Noname manuscript No.
(will be inserted by the editor)

Intelligent Data Analysis Approaches to Churn as a Business Problem: a Survey

David L. García · Àngela Nebot · Alfredo Vellido

Received: date / Accepted: date

Abstract Globalization processes and market deregulation policies are rapidly changing the competitive environments of many economic sectors. The appearance of new competitors and technologies leads to an increase in competition and, with it, a growing preoccupation among service providing companies with creating stronger customer bonds. In this context, anticipating the customer's intention to abandon the provider, a phenomenon known as churn, becomes a competitive advantage. Such anticipation can be the result of the correct application of information-based knowledge extraction in the form of business analytics. In particular, the use of intelligent data analysis, or data mining, for the analysis of market surveyed information can be of great assistance to churn management. In this paper, we provide a detailed survey of recent applications of business analytics to churn, with a focus on computational intelligence methods. This is preceded by an in-depth discussion of churn within the context of customer continuity management. The survey is structured according to the stages identified as basic for the building of the predictive models of churn, as well as according to the different types of predictive methods employed and the business areas of their application.

Keywords Churn analysis · intelligent data analysis · computational intelligence · customer continuity management · literature survey

This research was partially supported by Spanish MINECO TIN2012-31377 research project.

David L. García
Department of Computer Science, Universitat Politècnica de Catalunya, Spain
Tel.: +34-93-4137796
Fax: +34-93-4137833

Àngela Nebot
Department of Computer Science, Universitat Politècnica de Catalunya, Spain
Tel.: +34-93-4137783
Fax: +34-93-4137833

Alfredo Vellido
Department of Computer Science, Universitat Politècnica de Catalunya, Spain
Tel.: +34-93-4137796
Fax: +34-93-4137833

1 Introduction

The ongoing processes of globalization and deregulation are changing the competitive framework in the majority of economic sectors. The appearance of new competitors and technologies entails a sharp increase in competition. It also entails a growing preoccupation among service providing companies with the creation of stronger bonds with customers in a context in which they are required to fight over their customer portfolios in increasingly shifting markets. Many of these companies are diverting resources away from the goal of capturing new customers and are instead focusing on retaining the existing ones, particularly those who represent a higher return on investment. In this context, anticipating the customer's intention to abandon, a phenomenon also known as churn, and facilitating the launch of retention-focused actions both may become elements of competitive advantage.

Data mining (DM) was originally conceived as a process for business analytics, organized as a succession of stages defined according to a methodology [86]. Intelligent data analysis (IDA), in the form of pattern recognition (PR), machine learning (ML), statistics and related approaches, remains at the core of DM. IDA approaches are key in several stages of DM, including data exploration for knowledge discovery [86, 26] and data modelling for tasks of prediction, classification, segmentation, or rule extraction, amongst others.

Despite the fact that DM has outgrown its original purpose, IDA techniques have proven useful over de last decades in their application to business problems [97, 65]. We should therefore expect these techniques, applied to market surveyed information, which is often readily available for service providing companies, to be of assistance in churn management processes.

This paper aims to provide the reader, be it a business analyst or a data scientist, with a thorough survey and review of churn analysis applications of IDA techniques reported in recent academic literature.

In Section 2, we first describe the concept of churn or supplier abandonment by the customer as a business problem in some detail. In order to do so, we discuss business concepts such as customer life-cycle, customer loyalty and customer continuity management.

Section 3 shifts the focus of attention to different aspects of the use of IDA techniques in churn management. We explore them from the point of view of DM, differentiating the consecutive stages of the mining process for building predictive models of abandonment, which include: data gathering and understanding, feature selection, design and development of the predictive model, and validation and evaluation of this model.

This is followed by a compilation of the reviewed literature in the form of tables, according to two main criteria: the predictive method used in the application and the business area to which churn analysis is applied. Then, a discussion of the surveyed studies is provided in Section 4. It is again structured according to the four stages of the mining process for building predictive models of abandonment. All stages are discussed using examples from recent churn analysis literature. General patterns for each of the stages are commented and some recommendations, both general and business area-specific, are made. The paper wraps up with some concluding remarks.

2 Churn as a business problem

In the scenario of global competitive pressure described in the introduction, the possibilities of commercial development and, consequently, of adding value to a company require prolonging the useful life of its existing customers and increasing their average consumption. The management of the different commercial development alternatives in mature markets is illustrated in Figure 1.

Customer retention requires understanding how customer loyalty construction mechanisms work, in order to anticipate customers' intention to abandon and facilitate the launch of retention-focused actions as a result. This understanding enables defensive commercial strategies oriented

to retain and create loyalty bonds in existing customers, an approach that is likely to be more effective and less costly than an aggressive strategy to expand the overall size of the market by attracting potential customers. The cost of losing profitable customers to fiercely competitive markets is making many companies shift their target from the massive capture of new customers to the preservation of existing ones.

However, this struggle for achieving customer loyalty collides with the grinding exposure to advertised offers from competitors that customers face on a regular basis. Furthermore, customers' market awareness is constantly on the increase and, as a result, so are their expectations.

Companies have their customers as their main assets and they are responsible for the definition and implementation of policies that allow them to reach and prolong their maximum commercial development potential. In other words, they must prolong the life expectancy of their customer portfolio as much as possible, ensuring its adequate development in terms of value through the implementation of suitable commercial actions for each one of the stages of their life-cycle[1] (See an illustration of the stages of a customer life-cycle in Figure 2). Despite the fact that both dimensions in this figure, *generated value* and *time of customer-company relationship*, are strongly related in a unique *Customer Value Management* model, we understand that an appropriate development of customer's commercial value needs to guarantee its continuity first [28], although it is equally true that a proactive customer's value development has a positive influence on its relationship with the company: high purchase and high value customers use to have a longer lifetime value. Thereby, increasing customer's life expectancy should be the primary aim that determines and guides any posterior commercial action of value development.

The final objective is self-explanatory: the commercial relationship with customers -the valuable ones- must be kept and reinforced. For that purpose, companies should build strong customer defection-avoiding schemes. This is not to say that unprofitable customers should not be cared for. As supported by investigations in [36], companies should be aware of the fact that the abandonment of unprofitable customers by the supplier might lead to unintended negative impacts on the loyalty and attitudes of the retained valuable customers. As shown in this study, the increase of switching costs and satisfaction among the latter might not be enough to compensate for those negative impacts. Moreover, it may often be the case that the abandonment of unprofitable customers may have a significant economic impact that should be accounted for when quantifying lifetime customer value [37].

It should be borne in mind that companies and their customers are in a constant evolution that may lead to natural and unpredictable disruptions in the commercial relationship (change of home address, family life-cycle, change in interests, payment type, etc.). Thus, final success will not be based on lengthening customers' life-cycle in an unnatural fashion, but on ensuring that good customers do not leave prematurely.

2.1 Customer continuity management

The creation of loyalty bonds in customers requires a systematic approach to its management. The possible symptoms that might alert of a possible defection to the competition must be preventively anticipated, and this requires evaluation of their evolution over time. Therefore, the adoption of a suitable *Customer Continuity Management* (CCM) model [28] should make it easier for companies to critically review all aspects that might affect the construction of true loyalty bonds with customers, including policies for everyday management, both for the customer life-cycles and for the predicted and declared cases of customer loss (see Figure 3).

[1] Through the increase in services used (up-selling); the increase in consumption or wallet share (cross-selling); the construction of stronger loyalty bonds; the proactive retention actions on customers who intend to leave the actual provider; the launch of new products and services (innovation) and/or the adjustment of commercial costs, giving each costumer as expected.

The level of customer bonding and, as a result, their life expectancy is intimately bound to the level of customer satisfaction relative to the quality of the service provided by a company. The higher the level of quality of service that customers perceive, the stronger the loyalty bonds [20, 48, 49, 80, 84, 54]. Thus, consumers who experience high levels of satisfaction about the service usually stay with their current provider.

However, and despite customer satisfaction having a positive influence on the level of bonding, it does not always suffice. There are numerous situations in which better service quality does not have a significant impact on consumer loyalty: for instance, customers that change mobile phone operators in spite of the fact that their current provider offers greater coverage; customers that fill up in slow petrol stations, with bad accesses and no additional services for the driver; customers who prefer to travel with certain airlines despite the continuous delays to their flights, etc. In consequence, there must be other factors, beyond satisfaction with service, influencing customer loyalty.

At this point, we introduce the concept of *barriers to change* as a construct that should mediate the satisfaction-loyalty relationship [20, 48, 49]. When the level of customer satisfaction with respect to different providers is similar, the level of bonding should be expected to depend, to a large extent, on the nature and strength of the *barriers to change* in place. The existing literature usually portrays the *barriers to change* in a negative sense, as difficulties and burdens -emotional, social, or financial- that the customer must overcome when making the decision to change providers [20, 48, 49, 80, 56, 29]. However, in a market place that is becoming increasingly deregulated and competitive, the understanding of the construction of *barriers to change* as bureaucratic, contractual and/or in some cases, as the result of the abuse of a dominant position, is a limited viewpoint. Barriers such as penalties when cancelling a given service; problems in the portability of mobile phone numbers; delays in the provision of the new service that are the fault of the old provider: all these are actions that are becoming steadily more regulated and penalized by the market and are not sustainable in the medium term.

To be sustainable, barriers to change must be built, like satisfaction, on customer perception. In this way, the active development of barriers to change becomes an excellence factor, in addition to satisfaction with the service, that is difficult to overcome by competitors in their attempt to attract the best customers. The construction of policies and procedures that maintain and improve excellence in both dimensions (satisfaction and barriers to change) should act as tools of prevention that protect customers from being lured by competitors.

However, not all customers need the same level of service, nor are they all prepared to pay the same for it, or to obtain it in the same way. Common sense tells us that it is not possible to fulfill completely, in an increasingly heterogeneous environment, the difficult task of developing the loyalty of all customers. For this reason, starting from the certainty that dissatisfied customers will always exist, companies must concentrate their efforts on the development of a broad-spectrum retention program, maintaining and improving those dimensions of the offer and barriers to change that most and best impact on the overall bonding of customers as a group. The objective is not to protect all customers, but rather as many of them as possible and, in particular, those who are most valuable to a given company.

It has to be born in mind that the effect of prevention procedures is never foolproof. Over the natural life-span of the customers, it is possible that external changes such as the appearance of new products, variations in competitors' offers, technological changes and/or internal changes (improvement in the customer's knowledge level or increase in his exigency, socioeconomic changes, etc.) occur that might affect customers' expectations and, as a result, their level of satisfaction. Companies must watch out for these changes to adapt their policies and procedures so that they can maintain and improve customers' opinions about the service on offer. The process of analyzing the dimensions with most impact on the satisfaction and the subsequent adjustment in commercial procedures and policies should become an ongoing process over time.

On the other hand, and from an operational perspective, it is not possible, given the high cost involved, to ask all customers from time to time for their opinions (even more so in the case of companies with hundreds of thousands or even millions of active clients) on the satisfaction perceived of the service they are being offered and/or their level of bonding. Companies must therefore work with representative enough samples and develop, based on their analysis, appropriate commercial policies.

Customers' evolution must be tracked and the number of customers at risk of churning must be estimated. That is why companies must have a reliable prediction model (adapted to the market research and based on behavioural information systematically gathered by the company) that allows them to identify -with enough anticipation- those clients that show symptoms of propensity to switch service providers and, thus, launch efficient retention actions. Following with the medical metaphor, early diagnosis of the propensity to churn will reduce considerably the aggressiveness of the required loyalty bonding *treatment* and will increase the customer's *recovery* possibilities. In this context, the client's value (understanding as client's value the sum of his actual recurrent value and his potential value) becomes the fundamental dimension that will determine which type of *therapy*, proactive and/or reactive, should be applied at any time.

This business effort -measured in the form of discounts, benefits and privileges that are offered to the client so that she or he will dismiss the idea of changing providers- should be balanced against the customer's expected value. This means that there may be clients that the company will decide not to retain even if their intention to change is identified in advance, since the expected return on the prolongation of their customer life does not justify the cost of the necessary commercial action. Identical criteria can be applied when deciding recovery policies and actions for already lost clients. Note though that, as reported in [36], careless abandonment of the less profitable customers may lead to unexpected negative reactions from the valuable ones that companies aim to retain.

3 Intelligent Data Analysis for churn management

The design of CCM models has been suggested in the previous section as an adequate approach to the task of prolonging the life-cycle of company customers. Anticipating a customer's intention to abandon their current provider company should be considered a key element of any *therapeutic* strategy in churn management.

In this context, DM techniques, including IDA in the form of PR, ML, statistics and related approaches, as applied to market surveyed information, could and should play a key role in helping to understand how customer loyalty construction mechanisms work and helping to analyze customers' intention to abandon from different angles, from exploratory characterization to outright prediction.

However, not all cases of churn are equally important, nor are they all predictable. According to the reasons behind it, abandonment can be classified in different typologies:

- ***Involuntary cancellation***: It affects customers from which their current company withdraws the service (e.g., fraud, arrears). Generally, companies do not even consider these cancellations as abandonment for their records.
- ***Voluntary cancellation***: It corresponds to customers who consciously decide to change provider. Two variants of this type can be considered:
  - ***Circumstantial***: Due to changes in the customer's circumstances which do not allow them to continue (change of address, inclusion in the company's social benefit plans, change of marital status, children, etc.). This cancellation is intrinsically unpredictable.
  - ***Deliberate***: It occurs when the customer voluntarily decides to abandon their current supplier for a competitor.

Here, we only consider this last scenario: voluntary and deliberate churn. In the remaining of the section we describe and review the different stages of the standard process of design and

development of a predictive model of supplier abandonment (customer abandoning the supplier). This literature will then be summarily organized in Section 3.2 and in the Appendix in the form of tables created according to two main grouping criteria: the type of predictive model used in the study and its particular business area of application.

3.1 Building predictive models of abandonment

The design and development of predictive models of supplier abandonment or churn can be divided into four general stages [21], as seen in Figure 4. The last three stages of this process form a cycle that is completed only if and when adequate prediction results are achieved. We will now take a closer look at each of these stages in turn.

3.1.1 Stage 1: Identifying and obtaining the best data.

This might arguably be the most relevant stage in the construction of a predictive abandonment model. Experience shows that the quality and suitability of the available data determines the accuracy and predictive power of the resulting model. Different data combinations may be better or worse indicators for different problems and for different areas of business. Ultimately, it is a question of identifying the data that best fit the type of analysis being carried out. Only in this manner could useful and usable knowledge (in business terms) be extracted in subsequent stages of analysis.

This stage in the building of predictive models of abandonment would fit, from a DM process point of view, the phases of *problem* and *data understanding*. Bearing this in mind, and from a practical point of view, it is important to note that the predictive model should ideally be constructed on the basis of the available data gathered routinely by a company from its whole customer base, which can be an extremely costly process. Consequently, those data bearing most of the predictive power may not always be available. Companies, therefore, may often be faced by the trade-off problem of identifying the best possible data from what is available to them.

The process of understanding and interpreting the data is often difficult. Even though the data in each field of a database may seem self-explanatory and unambiguous, interpretation may be hampered by the use of specific and *ad hoc* company lingo, different numerical formats, or simply because their meaning is different from the apparently obvious. Given the usual lack of standards to facilitate this process at the company level, its success is largely based on good communication between database managers and the data analysts. In fact, these DM stages have not been duly documented in the majority of investigations carried out in recent years [35].

A selection of different data requirements and motifs for the analysis of churn can be drawn from the recent literature. The most relevant are detailed next.

A large group of studies base their models of abandonment prediction on customer use/consumption variables: Madden and colleagues [69], in their customer retention model for the Australian ISP (Internet Service Provider) industry, classified and used four categories of variables: economic, use, ISP choice and demographics. Ng and Lin suggested in [74] the use of customer consumption for identifying churn in the telecommunications market. In [101], it was concluded that the purchase of products and services can be better predicted using historic purchasing data. This view was backed in [40], where it was proposed that the analysis of transaction data, through historic account and customer data, could provide us with clues to identify the best incentives for a bank to offer its customers and to improve the marketing strategy. Data on customer usage have also been used to identify the behaviour of website-using customers [47] and to predict repeat purchasing by mail [95].

More recently, customer usage/consumption data have been complemented with other variables as key elements in identifying abandonment. For instance, in a study of customer deflection in the

wireless telecommunications market [88], customer data were grouped into four types: demographics, usage level, quality of service and marketing features. This method was supported in a more recent study in the same sector [110]. Still in the field of telecommunications [73], Neslin and Gupta classified the selected variables into three main categories: customer behavior (minutes of use, revenue, handset equipment, trends in usage), company interaction data (calls to customer service) and customer household demographics (age, income, geographic location, home ownership), similar to the ones used in [44]. Hung and colleagues considered in [45] that the most significant variables for churn prediction in the mobile telephone industry are: demographic data (age, penetration rate, and gender), payment and account data (monthly quota, billing amount, arrears account), call details (call duration, call type) and customer service data (number of PIN number changes, number of blocks and suspensions). In their research about abandonment of the subscribers of a newspaper publishing company [17], Coussement and Van den Poel grouped customer data into four groups: subscription data (time since last renewal, monetary value, product), socio-demographics (age, gender), client/company interactions (number of complaints, time since the last complaint, responses to marketing actions) and renewal-related variables (days between subscription renovation and expiry date). More recently, data from credit card holders in a Chinese bank were used in [76] and [103] to predict their abandonment, with a combination of usage variables (daily balance, abnormal usage, limit usage, revoking pays, transactions, etc.) and customer personal information.

3.1.2 Stage 2: Selection of attributes.

In this stage, the most appropriate attributes or features for prediction must be selected from those available to the analyst, which, in a supervised PR setting, which is the most common in the literature of churn IDA, would be those that minimize the classification or prediction error; in an unsupervised learning setting, which might address churn analysis as a market segmentation problem, would be those which best reflect the grouping or cluster structure of the data. From a DM process point of view, this stage would correspond to the phase of *data pre-processing*.

This pre-processing is paramount as it helps to reduce the dimensionality of the data so that only the important attributes are included for analysis, whereas the redundant, noisy and/or irrelevant ones are excluded [108]. As a result, the selected features are also much more likely to provide an easier interpretability, in practical terms, of the obtained solutions.

Feature selection in supervised settings is a problem that has been thoroughly studied throughout the years [33, 90, 68], and providing a survey of selection methods is beyond the scope of this review. There are different feature selection criteria for the less frequently explored problems in unsupervised learning (e.g., clustering), where the goal is "to find the smallest feature subset that best uncovers *interesting natural* groupings (clusters) from data" [24], although a number of variable ranking criteria are useful across applications, including saliency, entropy, smoothness, density and reliability [33].

Only a limited number of the studies reviewed in this paper resorted to quantitative methods of feature selection (as opposed to qualitative methods mostly based on expert or domain knowledge). Out of these there are a few that followed a basic approach (e.g., R^2 statistic coefficient [46], z-test [45], collinearity measure [77], Fisher score [98], etc.) that is completely unrelated to the modeling technique(s) of choice. In the feature selection literature, these approaches are commonly known as *filter* methods and can be particularly practical in datasets with a large number of features. One of their drawbacks is that they usually fail to remove redundant features.

Only in a few of the reviewed studies the attribute selection technique is intrinsically associated to the modelling technique of choice. These approaches also fall into a category of their own in the feature selection literature, namely that of *wrapper* methods.

We find diverse methods for DTs in [21, 74, 57]. In [21], feature selection is a two-step process that involves forward selection with error rate-based ranking followed by feature group selection using genetic algorithms. Feature selection via induction was used in [74] to choose salient customer

loyalty features. In [57], and despite the fact that prediction was performed on the basis of majority voting in an ensemble learning setting, feature selection was performed using the CART DT. Related to the previous, methods for the ensemble-of-DTs RF method can be found in [17].

A very sophisticated feature selection approach for ANNs is presented in [107]. In it, a two-step (two-domains: target and related source) procedure is presented: first, an initial feature subset is selected by the ANN only in the target domain. Then, the target domain is enriched using the source domain. This second step is repeated to generate several new feature subsets and a base classifier in each one. A best base classifier is selected dynamically for each test data case.

For SVMs in [25], a recursive feature elimination (RFE) algorithm is proposed. In it, nested feature subsets are selected through sequential backward elimination, removing a feature at each step according to a ranking score based on the feature weights.

3.1.3 Stage 3: Development of a predictive model.

Once the best data available for analysis have been selected, the next stage of predictive model development entails the choice of the most suitable methods and techniques for building such model. This corresponds to the *modeling* phase in DM. In a simplified manner, a predictive model can be defined as one that extracts patterns from the available data in order to make inferences for previously unseen data or future situations [85].

In the area of abandonment prediction, the most commonly used modelling techniques, as reflected in the literature, include decision trees (DT), regression analysis [64] and artificial neural networks (ANN) [18], while in more recent years new methods such as support vector machines (SVM) have proven their adequacy [17, 13, 107].

As later explained in Section 3.2, one of the criteria according to which our review of churn IDA literature is organized is the type of predictive model used for analysis. For the sake of simplicity, three general categories were considered: *standard methods*, *computational intelligence* (CI) *methods* and *other alternative methods*.

Standard methods

In this category, we consider two families of techniques, namely regression analysis from standard multivariate statistics and DTs, which are algorithmic methods with close ties to statistics and which are considered here as standard due to their widespread acceptance for churn analysis applications.

– **Regression analysis**: A popular standard multivariate statistics family of techniques used by researchers dealing with the prediction of abandonment. It is stated in [73] that logistic regression's popularity in particular is due to its quick and robust results as compared to other classification techniques, added to its conceptual simplicity and its closed-form solution available for posterior probabilities.

  Regression analysis was used in [72] and [71] to link customer retention with satisfaction and its attributes in the fields of wireless telecommunications and private banking, while Kim and Yoon [52] used a logistic regression (logit) model to determine subscriber churn in the telecommunications industry, based on discrete choice theory (study of behaviour in situations where decision makers must select from a finite set of alternatives). In the same field, logistic regression was also used in [61], [63] and [94] to predict abandonment.

  There are regression models specifically suited to the analysis of longitudinal data such as Cox regression, from the area of survival analysis, which was recently used in [32] as a natural framework for the investigation of suppler abandonment as a dynamic process.

  In other studies [11, 17], logit models were used to predict abandonment of a Pay-TV operator and a newspaper publisher, comparing them with more novel methods: Markov chains in the first one and SVMs in the second. In the first case, logit models showed better accuracy while, in

the second case, SVM only showed a better performance when an optimal parameter-selection procedure was applied. Huang *et al.* [46] and Lee *et al.* [60] proved that logistic regression outperformed ANNs, DTs and other methods in their churn prediction studies.

More recently, multiple criteria decision models were used in [103] to evaluate the accuracy of 12 different algorithms -including logistic regression, multiple DT algorithms and different Bayesian networks - when predicting churn on a bank's credit card holders. It was found that logistic regression yielded the highest predictive accuracy. Logistic regression and DTs were later used in [77] to forecast credit card customer defection, reporting a better performance of logistic regression.

– **Decision Trees**: Arguably, the most popular type of predictive models in business applications is the DT. In its different forms, it has become an important knowledge extraction method, used for the classification of markets from static data or for the classification of future events [78].

Popular choices of DT in churn data analysis include the C5.0 classification tree -a variant of the well-known C4.5-, which assembles classification trees by recursively splitting the instance space into smaller subgroups, according to an information entropy criterion, until only instances from the same class remain known as a pure node, or a sub-group containing occurrences from different classes known as impure nodes. The tree is allowed to grow to its full potential before it is pruned back in order to increase its power of generalisation on unseen data. The C4.5 method was chosen in [74] to automatically generate classification rules for the purpose of identifying potential defectors, while C5.0 was recently applied to churn analysis in telecommunication markets in [94].

Another frequently used DT is the classification and regression tree (CART), constructed by recursively splitting the instance space into smaller sub-groups until a specified criterion has been met. The decrease in impurity of the parent node against the child nodes defines the goodness of the split. The tree is only allowed to grow until the decrease in impurity falls below a user-defined threshold. At this time the node becomes a terminal, or leaf node [5].

The literature contains quite a few examples of DTs for the construction of predictive models of abandonment. Datta and colleagues [21] developed a model called Churn Analysis Modelling and Prediction (CHAMP). CHAMP also uses DTs to predict customer churn in the telecommunications industry.

In the analysis of wireless telecommunications markets, the performance of DT, ANN and logistic regression was compared in [46]. This study stated that the DT showed slightly better accuracy over the other methods (however, the authors affirmed that these results do not prove DT to be the best choice in all cases). This conclusion is supported by [72], [27] and [73]. These models can also been used for the purpose of rule extraction. An example of the use of Naïve Bayes Trees for rule extraction in the credit card business of the banking sector can be found in [25]. DTs have also been successfully applied in recent years to problems such as email users churn prediction [75], supplier selection [106], broadband internet users churn [42], churn in telecommunications companies [45, 63] and credit card users churn [103, 57, 77, 25].

Computational Intelligence methods

CI methods provide, in one form or another, flexible information processing capabilities for handling real life problems. Exploiting the tolerance for imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness, low solution cost and close resemblance with human-like decision making, is the overall objective of CI methods [79]. Techniques that fall into this category include evolutionary computation (EC), ANNs and other ML techniques, fuzzy logic (see, for instance, [7] for an interesting clustering application) and their combinations, such as neuro-fuzzy systems [35]. We shall briefly describe and review them next.

– **Artificial Neural Networks**: ANNs are a family of ML models that have successfully been used to estimate complex non-linear functions and have been applied to many types of churn problems with high predictive accuracy, mostly related to classification, and prediction [54, 46,

3]. An important factor when considering the practical use of ANN is that they do not necessarily uncover patterns in an easily understandable and interpretable form [1], which may limit the scope of their implementation in practice unless they are accompanied by sound strategies of feature selection from *stage 2*.

Although ANNs are reported [21] to be still scarcely being used by companies in their day-to-day operations, our review reveals that many authors have used them in an entrepreneurial setting for churn analysis [65, 45, 85, 57, 10, 34, 93, 43, 110, 73, 96, 30] due to their high predictive accuracy. In a recent study, Tiwari and colleagues [92] described a novel ANN method which predicted customers that were likely to be churn in the future with less time margin than previous models.

- **Support Vector Machines**: This ML method, based on statistical learning theory, is able to optimally separate two class of objects (e.g., churners and retained customers) through the generation of a multivariate maximally separating hyperplane. Its theoretical basis was established in [8] and [15]. SVMs have been widely used in recent studies due to notable advantages such as a lower number of controlling parameters and good generalization capability [12, 41]. This method, though, remains difficult to interpret in terms of the input attributes [87] and may perform poorly in problems with strongly overlapping classes.

  Several studies [91, 98, 62, 94] have used SVM methods to predict churn in the telecommunications sector. It has also been applied to the analysis of churn of newspaper subscribers [17], concluding that SVMs outperform logistic regression as a predictive method. This technique was also used in [107] as base classifier in an ensemble model applied to the analysis of churn in both the telecommunications and banking sectors.
- **DMEL (Data Mining by Evolutionary Learning)**: This EC algorithm aims to overcome the limitations of interpretation and understanding of the results obtained through some CI techniques, in contrast, for instance, with the clarity of the if-then-rules obtained through DT. DMEL uses a non-random initial population based on first order rules. Higher-order rules are then obtained iteratively using a genetic algorithm (GA) type process. The fitness value of a chromosome uses a function that defines the probability that the attribute value is correctly determined using the rules it encodes. The likelihood of prediction is estimated and the algorithm handles missing values.

  DMEL was used to predict churn in the telecommunications industry in [1]. More recently, Yeshwanth *et al.* [109] used a hybrid model to predict churn of mobile networks customers, combining data pre-processing based on DT algorithms with a GA classification process.
- **Bayesian networks**: These are probabilistic graphical models that lay somewhere in between multivariate statistics and CI. A Bayesian network can be seen as a probabilistic "white box" that represents conditional dependencies over a set of discrete stochastic variables as a directed acyclic graph. As such, this type of model could nicely represent the possibly complex interdependencies between variables leading to the churn phenomenon.

  An attempt to estimate whether a new customer will increase or decrease future spending using Bayesian networks was reported in [2]. Verbraken and colleagues [100] have recently continued this work by testing the predictive power of a number of Bayesian network algorithms in churn analysis and proposing a feature selection method based on the concept of the Markov Blanket. Bayesian networks have been successfully applied in recent studies of churn forecasting in the field of telecommunications [55] and in banking [103].

Other alternative methods

This category does in fact include methods that could at least partially fit into the previous categories, but which have scarcely been used in the context of churn analysis. They can thus be considered as somehow exotic in this context and include: Ensemble Learning in different forms, semi-Markov processes, mixture transition distribution models and goal-oriented sequential pattern algorithms.

- **Semi-Markov processes**: They were used in [47] to create a model that considers e-customer behaviour. The discrete-time semi-Markov process was designed as a probabilistic model for use in the analysis of complex dynamic systems. It has also been used in [89] to study customer retention.
- **Mixture transition distribution**: Prinzie and Van den Poel [82] introduced a mixture transition distribution (MTD) to investigate purchase-sequence patterns. The MTD was designed to allow estimations of high order Markov chains, providing a smaller transition matrix facilitating managerial interpretation.
- **Goal-oriented sequential pattern**: a novel algorithm for identifying potential churners using association rules that identify relationships amongst variables was introduced in [14]. The authors defined a two-step process for finding out association rules. In the first step, the large item set (attribute-value pairs) is defined, requiring compliance with certain minimum conditions of support and minimum confidence defined by the researcher. In the second stage, an *A Priori* algorithm is used to explore the rules of association.

3.1.4 Stage 4: Validation of results.

This is again a key to the success of any data-based churn analysis and corresponds to the *evaluation* phase in DM. Some of the most commonly used methods for model validation in the churn analysis literature are:

- **Cross-validation**: Most suitable in those cases in which data are scarce. In its most simple version, a single split of the data is generated (such as the 70/30 used in [46]; the 70% of cases used as training set and the 30% remaining as validation set). Cross-validation is based on the principle of using the available data for both training and validation. Several more sophisticated cross-validation methods have been proposed in the literature [35], including:
  - **K-fold cross-validation**: The learning set is randomly partitioned into K subsets of equal size. Each individual subset is then used in turn for validation, while the rest of the data are used for training. In the extreme case of choosing single case folds, the procedure is called leave-one-out cross-validation.
  - **Monte Carlo (or repeated random sub-sampling) cross-validation**: The learning set is repeatedly divided into two random sets, one of which is used for training and the other for validation. Not all cases are necessarily chosen at any point for validation. See, for instance, [94] for an application of the method to churn analysis.
- **Separate validation dataset**: Several authors [21, 5, 82] have successfully used single validation sets separated from the training sets in the validation of their predictive models of abandonment. This method should only be acceptable in those cases in which data availability is not an issue.

When testing the validity of a model, or comparing the results of different methods, the following set of indicators are the most commonly used [64, 22]:

- **Accuracy, Sensitivity and Specificity**: For classification models with a binary target variable. ***Accuracy*** measures the ratio of correctly classified observations (churn or not-churn) to the overall number of cases. ***Sensitivity*** instead measures the ratio of correctly predicted events (i.e., correct churn predictions) to the total number of events (total of churn cases), whereas ***specificity*** measures the ratio of correctly predicted non-events (i.e., not-churn) to the total number of non-events (total of not-churn cases). Although accuracy is intuitive and commonly used to compare prediction methods [64, 103, 77], it is not considered to be an optimum figure of merit for churn modelling because it is unreliable in a situation of class imbalance [61], which is by far the most common in this application area [83].

- **Area under the Receiver Operating Characteristic (ROC) curve**: ROC is a function of the *sensitivity* versus $1 - specificity$ for all values of the classification threshold and it has been widely used in the churn prediction literature [44, 17, 70, 32, 77, 61, 22, 105, 107]. Its Area Under the Curve (AUC), unlike accuracy, evaluates the ability of a classifier to distinguish between classes based on the predicted class membership probabilities and is therefore suitable for imbalanced classification problems [58].
- **Lift Chart**: This technique focuses on the segment of highest-risk customers, arranging them into deciles based on their predicted probability to churn and comparing its results with the rest of the cases. The Lift Chart has also been commonly used in recent studies [70, 32, 77, 16, 22]. It can be found in two different forms:
  - **Top decile Lift** (TDL): Is the churn rate in the top decile of ordered posterior churn probabilities over the churn rate in the total customer population [61].
  - **Lift Index** (LI): Is the weighted index of the correctly predicted churners, ranked by its posterior churn probability [19].
- **Loss function**: Calculated on the basis of customers' Life Time Value (LTV), this method indicates the loss caused by the error of the model, considering the effect of misclassified customers. Some examples can be found in recent work [77, 30].

3.2 A summarized review of the literature

We have compiled the reviewed literature in a number of detailed summary tables: 2 to 14, which can be found in the Appendix. They list the main references in recent literature (roughly over the last 15 years and including mainly peer-reviewed journal publications, although some conference publications and PhD thesis have also been included for their specific interest) that address the problem of building predictive models of abandonment. Following the same scheme of stages proposed for the previous section as guiding index, these tables show: the references to the articles; the type of data used in the analysis; the source from which those data were obtained; the attribute selection technique employed; the possible use of time series data in its definition; the techniques used to develop the predictive models and, finally, the methods used for validation, if any.

These detailed tables are organized according to two main criteria:

- According to the predictive methods used: Tables 2, 3 and 4 for standard techniques; tables 5, 6 and 7 for CI methods; and tables 8 and 9 for alternative ones.
- According to fields of application: Tables 10 and 11 (telecommunications), 12 (banking), and 13 and 14 (other areas of application).

Here, we show a single summary table structuring the reviewed publications according to, first, business area and, subsequently, type of analytical method (*standard methods*, CI *methods* and *other alternative methods*).

[Table 1 here]

4 Discussion

This section of discussion of the reviewed literature is articulated, first, according to the four stages for building predictive models of abandonment described in section 3.1. Comments and recommendations are made for each of these stages, followed by specific comments and recommendations for, in turn, the two main blocks of industries in which the tabular results are organized in the appendix, namely Telecommunications and Banking.

4.1 *Stage 1*: Identifying and obtaining the best data.

4.1.1 Main issues:

Some interesting conclusions can be drawn from the way the reviewed studies addressed stage 1 of the prediction model building process. The types of data gathered for analysis are quite varied and, to some extent, depend on the area of application. Overall, they do not seem to depend on the type of analysis technique applied in the study. Usage, socio-economic and demographic variables predominate, and the use of RFM attributes is quite common. Curiously, given their likely predictive power, very few studies resort to analysis of standard data concerning quality of service or customer satisfaction, such as the SERVQUAL (e.g. in [3]). This is probably due to the difficulty of obtaining this type of information from customers (in comparison with, for instance, usage or socio-demographics), but perhaps also due to a lack of awareness of their usefulness as predictors of churn.

Another key issue is the fact that churn should be treated as a dynamic process that naturally evolves over time. For this reason, it is surprising that most of the studies are either static from the point of view of the data gathered for analysis (that is, most of the investigated data are market *snapshots* that lack information on its evolution), or unclear about the periodicity of the surveyed data, or used data that were averaged over time. Even among those studies that used data collected over time, many just cover a few months, making it difficult to discount seasonality effects. Some exceptions include the impressive 77 years of data from a financial services company that were analysed in [96]; the 30 months of data used to create the model plus 12 months of data for prediction used in [17]; the 17 + 4 months in [16]; the 18 months in [39]; and the 12 + 12 months periods used in [103], [77] and [13]. Again, this could be at least partially justified by the cost and difficulty of gathering consistently homogeneous customer data over long periods of time in rapidly changing scenarios.

Beyond the problem of choosing the most adequate set of data attributes, the paradox is that, as remarked in [96] and corroborated in [32] a decade later, few studies, even if using longitudinal surveys, make use of dynamic analysis models. Little work in churn analysis adventures beyond limited time windows both for the data to be modeled and for the predictions. It has been proposed [32] that the most natural framework for this problem is longitudinal data analysis and its related methods such as survival analysis (in their study, using Cox regression), that is, a dynamic analytical framework that allowed to allocate customer retention efforts across time and identify early indicators of attrition. The design and development of such dynamic models is one of the most important pending challenges in churn analysis. A very interesting example of a proper dynamic analysis of churn can be found in recent work by Bose and Chen [7].

4.1.2 Recommendations per industry.

Telecommunications: Usage data seem to be the common denominator of recent studies in this business domain [74, 107, 94, 7] and, as seen from the previous paragraphs, these types of data are often complemented and enriched using demographics, socio-economic and marketing information.

Beyond that, recent technological advances have allowed to add a new type of data to the churn prediction mix that are of special relevance to the telecommunications sector: those generated by the customers' social communication networks, which could be analyzed under the conceptual framework of *social commerce* [31]. Sharing a somehow similar outlook on the problem, Kim and colleagues [53] developed procedures for churn prediction by examining the communication patterns among service subscribers, whereas Verbeke and co-workers [99] developed relational classifiers based on intercommunication graphs that allow to incorporate social network effects within a churn prediction model. This social communications network information could be easily added

to the data mix in other industries beyond telecommunications, mostly those with strong internet presence.

Banking and financial services: Interestingly, the banking sector seems to take a much narrower approach to the identification of the best data for churn analysis, at least as compared to telecommunications and other industries. An example of this is the fact that a number of churn studies in this review, mostly not related to banking and finances and including [40, 95, 81, 39, 102, 50, 66, 67, 13, 70] agree in suggesting the use of three groups of variables, globally known as RFM (*Recency*, *Frecuency* and *Monetary* variables):

- length of time since last purchase,
- frequency of use,
- economic expense effected over a certain time period,

as a source for the prediction of the churn probability. Note that most of these studies are in the areas of direct marketing and retailing, with the only exception of [40], which is an application to banking. Bose and Chen [6], when discussing direct marketing, stated that RFM variables are amongst the strongest performing variables in explaining future customer behavior. Nothing prevents the banking sector from using variations of RFM to enrich their churn analysis processes, especially given that modern banking often provides their customers with extra services that are not that far removed from the retail sector.

4.2 *Stage 2*: selection of attributes

4.2.1 Main issues

The investigation of stage 2 of the prediction model building process provides a really mixed picture in the reviewed literature. Many studies do not even consider attribute selection, while quite a few others justify the selection not on the quantitatively demonstrated impact of the data on the prediction but, instead, on domain knowledge (using attributes that have shown relevance in previous similar studies) or expert knowledge in the field (again, trusting prior experiences). As mentioned in the previous section, even among those investigations using some form of quantitative attribute selection, this often takes a very simple form, far from the state-of-the-art in the field and unrelated to the modelling technique itself and only in a few cases the attribute selection technique is intrinsically associated to the modelling technique of choice.

This is a worrisome picture for the following reasons:

- Almost any business domain is becoming data-dependent in most of its processes and, therefore, more and more data attributes are likely to become available for a process such as abandonment prediction. This could be particularly true for those industries whose environment is the internet, such as web retailers. With the increase of information availability, comes a harder problem of selection of the adequate attributes to include in the analysis of supplier abandonment by the customer. For these reasons, the use of adequate procedures of attribute selection should become compulsory.
- There is nothing intrinsically wrong with the idea of using domain knowledge or expert knowledge in the specific business area. Much is to be learnt from prior experiences, but excessive reliance on prior knowledge may come accompanied by a higher risk of negative bias on the selection of attributes for the churn problem, as it might prevent from using attributes whose relevance might not be obvious to the human expert, but could still be found by data-based quantitative attribute selection methods.

– It is particularly unusual that such a small number of the reviewed studies using CI methods resort to attribute selection procedures, given the huge effort in this direction that has been made in the CI field. There are countless selection procedures specifically devised for each of the many CI methods that have been employed in the reviewed supplier abandonment studies [33, 68].

4.2.2 Recommendations per industry

Telecommunications: As seen in the previous section, the reviewed studies of churn analysis in the telecommunications sector are precisely characterized by a rich usage of data attributes, including usage data and often also using demographics, socio-economic and marketing information. This complex mixture of attributes would be enough motivation for the use of attribute selection procedures. Not using such procedures would entail running a few risks, including:

– The combination of datasets with a large number of cases (becoming more and more common in this field) *and* a large number of attributes may become burdensome from a computational viewpoint, specially for complex modeling techniques.
– Many of the available attributes maybe uninformative for prediction purposes, or redundant, negatively affecting the prediction capabilities of the churn prediction models.
– Without appropriate attribute selection, it becomes impossible to ascertain which attributes are the most relevant for churn classification and prediction. This becomes an easily avoidable limitation for knowledge extraction in the domain and precludes experts from saving unnecessary data gathering efforts.

For all these reasons, it is strange that a majority of the reviewed studies in the area of telecommunications either do not report any attribute selection at all, or base the selection on expert and domain knowledge. In both cases, this is tantamount to not performing any data-based quantitative attribute selection. Only a minority of papers report the use of either statistical techniques (such as the R^2 or the Fisher score), or methods associated to non-traditional models. The obvious recommendation here would be to encourage researchers in the area to take attribute selection seriously as an almost compulsory part of the DM process oriented towards novel knowledge discovery. An example of a proper and intensive attribute selection process using CI methods can be found in [107].

Banking and financial services: The banking and financial services sector provides us with the opposite side of the coin in terms of attribute selection, when compared to telecommunications. Even though there are still too many papers with either no attribute selection at all or a selection based only on expert and domain knowledge, the majority of the reviewed publications have used quantitative selection methods based on either statistics or embedded in the own modeling process. Further good news is that the most recent publications of churn analysis in banking seem to be paying adequate attention to this issue; examples of these are the detailed methods that can be found in [25] and, again, in [107].

4.3 *Stage 3*: Development of a predictive model.

4.3.1 Main issues:

The key issue in this stage is, precisely, the choice of the most adequate model or models (separately or in combination) for churn analysis.

From the summary Figures 5 and 6, it is clear that two of the *standard* methods are, by far, the most popular ones in the reviewed literature: DTs and Regression Analysis. Their reasonable

predictive accuracy together with their wide acceptance and ease of implementation are arguably the main reasons behind the fact that more than 46% of the reviewed methods (see Figure 5, left) belong to these categories. They are mentioned in more than 70% of the analysed literature, acting as a benchmark to compare the results of other methods, or as the main predictive method. These methods have remained popular over the years, even increasing to a 51% share of the references published in the last 5 years (see Figure 5, right).

A further and perhaps more important reason to explain this popularity is their transparency: the outcomes of both families of methods can easily be interpreted in terms of the attributes used to create the model; as a result, business rules to explain the churn phenomenon are reasonably easy to come by. In the particular case of DT, such rules take an attractive hierarchical form resulting from the tree-like model that branches according to specific attributes.

For their part, *Computational Intelligence* methods are also reasonably well accepted but still a bit of a novelty despite their long-standing record in many other application fields, including business [97, 65]. ANNs are the most commonly used ones (in 16% of the reviewed literature), although in decline over the last few years. Their proficient predictive ability turns them into very attractive methods for researchers, but their lack of straightforward interpretability may slow their usage down in entrepreneurial environments [104, 2]. Such lack of interpretability is often the result of the nonlinear nature, which confers them with great modeling flexibility at the price of obscuring the relationship between the outcome and the data attributes. The adoption of CI methods in this area may also be slowed down by the lack of clear standards for their implementation.

Finally, the denominated *alternative* methods, less specific and constantly improving, have only a marginal role ($17,2\%$ of the studied literature in recent years) as methods of choice in supplier abandonment modelling.

Despite the obvious importance of such approach, comparative analyses in which the more traditional statistical methods are exhaustively compared with CI techniques are somehow still uncommon in this application field [51, 94]. Such lack of comparisons limits the possibility of properly assessing the relative virtues and limitations of each of the methods.

A sound alternative to model comparison is Ensemble Learning, in which the predictions yielded by each of the models (learners) in the ensemble are combinations of the individual predictions of multiple algorithms. They have shown strong and robust prediction performance in many application areas. According to the combination of algorithms, the most popular ensemble learning methods in the field (although not the only ones; see for instance [107] and [94]) are:

- **Random Forest (RF)**: It is a combination of Bagging [9], Random Subspace Method [38] and CART DTs [5]. RFs solve the high instability that hampers the use of DT and they have been used in several marketing research studies [11, 10, 59] due to their high predictive performance and robustness to outliers and noise. Research about churn in newspaper subscribers [17] found that RF outperformed SVMs and logistic regression, and its usefulness was recently confirmed when predicting abandonment in the online gaming industry [16].
- **GAMens**: A combination of Bagging and Random Subspace Method (RSM) with Generalized Additive Models (GAM). The latter [17, 4] is a flexible technique for nonparametric regression. Recent work [23] has proven that GAMens can be competitive with RF in accuracy. More recently, they proved the predictive ability of this method applied to problems in several industries such as supermarkets, banking and telecommunications [22].

4.3.2 Recommendations per industry

Telecommunications: From an industry of application viewpoint, the reviewed literature reveals (see Figure 6, left) that telecommunications is the more active sector in terms of churn prediction modelling research (46% of the publications apply to this sector), followed by banking, found in 23% of literature. This makes sense, as churn is a much more pressing concern in the very volatile telecommunications market.

The profile of analytical model choice revealed by the graphics in Figure 6 (right) is quite interesting, as it indicates that telecommunications is, out of the reviewed sectors, the one in which less traditional models are being tested. This could perhaps be explained by the fact that banking is, by comparison, a more mature type of business with a long-standing tradition of use of statistical methods (which is a similar situation to the one observed, for instance, in clinical medicine, in which statistical methods are still very prevalent). It is also curious to note that telecommunications includes a majority of studies in which several different techniques are compared or combined, which is a very positive approach. Combining both previous thoughts, an specific recommendation for this sector would be the integration of multivariate statistical approaches in the comparison and combination of methods for churn analysis.

Banking and financial services: It can be observed from Figure 6 (right) that in banking and financial services, *alternative* methods are extraordinarily prevalent, much in contrast with the figures for the telecommunications sector. The suspicion that such contrast in the type of model choice is mostly motivated by the internal inertias of each industry is impossible to avoid, given that the churn problem itself is not that industry-dependent and the type of analyzed data does not justify the use of radically different methods.

The relative scarcity of the use of CI methods in this sector is somehow surprising, given that it was precisely in banking and financial services that CI scored some its first extremely successful real-world implementations, pioneering commercial software in the 1990's for problems such as bankruptcy prediction and credit scoring, amongst others [97]. Churn analysis in banking and financial services could, and perhaps should, benefit from the use of this accumulated knowledge and the resulting best-practice guidelines. This would also benefit from a more integral DM approach to data analysis that embedded the attribute selection procedures of *stage 2* and robust validation techniques from *stage 4*, an integration that is usually straightforward in CI methods.

4.4 *Stage 4*: Validation of results

4.4.1 Main issues

The last stage of the prediction model building process, the *validation of results*, reveals that the literature in the field is far from agreeing on any standards. The dominating evaluation metric is the accuracy, which is often inadequate in situations of class unbalance, that is, when the prevalence of one of the classes is much higher than that of the rest. Needless to say, this is the most common scenario in churn analysis, where the number of churning cases is often much lower than the number of non-churning cases. Even though, sensitivity, specificity and the related precision, recall and AUC measures, which are much more adequate metrics in situations of class unbalance, are also used in quite a few studies. Lift charts are also quite commonly used. Note that sensitivity should be given special relevance in this context, as it measures the ratio of correctly predicted events (correct churn predictions) to the total number of events (total of churn cases). In most real scenarios, service providers should strive to maximize this metric, because correctly detecting the propensity of abandonment, which is a relatively uncommon event, is the most difficult task in churn analysis, but also the most important one for CCM.

Surprisingly, many studies do not seem to include any form of validation explicitly, something that should be compulsory if we aim to assess the ability of the model to generalize its results to unseen data. Such requirement should be specially relevant in churn analysis, in which the main goal is inferring the potential supplier abandonment of out-of-sample customers. This is a surprising finding, given that data scarcity is not a problem for any of the problems under analysis and, therefore, there is no barrier for the definition of a proper training-validation-test setting.

Out of those studies using validation, quite a few make use of different forms of cross-validation. The use of separate subsets for the three-way process of building, validating, and testing the model

(which is the most appropriate and robust approach of validation) seems to be rarely considered in this area. This means that many of the reported results risk being, at best, over-optimistic estimations, or, at worst, or strongly biased and unreliable ones.

4.4.2 Recommendations per industry

Telecommunications: The use of appropriate validation seems specially important in churn analysis for the telecommunications sector. This is because of its higher levels of customer volatility and its lower switching barriers, which should make the adequate assessment of out-of-sample customers a more pressing concern. From Tables 10 and 11 in the appendix, it is clear that cross-validation is the most favoured strategy, but still a few studies only report a single training/test data split, which is a clearly sub-optimal validation strategy that risks biasing the results, undermining our confidence on their reliability. Even worse, quite a few studies in this field completely ignore the validation strategy, which means that their results could easily be over-optimistic as they might just reflect an over-fitting of a very specific and not necessarily representative data sample.

Another issue that demands attention in this area, related to one of the previous general recommendations, is that only a handful of the reviewed papers pay attention to the calculation of the sensitivity or related performance metrics.

Banking and financial services: Customers in banking and financial services might be far less volatile than those of telecommunication service providers. This, in principle, could mean that population samples are likely to be more stable and, therefore, models of such populations are less likely to be affected by data over-fitting; that is, out-of-sample estimations are bound to be more similar to the in-sample ones. Having said that, it is still surprising the extent to which many churn studies in this area do not provide any proper validation strategy for their models. Only one of the reviewed papers [25] resorts to a standard complete training-validation-test strategy with 10-fold cross-validation. On a more positive note, most studies (specially the most recent ones) use adequate evaluation metrics, including sensitivity, specificity and AUC.

5 Conclusions

Supplier abandonment is one of the main problems faced by service-providing companies in rapidly changing and extremely competitive mature markets. This is a problem that can only be addressed through personalization strategies oriented to retain valuable customers. When the market base is large, such personalization is only feasible as a data-intensive business analytics problem.

In such a scenario, a DM methodology based on IDA may become a useful knowledge and information management tool for knowledge extraction from domain data. This paper has surveyed and reviewed recent literature in which the use of IDA to build predictive models has been proposed to address the churn problem, with a (non-exclusive) focus on the use of CI techniques. Importantly, this literature has been reviewed according to the different stages of DM implementation, from data gathering and understanding, to predictive model validation. Different business areas of application have also been independently considered, with a focus on telecommunications, banking and financial services. Publications in the field are abundant and we have tried to narrow the scope of the review by prioritizing journal publications and covering the period 2000-2015. The reviewed studies are disseminated throughout a large number of international journals (over 20), although two of them stand out for the special attention paid to the problem of churn analysis, namely Expert Systems with Applications and the European Journal of Operational Research.

From the existing literature, the relevance of adequate data gathering and attribute selection procedures becomes evident. It is also clear that no particular IDA method has the upper hand in terms of results, which means that the choice of method is very problem-dependent. It must

be noted, though, that there are trends in the choice of methods that seem to be motivated by nothing more than their popularity in a given application area. Given that, with few exceptions, data characteristics do not substantially vary over areas of application, similar methods or their combinations could be used in any of them. We argue in favor of the use of several methods for performance comparison purposes, where the simplest amongst them could be used to obtain a performance baseline. Alternatively, model combination methods such as those of the Ensemble Learning family are worth investigating.

References

1. Au WH, Chan KC, Yao X (2003) A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Transactions on Evolutionary Computation 7(6):532–545
2. Baesens B, Verstraeten G, Van den Poel D, Egmont-Peterson M, Van Kenhove P, Vanthienen J (2004) Bayesian network classifiers for identifiying the slope of the customer lifecycle of long-life customers. European Journal of Operational Research 156(2):508–523
3. Behara RS, Fisher WW, Lemmink JG (2002) Modelling and evaluating service quality measurement using neural networks. International Journal of Operations and Production Management 22(10):1162–1185
4. Berg D (2007) Bankruptcy prediction by generalized additive models. Applied Stochastic Models in Business and Industry 23(2):129–143
5. Bloemer JM, Brijis T, Vanhoof K, Swinnen G (2003) Comparing complete and partial classification for identifying customers at risk. International Journal of Research in Marketing 20(2):117–131
6. Bose I, Chen X (2009) Quantitative models for direct marketing: A review from systems perspective. European Journal of Operational Research 195(1):1–16
7. Bose I, Chen X (2015) Detecting the migration of mobile service customers using fuzzy clustering. Information & Management 52(2):227–238
8. Boser BE, Guyon IM, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: Fifth Annual Workshop on Computational Learning Theory, Pittsburg, pp 114–152
9. Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140
10. Buckinx W, Van den Poel D (2005) Customer base analysis: Partial defection of behaviorally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research 164(1):252–268
11. Burez J, Van den Poel D (2007) CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications 32(2):277–288
12. Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2):121–167
13. Chen ZY, Fan ZP, Sun M (2012) A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. European Journal of Operational Research 223(2):461–472
14. Chiang D, Wang Y, Lee S, Lin C (2003) Goal-oriented sequential pattern for network banking and churn analysis. Expert Systems with Applications 25(3):293–302
15. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273–297
16. Coussement K, De Bock KW (2013) Customer churn prediction in the online gambling industry: the beneficial effect of ensemble learning. Journal of Business Research 66(9):1629–1636
17. Coussement K, Van den Poel D (2008) Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34(1):313–327

18. Crespo F, Weber R (2005) A methodology for dynamic data mining based on fuzzy clustering. Fuzzy Sets and Systems 150(2):267–284
19. Crone SF, Lessmann S, Stahlbock R (2006) The impact of preprocessing on data mining. an evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research 173(3):781–800
20. Cronin JJ, Brady MK, Hule GT (2000) Assessing the effects of quality, value, and customer satisfaction on customer behavioural intentions in service environments. Journal of Retailing 76(2):193–218
21. Datta P, Masand B, Mani PR, Li B (2000) Automated cellular modelling and prediction on a large scale. Artificial Intelligence Review 14(6):485–502
22. De Bock K, Van den Poel D (2012) Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. Expert Systems with Applications 39(8):6816–6826
23. De Bock K, Coussement K, Van den Poel D (2010) Ensemble classification based on generalized additive models. Computational Statistics & Data Analysis 54(6):1535–1546
24. Dy J, Brodley C (2004) Feature selection for unsupervised learning. The Journal of Machine Learning Research 5(1):845–889
25. Farquad M, Ravi V, Raju S (2014) Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing 19:31–40
26. Fayyad U, Piatetski-Shapiro G, Smith P (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3):37–54
27. Ferreira JB, Vellasco M, Pacheco MA, Barbosa CH (2004) Data mining techniques on the evaluation of wireless churn. In: Proceedings of the 12^{th} European Symposium on Artificial Neural Networks (ESANN), pp 483–488
28. García DL, Vellido A, Nebot À (2007) Customer continuity management as a foundation for churn data mining. Technical report LSI-07-2-R, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
29. Gilbert R (1989) Mobility barriers and the value of incumbency. Handbook of Industrial Organisation 1:475–535
30. Glady N, Baesens B, Croux C (2008) Modeling churn using customer lifetime value. European Journal of Operational Research 197(1):402–411
31. Gonçalves Curty R, Zhang P (2011) Social commerce: Looking back and forward. Proceedings of the American Society for Information Science and Technology 48(1):1–10
32. Gür-Ali O, Aritürk U (2014) Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Systems with Applications 41(17):7889–7903
33. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. The Journal of Machine Learning Research 3(1):1157–1182
34. Hadden J (2008) A customer profiling methodology for churn prediction. PhD thesis, Cranfield University, UK
35. Hadden J, Tiwari A, Roy R, Ruta D (2007) Computer-assisted customer churn management: State-of-the-art and future trends. Computers and Operations Research 34(10):2902–2917
36. Haenlein M, Kaplan AM (2012) The impact of unprofitable customer abandonment on current customers exit, voice, and loyalty intentions: an empirical analysis. Journal of Services Marketing 26(6):458–470
37. Haenlein M, Kaplan AM, Schoder D (2006) Valuing the real option of abandoning unprofitable customers when calculating customer lifetime value. Journal of Marketing 70(3):5–20
38. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8):832–844
39. Ho Ha S, Min Bae S, Chan Park S (2002) Customers' time-variant purchase behaviour and corresponding marketing strategies: an online retailer's case. Computers and Industrial Engi-

neering 43(4):801–820
40. Hsieh N (2004) An integrated data mining and behavioural scoring model for analysing bank customers. Expert Systems with Applications 27:623–633
41. Hsu C, Chang C, Lin C (2008) A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taiwan
42. Huang BQ, Kechadi MT, Buckley B (2009) Customer churn prediction for broadband internet services. In: T B Pedersen MKM, Tjoa M (eds) DaWaK 2009, LNCS 5691, Springer-Verlag, Berlin, pp 229–243
43. Huang BQ, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. Expert Systems with Applications 39(1):1414–1425
44. Huang Y, Kechadi T (2013) An effective hybrid learning system for telecommunication churn prediction. Expert Systems with Applications 40(14):5635–5647
45. Hung S, Yen DC, Wang HY (2006) Applying data mining to telecom churn management. Expert Systems with Applications 31(3):512–524
46. Hwang H, Jung T, Suh E (2004) An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Systems with Applications 26(2):181–188
47. Jenamani M, Mohapatra PK, Ghose S (2003) A stochastic model of e-customer behaviour. Electronic Commerce Research and Applications 2(1):81–94
48. Jones MA, Mothersbaugh DL, Beatty SE (2000) Switching barriers and repurchase intentions in services. Journal of Retailing 70(2):259–270
49. Jones MA, Mothersbaugh DL, Beatty SE (2002) Why customers stay: measuring the underlying dimensions of services switching costs and managing their differential strategic outcomes. Journal of Business Research 55(6):441–450
50. Jonker J, Piersma N, Van den Poel D (2004) Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert Systems with Applications 27(2):159–168
51. Keramatia A, Jafari-Marandi R, Aliannejadi M, Ahmadian I, Mozaffari M, Abbasi U (2014) Improved churn prediction in telecommunication industry using datamining techniques. Applied Soft Computing 24:994–1012
52. Kim H, Yoon C (2004) Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications Policy 28(9-10):751–765
53. Kim K, Jun C, Lee J (2014) Improved churn prediction in telecommunication industry by analyzing a large network. Expert Systems with Applications 41(15):6575–6584
54. Kim MK, Park MC, Jeong DH (2004) The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy 28(2):145–159
55. Kisioglu P, Topcu YL (2010) Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey. Expert Systems with Applications 38(6):7151–7157
56. Klemperer P (1987) Markets with consumer switching cost. The Quarterly Journal of Economics 102(2):375–394
57. Kumar D, Ravi V (2008) Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies 1(1):4–28
58. Langley P (2000) Crafting papers on machine learning. In: Langley P (ed) 17th International Conference on Machine Learning (ICML 2000), Stanford University, pp 1207–1216
59. Larivière B, Van den Poel D (2005) Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications 29(2):472–484
60. Lee TS, Chiu CC, Chou YC, Lu CJ (2006) Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics and

Data Analysis 50(4):1113–1130
61. Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. Journal of Marketing Research 43(2):276–286
62. Lessmann S, Voß S (2009) A reference model for customer-centric data mining with support vector machines. European Journal of Operational Research 199(1):520–530
63. Lima E, Mues C, Baesens B (2010) Monitoring and backtesting churn models. Expert Systems with Applications 38(1):975–982
64. Lima EO (2009) Domain knowledge integration in data mining for churn and customer lifetime value modelling: New approaches and applications. PhD thesis, University of Southampton, Faculty of Law, Arts and Social Sciences, Southampton, UK
65. Lisboa PJ, Edisbury B, Vellido A (2000) Business Applications of Neural Networks. World Scientific Publishing Co
66. Liu D, Shih Y (2005) Hybrid approaches to product recommendation based on customer lifetime value and purchase references. The Journal of Systems and Software 77(2):181–191
67. Liu D, Shih Y (2005) Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management 42(3):387–400
68. Liu H, Motoda H (2007) Computational Methods of Feature Selection. Chapman and Hall / CRC Data Mining and Knowledge Discovery Series
69. Madden G, Savage SJ, Coble-Neal G (1999) Subscriber churn in the australian ISP market. Information Economics and Policy 11:195–207
70. Miguéis V, Camanho A, Falcão e Cunha J (2013) Customer attrition in retailing: An application of multivariate adaptive regression splines. Expert Systems with Applications 40(16):6225–6232
71. Mihelis G, Grigoroudis E, Siskos Y, Politis Y, Malandrakis Y (2001) Customer satisfaction measurement in the private bank sector. European Journal of Operational Research 130(2):347–360
72. Mozer MC, Wolniewicz R, Grimes DB, Johnson E, Kaushansky H (2000) Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Transactions on Neural Networks 11(3):690–696
73. Neslin SA, Gupta S (2006) Defection detection: Improving predictive accuracy of customer churn models. Journal of Marketing Research 43(2):204–211
74. Ng K, Liu H (2000) Customer retention via data mining. Artificial Intelligencie Review 16(4):569–590
75. Nie G, Zhang L, Li X, Shi Y (2006) The analysis on the customers churn of charge email based on data mining. In: Sixth IEEE International Conference on Data Mining (ICDM), Springer-Verlang, Hong Kong, China, pp 843–847
76. Nie G, Wang G, Zhang P, Tian Y, Shi Y (2009) Finding the hidden pattern of credit card holder's churn: a case of China. In: Computational Science - ICCS 2009, Springer-Verlag, Berlin, pp 561–569
77. Nie G, Rowe W, Zhang L, Tian Y, Shi Y (2011) Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications 38(12):15,273–15,285
78. Osei-Bryson KM (2004) Evaluation of decision trees: a multi criteria approach. Computers and Operations Research 31(11):1933–1945
79. Pal SK, Ghosh A (2004) Soft computing data mining. Information Sciences 163(1-3):5–12
80. Patterson MP, Smith T (2003) A cross-cultural study of switching barriers and propensity to stay with service providers. Journal of Retailing 79(2):107–120
81. Pfeifer PE, Carraway RL (2000) Modeling customer relationships as Markov chains. Journal of Interactive Marketing 14(2):43–55
82. Prinzie A, Van den Poel D (2006) Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. European Journal of Operational research 170(3):710–734

83. Provost P, Fawcett T, Kohavi R (2000) The case against accuracy estimation for comparing induction algorithms. In: 15th International Conference on Machine Learning (ICML 1998), Morgan Kaufman, Madison, Wisconsin, pp 445–453
84. Ranaweera C, Neely A (2003) Some moderating effects on the service quality-customer relation link. International Journal of Operations and Production Management 23(2):230–248
85. Rygielski J, Wang J, Yen DC (2002) Data mining techniques for customer relationship management. Technology in Society 24(4):483–502
86. Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing 5(4):13–22
87. Shin HW, Sohn SY (2004) Segmentation of stock trading customers according to potential value. Expert Systems with Applications 27(1):27–33
88. Slater SF, Narver JC (2000) Intelligence generation and superior customer value. Journal of the Academy of Marketing Science 28(1):120–127
89. Slotnick SA, Sobel MJ (2005) Manufacturing lead - time rules: Customer retention versus tardiness costs. European Journal of Operational Research 163(3):825–856
90. Sun Z, Bebis G, Miller R (2004) Object detection using feature subset selection. Pattern Recognition 37(11):2165–2176
91. Suryadi K, Gumilang S (2008) Actionable decision model in customer churn monitoring based on support vector machines technique. In: 9^{th} Asia Pacific Industrial Engineering and Management Systems Conference, Bandung, Indonesia
92. Tiwari A, Hadden J, Turner C (2010) A new neural network based customer profiling methodology for churn prediction. In: ICCSA 2010, Springer-Verlag, Berlin, chap IV, pp 358–369
93. Tsai CF, Lu YH (2009) Customer churn prediction by hybrid neural networks. Expert Systems with Applications 36(10):12,547–12,553
94. Vafeiadis T, Diamantaras K, G S, Chatzisavvas K (2015) A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory 55:1–9
95. Van den Poel D (2003) Predicting mail-order repeat buying: which variables matter? Tijdschrift voor Economie and Management 48(3):371–403
96. Van den Poel D, Laraviére B (2004) Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational research 157(1):196–217
97. Vellido A, Lisboa PJ, Vaughan J (1999) Neural networks in business: a survey of applications (1992-1998). Expert Systems with Applications 17(1):51–70
98. Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2011) New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218:211–219
99. Verbeke W, Martens D, Baesens B (2014) Social network analysis for customer churn prediction. Applied Soft Computing 14, part C:431–446
100. Verbraken T, Verbeke W, Baesens B (2014) Profit optimizing customer churn prediction with Bayesian network classifiers. Intelligent Data Analysis 18(1):3–24
101. Verhoef PC, Donkers B (2001) Predicting customer potential vualue, an application in the insurance industry. Decision Support Systems 32(2):189–199
102. Verhoef PC, Spring PN, Hoekstra JC, Leeflang PS (2003) The commercial use of segmentation and predictive modelling techniques for database marketing in the Netherlands. Decision Support Systems 34(4):471–481
103. Wang G, Liu L, Peng Y, Nie G, Kou G, Shi Y (2010) Predicting credit card holder churn in banks of China using data mining and MCDM. In: IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp 215–218
104. Wei CP, Chiu LT (2002) Turning telecommunications call details to churn preduction: a data mining approach. Expert Systems with Applications 23(2):103–112
105. Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques, 2^{nd} edn. Morgan Kaufman, San Francisco

106. Wu D (2009) Supplier selection: A hybrid model using DEA, decision tree and neural network. Expert Systems with Applications 36(5):9105–9112
107. Xiao J, Xiao Y, Huang A, Liu D, Wang S (2015) Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowledge and Information Systems 43(1):29–51
108. Yan L, Wolniewicz RH, Dodier R (2004) Predicting customer behaviour in telecommunications. IEEE Intelligent Systems 19(2):50–58
109. Yeswanth V, Vimal Raj V, Saravanan M (2011) Evolutionary churn predition in mobile networks using hybrid learning. In: Proceedings of the 24^{th} International Florida Artificial Intelligence Research Society Conference (FLAIRS), Florida, USA, pp 471–476
110. Zhao Y, Li B, Li X, Liu W, Ren S (2005) Customer churn prediction using improved one-class support vector machine. In: Li X, Wang S, Yang-Dong Z (eds) Advanced Data Mining and Applications, Springer, Lecture Notes in Computer Science, Vol.3584, pp 300–306

6 Appendix

The detailed tables of publications are reported in this appendix, following the description in section 3.2.

[Tables 2 to 14, to be placed here]

Figure Captions

Fig 1 Commercial development alternatives in mature markets. Left) the figure shows the main axis of the development of commercial value in a mature market environment: on one side, Customer Continuity Management and Customer Development -aspects that complement and configure the so-called Customer Value Management- and, on the other side, the development of selective strategies of high value customers acquisition. Right) the figure shows the customer-focused policies that can be developed for each of the defined strategic axes.

Fig 2 Stages in a client's life-cycle. The figure shows the generated value -ordinate axis- of three illustrative customer profiles -*gold*, *silver*, *bronze*- during their time of relationship with the company (abscissa axis). Moreover, the figure shows the stages of customer-company interactions and the basic commercial aspects to solve in each one of the stages.

Fig 3 Customer life-cycle management model: Customer Continuity Management. The figure shows both a) prevention policies integrated in the ordinary customer management (upper part of the diagram), formed by the excellence in service quality, the creation of positive switching barriers and the development of proactive bonding policies; and b) the specific therapy policies for customer deflection (lower part of the diagram), which requires the creation of reactive retention policies, proactive recovery policies and a segmented retention island.

Fig 4 Stages of the predictive model building process ([21]).

Fig 5 Abandonment prediction methods in recent literature. *Left*: overall summary. *Right*: summary of literature published over the period 2009-2015.

Fig 6 Abandonment prediction methods by field of application. *Left*: distribution by field of application. *Right*: detail on the type of applied methodology.

Figure 1

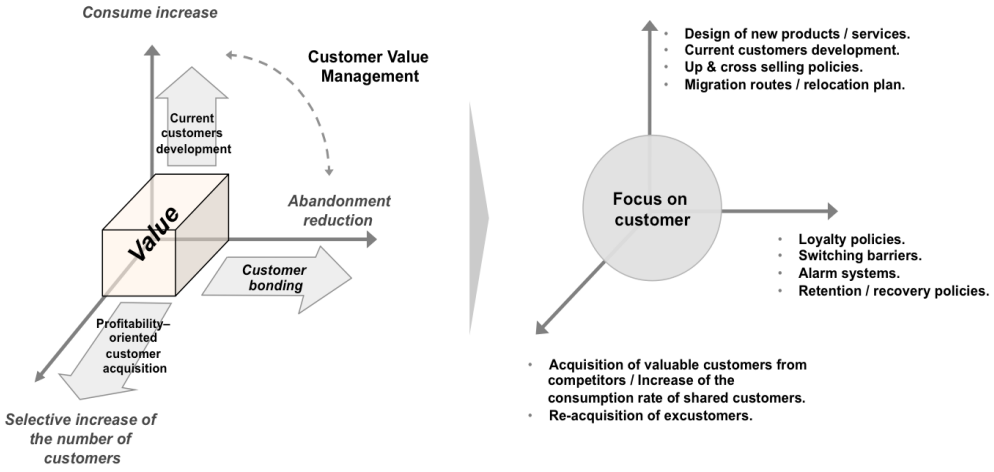
Consume increase
Customer Value Management
Current customers development
Abandonment reduction
Value
Customer bonding
Profitability–oriented customer acquisition
Selective increase of the number of customers
Increasing value axes
Design of new products / services.
Current customers development.
Up & cross selling policies.
Migration routes / relocation plan.
Focus on customer
Loyalty policies.
Switching barriers.
Alarm systems.
Retention / recovery policies.
Acquisition of valuable customers from competitors / Increase of the consumption rate of shared customers.
Re-acquisition of excustomers.
Guided commercial decisions design

Figure 2

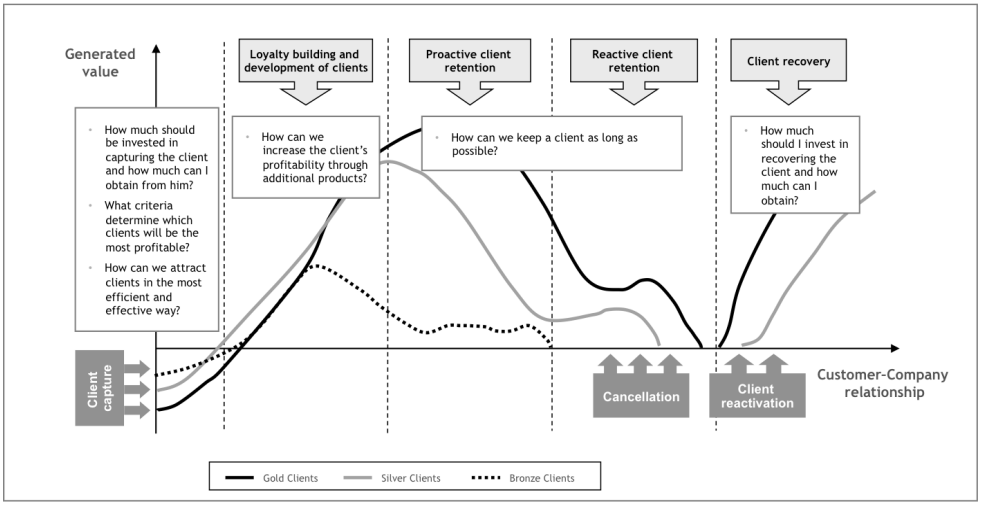
Generated value
Loyalty building and development of clients
Proactive client retention
Reactive client retention
Client recovery
How much should be invested in capturing the client and how much can I obtain from him?
What criteria determine which clients will be the most profitable?
How can we attract clients in the most efficient and effective way?
How can we increase the client's profitability through additional products?
How can we keep a client as long as possible?
How much should I invest in recovering the client and how much can I obtain?
Client capture
Cancellation
Client reactivation
Customer-Company relationship
Gold Clients
Silver Clients
Bronze Clients

Figure 3

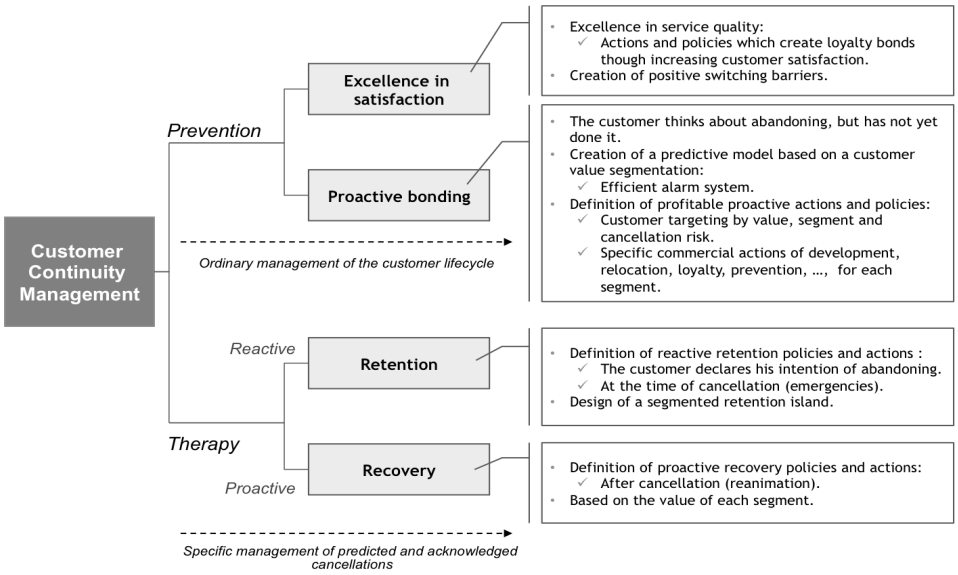
Customer Continuity Management
Prevention
Excellence in satisfaction
Proactive bonding
Ordinary management of the customer lifecycle
Therapy
Reactive
Retention
Proactive
Recovery
Specific management of predicted and acknowledged cancellations
Excellence in service quality:
Actions and policies which create loyalty bonds though increasing customer satisfaction.
Creation of positive switching barriers.
The customer thinks about abandoning, but has not yet done it.
Creation of a predictive model based on a customer value segmentation:
Efficient alarm system.
Definition of profitable proactive actions and policies:
Customer targeting by value, segment and cancellation risk.
Specific commercial actions of development, relocation, loyalty, prevention, ..., for each segment.
Definition of reactive retention policies and actions :
The customer declares his intention of abandoning.
At the time of cancellation (emergencies).
Design of a segmented retention island.
Definition of proactive recovery policies and actions:
After cancellation (reanimation).
Based on the value of each segment.

Figure 4

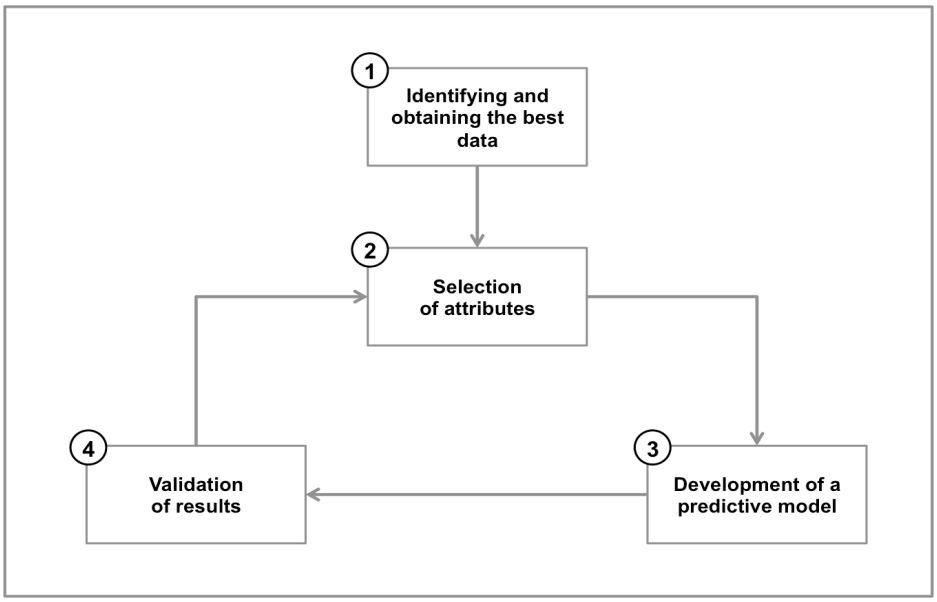
1
Identifying and obtaining the best data
2
Selection of attributes
3
Development of a predictive model
4
Validation of results

Figure 5


Alternative
20,9%
Regression
19,8%
Others CI
16,3%
DT
26,7%
ANN
16,3%
Alternative
17,2%
Regression
27,6%
Others CI
20,7%
DT
24,1%
ANN
10,3%

Figure 6

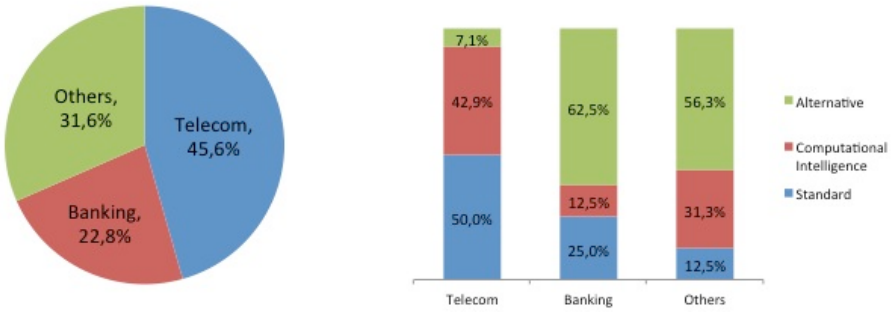
Others, 31,6%
Telecom, 45,6%
Banking, 22,8%
7,1%
42,9%
50,0%
62,5%
12,5%
25,0%
56,3%
31,3%
12,5%
Telecom
Banking
Others
Alternative
Computational Intelligence
Standard

Table 1 Summary of Tables 2 to 14, located in the Appendix section. From left to right, it is structured first by application area and then by category of predictive model (as in stage 3 of Section 3; the "mixed" category includes studies in which methods from different categories have been used). References are listed in the right-hand column. Studies that applied quantitative attribute selection (stage 2) methods are identified with symbol "* ", while studies that applied explicit validation (stage 4) methods are identified with symbol "† ".

Application Area	Method Category	References
Telecommunications	Standard	[69]*,[72]†,[74]*†,[104],[1],[27]†,[46]*†,[52],[73]*,[110],[45]*, [61],[106]†,[64]†,[63]*†,[98]*,[109],[51]*,[94]†
	CI	[69]*,[72]†,[1],[27]†,[46]*†, [73]*,[110],[45]*,[91],[34],[93]†, [106]†,[64]†,[55],[98]*,[109],[13], [51]*,[107]*,[94]†,[7]
	Alternative	[106]†,[23]†,[98]*,[109],[107]*, [94]†
	Mixed	[69]*,[72]†,[1],[27]†,[46]*†, [73]*,[110],[45]*,[106]†,[64]†,[98]*, [109],[51]*,[94]†
Banking	Standard	[101],[60]*,[30]†,[57]*†,[103]†,[77]†,[32]*†,
	CI	[87],[40],[30]†,[57]*†,[103]†,[25],*†,[107]*
	Alternative	[14]*,[96],[59]†,[82]†,[60]*,[57]*†,[107]*
	Mixed	[60]*,[30]†,[57]*†,[103]†,[107]*
Others	Standard	[10]†,[70]*†,[95]*,[75],[17]*†,[11]†,[42],[62]*†
	CI	[39],[2],[10]†,[92],[3],[50],[17]*†,[42],[62]*†,[13]†
	Alternative	[10]†,[67]*,[23]†,[17]*†,[23]†,[81],[47],[89],[11]†,[16]*†
	Mixed	[10]†,[17]*†,[11]†,[42],[62]*†

Table 2 Literature on abandonment prediction modelling, listed in chronological order (1999-2005) and corresponding to the use of standard methods (1 out of 3, continues in the next table).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[69]	4 categories: use, economical, choice of provider and demographical.	Web Survey.	Not specified.	No.	Logistic Regression.	Accuracy on separate validation sets.
[21]	Billing, service and socio-demographic data on cell phone users.	Database.	DTs- and GAs-based.	Data collected on a monthly basis and predicting future churn.	CHART own method, as a combination of DTs and ANNs.	Accuracy on a separate validation data set.
[72]	Call details, quality (interferences and signal coverage), financial and service application (contract details, rate plan, handset type and credit report) and demographic information.	Database.	Not specified.	3 months observation and 2 months prediction.	Logistic Regression and DTs (C5.0).	Lift Chart and Cross-validation.
[74]	Usage data on telecom customers.	Database.	Decision Trees.	Not specified.	DTs (C4.5).	Cross-validation.
[101]	Historical purchase behavior, socio-demographics and actual purchase of banking customers.	Survey.	Domain Knowledge.	No (survey at the moment of last purchase).	Linear Regression.	Accuracy.
[104]	Variables related to the contract (length of service, payment type, contract type) and consumption variables (minutes of use, frequency and sphere of influence).	Database.	Interviews with experts.	3 periods: Observation, retention and prediction.	DTs (C4.5).	Accuracy and sensitivity.
[95]	RFM, behaviourals (specifics of the company, prediction of whether purchase by post will be repeated or not) and non-behaviourals (satisfaction).	Survey and Database.	Sequential Search Algorithm.	4 year historical data plus one survey. 6 months prediction.	Logistic Regression.	Accuracy and AUC.
[1]	Customer localisation, customer type, payment method, service plan, monthly use, number of calls made and number of calls abnormally ended (251 variables)	Database.	Interviews with experts.	2 months for training and 1 month for prediction.	DTs (C4.5).	Lift Chart and Accuracy values.
[27]	Billing, consumer usage, demographics, customer relationship and market data from a wireless telco company.	Company database.	Domain knowledge.	No (data collected during 9 months).	DTs.	Ten-fold cross validation.
[46]	Socio-demographic and usage variables.	Database.	R^2 method.	No (6 months data).	DTs and Logistic Regression.	Lift Chart and cross-validation.
[52]	Company service, demographic and service use characteristics.	Survey.	Not specified.	No.	Logistic Regression.	Log likelihood.
[110]	Demographics, usage level, quality of service and marketing features from wireless telecom customers.	Company database.	No.	Training set: 3 months data; churn measured after 5 months.	DTs.	Accuracy.
[10]	Past purchase behaviour, modeled with RFM variables from customers of Fast-Moving Consumer Goods (FMCG) retail company, complemented with: payment method, length of customer-supplier relationship, shopping behaviour, promotional behaviour and brand purchase.	Company database.	Allusion to previous research.	5-months observation period and 5-months validation period.	DTs.	Accuracy and AUC, using separate validation sets.

Table 3 (Continues from the previous table) References on abandonment prediction modelling, listed in chronological order (2006-2009) and corresponding to the use of standard methods (2 out of 3, continues in the next table).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[73]	Customer behavior (minutes of use, revenue, handset equipment, trends in usage), company interaction data (calls to customer service) and demographics (age, income, location, house ownership) of wireless telecommunication users.	Data provided by Teradata Center for CRM.	Exploratory data analysis, domain knowledge and stepwise.	Data collected on a 3-month period, churn evaluated on the 5th month.	Logistic Regression and DTs.	Top decile lift and Gini coefficient, with two validation data sets.
[45]	Transaction and contract data: demographic, payment, call details and customer service data.	Database.	Interviews with experts and customers; z-test.	Data collected on a 6 month period. 1 month prediction.	DTs (C5.0).	5% Top-Decile Lift on separate validation data sets.
[60]	Gender, age, marriage status, educational level, occupation, job, position, annual income, residential status and credit limits, from banking customers.	Database.	Stepwise discriminant analysis.	Not specified.	CART and Logistic Regression.	Accuracy.
[61]	RFM variables, minutes of customer care calls, number of adults in the household and education level, for costumers of a wireless telecommunications company.	Company database, provided by CRM Centre.	Allusion to previous research.	Training dada from 4 non-consecutive months; validation data from a future point in time.	DTs.	Top-decile lift and Gini coefficient.
[75]	On a charge email service: customer account data (storage bought, length of using and service), usage (number of payments in the last 3 months, total paid amount, complaints) and personal information (age, phone number provided).	Company database.	Advised by employees familiar with the database.	Historical training data tested with previously known churn data.	DTs.	Accuracy.
[17]	Client-company interactions, renewal-related information, socio-demographics and subscription-describing information related to subscribers of a newspaper publishing company.	Database.	Based on RF relevance measures.	30 months data collection and 1 year prediction period.	Logistic Regression.	Cross-validation with accuracy and AUC.
[30]	Evolution in time of the RFM variables (number of invoices last month, amount invoiced, number of withdrawals) related to transactions of a financial company customers.	Company database.	Allusion to previous research.	Training set: 6 consecutive months; validation: next 3 months.	Logistic Regression and DTs.	Accuracy, Loss function and AUC.
[57]	Sociodemographic (age, level of studies, income) and behavioural (monthly credit, number of cards, web transactions, margin) data from credit card users.	Dataset from the Business Intelligence Cup, Univ. of Chile.	CART-based.	Data from 3 consecutive months; results tested during the subsequent year.	DTs (J48) and Logistic Regression.	Sensitivity and accuracy values, on cross-fold validation data set.
[42]	Henley segments, broadband usage, dial types, spend of dial-up, line-information, billing, payment and account information on broadband internet services customers.	Company database.	Domain knowledge.	No.	Logistic Regression and Decision Trees.	Accuracy.
[62]	Usage data from 9 different data sets corresponding to different industries: 3 credit companies, 2 from the mail-order industry, and 1 from the energy industry, direct marketing, fraud detection and e-commerce.	Datasets from UCI ML Repository and the annual DM Cup.	Recursive Feature Elimination.	No (both training and validation on historical data).	Logistic Regression and DTs.	Ten-Fold Cross validation on AUC values.
[64]	Three independent datasets related to telecom industry containing usage, billing and sociodemographic data.	Provided by CRM Centre at Duke University.	Own method, integrating Domain Knowledge.	Data collected during 4 months; churn calculated on the 31-60 days after sampling.	Logistic Regression and Decision Trees.	K-fold CV, accuracy, sensitivity, specificity and AUC.

Table 4 (Continues from the previous table) References on abandonment prediction modelling, listed in chronological order (2010-2015). Standard methods (3 out of 3).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[63]	Usage (average monthly minutes and revenue, days of current equipment, failed voice calls) and sociodemographical (area, ethnicity, presence of child in household) data from a mobile telecommunications company.	Dataset used in the churn tournament by Duke Univ.	Domain knowledge and stepwise selection.	Four non-consecutive months; churners leave 1 to 2 months after being sampled.	Logistic Regression and DTs.	Accuracy, sensitivity, specificity and AUC. Cross-fold validation.
[103]	Customer personal information, credit card basic data, transaction and abnormal usage information.	Company database.	Domain knowledge.	12 months observation period and 12 months testing period.	Logistic Regression and DTs.	Accuracy, AUC, Precision, Recall, Mean Absolute Error and computing time.
[77]	Customer personal information, credit card basic data, transaction and abnormal usage information.	Company database.	Delete variables with high multicollinearity.	12 months observation period and 12 months testing period.	Logistic Regression and DTs.	Accuracy and AUC.
[98]	Sociodemographics, call behaviour, financial and marketing related variables from eleven wireless telecom operators.	Company and public database.	Based on Fisher score.	No (averages over 1 year data).	DTs.	Accuracy, sensitivity, specificity, AUC.
[70]	Historical purchase data, RFM and demographics	Database	Stepwise selection	Two years of data	Logistic regression and Multivariate Adaptive Regression Splines	AUC and Top-Decile Lift. 10-fold cross-validation.
[32]	Customer behavior, demographics, customer-company interactions, and economic indicators	Banking company database	Stepwise selection	Longitudinal data	Cox regression, logistic regression and DTs	AUC and top decile lift. Bootstrapping.
[51]	Customer behavior, demographics and usage data	Telecommunications database of unreported origin	Feature selection through exhaustive recombination of variables	Not reported	Multiple DT models and k-nearest neighbours, compared with alternative methods as stand-alone models or hybridized through an *ad hoc* method	accuracy, precision, recall and F-score. Single training-test split.
[25]	Customer behavior, demographics and usage data	Credit card original database from Chilean bank	Recursive feature elimination associated to SVM classifier	Not reported	Naïve Bayes Tree rule extraction on top of a SVM classifier	accuracy, sensitivity, specificity. Training-test split with 10-fold CV for training.
[94]	Usage data in the telecommunications sector	Standard database from the UCI machine learning repository	No	Not declared	C5.0 DT and Logistic regression, compared with alternative methods	Precision, recall, accuracy and F-measure with Monte Carlo-based cross-validation.

Table 5 References on abandonment prediction modelling, listed in chronological order (2000-2005) and for the use of CI methods (1 out of 3, continues in the next table).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[72]	Call details and quality, financial and service information and demographics.	Database.	Not specified.	3 months observation and 2 months prediction.	ANNs.	Lift Chart and cross-validation.
[21]	Billing, service and socio-demographic data on cell phone users.	Database.	Decision Trees and Genetic Algorithms.	Data collected monthly, predicting future churn.	CHART own method, as a combination of DTs and ANNs.	Accuracy on a separate validation data set.
[3]	41 SERVQUAL-based service quality variables, grouped as tangibles, reliability, responsiveness, assurance and empathy, at an auto-dealership network.	Survey.	Allusion to previous research.	No (study developed at one single point in time).	ANNs.	Least average error, least root mean square error and accuracy.
[39]	RFM variables in the online retail industry.	Database.	Not specified.	18 months.	SOM networks.	Not Specified.
[1]	Customer localisation, customer type, payment method, service plan, monthly use, number of calls made and number of calls abnormally ended (251 variables)	Database.	Interviews with experts.	2 months for training and 1 month for prediction.	DMEL algorithm and ANNs.	Lift Chart and Accuracy values.
[2]	Purchasing behaviour: volume of purchases during first 6 months as customer; "broadness" of purchases; "bargaining tendency" and "price sensitivity".	Database.	Not specified.	Yes (8 weekly periods).	Bayesian Networks.	Accuracy and AUC.
[27]	Billing, consumer usage, demographics, customer relationship and market data from a wireless telco company.	Company database.	Domain knowledge.	No (data collected during 9 months).	MLP ANN, Neuro-Fuzzy Systems and GAs.	Ten-fold cross validation.
[40]	Customer attributes and credit card usage from banking customers.	Company database.	*A priori* association over the different customer segments.	Data collected in a 12 months period; no prediction period.	SOM networks.	Segmentation quality.
[46]	Socio-demographic and usage variables.	Database.	R^2 method.	No (6 months data).	Artificial Neural Networks.	Lift Chart and cross-validation.
[87]	Past transactions carried by the costumer.	Database.	Not specified.	No (3 month: averages and totals).	K-means, SOM networks and Fuzzy K-means.	*Intra-class* method for cluster compactness.
[110]	Demographics, usage level, quality of service and marketing features from wireless telecom customers.	Company database.	All the features contained in the company data set are included.	Three months training data; churn measured after 5 months.	SVMs, ANNs and Bayesian Networks.	Accuracy.
[10]	Past purchase behaviour, modeled with RFM variables from customers of Fast-Moving Consumer Goods (FMCG) retail company, complemented with: payment method, length of customer-supplier relationship, shopping behaviour, promotional behaviour and brand purchase.	Company database.	Automatic Relevance Determination.	5-months observation period and 5-months validation period.	Automatic Relevance Determination ANNs.	Accuracy and AUC, using separate validation sets.

Table 6 (Continues from the previous table) References on abandonment prediction modelling, listed in chronological order (2006-2009) and for CI methods (2 out of 3, continues in the next table).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[73]	Customer behavior, company interaction data and demographics from wireless telecommunication users.	Data provided by Teradata Center for CRM.	Exploratory data analysis, domain knowledge and stepwise.	Data collected on a 3-month period, churn evaluated on the 5th month.	ANNs.	Top decile lift and Gini coefficient, with two validation data sets.
[60]	Gender, age, marriage status, educational level, occupation, job, position, annual income, residential status and credit limits, from banking customers.	Database.	Stepwise discriminant analysis.	Not specified.	SVMs and ANNs.	Accuracy.
[45]	Transaction and contract data: demographic, payment, call details and customer service data.	Database.	Interviews with experts and customers; z-test.	Data collected on a 6-month period. 1-month prediction.	ANNs.	5% Top-Decile Lift on separate validation data sets.
[17]	Client-company interactions, renewal-related information, socio-demographics and subscription-describing information related to subscribers of a newspaper publishing company.	Database.	Based on Random Forest importance measures.	30 months data collection and 1-year prediction period.	SVMs.	Cross-validation with accuracy and AUC.
[34]	Data related to customer repairs and complaints, applied to three different cases related to telecommunications (residential mobile phone, broadband mobile phone, business landline).	Database	Interviews with company experts.	1 month for training and 12 for prediction.	ANNs	Accuracy.
[91]	Usage and relationship variables, plus demographics from wireless telecommunications customers.	Database provided by the Center for CRM at Duke University.	Interview with experts.	Not specified.	SVMs.	Accuracy, with a 80% training data set and 20% validation dataset.
[30]	Evolution in time of the RFM variables (number of invoices last month, amount invoiced, number of withdrawals) related to transactions of a financial company customers.	Company database.	Allusion to previous research.	Six consecutive months training data; validation set: next 3 months.	ANNs.	Accuracy, loss function and AUC.
[57]	Socio-demographic (age, level of studies, income) and behavioural (monthly credit, number of cards, web transactions, margin) data from credit card users.	Data set from the Business Intelligence Cup, Univ. of Chile.	SVMs.	Three consecutive months data; results tested over subsequent year.	SVM-based.	Sensitivity and accuracy values, on cross-validation data set.
[42]	Henley segments, broadband usage, dial types, spend of dial-up, line-information, billing, payment and account information on broadband internet services customers.	Company database.	Domain knowledge.	No.	SVMs and ANNs.	Accuracy.
[62]	Usage data on 9 different data sets corresponding to different industries: 3 credit companies, 2 from the mail-order industry, and 1 from the energy industry, direct marketing, fraud detection and e-commerce.	Data sets from UCI ML Repository and the annual DM Cup.	Recursive Feature Elimination.	No (both training and validation on historical data).	SVM.	Ten-Fold Cross validation on AUC values.
[93]	CRM variables from a telecommunications company.	Company database.	No.	Six months data; churners leave 1-2 months after being sampled.	SOMs and ANNs.	Five-fold cross validation.
[64]	Three independent data sets related to telecom industry containing usage, billing and socio-demographic data.	Provided by CRM Centre at Duke Univ.	Integration of domain knowledge in the FS process.	Four months data; churn calculated on the 31-60 days period after sampling.	ANNs.	K-fold CV; accuracy, sensitivity, specificity and AUC.

Table 7 (Continues from previous table) References on abandonment prediction modelling, listed in chronological order (2010-2015) and using CI methods (3 out of 3).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[103]	Customer personal information, credit card basic data, transaction and abnormal usage information.	Company database.	Domain knowledge.	12 months observation period and 12 months testing period.	Bayesian networks and clustering-based classifiers.	Accuracy, AUC, precision, recall, mean absolute error and computing time.
[55]	Customer average and trends in call minutes, frequency of calls and billing, added to demographics as place of residence, age and tariff type, for a telecommunications company.	Company Database.	Domain knowledge.	Not specified.	Bayesian Networks.	Accuracy.
[92]	35 variables from a real industry retailer, with no specification on the kind of features included.	Company Database.	Not specified.	Data from 13 consecutive months.	ANNs.	Accuracy, sensitivity and specificity.
[98]	Sociodemographics, call behaviour, financial and marketing related variables from eleven wireless telecom operators.	Company and public database.	Based on Fisher score.	No (averages of 1 year data).	SVMs, ANNs, Bayesian Networks.	Accuracy, sensitivity, specificity, AUC.
[13]	Purchase history (longitudinal) data and socio-demographic (static) variables on three different fields: supermarkets, leisure and telecom	Database.	Interviews with experts.	1-year data collection and 1-year prediction, with seasonal overlapping.	SVMs.	AUC.
[51]	Customer behavior, demographics and usage data	Telecommunications database of unreported origin	Feature selection through exhaustive recombination of variables	Not reported	ANNs and SVMs with different kernels, as stand-alone models or hybridized through an *ad hoc* method	accuracy, precision, recall and F-score. Single training-test split.
[25]	Customer behavior, demographics and usage data	Credit card original database from Chilean bank	Recursive feature elimination associated to SVM classifier	Not reported	Naïve Bayes Tree rule extraction on top of a SVM classifier	accuracy, sensitivity, specificity. Training-test split with 10-fold CV for training.
[107]	Usage data from a telecom company (UCI dataset) and from a credit card business of a Chinese bank.	Two datasets, one from the UCI machine learning repository; the other an original database from Chinese bank.	Intensive two-stage feature selection based on ANNs, transfer learning and Mutual Information.	Data spanning from May to December, 2010 (Chinese bank).	Ensemble learning with SVMs as base classifiers. Transfer learning.	sensitivity, specificity, AUC.
[94]	Usage data in the telecommunications sector	Standard database from the UCI machine learning repository	No	Not declared	ANNs and SVMs combined by boosting, compared with alternative methods	Precision, recall, accuracy and F-measure with Monte Carlo-based cross-validation.
[7]	Revenue and usage data in the mobile telecommunications sector	Original database from AAA Honk Kong-based telecommunications company.	No	1-year data, from September 2004 to August 2005.	Adaptation of the Fuzzy c-Means clustering algorithm.	Adjusted SC index for cluster validity.

Table 8 References on abandonment prediction modelling, listed in chronological order (2000-2005) and concerning the use of alternative methods (1 out of 2, continues in the following table).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[81]	RFM variables to model customer-firm relationship and optimize marketing expenditure, which can be applied to many industries.	Artificial data.	RFM variables in different examples.	Not Specified.	Markov Chain Modelling.	No validation stage: theoretical study.
[14]	Transaction data (frequency of transaction of banking customers are analysed).	Database.	Rule extraction-based selection.	Data from the last operative month from the last 6 months.	Goal-oriented sequential pattern.	Computational efficiency against existing algorithms.
[47]	E-customer behaviour, defining customer usage through the website as a mixture of states, transitions between states, holding time, waiting time and total time spent; using web analytics.	Visitor's navigational data collected from the website.	Domain knowledge.	No (data collected during 3 months with no validation stage).	Semi-Markov Processes.	No validation stage.
[50]	RFM variables to segment customers from a direct-mail setting for a charity organization: number of non-answered mails, % response in the last 2 years and overall, size of responses in the last 2 years and overall.	Database.	Domain knowledge.	No.	GAs and K-means clustering.	Validation based on the real effects of the improved segmentation on the customer base.
[96]	Customer behaviour information, socio-demographic information, merger and prosperity index.	Database.	Not specified.	77 year database. Length of sub-periods are not specified.	Survival analysis.	Statistical relevance.
[10]	Past purchase behaviour, modeled with RFM variables from customers of Fast-Moving Consumer Goods (FMCG) retail company, and complemented with: payment method, length of customer-supplier relationship, shopping behaviour, promotional behaviour and brand purchase.	Company database.	Allusion to previous research.	5-months observation period and 5-months validation period.	RFs.	Accuracy and AUC, using separate validation sets.
[59]	Past customer behavior, demographics and interaction with intermediaries from banking and insurance customers.	Company database.	Domain knowledge.	Data collected at a given date, evaluated on the following 9 months.	RFs.	AUC with separate validation datasets.
[67]	RFM variables from customers of a hardware retail company.	Company database.	Analytical Hierarchical Process and interviews with experts.	No (data from the 2 previous years).	Preference-based Collaborative Filtering.	Validation focussed on the goodness of the segmentation.
[89]	Model for optimizing a manufacturing firm's profit, based on customer specifications and processing time, backlog and due-date quotation.	Statistically modeled with artificial data.	Previous domain knowledge.	Not Specified.	Discrete-time Semi-Markov Decision Processes.	Fractional error, weighted absolute value and weighted sum of differences on the lead times.

Table 9 (Continues from the previous table) References on abandonment prediction modelling, listed in chronological order (2006-2013) and concerning alternative methods (2 out of 2).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[82]	Purchase behaviour, seen as a sequence of the historical purchased products (grouped in 9 sets of products) of the customers of a Financial Services company.	Company database.	Domain Knowledge.	Not specified.	Mixture Transition Distribution and Markov chains.	Log-likelihood. Separate validation data sets.
[11]	Usage variables on customers from a Pay-TV company.	Company database.	Undersampling.	Data collection at a specific date, evaluation during the subsequent year.	Markov Chains and RFs.	Accuracy, AUC and Lift Chart with cross-validation.
[57]	Socio-demographic (age, level of studies, income) and behavioural (monthly credit, number of cards, web transactions, margin) data from credit card users.	Data set from the Business Intelligence Cup, Univ. of Chile.	SVMs.	Data from 3 consecutive months; results tested during the subsequent year.	RFs.	Sensitivity and accuracy. Cross-validation.
[17]	Client-company interactions, renewal-related information, socio-demographics and subscription-describing information related to subscribers of a newspaper publishing company.	Database.	Based on RF importance measures.	30 months data collection and 1 year prediction period.	RFs.	Cross-validation. Accuracy and AUC.
[106]	Variables defining the performance of suppliers of a telecommunications company. Quality management practices and systems, documentation and self-audit, process and manufacturing capability, management of the firm, etc.	Company database, available in previous studies.	Primary features in the data set.	Validation and test data sets correspond to different suppliers on different periods.	Hybrid model combining Data Envelopment Analysis, DTs and ANNs.	Four-fold cross validation; accuracy.
[23]	Demographics, historical transactional information and financial variables on different cases (supermarkets, banking, telecom and mailing services).	Database.	Undersampling.	1 month data collection and 1 year performance period.	Ensemble Classifiers (GAMens).	Accuracy, AUC, Top Decile Lift and Lift Index.
[98]	Socio-demographics, call behaviour, financial and marketing related variables from eleven wireless telecom operators.	Company and public database.	Based on Fisher score.	No (averages of 1 year data).	Ensemble methods.	Accuracy, sensitivity, specificity, AUC.
[109]	Usage time, location and customers' underlying social network to predict their churn from a mobile operator.	Company Database.	Research on previous literature.	Training data: 3 consecutive months; churn predicted comparing variation between periods.	Hybrid models, combining DTs, GAs and Game Theory.	Accuracy.
[16]	RFM variables to define past customer behaviour for an online-gaming website.	Company database.	Variable importance score, related to its predictive accuracy.	17 months training data, 4 months for churn measurement.	Ensemble models: RFs and GAMens.	Top-Decile Lift and Lift Index.

Table 10 References on abandonment prediction modelling, listed in chronological order (2000-2009) for the telecommunications application area (1 out of 2, continues in the next table).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[21]	Billing, service and socio-demographic data from cell phone users.	Database.	DTs and GAs.	Monthly data.	Combination of DTs and ANNs.	Accuracy on validation set.
[72]	Call details and quality, financial and service information and demographics.	Database.	Not specified.	3 months observation and 2 months prediction.	Logistic Regression, DTs (C5.0) and ANNs.	Lift Chart and Cross-validation.
[74]	Usage data on telecom customers.	Database.	DTs.	Not specified.	DTs (C4.5).	Cross-validation.
[104]	Variables related to the contract and consumption.	Database.	Interviews with experts.	Observation, retention and prediction periods.	DTs (C4.5).	Accuracy and sensitivity.
[1]	Customer localisation and type, payment method, service plan, monthly use, number of calls made and abnormally ended.	Database.	Interviews with experts.	Two months training; one month prediction.	DMEL, DTs (C4.5) and ANNs	Lift Chart and accuracy.
[27]	Billing, consumer usage, demographics, customer relationship and market data from a wireless telco company.	Company database.	Domain knowledge.	No (data collected over 9 months).	DTs, ANNs, Neuro-Fuzzy Systems and GAs.	Ten-fold cross validation.
[46]	Socio-demographic and usage variables.	Database.	R^2 method.	No (6 months data).	ANNs, DTs and Logistic Regression.	Lift Chart and cross validation.
[52]	Company service, demographic and service use characteristics.	Survey.	Not specified.	No.	Logistic Regression.	Log likelihood.
[110]	Demographics, usage level, quality of service and marketing features from wireless telecom customers.	Company database.	No.	3-months training; churn measured after 5 months.	SVMs, ANNs, DTs and Bayesian Networks.	Accuracy.
[73]	Customer behavior, company interaction data and demographics from wireless telecommunication users.	Data provided by Teradata Center for CRM.	Domain knowledge and stepwise.	Three-month data; churn evaluated on the 5^{th} month.	Logistic Regression, DTs and ANNs.	Top decile lift and Gini coefficient; two validation sets.
[45]	Transaction and contract data: demographic, payment, call details and customer service data.	Database.	Interviews with experts and customers; z-test.	6-months data; 1-month prediction.	DTs (C5.0), ANN.	5% Top-Decile Lift on separate validation sets.
[61]	RFM variables, minutes of customer care calls, number of adults in the household and education level, from wireless telecom customers.	Company database, provided by CRM Centre at Duke U.	Allusion to previous research.	Training data: 4 non-consecutive months; validation with future data.	DTs.	Top-decile lift and Gini coefficient.
[91]	Usage variables, relationship and demographics from wireless telecommunications customers.	Database provided by the Center for CRM at Duke U.	Interview with experts.	Not specified.	SVMs.	Accuracy; 80% training / 20% validation.
[34]	Customer repairs and complaints data from 3 telecommunications scenarios (residential mobile, broadband mobile, business landline).	Database	Interviews with company experts.	1-month training and 12-month prediction.	ANNs	Accuracy.
[93]	CRM variables from a telecommunications company.	Company database.	Primary features provided by the company.	Six months data; churners leave 1-2 months after sampled.	ANNs and SOM.	Five-fold cross validation.
[106]	Variables defining telecommunications company suppliers' performance. Quality management practices, documentation and self-audit, process and manufacturing capability, etc.	Company database, available in previous studies.	Primary features in the data set.	Validation and test sets correspond to different suppliers on different periods.	Hybrid model: Data Envelopment Analysis, DTs and ANNs.	Four-fold cross validation; accuracy.
[64]	Three independent data sets related to telecom industry containing usage, billing and socio-demographic data.	Provided by CRM Centre at Duke Univ.	Domain knowledge-based.	4-months data; churn calculated 1-2 months after sampling.	Logistic Regression, DTs and ANNs.	K-fold CV; accuracy, sensitivity, specificity and AUC.

Table 11 (Continues from previous table) References on abandonment prediction modelling, listed in chronological order (2010-2015) for the telecommunications application area (2 out of 2).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[23]	Demographics, historical transactional information and financial variables on different scenarios (supermarkets, banking, telecom and mailing services).	Database.	Undersampling.	1 month data collection and 1 year performance period.	Ensemble Classifiers (GAMens)	Accuracy, AUC, Top Decile Lift and Lift Index.
[55]	Customer usage and demographics from a telecom company.	Company Database.	Domain knowledge.	Not specified.	Bayesian Networks.	Accuracy.
[63]	Usage and socio-demographic data from a mobile telecommunications company.	Data set used in the churn tournament by Duke Univ.	Domain Knowledge and Stepwise selection.	4 non-consecutive months data; churners leave the company 1-2 months after sampled.	Logistic Regression and DTs.	Accuracy, sensitivity, specificity and AUC. Cross-validation.
[98]	Socio-demographics, call behaviour, financial and marketing related variables from eleven wireless telecom operators.	Company and public database.	Based on Fisher score.	No (averages of 1 year data).	DTs, ensembles, ANNs, Bayesian Networks, SVMs.	Accuracy, sensitivity, specificity, AUC.
[109]	Usage time, location and customers' underlying social network to predict their churn from a mobile operator.	Company Database.	Research on previous literature.	3 consecutive months data; churn predicted comparing period variations .	Hybrid models, combining DTs, GAs and Game Theory.	Accuracy.
[13]	Purchase history (longitudinal) data and socio-demographic (static) variables on three different fields: supermarkets, leisure and telecom	Database.	Interviews with experts.	1-year data collection and 1-year prediction, with seasonal overlapping.	SVMs.	AUC.
[51]	Customer behavior, demographics and usage data	Telecommunications database of unreported origin	Feature selection through exhaustive recombination of variables	Not reported	Multiple DT models, ANNs, k-nearest neighbours and SVMs, as stand-alone models or hybridized through an *ad hoc* method	accuracy, precision, recall and F-score. Single training-test split.
[107]	Usage data from a telecom company (UCI dataset).	Standard dataset from the UCI machine learning repository.	Intensive two-stage feature selection based on ANNs, transfer learning and Mutual Information.	Data spanning from May to December, 2010 (Chinese bank).	Ensemble learning with SVMs as base classifiers. Transfer learning.	sensitivity, specificity, AUC.
[94]	Usage data in the telecommunications sector	Standard database from the UCI machine learning repository	No	Not declared	Logistic Regression, Naïve Bayes, DTs, ANNs and SVMs combined by boosting	Precision, recall, accuracy and F-measure with Monte Carlo-based cross-validation.
[7]	Revenue and usage data in the mobile telecommunications sector	Original database from AAA Honk Kong-based telecommunications company.	No	1-year data, from September 2004 to August 2005.	Adaptation of the Fuzzy c-Means clustering algorithm.	Adjusted SC index for cluster validity.

Table 12 References on abandonment prediction modelling, listed in chronological order (2001-2015) for the Banking and Financial Services field.

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
[101]	Historical purchase behavior, socio-demographics and actual purchase on banking customers.	Survey.	Domain Knowledge.	No (survey at the moment of last purchase).	Regression.	Accuracy.
[14]	Transaction data (frequency of transaction of banking customers).	Database.	Rule extraction-based.	Last operative month from the last 6 months.	Goal oriented sequential pattern.	Computational efficiency against existing algorithms.
[87]	Past transactions carried by the costumer.	Database.	Not specified.	No (3 month: averages and totals).	K-means, SOM and Fuzzy K-means.	Intra-class method for cluster compactness.
[96]	Customer behaviour information, socio-demographic information, merger and prosperity index.	Database.	Not specified.	77 year database.	Survival analysis.	Statistical relevance.
[40]	Customer attributes and credit card usage from banking customers.	Company database.	A priori association.	12 months data; no prediction period.	SOM.	Goodness of segmentation.
[59]	Past customer behavior, demographics and interaction with intermediaries for banking and insurance customers.	Company database.	Domain knowledge.	Data collected at given date, evaluated over next 9 months.	RFs.	AUC with separate validation data sets.
[82]	Purchase behaviour, seen as a sequence of the historical purchased products (grouped in 9 sets of products).	Company database.	Domain Knowledge.	Not specified.	Mixture transition distribution and Markov chains.	Log-likelihood. Separate validation data sets.
[60]	socio-demographics and credit limits from banking customers.	Database.	Stepwise discriminant analysis.	Not specified.	CART, Logistic regression, ANNs and SVMs.	Accuracy.
[30]	Evolution in time of the RFM variables related to transactions of a financial company customers.	Company database.	Allusion to previous research.	6 consecutive months training data; next 3 months validation.	Logistic Regression, DTs and ANNs.	Accuracy, Loss function and AUC.
[57]	Socio-demographic and behavioural data from credit card users.	Dataset from the Business Intelligence Cup, Univ. Chile.	CART.	3 consecutive months data; tested during the subsequent year.	DTs (J48), Logistic Regression, RFs and SVMs.	Sensitivity and accuracy; cross-validation.
[103]	Customer personal information, credit card basic data, transaction and abnormal usage information.	Company database.	Domain knowledge.	12-months training and 12-months test data.	DTs, Logistic regression, Bayesian Networks and clustering-based classifiers.	Accuracy, AUC, precision, recall, mean absolute error and computing time.
[77]	Customer personal information, credit card basic data, transaction and abnormal usage information.	Company database.	multicollinearity-based selection.	12-months training and 12-months testing.	Logistic Regression and DTs.	Accuracy and AUC.
[32]	Customer behavior, demographics, customer-company interactions, and economic indicators	Banking company database	Stepwise selection	Longitudinal data	Cox regression, logistic regression and DTs	AUC and top decile lift. Bootstrapping.
[25]	Customer behavior, demographics and usage data	Credit card original database from Chilean bank	Recursive feature elimination associated to SVM classifier	Not reported	Naïve Bayes Tree rule extraction on top of a SVM classifier	accuracy, sensitivity, specificity. Training-test split with 10-fold CV for training.
[107]	Usage data from a credit card business of a Chinese bank.	Original dataset from credit card business of a Chinese bank.	Intensive two-stage feature selection based on ANNs, transfer learning and Mutual Information.	Data spanning from May to December, 2010 (Chinese bank).	Ensemble learning with SVMs as base classifiers. Transfer learning.	sensitivity, specificity, AUC.

Table 13 References on abandonment prediction modelling, listed in chronological order for the Retail (2002-2013), Mail and Delivery Services (2002-2004) and Other (next table) fields of application (1 out of 2).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
Retail						
[39]	RFM variables in the online retail industry.	Database.	Not specified.	18 months.	SOM networks.	Not Specified.
[2]	Purchasing behaviour: volume of purchases during first 6 months as customer; "broadness" of purchases; "bargaining tendency" and "price sensitivity".	Database.	Not specified.	Yes (8 weekly periods).	Bayesian Networks.	Accuracy and AUC.
[10]	Past purchase behaviour, modeled with RFM variables from customers of Fast-Moving Consumer Goods (FMCG) retail company, and complemented with: payment method, length of customer-supplier relationship, shopping and promotional behaviour and brand purchase.	Company database.	Allusion to previous research.	5-months observation period and 5-months validation period.	RFs, Automatic Relevance Determination ANNs and Logistic Regression.	Accuracy and AUC, using separate validation sets.
[67]	RFM variables from customers of a hardware retail company.	Company database.	Analytical Hierarchical Process and interviews with experts.	No (data from the 2 previous years).	Preference-based Collaborative Filtering for customer clustering.	Validation focussed on the goodness of the segmentation.
[23]	Demographics, historical transactional information and financial variables on different cases (supermarkets, banking, telecom and mailing services).	Database.	Undersampling.	1 month data collection and 1 year test period.	Ensemble Classifiers (GAMens)	Accuracy, AUC, Top Decile Lift and Lift Index.
[92]	35 variables from a real industry retailer, with no specification on the kind of features included.	Company Database.	Not specified.	Data from 13 consecutive months.	ANNs.	Accuracy, sensitivity and specificity.
[70]	Historical purchase data, RFM and demographics	Database	Stepwise selection	Two years of data	Logistic regression and MARS	AUC and Top-Decile Lift. 10-fold cross-validation.
Mail and delivery services						
[3]	41 SERVQUAL-based service quality variables, grouped as tangibles, reliability, responsiveness, assurance and empathy, at an auto-dealership network.	Survey.	Allusion to previous research.	No (study developed at one single point in time).	ANNs.	Least average and root mean square errors; accuracy.
[95]	RFM, behavioural (specifics of the company, prediction of whether purchase by post will be repeated or not) and non-behavioural (satisfaction).	Survey and Database.	Sequential Search Algorithm.	4 year historical data plus one survey. 6 months prediction.	Logistic Regression.	Accuracy and AUC.
[50]	RFM variables to segment customers from a direct-mail setting for a charity organization.	Database.	Domain knowledge.	No.	GAs and K-means clustering.	Validation based on improvement of customer segmentation.

Table 14 (Continues from previous table) References on abandonment prediction modelling, listed in chronological order for Retail (previous table) , Mail and Delivery Services (2006-2010) and Other (2000-2013) application areas (2 out of 2).

Ref.	Data type	Data gathering	Feature selection	Time periods	Modelling techniques	Validation method
Mail and delivery services						
[75]	On a charge email service, customer account data, usage and personal information.	Company database.	Advised by company employees.	Historical training data tested with churn data.	DTs.	Accuracy.
[17]	Client-company interactions, renewal-related information, socio-demographics and subscription-describing information related to subscribers of a newspaper publishing company.	Database.	RF-based importance measures.	30 months data collection and 1 year prediction period.	SVMs, Logistic Regression, RFs.	Cross-validation; accuracy and AUC.
[23]	Demographics, historical transactional information and financial variables on different scenarios (supermarkets, banking, telecom and mailing services).	Database.	Undersampling.	1 month data collection and 1 year performance period.	Ensemble Classifiers (GAMens)	Accuracy, AUC, Top Decile Lift and Lift Index.
Others						
[81]	RFM variables to model customer-firm relationship and optimize marketing expenditure, which can be applied to many industries.	Artificial data.	RFM variables.	Not Specified.	Markov Chain Modelling.	No validation: theoretical study.
[47]	E-customer behaviour, defining its customer usage through the website as a mixture of states, transitions between states, holding time, waiting time and total time spent.	Visitor's navigational data collected from website.	Domain knowledge.	Three months data with no validation.	Semi-Markov Processes.	No validation stage.
[89]	Model for optimizing a manufacturing firm's profit, based on customer specifications and processing time, backlog and due-date quotation.	Artificial data.	Previous domain knowledge.	Not Specified.	Discrete-time Semi-Markov Decision Processes.	Fractional error, weighted absolute value and weighted sum of differences on the lead times.
[11]	Usage variables on customers from a Pay-TV company.	Company database.	Undersampling.	Data collection at a specific date, evaluation during the subsequent year.	Markov Chains, RFs and Logistic Regression.	Accuracy, AUC and Lift Chart; cross validation.
[42]	Henley segments, broadband usage, billing, payment, etc. from internet services customers.	Company database.	Domain knowledge.	No.	Logistic Regression, DTs, ANNs and SVMs.	Accuracy.
[62]	Usage data from 9 different industries: 3 credit companies, 2 from the mail-order industry, and 1 from the energy industry, direct marketing, fraud detection and e-commerce.	Datasets from the UCI ML Repository and the annual DM Cup.	Recursive Feature Elimination.	No (training and validation on historical data).	SVM, Logistic Regression and DTs.	Ten-fold cross validation. AUC.
[13]	Purchase history (longitudinal) data and sociodemographic (static) variables on three different fields: supermarkets, leisure and telecom	Database.	Interviews with experts.	1-year data collection and 1-year prediction, with seasonal overlapping.	SVMs.	AUC.
[16]	RFM variables to define past customer behaviour for an online-gaming website.	Company database.	Predictive accuracy-related importance score.	17 months training data; 4 months for churn measurement.	RFs and GAMens.	Top-Decile Lift and Lift Index.