - # of renewals 0	
Change X5 - Weekly contracted frequency (from 1 to 7) □	X5 - Weekly contracted frequency (from 1 to 7) 0	
Change X6 - Fitness activity attendance □	X6 - Fitness activity attendance <20	

FIGURE 5 Scenario-simulator form

To evaluate the statistical relevance, we will use chi-square to obtain the confidence interval that will allow us to conclude whether or not there is a causal relationship between the application of the actions to the actionable rules and the reduction of the number of withdrawals.

To obtain the cost involved in each form of communication used in the campaigns, the following factors were considered: (a) any form of communication (Email, SMS or personal contact) requires preparation of the model by a human, so it was considered that there was an expenditure of 4 hrs of work, whose cost was calculated from the average salary in 2017 (PORDATA, 2019); (b) the cost associated with sending an email is practically non-existent; (c) the average market cost of SMS; (d) and finally that a telephone or personal contact could take an average of 10 min, and the corresponding value was also calculated based on the average salary in 2017. The preparation and unit costs considered are shown in Table 13.

To calculate the business utility, we use the average market value of a monthly fee, which was, in 2018, 39.50€ (Pedragosa & Cardadeiro, 2019).

4.2 | Scenario 1: Communication scenario with BPMN

With the generator of scenarios tool, we use the users defined by Actionable Rule A and work with this subset of users through the entire simulation of the campaign.

This scenario aims to simulate a sequence of three communication actions on users who have not visited the sports facility for more than 30 days and less than 60 days and have never renewed their registration. This retention strategy includes a communication process in three stages, starting with a personalized e-mail sent to the user, followed by the use of an SMS and finally a personal contact, creating a funnel workflow. It is expected that after sending each communication, the behaviour of a part of the target users will change, so for this subset, it will no longer be necessary to send them the following communications. Figure 6 presents a business process model and notation (BPMN) diagram, which illustrates the proposed communication scenario.

In this simulation, the subset is divided into a test and control group, where the first group has 594 users and the second 593, and both contain, at the beginning, three users who have not given up.

The simulation of the communication strategy occurs in three different instants, t_1, t_2 and t_3, using e-mail, SMS and personal contacts, respectively. It was considered by definition that it would be possible to recover 5% of users after t_1, 10% after t_2 and 15% after t_3, making up the 30% of annual recoveries indicated by Frota (2011).

TABLE 13 Preparation and unit costs

Media	Preparation cost	Unit price
Email	33.75€	0.000€
SMS	33.75€	0.035€
Personal contact	33.75€	1.406€

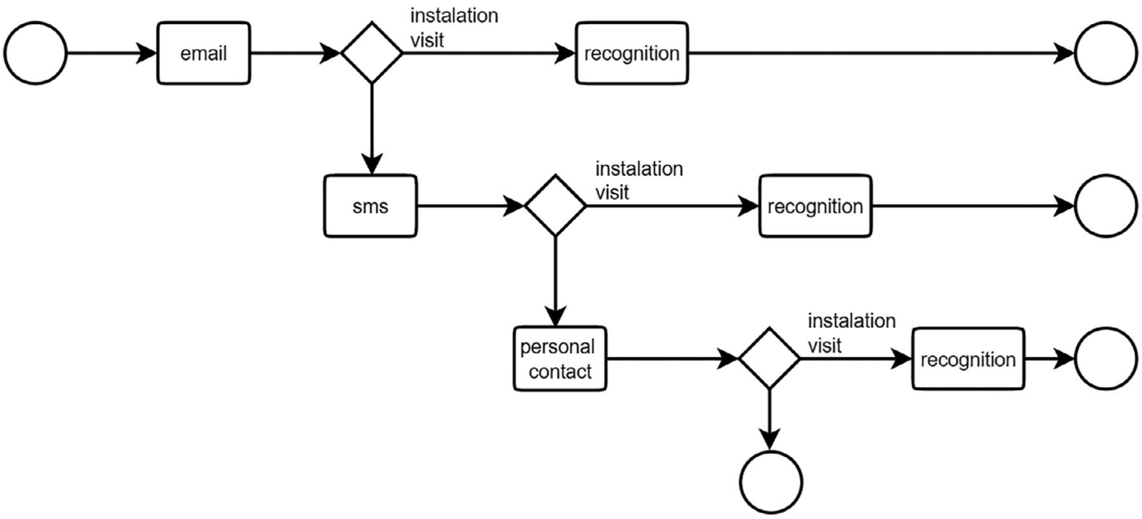


FIGURE 6 BPMN diagram for scenario 1

Since these figures are annual, and despite the seasonal nature of sports activities, we consider an average recovery of 0.42% per month in response to the email campaign, 0.83% per month in response to the SMS campaign and 1.25% per month in response to personal contacts.

It was also considered that it would be possible to spend 1 hr a day, at least on working days, to make personal contacts, which make a total of 120 monthly contacts per month.

Table 14 illustrates the simulation, where it was avoided the dropout of 17 users.

With the result obtained for chi-square (χ^2 = 9.944), we can conclude, through the query of the distribution table, with one degree of freedom and a confidence level of 99%, that the loyalty actions carried out with the proposed campaign had the effect wanted.

In this scenario, as in all of the following, we consider the initial three non-quitters of the test group as recovered from the first retention action, as expressed in the BPMN diagram, these non-quitters will no longer be the target of subsequent actions.

In addition to the significance obtained through the A/B test, we also calculate the cost and the business utility expressed in euros. To obtain this value, we consider the average monthly fee for a regular inscription as referred before (39.50€). For the number of users recovered from the simulation, the business utility is 671,50€ for each month in which the recovered users maintained their registration with a total cost of 256,97€.

4.3 | Scenario 2: Evaluation with a personal trainer

In this scenario, while considering the same actionable rule A as a starting point, only one communication action is performed, inviting the user to make an assessment of the physical condition or sports practice, according to the activity they attend.

On the other hand, we consider recovering a different percentage of users. While in scenario 1, we assume that the motivation of pre-dropout users was undifferentiated; in this scenario, we consider that the motivation for quitting is due to club-related problems, which according to Frota (2011) leads to 25% of dropouts, of which 45% can be recovered, which results in an average monthly percentage of 0.94%.

However, two different ways of communicating were considered: one by email (a) and one by SMS (b). Table 15 illustrates the two simulations, where it was avoided the dropout of 10 users who came to the sports facilities and made the assessment, the cost and the business utility associated with each option.

4.4 | Scenario 3: Aquatic activities in group/coaching effect

Finally, in the third scenario, using actionable rule C, that is, users that do not visit the facility between 15 and 30 days and do not attend aquatic activities, are invited to attend a swimming class in order to increase group/coaching effect.

Since one of the actionable attributes is related to the frequency or not of water activities, we sought to find a relationship between (a) those who practice water activities and those who give up, and also (b) between all those who practice fitness and those who give up. The difference between these two relationships points to a percentage of users who can be recovered by inviting them to try and subscribe to water activities.

TABLE 14 Scenario 1

	Control group		Test group				
	No. of potential dropout	No. of dropouts	No. of potential dropout	No. of dropouts	No. of avoided dropouts	Cost	Business utility
t_1 (email)	593	590	594	588	5	33.75€	197.50€
t_2 (SMS)	593	590	588	576	5	54.33€	197.50€
t_3 (Personal contact)	593	590	576	556	7	168.750€	276.50€
Totals					17	256.97€	671.50€

TABLE 15 Scenario 2

	Control group		Test group				
	No. of potential dropout	No. of dropouts	No. of potential dropout	No. of dropouts	No. of avoided dropouts	Cost	Business utility
t_1 (email)	593	590	594	584	9	75.75€	237.00€
t_1 (SMS)	593	590	594	584	9	96.54€	237.00€

In Table 5, where we can find the number of users who abandon and do not abandon water and fitness activities, we obtain a = 58.06% and b = 85.98%. Due to the difference between these two quantities, we consider that it is possible to recover 27.92% (approximately 28%) of users.

We also consider that the frequency for water activities experience does not represent additional costs beyond the preparation of the contact form.

Simialr to previous scenario, communications by email and SMS were simulated and the results, including costs and business utility of each simulation, are presented in Table 16.

4.5 | Bi-dimensional approach of the three scenarios

The scenarios detailed in the previous sections allowed to simulate the recovery of users through the actionable attributes that define their behaviour profile. In each scenario, the cost and business utility associated with each retention action were calculated.

In this section, starting from the relationship between these two quantities, we determine the efficiency ratio (Business utility/cost) of each retention action, which is presented in Table 17. We also find the efficiency ratio obtained after the sequence of the three actions of scenario 1.

From previous data, we can draw the graph in Figure 7, where we can see that there are two Pareto non-dominated solutions: retention action via email in scenario 3, and the sequence of the three retention actions of scenario 1.

For these scenarios, the Pareto optimal set P* = {3-Email, 1-All}. In Figure 7, Pareto front, PF* = {$F(x)$, $x \in P^*$}, is represented by the line that connects the two Pareto non-dominated solutions. The decision maker can choose one of the two non-dominated solutions, either the 3-Email solution, which is the most efficient with a lowest cost, or the solution 1-All, which returns the largest business utility.

4.6 | Discussion of the new decision tree

The above scenarios correspond to the simulation of retention actions applied over a certain period of time. Upon completion of a retention action or retention action sequence, the behaviour of the target users of the retention actions is expected to change. As a result, the predictive model is expected to generate different decision trees, create or eliminate previous actionable rules, and eventually some of the previous actionable attributes, loss or gain relevance, or even other attributes become relevant and actionable.

The most efficient scenario is the 3-mail, from the third scenario, using actionable rule C. Now, we discuss the impact of the retention interventions in the whole decision tree by changing only actionable attributes of rule C during the first, second, sixth and twelfth months.

TABLE 16 Scenario 3

	Control group		Test group				
	No. of potential dropout	No. of dropouts	No. of potential dropout	No. of dropouts	No. of avoided dropouts	Cost	Business utility
$t_{1\ (\text{email})}$	263	252	264	245	9	33.75€	237.00€
$t_{1\ (\text{SMS})}$	263	252	264	245	9	42.99€	237.00€

TABLE 17 Efficiency ratio (Business utility/cost) from each scenario

Scenario	Communication action	Cost	Business utility	Efficient ratio
1	Email	33.75€	197.50€	5.852
1	SMS	54.33€	197.50€	3.633
1	Personal contact	168.75€	276.50€	1.639
1	All (3 retention actions)	256.87€	671.50€	2.614
2	Email	75.75€	237.00€	3.129
2	SMS	96.54€	237.00€	2.455
3	Email	33.75€	237.00€	7.022
3	SMS	42.99€	237.00€	5.513

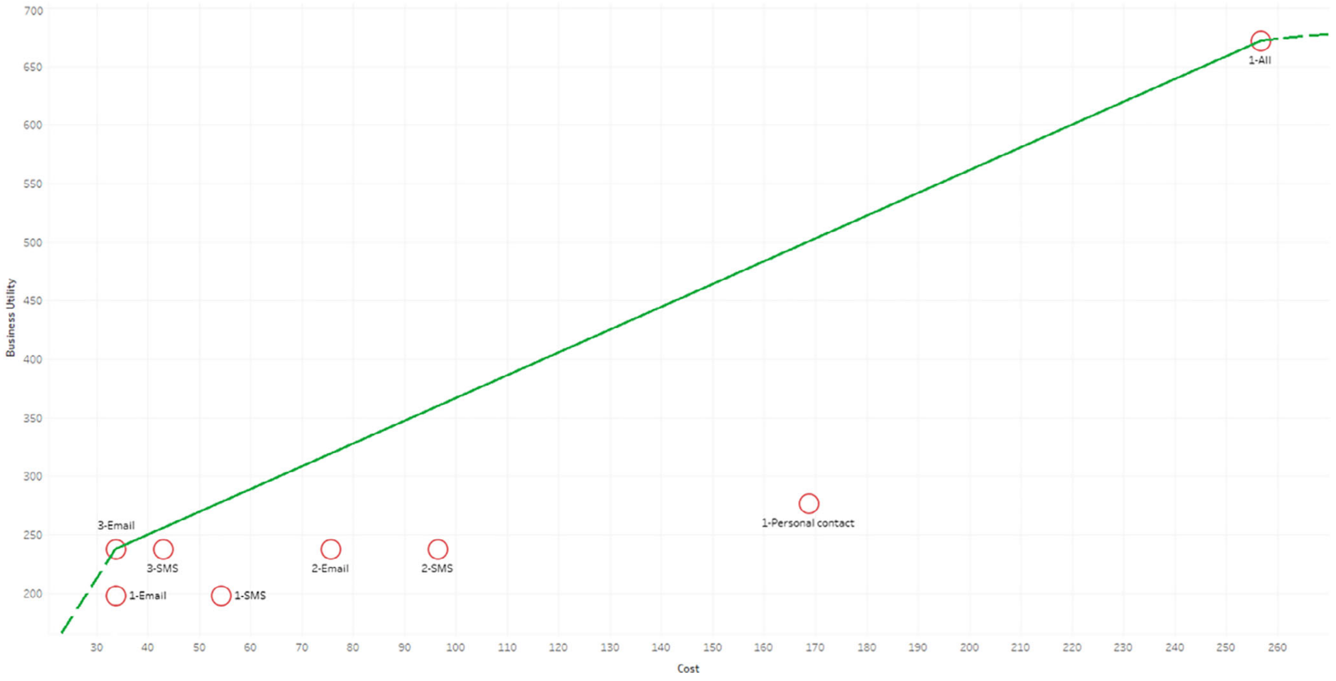
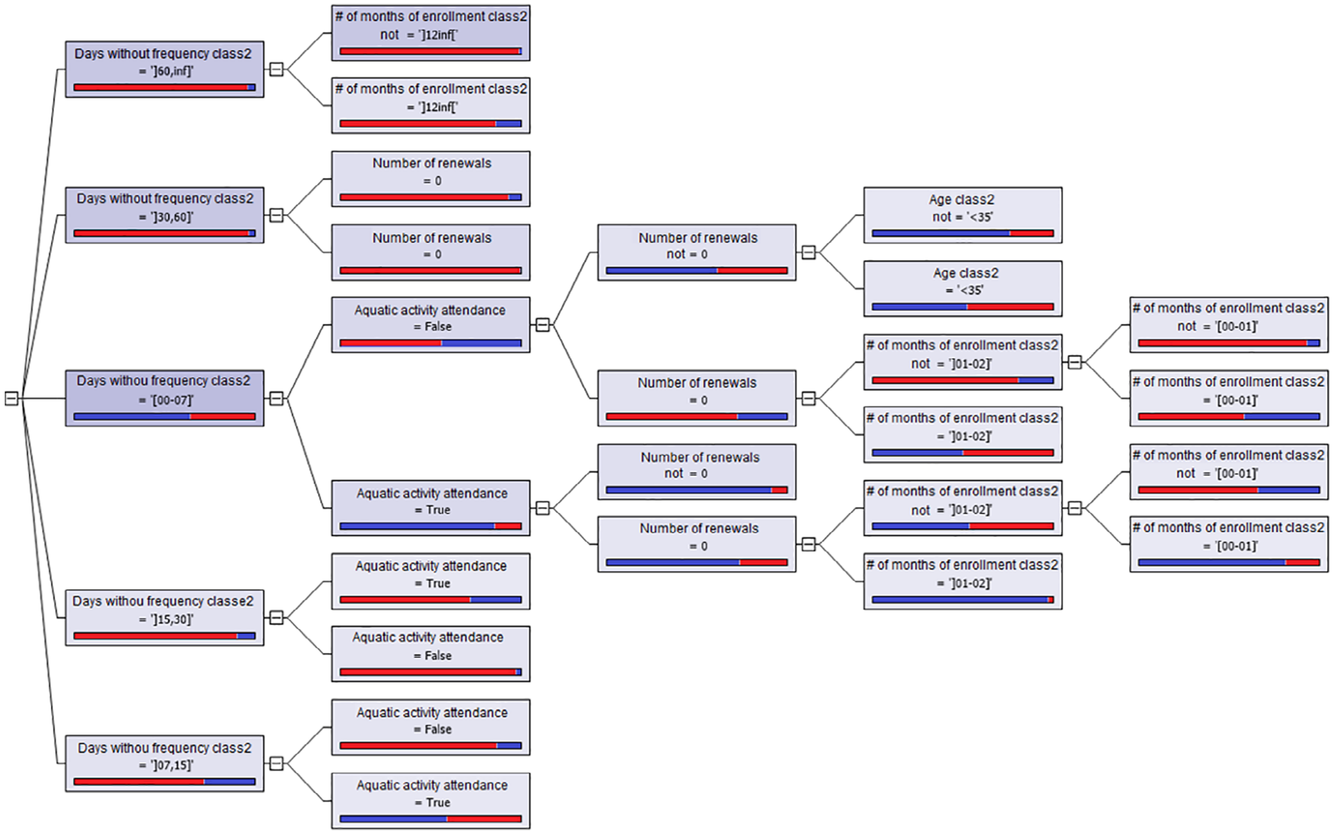


FIGURE 7 Business utility versus cost



Days without frequency class2 = ']60,inf]'
# of months of enrollment class2 not = ']12inf['
# of months of enrollment class2 = ']12inf['
Days without frequency class2 = ']30,60]'
Number of renewals = 0
Number of renewals = 0
Days withou frequency class2 = '[00-07]'
Aquatic activity attendance = False
Number of renewals not = 0
Age class2 not = '<35'
Age class2 = '<35'
Number of renewals = 0
# of months of enrollment class2 not = ']01-02]'
# of months of enrollment class2 = ']01-02]'
# of months of enrollment class2 not = '[00-01]'
# of months of enrollment class2 = '[00-01]'
Aquatic activity attendance = True
Number of renewals not = 0
Number of renewals = 0
# of months of enrollment class2 not = ']01-02]'
# of months of enrollment class2 = ']01-02]'
# of months of enrollment class2 not = '[00-01]'
# of months of enrollment class2 = '[00-01]'
Days withou frequency classe2 = ']15,30]'
Aquatic activity attendance = True
Aquatic activity attendance = False
Days withou frequency class2 = ']07,15]'
Aquatic activity attendance = False
Aquatic activity attendance = True

FIGURE 8 Decision tree after applying retention actions to the rule for 12 months

To illustrate the possible impact on the predictive model and, consequently, on the definition of new actionable attributes, profiles and rules, we selected the most efficient retention action from Table 17 and its actionable rule C and simulated the application of the predictive model during 12 months, as described in Scenario 3. We recall that, in the scenario considered, the retention actions are applied to users who present the

No. of nodes	Depth	Accuracy	Sensitivity	Specificity	Precision	F score
29	6	88.18%	92.44%	75.20%	91.90%	92.17%

TABLE 18 Evaluation metrics of predictive model with holdout method after applying retention actions to rule C for 12 months

non-attendance profile of the facilities between 15 and 30 days and do not attend water activities. Given that the recovery rate, calculated in Section 4.4, is 28% per year, the monthly recovery rate is 2.33% (0.28/12). This percentage of recovered users return to the facilities weekly and become users of water activities after the application of the interventions.

The decision tree that results from the application of actions during 1 month shows no change from the initial decision tree, which results from the reduced number of users recovered.

By the end of the second month, differences are already visible in the resulting tree. Although actionable rule C remains unchanged, since the value of the "Days without frequency" attribute goes from [15–30] to [0–7] and the "Aquatic activities attendance" attribute goes from False to True, a new split emerges.

After 6 months, the "Weekly contracted frequency" attribute is no longer in the decision tree and is, therefore, no longer a relevant attribute.

From the seventh to the twelfth month no new changes in the decision tree were detected. However, it is expected that, as these retention actions continue to apply, further changes will occur as soon as the number of users who fall within the profile, defined by actionable rule C, is reduced to a level at which this profile is no longer statistically significant.

It is also expected that the diversity of user characteristics retrieved at the level of the other actionable attributes will also cause changes in the decision tree structure. The decision tree obtained after 12 months is the one presented in Figure 8 and the corresponding metrics are shown in Table 18.

5 | CONCLUSIONS

The annual dropout in sports services is around 40%, which reveals the importance of customer retention interventions that require concrete and personalized actions. Most of the data mining techniques achieve interesting patterns for churning. However, the implementation in fieldwork is still in development.

The proposed framework, presented in Cavique (2006)), Pinheiro and Cavique (2018, 2019a, 2019b), works in a multiple-step pipeline, which includes data preparation, knowledge discovery using predictive models and retention interventions with evaluation.

In the data collection phase, real data from databases of the sports facilities are used, extracted from a period of 3 years. The original database contained 21,755 users, 122,805 subscriptions and 3,344,947 card pass. After the data-cleaning process, the information comprehends more than 8,300 users with 51 attributes.

In the predictive model phase, actionable attributes are identified and actionable knowledge rules, extracted with the decision trees, with a dropout threshold above 90%, are detailed. Actionable attributes are identified and the domain-driven approach presented by Cao (2007, 2010), where the predictive model delivers actionable knowledge is used.

Pearl and Mackenzie (2018) see data mining in the first step of causality rather than in the last step. Given that causality is the key in implementation of actionable knowledge; in this work, the effort to operationalize actionable knowledge has led us to combine the fields of machine learning and causal inference.

Retention intervention can be implemented with an A/B test or with a counterfactual impact evaluation. However, no companies are at the moment available to apply retention strategies, given the new constraints imposed by the General Data Protection Regulation. Given these restrictions, we developed a bi-objective scenario-based study oriented to estimate the incremental effect of the retention intervention campaign. For each scenario, the business utility and associate cost are measured and the most efficient solutions are obtained using the Pareto front.

In this work, we combined actionable rules with the simulation of what-if scenarios in order to obtain results in the real-world of sports services. We believe that the combination of machine learning and causal inference has an effective impact in organizations; therefore, in future work, counterfactual interventions will be used instead of scenarios.

ORCID

Paulo Pinheiro https://orcid.org/0000-0002-8912-2244
Luís Cavique https://orcid.org/0000-0002-5590-1493

REFERENCES

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Avourdiadou, S., & Theodorakis, N. D. (2014). The development of loyalty among novice and experienced customers of sport and fitness centres. *Sport Management Review*, *17*(4), 419–431. https://doi.org/10.1016/j.smr.2014.02.001

Bedford, P. (2009). *Retain and gain*: *Keeping your members engaged*. Challenges and Opportunities for Membership Growth and Retention, San Francisco.

Brito, C. M., & Lencastre, P. de. (2000). Os Horinzontes do Marketing. EDITORIAL VERB>O.

Cao, L. (2007). *Domain-driven*, *actionable knowledge discovery*. IEEE Intelligent systems. Sydney: IEEE Computer Society. pp. 78–79

Cao, L. (2010). Domain-driven data mining: Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 755–769. https://doi.org/10.1109/TKDE.2010.32

Cavique, L. (2006). *Relatório da Unidade Curricular de Database Marketing*, *2005–2006*. (unpublished). Escola Superior de Comunicação Social, Instituto Politécnico de Lisboa.

Collette, Y., & Siarry, P. (2011). *Multiobjective optimization, principles and case studies: Decision Engineering Series*, Berlin Heidelberg New-York: Springer.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). New York, NY: Hafner Publishing Company.

Frota, M. (2011). Gestão da Retenção. In *Manual de Gestão de Ginásios e Health Clubs-Excelência no sector do Health & Fitness* (pp. 103–148). Portugal: AGAP.

Gama, J., Carvalho, A. P. d. L., Faceli, K., Lorean, A. C., & Oliveira, M. (2017). In E. Silabo (Ed.), *Extração de Conhecimento de Dados* (3rd ed.). Portugal: Silabo.

Gonçalves, C. (2012). Variáveis Internas e Externas ao Indivíduo que influenciam o Comportamento de Retenção de Sócios no Fitness. *PODIUM Sport, Leisure and Tourism Review*, *1*(2), 28–58.

Gorgoglione, M. U. P. (2011). Beyond customer churn: Generating personalized actions to retain customers in a retail bank by a recommender system approach. *Journal of Intelligent Learning Systems and Applications*, *03*(02), 90–102. https://doi.org/10.4236/jilsa.2011.32011

Howat, G., & Assaker, G. (2016). Outcome quality in participant sport and recreation service quality models: Empirical results from public aquatic centres in Australia. *Sport Management Review*, *19*(5), 520–535. https://doi.org/10.1016/j.smr.2016.04.002

Hughes, A. M. (2005). *Strategic database marketing* (3 ed.). Professional Publishing, New York: McGraw-Hill.

IHRSA. (2012). IHRSA Member retention report. Retrieved November 4, 2019 from http://download.ihrsa.org/pubs/2012_IHRSA_Retention_Guide.pdf

Kotler, P., & Keller, K. L. (2012). *Marketing management*. Paris: Pearson Education.

Microsoft. (2017). *Microsoft Decision Trees Algorithm Technical Reference*. Retrieved April 18, 2019 from https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm-technical-reference?view=sql-server-2014

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer* (1 ed.). Wiley.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1 ed.). Hachette Book Group, New York: Basic Books.

Pedragosa, V., & Cardadeiro, E. (2019). Visão Geral dos Principais Indicadores do Mercado do Fitness em Portugal para 2018. Retrieved November 14, 2019 from AGAP Portugal Activo website: https://issuu.com/addmore10/docs/portugal_activo_n2/1?ff

Peppers, D., & Rogers, M. (2004). Managing customer relationships: a strategic framework. John Wiley & Sons.

Pinheiro, P., & Cavique, L. (2015). Determinação de Padrões de Desistência em Ginásios. *Revista de Ciências Da Computação*, *10*, 33–60.

Pinheiro, P., & Cavique, L. (2018). *Models for increasing retention in regular sports services*: *Predictive analysis and loyalty actions*. https://doi.org/10.23919/CISTI.2018.8399160

Pinheiro, P., & Cavique, L. (2019a). An actionable knowledge discovery system in regular sports services. In Á. Rocha, H. Adeli, L. P. Reis, & S. Constanzo (Eds.), *Advances in intelligent systems and computing* (Vol. 2, pp. 461–471). Portugal: Univ. Aberta. https://doi.org/10.1007/978-3-030-16184-2_44

Pinheiro, P., & Cavique, L. (2019b). *Extracting actionable knowledge to increase business utility in sport services*. 19th EPIA conference on artificial intelligence, EPIA 2019, Proceedings, 2. pp. 397–409. https://doi.org/10.1007/978-3-030-30244-3_33

PORDATA. (2019). Salário médio mensal dos trabalhadores por conta de outrem: remuneração base e ganho. Retrieved November 14, 2019 from https://www.pordata.pt/Portugal/Salário+médio+mensal+dos+trabalhadores+por+conta+de+outrem+remuneração+base+e+ganho-857-6932

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331.

San-Emeterio, I. C., Iglesias-Soler, E., Gallardo, L., Rodriguez-Cañamero, S., & García-Unanue, J. (2016). A prediction model of retention in a Spanish fitness center. *Managing Sport and Leisure*, *21*(5), 300–318.

Sobreiro, P., Pinheiro, P., & Santos, A. (2018). *Performance in the prediction of dropout using the machine learning in sport services*. 18a Conferência Da Associação Portuguesa de Sistemas de Informação.

Sobreiro, P., & Santos, A. (2017). Approach for predicting dropout in a Health Club. *Revista Intercontinental de Gestão Desportiva -RIGD*, 7(3). Brazil: AIGD - Aliança Intercontinental de Gestão do Desporto.

Sobreiro, P., Silva, A., Conceição, A., Louro, H., Santos, A., Pinheiro, P., & Carvalho, P. G. (2019). *Customer dropout in aquatic centres*: *A survival analysis*. Submitted to European Sport Management.

Su, P., Zhu, D., & Zeng, D. (2014). A new approach for resolving conflicts in actionable behavioral rules. *Scientific World Journal*, *2014*, 530483.

Surujlal, J., & Dhurup, M. (2012). Establishing and maintaining customer relationships in commercial health and fitness centres in South Africa. *International Journal of Trade, Economics and Finance*, *3*(1), 14–18. https://doi.org/10.7763/IJTEF.2012.V3.165

Talbi, E. G. (2009). *Metaheuristics, from design to implementation*, Hoboken, New Jersey: John Wiley & Sons, Inc.

Yang, Q., Yin, J., Ling, C., & Pan, R. (2007). Extracting actionable knowledge from decision trees. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 43–55. https://doi.org/10.1109/TKDE.2007.250584

AUTHOR BIOGRAPHIES

Paulo Pinheiro He received bachelor and master degree in Computer Science from Universidade Aberta. Currently it is Executive Director of CEDIS, of which he is also a founding partner, having as main responsibility for the direction of the Development Department and responsible for guiding projects in the areas of management of sports facilities, education and ticketing. He was a freelance programmer for IEFP and a trainer at LISNAVE. He was the author of several articles for magazines, conferences and journals.

Luís Cavique He is a tenured Assistant Professor at the Computer Science Section in the Department of Sciences and Technology at Universidade Aberta. He worked in the Polytechnic Education System from 1991 to 2008, namely as Adjunct Professor in the Setubal and in the Lisbon Polytechnic Institute. He received the degree in Computer Science Engineering from the New University of Lisbon (FCT-UNL) in 1988, the MSc degree in Operational Research and Systems Engineering from the Technical Lisbon University (IST-UTL) in 1994 and the PhD degree in Systems Engineering from the Technical Lisbon University (IST-UTL) in 2002. His research areas are in the intersection of Computer Science and Systems Engineering, namely the Heuristic Optimization and the Data Mining.

How to cite this article: Pinheiro P, Cavique L. A bi-objective procedure to deliver actionable knowledge in sport services. *Expert Systems*. 2020;e12617. https://doi.org/10.1111/exsy.12617