

Received January 9, 2022, accepted January 17, 2022, date of publication January 26, 2022, date of current version February 9, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3146397

A SLR on Customer Dropout Prediction

PEDRO SOBREIRO¹, DOMINGOS DOS SANTOS MARTINHO²,
JOSÉ G. ALONSO³, (Member, IEEE), AND JAVIER BERROCAL³, (Member, IEEE)

¹Quality of Life Research Centre, Polytechnic Institute of Santarém, 2040-413 Rio Maior, Portugal

²ISLA Santarém, 2000-241 Santarém, Portugal

³Quercus Software Engineering Group, University of Extremadura, 06006 Badajoz, Spain

Corresponding author: Pedro Sobreiro (sobreiro@esdrm.ipsantarem.pt)

This work was supported in part by the FEDER/Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación under Project RTI2018-094591-B-I00, in part by the 4IE+ Project by the Interreg V-A España-Portugal (POCTEP) (2014–2020) Program under Grant 0499-4IE-PLUS-4-E, in part by the Department of Economy, Science and Digital Agenda of the Government of Extremadura under Grant GR18112 and Grant IB18030, in part by the Valhondo Calaff Institution, and in part by the European Regional Development Fund.

ABSTRACT Dropout prediction is a problem that is being addressed with machine learning algorithms; thus, appropriate approaches to address the dropout rate are needed. The selection of an algorithm to predict the dropout rate is only one problem to be addressed. Other aspects should also be considered, such as which features should be selected and how to measure accuracy while considering whether the features are appropriate according to the business context in which they are employed. To solve these questions, the goal of this paper is to develop a systematic literature review to evaluate the development of existing studies and to predict the dropout rate in contractual settings using machine learning to identify current trends and research opportunities. The results of this study identify trends in the use of machine learning algorithms in different business areas and in the adoption of machine learning algorithms, including which metrics are being adopted and what features are being applied. Finally, some research opportunities and gaps that could be explored in future research are presented.

INDEX TERMS Customers, dropout prediction, machine learning, systematic review.

I. INTRODUCTION

Customer analysis is fundamental to developing business and marketing intelligence [1], thereby supporting understandings of historical data and identifying trends and patterns [2]. This process is also known as data mining, i.e., the extraction of knowledge from data [3], and fitting models are required to determine patterns in observed data [4] to provide a means to answer business questions that have been traditionally time-consuming to solve [5]. Usually, techniques such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, and high-performance computing [6] are employed to extract meaningful and useful information or patterns through automated or semiautomated methods [2], [3], [7]. The information that many organizations accumulate can support managers in making decisions by providing more insights about customers. The developmental analysis of existing data in organizations allows accurate targeting of customers [8] and a better understanding of how the loyalty

mechanisms work to predict the customer intention to drop out [9]. These perspectives indicate that companies are realizing that their customer database is the most valuable asset that possess [10], [11]. This realization provides opportunities for organizations to explore their existing data to gain a competitive advantage.

The understanding of customer data allows customers to be retained through the exploration of existing information. According to Glady *et al.* [12], churn is a marketing-related term that represents a consumer who is switching from one company to a competitor in the near future, and, according to Verbeke *et al.* [13], is a management science problem that adopts a data mining approach to try to solve the problem related to the lower costs for retaining customers versus the costs of attracting new customers [14]. This approach requires determining which customers have a higher propensity to attrite [15], but the reasons for dropping out could be attributed to different events, Berry and Linoff [2] thus divide customer dropout in voluntary and involuntary categories; voluntary dropout represents a decision by the customer to end the relationship, while involuntary dropout occurs when the company ends the customer relationship because the

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca¹.

customer does not fulfil their obligations, via breaches such as lack of payment or abuse of service. This method creates two scenarios leading to an end event that is referred to as dropout. Another factor referred to by Neslin *et al.* [16] is that the end of the relationship occurs in a given time period. The identification of dropout can be developed in different contexts. There are two main scenarios, as follows [17], [18]: (1) contractual settings, where customers pay a fee, such as a subscription, and the customer informs the company that they are ending the relationship, and (2) noncontractual settings, for example, a customer buying a book, where a firm has to infer if the customer is still active (did not drop out). The contractual business concept implies that usage and retention are interconnected processes; i.e., customers need to renew their contracts/memberships/subscriptions to continue that usage [19]. The main characteristic of a contractual setting is the interaction in which the customer ends the relationship [20]. Dropout prediction requires proper approaches that consider the environment where it is developed. In contractual setting scenarios, such as insurance, telecommunications, and magazine subscriptions, firms can accurately understand the cash flow generated by their customers, as customers usually sign long-term contracts with firms [21]. The customer must choose whether to opt-in or to opt-out [22], i.e., (1) customers will choose to opt-in if they want to enter into a contract with a particular form (e.g., renewal form) or (2) customers will choose to opt-out if they prefer not to renew.

Depending on the contractual relationships and the knowledge of the exact time of churn, contracts are not fully applicable in the context of catalog retailers or other non-membership businesses [23]; moreover, contractual settings provide subscription records and usage logs [24].

By anticipating possible churners, companies can develop countermeasures, such as concessions, to retain as many customers as possible [25] and develop predictions regarding future intentions to facilitate their segmentation to support actions according to their likelihood of dropping out. The prevention of customer dropout is a problem that needs to be addressed that, according to [12], requires a retention action if the company wants to avoid dropout. According to Amin *et al.* [26], this problem impacts organizational performance, causes reduced sales, allows competitors to gain new customers, increases the cost of attracting new customers in relation to the lower cost of retaining them, and is a risk to the company image with the loss of the market and customer base. The quantification of dropout could be measured as a percentage rate that represents the customers ending a subscription or terminating the business relation [27] or as a churn probability calculated for each customer using historical data to predict their future behaviors [28] to calculate the customers who have already churned or to determine the churn risk.

In a contractual setting scenario, customer churn is more damaging than just selling less product, meaning that there is a well-defined termination of a relationship [29] that

should be properly addressed. Companies are more profitable when they retain more customers due to the lower marketing costs and improved sales [26]. If an organization can predict a possible dropout and develop countermeasures to avoid desertion, they can avoid customer defections that lead to a loss of money. Reichheld [30] evidenced that reducing dropout rates by 5% (e.g., from 15% per year to 10% per year) could represent an increase in profits up to double, as acquiring new customers costs 5 to 6 times more than retaining existing customers [31]. Existing organizations are addressing this problem by shifting their target from capturing new customers to preserving existing customers [9], as investments in retention strategies have higher returns than acquisitions [32]. The importance of customer retention to maintain organizational profitability [33] causes the problem of how to quantify the financial impact of customer retention actions under the assumption that the organization goal should be related to an increase in the lifetime of the customer to increase their profits. The customer lifetime value (CLV) allows us to measure this impact, as it represents a monetary value resulting from the duration of a relationship with a customer [16]. Targeting existing customers according to their risk of dropout to increase lifetime expectancy should be addressed considering the likelihood of churn [32] and not wasting incentives on customers who are not profitable in the development of retention actions [16]. CLV supports the determination of a budget for marketing decisions and strategic and tactical decisions in companies [34]. Using this assumption, Verbeke *et al.* [13] suggested calculating the maximum profit that can be generated by including the optimal fraction of customers with the highest predicted probabilities to attrite in retention campaigns to optimize the process. The problem is addressed in different organizations, such as the publishing, financial services, insurance, electric utilities, health care, banking, internet, telephone, and cable service industries [16]. The objective is to maintain a relationship with customers. Valuable relationships must be retained and reinforced by building strong customer defection-avoiding schemes [9].

Predictive learning could be adopted to anticipate customer dropout; however, after the emergence of the fields of artificial intelligence (machine learning) [35], machine learning is considered a modern extension of predictive analytics [36], understood as an automated process to extract patterns from the data [37] generalizing from the examples in the training set [38]. More recently, deep learning has emerged, which is considered a part of machine learning [39], mimics the behavior of the structure of the human brain [40] and relies on machine learning algorithms that model nonlinear high-level abstractions [39]. Machine learning encompasses deep learning and is understood to be a consequence of predictive analytics.

Machine learning extracts patterns from the data [37] by generalizing from the examples in the training set [38]. Machine learning could be utilized to extract knowledge to understand dropout for the development of effective

retention strategies [41] and to allow the discovery of patterns to support the hypothesis for addressing existing problems. Machine learning is applied to develop churn prediction models that generalize the relationship between churn behavior and historical data to produce predictions about the future behaviors of a company's customers, which is influenced by the input data and algorithm selected to model [42]. Using historical data, a model could be trained to classify a future dropout or nondropout. This perspective could be realized by predicting two possible events, either dropout or nondropout, which allow us to identify ways to induce customers to stay [16] by estimating the probability of customer churning in a given eperiod of time [33]. Machine learning can be used to develop retention strategies based on existing data [41] by extracting patterns from data [37] that support the development of counteractions before an event occurs.

Wai-Ho Au *et al.* [43] developed an approach to identify several rules to reduce customer dropout, such as adding bonus schemes, incentives to service renovation or supporting the exploration of existing situations to understand behavioral patterns. The idea is to prolong the life expectancy of customers relationships and to develop strategies that create value supported in the implementation of commercial actions in their life cycle [9]. Ng and Liu [44] suggested that deviations from normal patterns of behaviors are strong signs of likely defection, where deviation analysis is performed to detect customers who are at risk of dropout. There are several algorithms that can be adopted [42]: logistic regression, artificial neural networks, survival analysis, Markov chains, support vector machines, generalized additive models, decision trees, naive Bayes classifiers, K-nearest neighbor classifiers, random forests, cost-sensitive classifiers and evolutionary algorithms. In telecommunications, algorithms such as decision trees, neural networks, support vector machines, Bayesian belief networks, and regression, have been commonly employed [45].

Classification algorithms should fulfill the following two requirements: classification performance and higher interpretability. [46]. Some algorithms, such as neural network black box models [47] or support vector machines [48], lack of interpretability. Higher interpretability allows us to solve the existing problem of a lack of causality [15], where the principle of Occam's Razor, is usually interpreted to be simpler. Decision trees are very popular because of their interpretability [48], [49] and conceptual transparency [45]. Interpretability is an important aspect for the selection of an algorithm to classify dropout, as it allows the marketing department to extract valuable information from a model to design effective retention campaigns and strategies [50].

Risselada *et al.* [29] analyzed the performance of predicting models in several periods of time after the estimation period and identified a substantial decrease in the predictive quality. The following questions arise: what is the accuracy when considering a temporal factor, and do churn patterns exist when considering time-changing parameters? Survival analysis, or more generally, time-to-event analysis, allows us

to consider a dynamic perspective considering the duration of the relationship with the customers. Survival Analysis consists of a set of methods to describe the probability of surviving past a specified time point, or more generally, the probability that the event of interest has not yet occurred by this time point [51], and attempts to address the timings related to customer dropout. Survival analysis is a class of statistical methods that modelling the occurrence and timing of events, e.g., customer attrition aiming to establish descriptive or predictive models in which the risk of an event depends on covariates [52]. More recent machine learning algorithms for survival analysis provide additional advantages that surpass the existing limitations, such as Cox proportional hazard assumptions [53], and allow the use of time-dependent variables.

Increasing retention contributes to improving the lifetime of a customer, which requires a set of techniques that can be applied to learn how to model the relationship between a set of descriptive variables and a target variable using a set of historical examples [37]. Existing studies mainly address nontime-varying variables [54] and propose different time periods to predict churn. The approaches employed to develop dropout prediction should simultaneously consider interpretability [49], classification performance [46], and financial objectives related to business profitability to optimize the customer lifetime value [34].

To address these perspectives and to increase the understanding related to predicting dropout in the context of contractual settings, the prediction should be developed to simultaneously consider the dropout prediction accuracy, the timing when this dropout occurs, and its interpretability to support better addressing of customer churn. However, at this point, we lack an overview of the research related to the use of machine learning algorithms, to target the dropouts and address these issues.

Our main research question is given as follows: what is the current state of machine learning research studies on predicting dropout in contractual settings? This research analyzes the state-of-the-art techniques and identifies machine learning studies to predict customer dropout. This study was developed under a systematic literature review methodology of following the guidelines of Kitchenham and Charters [55].

Our paper is organized as follows: In the next section, we address our research methodology, considering the research questions, and the search strategy applied to pursue our research objectives. The results section presents the research results according to the previously identified questions, and the validity threats of how to mitigate research threats in software engineering. Finally, some open questions, conclusions and future work are presented.

II. RESEARCH METHODOLOGY

According to Fink [56], systematic literature review (SLR) is a systematic, explicit, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars,

and practitioners. The current research presents a SLR to provide guidance to researchers planning future studies and summaries of the literature on a particular issue [57].

The importance of understanding customer dropout and the diversity of employed algorithms requires an understanding of trends and existing problems to create a ground base of knowledge. To development of the systematic literature review, the methodology applied by Kitchenham & Charters [55] was adopted and developed in three stages, namely, planning, conducting and reporting, as described in Figure 1. The planning stage involves defining the research need, identifying the research questions and developing the review protocol. The conducting stage involves research identification, study selection, quality assessment, data extraction, and data synthesis. The last stage develops one activity, i.e., the report review.

A. SEARCH QUESTIONS

The first question addressed in this study, which is presented in the introduction, is presented as follows: what is the current state of machine learning research studies on predicting customer dropout in contractual settings? Based on this research question, the following questions were identified to determine the main aspects: related to customer dropout with contractual settings:

- RQ1: What is the current state of the research being developed?
- RQ2: What algorithms have been employed to predict dropout?
- RQ3: Which features are used to predict dropout?
- RQ4: When does dropout occur?
- RQ5: How is machine learning algorithm dropout prediction accuracy measured?

RQ1 aims to identify what studies have been published thus far in the area. This will allow the business area to be researched to be understood. The second and third research questions (RQ2 and RQ3, respectively) aim to identify which different types of machine learning algorithms are applied to predict dropout (RQ2) to deepen the knowledge of the approaches that are being utilized to address dropout. RQ3 explores the features that are selected to predict dropout. The research question RQ4 intends to understand if the timing related to customer dropout is considered; this question explores, for example, if any study explores the advantage of using survival analysis, as considering that allows us to examine not only if an event occurs but also how long it took to occur [58]. This perspective entails understanding the dropout event in the potential research to be explored and also the timing. Finally, RQ5 aims to identify how the accuracy of machine learning algorithms in predicting dropout is measured. The goal of dropout prediction is to develop a classification task to identify an optimal functional mapping between the input data X (describing the input pattern) with the class label Y to accurately predict the label of the unseen data, such as $Y=f(X)$ [59]. The prediction should consider characteristics of the data such as class imbalance [60].

B. SEARCH STRATEGY

This phase requires the identification of the search strategy to identify advances in the research of machine learning algorithms to predict dropout. The authors used the Petticrew and Roberts [57] population, intervention, comparison, outcomes and context (PICOC) method to define the research string, recommended by Kitchenham and Charters [55], as illustrated in Table 1. The keywords identified were machine learning, customer dropout, and contractual. Other alternative keywords, such as churn and membership, were also adopted considering that Verbeke et al. [50] used the term churn to define the end of a relationship with an organization and Ascarza [18] used membership to represent a relationship with a professional group. The adopted search criteria were (('customer dropout') OR ('customer churn')) AND 'machine learning' AND ('contractual' OR 'membership'), which were applied to the title, abstract, and keywords in the search period between January 2000 and June 2020.

TABLE 1. PICOC criteria.

PICOC	Description
Population	Research papers about dropout with contractual settings
Intervention Comparison	Machine learning algorithms to predict dropout Different types of algorithms that address machine learning to predict dropout
Outcome	Synthesis identifying research questions, algorithms employed, gaps in the research domain and best practices identified
Context	Academia and industry

The search, which identified 448 articles, was performed on the following databases: SpringerLink, Scopus, Science@Direct, ISI Web of Science, IEEE Digital Library, and ACM Digital Library (Table 2). The results from each database were exported to a bibtex file and merged into a single bibtex file, and the bib2df package [61] was used to parse the bibtex fields and collect the metadata.

The following inclusion and exclusion criteria were applied for the study selection:

- 1) *Inclusion:*
 - a) From 2000 to 2020
 - b) Studies peer-reviewed
- 2) *Exclusion:*
 - a) Books
 - b) Non-English articles
 - c) patents
 - d) thesis

The article filtering process is summarized in Fig. 2. The articles from the different sources were combined, and 28 duplicate articles were removed, which produced a total of 420 articles. Additionally, 5 entries were removed due to the exclusion criteria of books, and 13 entries were removed because of incomplete information (title and DOI). After this process was completed, 402 papers were identified.

Considering the time involved in the protocol development, searching and retrieval of references, paper screening, data

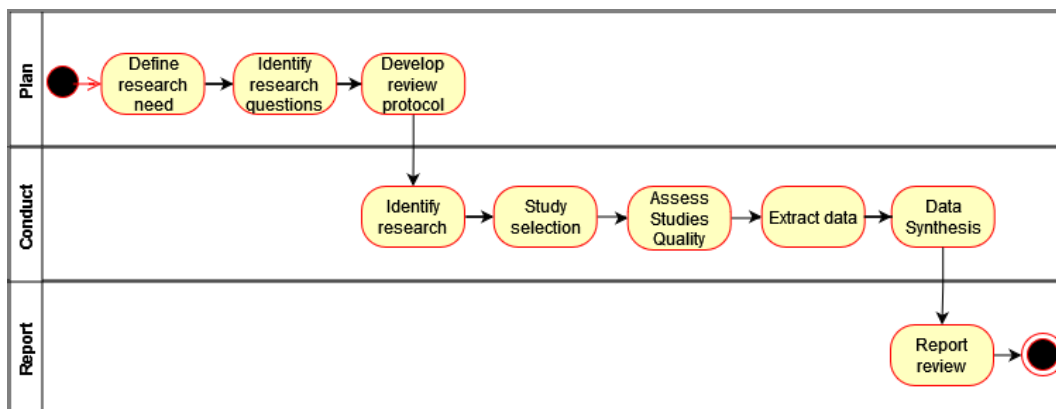


FIGURE 1. SLR phases based on Kitchenham and Charter [55].

TABLE 2. Articles identified for the initial dataset.

Source	Articles
Scopus	210
IEEE	20
SpringerLink	78
Science Direct	126
ISI Web of Knowledge	6
ACM	8
Total	448

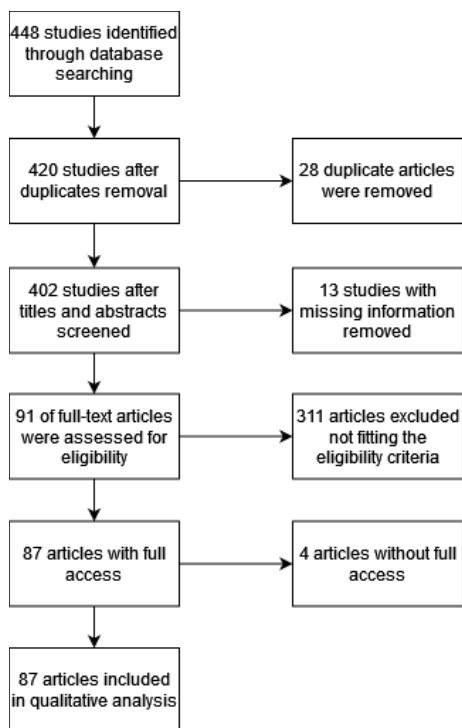


FIGURE 2. Flow diagram showing the selection process.

extraction and quality assessment, data entry, and tabulation, which according to Allen & Olkin [62] corresponds to approximately 51% of the time involved, we adopted a machine learning system that can estimate the probability

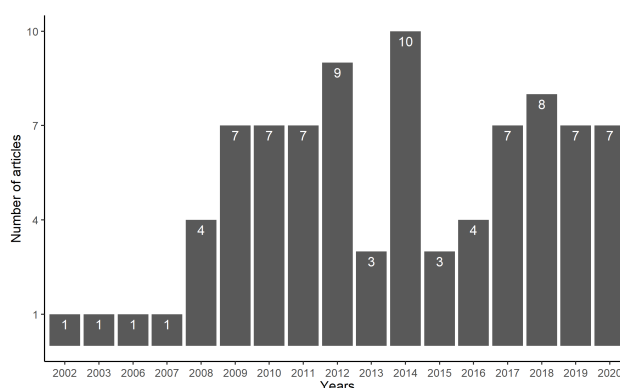


FIGURE 3. Distribution of the selected studies between 2000 and 2020.

that a document should be included and automatically rank the documents from most relevant to least relevant. This process allowed the human reviewer to identify the studies to include earlier in the screening process [63]. Following this assumption, we employed ASReview [64], which also allows us to identify 95% of the eligible studies after screening between 8% of the studies and 33% of the studies [64]. The naive Bayes classifier was adopted using the following steps in ASReview:

- 1) Select five relevant and irrelevant articles;
- 2) ASReview orders the publications in such a way that the most relevant publications are listed first, simplifying the analysis of the abstracts;
- 3) All articles were screened, although the last 178 articles were not determined to be relevant articles.

After the screening process, the final dataset included 87 articles (Figure 3), excluding 311 articles that did not fit the eligibility criteria and that did not address machine learning customer dropout with contractual settings after analyzing the abstract and 4 articles for which we did not have full text access to. The analysis was carried out with two approaches. The first focused on the bibliometric analysis and overall trends of the articles using natural language

TABLE 3. Articles identified from the initial dataset.

Index	Articles	Source	Area	Citations	Index	Articles	Source	Area	Citations
A1	[40]	Scopus	Telecom	3	A45	[67]	Scopus	Retail	19
A2	[68]	Springer	Telecom	2	A46	[69]	Science	Telecom	37
A3	[15]	Scopus	Telecom	0	A47	[70]	Scopus	Telecom	2
A4	[26]	Science	Telecom	103	A48	[71]	Springer	Telecom	0
A5	[72]	Scopus	Software	0	A49	[73]	Springer	Telecom	0
A6	[74]	Springer	Telecom	51	A50	[75]	Scopus	Gaming	5
A7	[18]	Scopus	Telecom and Media	91	A51	[76]	Springer	Software	0
A8	[77]	Springer	Telecom	24	A52	[78]	Scopus	Telecom	14
A9	[79]	Scopus	Media	70	A53	[80]	Scopus	Telecom	6
A10	[81]	Springer	Media	0	A54	[82]	Scopus	Energy	49
A11	[83]	Scopus	Telecom	0	A55	[84]	Scopus	Financial	0
A12	[49]	Scopus	Financial	58	A56	[85]	Scopus	Telecom	7
A13	[86]	Scopus	Media	0	A57	[87]	Scopus	Telecom	1
A14	[88]	Scopus	Media	476	A58	[8]	Scopus	Financial	162
A15	[89]	Scopus	Media	211	A59	[90]	Scopus	Financial and Telecom	10
A16	[91]	Scopus	Media	108	A60	[92]	Scopus	Gaming	54
A17	[93]	Scopus	Gambling	104	A61	[22]	Scopus	Software	11
A18	[28]	Science	Telecom	93	A62	[94]	Springer	Telecom	0
A19	[32]	Scopus	Media	162	A63	[29]	Science	Telecom and Financial	66
A20	[95]	Scopus	Media	454	A64	[96]	Scopus	Hospitality	0
A21	[97]	Scopus	Financial and Retail	0	A65	[98]	Springer	Telecom	0
A22	[42]	Science	Financial, Telecom and Retail	120	A66	[99]	Science	Logistic	5
A23	[46]	Scopus	Financial and Telecom	41	A67	[100]	IEEE	Fitness	2
A24	[101]	Science	Education	12	A68	[102]	Scopus	Financial	4
A25	[33]	Science	Financial	5	A69	[27]	Science	Financial	22
A26	[103]	Science	Telecom	82	A70	[25]	Springer	Telecom	6
A28	[104]	Scopus	Financial	0	A71	[105]	Scopus	Media	121
A27	[106]	Scopus	Telecom	5	A72	[107]	Scopus	Financial	1
A29	[108]	Science	Financial	110	A73	[50]	Scopus	Telecom	345
A30	[109]	Scopus	Financial	19	A74	[24]	Scopus	Telecom	120
A31	[9]	Scopus	Financial and Telecom	27	A75	[41]	Scopus	Telecom	303
A32	[12]	Scopus	Financial	215	A76	[110]	Scopus	Telecom	0
A33	[111]	ACM	Telecom	1	A77	[112]	Scopus	Telecom	10
A34	[54]	Scopus	Financial	81	A78	[113]	Springer	Telecom	0
A35	[114]	Scopus	Financial	69	A79	[43]	Scopus	Telecom	416
A36	[115]	Springer	Telecom	0	A80	[116]	Scopus	Financial	6
A37	[5]	Science	Telecom	606	A81	[47]	Scopus	Telecom	528
A38	[117]	Scopus	Financial	0	A82	[118]	Science	Telecom	148
A39	[119]	Springer	Telecom	64	A83	[120]	Scopus	Telecom	41
A40	[121]	Springer	Telecom	0	A84	[122]	Scopus	Financial and telecom	114
A41	[48]	Springer	Retail	0	A85	[123]	Scopus	Financial	343
A42	[124]	Scopus	Security	0	A86	[125]	Scopus	Telecom	78
A43	[126]	Scopus	Financial	0	A87	[127]	Scopus	Telecom	22
A44	[45]	Science	Telecom	124					

processing applied to the extracted data from the PDF [65]. The second approach developed the qualitative analysis of the articles, providing a greater transparency [66] and systematization and synthesis of the knowledge [57]. The data extraction process was developed at the same time that the papers were reviewed to identify the dropout domain, the type of organization (e.g., airline company, insurance company or telecommunication) and dropout prediction techniques (e.g., decision trees, logistic regression, or support vector machine). The domain identification was based on the article business area description and the context of the dataset (i.e., business area) used to predict the dropout.

III. RESULTS

This section presents the analysis that emerged from the systematic literature review to address the research questions, which provides the results from the bibliometric analysis and the quality scores and a qualitative analysis of the content of the articles.

The final list of the selected articles by source included 55 articles from Scopus, 1 article from IEEE, 78 articles from SpringerLink, 14 articles from Science Direct, and 1 article from ACM. To obtain an overall understanding of the existing studies, we developed a classification to identify the area of the study, which is represented in Table 3.

A. RQ1: WHAT IS THE CURRENT STATE OF THE RESEARCH BEING DEVELOPED?

In this research question, we present the answer to the question regarding the research being developed. Figure 4 shows the evolution of the research using machine learning to predict dropout since 2000. Between 2000 and 2008, only four studies were published: three studies were published in the telecommunications sector [5], [43], [47] and one study was published in the media business sector [89]. The number of publications started to increase in 2008, initially targeting the media, financial and telecom areas [86], [95], [102], [118].

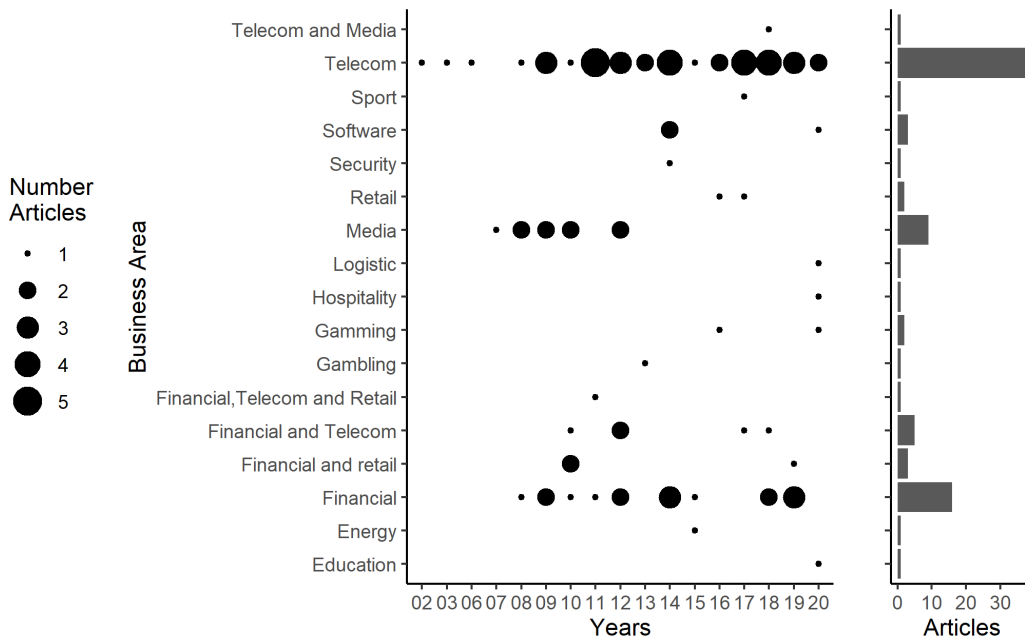


FIGURE 4. Distribution of the selected articles between 2000 and 2020 and by area.

The research growth in the prediction of dropout, namely, with contractual settings, focused on the business telecom, financial and media areas.

The dropout research mainly addresses telecommunications, which corresponds to approximately 44% of the publications, followed by the financial area (representing 18%); the remaining research focuses on the areas of software (3%), retail (2%) and games (2%), while energy, gambling, hospitality, logistics and security represent only 1% each (Figure 4). There are overlapping areas that target several business areas, such as the financial and telecom industries (6%) [9], [29], [46], [90], [122], financial and retail industries (3%) [68], [74], [97], financial, telecom and retail industries (1%) [42], and telecom and media industries (1%) [18]. These studies developed an analysis targeting more than one business area as proof of concept of the approach being tested to predict dropout. Certain areas only addressed the subject in 2020, such as logistics [99], hospitality [96] and education [101]; however, the problem was addressed using traditional statistical methods without machine learning approaches [128]. However, research expansion is underway.

Figure 5 shows that the majority of these studies are being published in the Journal of Expert Systems with Applications, followed by the European Journal of Operational Research, Knowledge and Information Systems, and Decision Support Systems.

B. RQ2 - WHAT ALGORITHMS HAVE BEEN EMPLOYED TO PREDICT DROPOUT?

After investigating the current state of the research being developed, we address the algorithms utilized to target customer dropout. Sixty different types of algorithms were

identified. To simplify the organization of these algorithms, they were organized into the following categories: (1) nature-inspired [129], which encompass nature-inspired algorithms, including artificial neural networks, fuzzy systems, evolutionary computing, and swarm intelligence; (2) ensemble, which use multiple classifiers that are built independently and that participate in a voting procedure to obtain a final class prediction, such as random forests, bagging and boosting [50]; (3) decision trees, which are one of the most popular classification techniques [45]; (5) statistical, such as logistic regression, naive Bayes, and Bayesian networks [50]; (6) rule based, which allows us to reduce the amount of information to understandable statistically supported statements [45]; (7) clusters to classify customers [69]; (8) survival analysis methods, which model the occurrence and timing of events [52]; (9) Markov chain [89]; (10) SVM, which classify the data using the maximal margin hyperplane [115]; and (11) other approaches not included in the previous categories.

The most commonly employed algorithms were decision trees (52 papers), logistic regression (35 papers), random forest (30 papers), support vector machines (29 papers), neural networks (20 papers), Bayesian networks (17 papers), bagging (10 papers), k-means (8 papers), AdaBoost (6 papers), genetic algorithms (5 papers), boosting (4 papers), multilayer perceptron (4 papers), survival analysis (3 papers), and general additive model (3 papers), followed by several different rule-based algorithms, such as decision rule (2 papers) or ripper (2 papers), extreme learning machines (2 papers), root-boost (2 papers), and other algorithms (with only one article).

Table 4 summarizes the surveyed works by the algorithms adopted to predict dropout; we organized the 60 different algorithms into 10 categories.

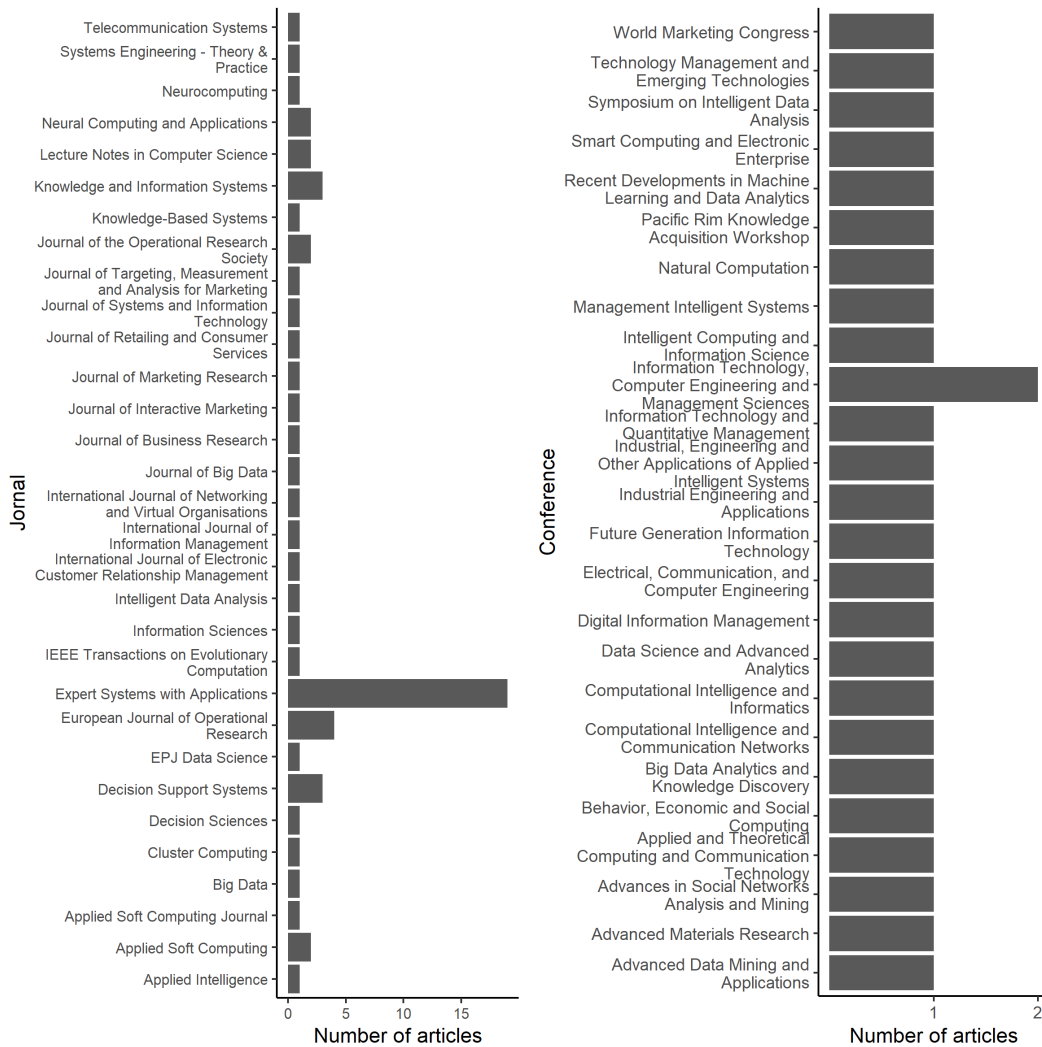


FIGURE 5. Distribution of the selected articles by journal and conference.

Algorithms that are inspired by nature [129], [130] include neural networks, fuzzy systems, evolutionary computing, and swarm intelligence [129]. The first paper identified using neural networks and genetic algorithms was [43]. Different algorithms that use neural networks were identified, and several algorithms based on the concept of an information processing mechanism inspired in the biological system that comprise interconnected elements (neurons) collaborating to solve problems were reported [45]. Different approaches were adopted: (1) counterpropagation neural networks [85] and multilayer perceptron [28], [50], [77], [115], [121]; (2) extreme learning machines [40], [84]; (3) long short term memory [15]; (4) convolutional neural networks [15]; (5) multilayer perceptron [28], [50], [77], [115], [121]; (6) self-organizing error-driven [121]; and (7) fuzzy ARTMAP [85]. Some studies addressed the use of genetic algorithms [26], [43], [70], [102], [122] and swarm intelligence, such as particle swarm optimization [112] and ant colony optimization [41].

Other studies adopted different approaches, such as ensemble methods, to address the problem or created new algorithms to employ this idea. Ensemble methods use multiple base classifiers to improve the predictive performance, which as stated by [50], is the idea that underlies its adoption by other studies [93], [120]. [119] considered ensemble methods to be one of the best approaches to model customer churn. The most commonly employed ensemble method was random forest, which overcomes the instability of decision tree subsetting n random predictors to grow trees using bootstrap sample data [95], which provides good predictive performance. The random forest was adopted for the following five main reasons [86], [95]: (1) predictive performance; (2) robustness to outliers and noise; (3) reasonable computing time and (5) ease of implementation. Another ensemble method was bagging (bootstrap + aggregating), which adopts a model averaging approach and boosting using weak learners to iteratively learn and create a strong classifier. The analyzed studies explored the use of boosting using other

TABLE 4. Articles by adopted algorithms.

Type	Algorithm	Cases	Articles
nature-inspired	neural networks	20	[5, 12, 22, 25, 40, 43, 45, 47, 50, 71, 75, 84, 90, 100, 105, 106, 118, 122, 123, 125]
	genetic	5	[26, 43, 70, 102, 122]
	multilayer perceptron	4	[28, 50, 77, 115, 121]
	extreme learning machines	2	[40, 84]
	particle swarm optimization	1	[112]
	counter propagation neural networks	1	[85]
	fuzzy ARTMAP	1	[85]
	long short term memory	1	[15]
	convolutional neural network	1	[15]
	self-organizing error-driven ant colony optimization	1	[121]
ensemble	random forest	30	[15, 18, 22, 24, 28, 32, 33, 42, 46, 49, 50, 72, 75, 76, 77, 78, 79, 86, 88, 89, 90, 95, 97, 99, 100, 106, 112, 119, 126, 127]
	bagging	6	[24, 28, 46, 50, 77, 120]
	adaboost	6	[77, 97, 99, 106, 116, 121]
	boosting	4	[28, 50, 77, 106]
	rootboost	2	[42, 119]
	rotation forest	2	[42, 119]
	extreme gradient boosting	1	[90]
	gradient boosting	1	[28]
	logistic model tree	1	[50]
	decorate	1	[119]
	random subspace method	1	[46]
feature-selection-based dynamic transfer ensemble	1	[120]	
decision tree	decision tree	52	[5, 8, 12, 25, 29, 40, 41, 42, 43, 45, 47, 50, 50, 54, 71, 72, 73, 74, 75, 77, 79, 85, 90, 92, 93, 94, 100, 105, 111, 112, 115, 117, 118, 121, 122, 123, 125]
	alternating decision tree	2	[24, 50]
	Chi-Square automatic interaction detection	2	[74, 94]
statistical	logistic regression	35	[8, 12, 24, 25, 29, 32, 33, 41, 46, 50, 54, 72, 74, 75, 77, 78, 79, 80, 82, 88, 89, 90, 91, 94, 95, 98, 100, 103, 110, 114, 115, 118, 124, 125, 127]
	Bayesian	17	[24, 25, 28, 50, 71, 77, 92, 101, 102, 104, 106, 108, 109, 110, 112, 117, 118]
	general additive models	3	[46, 91, 93]
	general linear model	1	[18]
rule based	decision rule	2	[47, 105]
	ripper	2	[41, 50]
	part	1	[50]
	exhaustive algorithm	1	[26]
	covering algorithm	1	[26]
	LEM2	1	[26]
clusters	k-means	8	[5, 25, 45, 50, 99, 106, 107, 112]
	fuzzy c-means	2	[25, 113]
	possibility c-means	1	[113]
	hierarchical	1	[69]
	cluster based local outlier factors	1	[107]
	local outlier factors	1	[107]
survival	survival analysis (Kaplan Meyer or cox regression)	3	[54, 86, 92]
	random survival forest	1	[96]
Markov	Markov chain	2	[89, 110]
	Markov logic network	2	[103, 110]
SVM	support vector machine	29	[18, 22, 25, 28, 32, 41, 45, 50, 70, 71, 72, 75, 77, 82, 90, 92, 95, 99, 104, 108, 109, 111, 112, 114, 115, 116, 117, 118, 127]
others	svm-rfe	1	[108]
	multivariate time series	2	[15, 99]
	transfer learning	1	[120]
	random walk	1	[75]
	active learning based approach	1	[41]
	social networks	1	[68]
	fuzzy classifier	1	[77]

approaches, such as (1) AdaBoost [77], [97], [99], [106], [116], [121]; (2) gradient boosting [28]; and (3) extreme gradient boosting [90].

Another category was the decision tree category, which includes the most commonly used algorithm, i.e., the decision tree, due to its easy use [8], interpretability [8], [86], [93],

and conceptual transparency [45]. Some studies identified algorithms such as CART [85], [93] or C4.5 [50], [77], [106], which define the metric for the homogeneity of the groups, where CART uses the Gini criterion and C4.5 uses the entropy measure [50].

The statistical category includes logistic regression. The logistic regression technique is a well-known and simple technique that was referenced by 35 articles. The simplicity and availability of logistic regression and its use as a benchmark against others were identified as advantages in several studies [32], [93]. Some researchers address the problem of tuning the data preparation and analyzing the performance of standard methods, such as logistic regression, against more advanced classifiers, such as bagging, boosting and random forests [28]. Other algorithms, such Bayesian classifier also included certain variants, such as: (1) naive Bayesian classifier [92], [110], [118], (2) naive Bayes tree [108], [109], [117], and (3) Bayesian belief network [101]. The naive Bayes tree combines decision trees and naive Bayes classifiers. The general additive model was also explored, allowing us to relax in the assumptions that considered a nonparametric technique [91], and the general linear model [18] was also explored.

The category rule encompasses several algorithms under the assumption that certain approaches are black box models, such as support vector machines [108]. These approaches allowed us to extract rules, such as association rules, to reduce large amounts of information to small and understandable amounts of information [45].

The cluster category reflects the employment of approaches based on the adoption of unsupervised algorithms to develop the segmentation of customers, to support the use of machine learning algorithms to predict the churn in each segment [5], [25], [112]. Jafari-Marandi *et al.* [121] explored a similar approach that combine clustering methods parallel to classification methods with the aim of creating more control in the decision-making process of churn management, but at the least, expect to increase the accuracy by exploring the individuality of each customer to optimize the classification decision process. Ullah *et al.* [107] also employed clustering by employing fuzzy c-means, possibility c-means and possibility fuzzy c-means. Other studies applied k-means [106], [107], local outlier factors [107], and cluster based local outlier factors [107] for classification.

Support vector machines are algorithms that attempt to obtain an optimal hyperplane to maximize the margin between positive examples and negative examples [32]. When linear transformation is not possible, support vector machines supports the transformation of the input space via nonlinear mapping into a higher dimensional space using a kernel function [95] to transform to a higher dimensional space. Verbeke *et al.* [41] suggested its use considering the higher performance due to its ability to capture nonlinearities. Several studies employed this technique by selecting different kernel functions, such as the radial basis function [95]. In certain cases, the survival approach was adopted to

model the dropout occurrence and timing of the events. The idea was clarified by Burez and Vandenoel [86] by identifying a static approach, where churn behavior was predicted at a moment in time using random forest and dynamic approaches over a period of time. The dynamic approach was explored in several studies identified in Table 4 with survival analysis. Routh *et al.* [96] utilized a random survival forest to explore the advantage of an algorithm based on random forests. The Markov concept explores a random process that is independent of the past state considering the state in a previous event. Predictions are established based on the present state. The algorithms selected include the Markov chain and Markov logic network. Some articles proposed different approaches, such as multivariate time series, transfer learning, random walk, active based learning, social networks and fuzzy classifiers. The exploration of these different approaches started in 2011 with Verbeke *et al.* [41], where active learning was employed in one study, followed by transfer learning in 2015 [120]. More recently, in 2017, fuzzy classifiers [77] were applied, and random walk [75] and multivariate time series [15], [99] were utilized in studies published in 2020.

The variety and exploration of different approaches to address dropout accuracy has led to the integration of several algorithms to increase performance. Vijaya and Sivasankar [112] suggested that works that adopt hybrid models combining more than one classifier can achieve increased performance compared with those using single classifiers. This idea is not new, and some studies have explored the combination of clusters with churn prediction [5], [25], [111], where if the customers are grouped into clusters, the prediction accuracy can be improved within each cluster. The hybrid approach using clustering and classification, which segments the customers before developing a classification, could be effective [121]. An approach that is different from ensemble methods uses several models that are built independently before developing a final step that combine different models to create a final class prediction; the approach applies a pipeline concept, where the output of one algorithm server as the input for another algorithm [108].

Generally, the majority of the algorithms identified decision trees, logistic regression, random forest, support vector machines, and neural networks. The interpretability and simplicity of the first three approaches justifies its wider adoption. The variety of the adopted approaches also denotes an increasing use of hybrid models, where several approaches are combined to improve the performance accuracy.

C. RQ3 - WHAT ARE THE FEATURES USED TO PREDICT DROPOUT?

The aim of this question is to explore which features are being employed to predict dropout. The main approach in all the addressed studies is the use of demographic data and behavioral data related to the use of the service that is being purchased (Figure 6). Some researchers improved this perspective. Risselada *et al.* [29] utilized customer characteristics to represent variables such demographic information,

socioeconomic information (income) and commitment and the relationship to the length (time in the relationship with the company), breadth (usage of other services), and depth (type of service purchased). Gür Ali and Arıtürk [54] proposed the use of new types of data, such as customer interactions with the organization, using a survey to address service quality, dropout and economic indicators reflecting external economic indicators that could influence dropout. The usage of external economic factors is unique in the research analyzed. Ballings *et al.* [81] explored the use of pictorial data to represent feelings and mindset, which represent variables that cannot be extracted from internal processes within the organization, such as purchase intentions, interests, satisfaction, and emotions. Benedek *et al.* [83] employed network topological properties to represent whether the calls were made to customers within the same telecommunication service or to a customer using another telecommunication service. Kaya *et al.* [126] adopted spatio-temporal data to explore the time and location as behavioral data. Moeyersoms and Martens [82] is the only study that uses high cardinality data, such as bank account numbers, family names and ZIP codes to predict customer dropout. Other studies addressed the problem using Recency, Frequency, and Monetary (RFM) model generated variables [32], [80], [93]. Al-Molhem *et al.* [68] conducted social network analysis to enhance the results of churn prediction models in the telecom domain using call detail records to construct a weighed graph representing the distance between two subscribers to calculate the centrality. The usage of marketing related variables, such as promotions offered to a customer, calls developed in a retention strategy, and helpdesk interactions, were applied by Verbeke *et al.* [50]

However, certain studies did not identify the features employed [9], [12], [18], [22], [25], [42], [48], [67], [70], [73], [85], [87], [88], [97]–[99], [107], [110], [112], [113], [116], [119], [122], [127], which in some cases are related to the usage of more than one database. Idris *et al.* [119] used two databases (orange and cell2cell). Burez and Van den Poel [88] used data from two banks with 117808 and 102279 customers, a mobile dataset ($n=100205$), newspaper ($n=122189$) and PayTV (143198). De Bock and Van den Poel [97] explored two datasets (bank1=23562 and bank2=42783), which made them difficult to describe in a research paper.

The selection of the features and how they are used are fundamental to dropout prediction. Feature selection is considered an important and critical step [77], where a previously established process exists for extracting data, such as the Cross Industry Standard Process for Data Mining (CRISP-DM), which develops six distinct stages, i.e., business understanding, data understanding, data preprocessing, modeling, evaluation, and deployment [28]. This underlies the assumption of initially addressing aspects related to the preparation of the data. The existing data analysis techniques use the extracted patterns to make churn decisions for all customers depending on the quality of the learning using the training

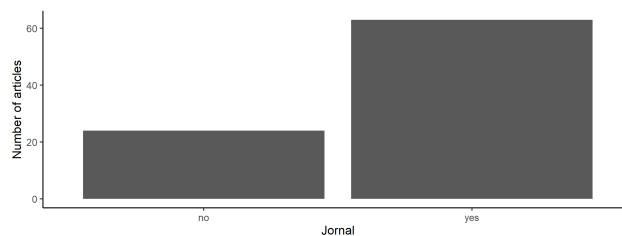


FIGURE 6. Articles identifying features.

set [121]. Coussement *et al.* [28] suggested that optimal optimization in data preparation provides competitive results against classifiers such as neural networks or ensemble methods (e.g., random forests), where optimization was achieved using decision-tree-based remapping for categorical variables, equal frequency binning for continuous variables, and weight-of-evidence conversion as the representation method, suggesting the following techniques for the preparation of the data:

- 1) Transformation of categorical and continuous variables in values;
- 2) Remapping categorical variables;
- 3) Discretization (binning) continuous variables;
- 4) Dummy encoding;
- 5) Weight of evidence for calculating the strength of a category to separate the churners $\frac{\text{proportionchurners}}{\text{proportionnon-churners}}$ in a category; and
- 6) Variable selection using heuristics, such as subsetting variables with higher correlation with dependent variable and low intercorrelation among independent variables.

Different types of features and approaches were adopted to predict customer dropout, such as demographic information, behavioral data and relationship length. Certain researchers have applied external variables to improve the prediction accuracy, such as economic indicators or high cardinality data. However, some studies did not identify which features were utilized, which in most cases was related to the use of more than one dataset in the study. The process related to the preparation of the dataset to be applied in the model prediction is fundamental, and certain studies employed several machine learning algorithms to prepare the data before the prediction model was actually developed.

D. RQ4 - WHEN DOES DROPOUT OCCUR?

Traditional dropout analysis is focused on a dependent variable (churn or not-churn), and different approaches are needed to understand when dropout occurs. Generally, a static perspective of the dropout problem is considered, discarding a temporal perspective of the problem. Few studies address this problem or do not consider dropout prediction as a static prediction, which should be considered. The models should be adapted regularly [29], or this problem should be addressed by considering “time” as an important feature to consider. The problem in the assumption lies in the dynamic behavior

of the customer, which assumes that decisions change over time and that a customer with the intention to drop out at the end of the month may not engage in this behavior at the end of the month [15]. To account for time to churn, survival models have been developed [96]. These models solve important limitations in traditional methods, such as regressions, which would be appropriate only when all customers end the relation, do not allow us to predict when it will occur and do not consider censored data (observations with incomplete information about churn time) to capture the temporal dimension of the churn prediction challenge [92].

Perianez *et al.* [92] approached this problem using survival analysis, with the goal to determine not only if a customer will dropout but also when dropout would occur. Burez and Vandenpoel [86] addressed this problem using a survival distribution to understand when customers stop a subscription, which provides an additional perspective regarding behavioral data. The duration of the usage of a service could be measured in the days (or another time unit) between the contract starting date and the contract termination date [47]. This idea is also explored by Liu *et al.* [75], who analyzed the duration of a player game at a specific timestamp determined by their interaction within a subscription model for a player platform by identifying the retention rate of mobile games as a function of time. Other studies attempted to address this issue by using multivariate time series [15], convolutional neural networks and recurrent neural networks to learn the representative features of customers, especially focusing on daily level dynamic changes; but they lacked the ability to target customers according to their churning causes [15]. The availability of behavioral data (such as those represented in the time series) allows us to improve the model performance [82] and complement the information existing in the form of structured data (e.g., sociodemographic data). Gök *et al.* [111] did not explore the timings of the dropouts but tried to use the information available to create clusters from time series to feed the algorithm to predict churn in a second phase as customers behave in the form of a time series combined with the available features.

The timings involved in dropout prediction should consider the periods related to the estimation period; as stated by Risselada *et al.* [29], the predictive performance deteriorates considerably within a few periods after the estimation period because the parameter estimates change over time and the significant variables differ between two periods. Xiao *et al.* [120] suggested that machine learning models consider the training dataset and that the test dataset is subject to the same distribution, which is not satisfied in many situations, considering that changes occur as time progresses. The data representing the behaviors of the customers should be considered, which allows improvement of the model performance if increased [131].

The duration of the relationship until dropout allows for the evaluation of the tenure relationship between the customer and the organization, indicating lower dropout risks in longer relationships and higher risks in the first 18 months

of the relationship [40]. This pattern was also identified by Burez and Vandenpoel [86] regarding churning after a year of subscription, where one out of three customers leave the company within one year, and half of the customers leave the company within two years. The timestamp gives an additional perspective of the dropout risk, allowing us to identify when it occurs and when to consider retention strategies considering the timing of the events. This approach enhances the understanding of the importance of the duration of the relationship between the customer and the organization as an important feature for predicting dropout, which was considered in fewer studies using survival analysis [54], [86], [92], [96].

However, as previously mentioned, few studies have addressed dropout considering a temporal perspective of the problem. The timing of the dropout is an important feature, which allows us to consider that the decision of the customers to make dropout changes over time, where the survival models allow us to solve the problem of traditional approaches to predict dropout or nondropout. The temporal perspective also allows us to identify when retention actions should be enacted to reduce dropout, providing indicators that should be explored to complement existing indicators that represent a research area opportunity.

E. RQ5 - HOW IS MACHINE LEARNING ALGORITHM DROPOUT PREDICTION ACCURACY MEASURED?

One important aim of this systematic review is to identify trends that can allow researchers to understand the types of metrics that are being employed to analyze the performance of machine learning algorithms. In this research, several metrics were identified. The literature shows the application of different cost-sensitive performance metrics, such as the area under the receiver operating curve (AUC), sensitivity, specificity, recall, precision, and F-score [121], considering that in the case of churn prediction, false negatives are five times as undesirable as false positives. The most commonly employed indicator to measure the performance was accuracy, which was utilized in 39 studies, followed by roc auc in 36 and sensitivity in 36, lift in 26, precision in 16, specificity in 9, and f-measure in 13.

Accuracy reflects the percentage of correctly identified cases [8], [45], [95] and represents an overall performance that is widely adopted [95], represent almost 50% of all occurrences. The accuracy is calculated using the ratio between the total number of correct predictions and the total number of predictions and has been widely employed in several studies [25], [26], [40], [45], [46], [54], [69], [71], [72], [76], [85], [98], [101], [103]–[105], [108], [109], [111]–[118], [118], [120]–[122], [124]. This finding could be attributed to the intuitive interpretation of the indicator and easy calculation.

Certain studies also addressed correct cases, thereby providing an overall performance indicator of the model performance without considering incorrect cases [8], [12], [32], [41], [50], [74], [95].

The sensitivity represents the positives (dropout) regarding their correct prediction and incorrect prediction (false

negatives). Sensitivity appears in several studies, and its higher adoption could be related to the importance of correctly predicting the customers who will drop out, i.e., the proportion of dropouts that truly dropout. This metric was referred to as sensitivity in certain studies [41], [41], [50], [85], [101], [103], [104], [106], [108], [109], [116], [117], [119], recall [26], [45], [69], [72], [75], [76], [105], [105], [107], [111], [114], [121], or true positive [12], [25], [77], [82], [102], [111]–[113], [115].

The studies that use ROC AUC (receiver operating characteristic area under the curve) provide a trade-off between the hit rate and the error rate. The AUC is a numerical value between 0.5 and 1 that combines the sensitivity and the specificity, which is the number of negatives (customers who did not drop out) against the total number of customers. Several studies adopted this metric [15], [28], [32], [42], [46], [49], [50], [54], [74], [75], [77]–[81], [86], [88]–[92], [94], [95], [100], [101], [103], [106], [110], [112], [116], [119], [120], [122], [125]–[127]. The main advantage in dropout is attributed to not placing more emphasis on one class over another class, meaning that it is not biased against the minority class [88] and that it is adopted to address the imbalance in the datasets [126].

The problem of determining customer dropout was also addressed using a perspective to target retention actions [93]; the lift metric allows us to focus on customers with a level of certainty that dropout will occur; and marketing efforts are directed to those who will almost definitively churn. Several studies adopted the lift [5], [15], [18], [43], [49], [54], [78], [80], [82], [89], [97], [118], [125] and accuracy considering that only 10% of customers are most likely to churn [8] in the studies [24], [28], [29], [42], [46], [90], [91], [93], [95], [123], [127]. This approach was expanded, and the maximum profit was considered to optimize the targeting action in retention strategies [110].

Precision indicates the fraction of the detected class members that are actually correct [69]; and was adopted in the studies [25], [26], [45], [69], [72], [75], [76], [82], [96], [103], [105], [107], [111], [112], [114], [121].

Other works employed specificity [85], [104], [106], [108], [109], [117], [119], which is also identified as the true negative [111], [112], and represent the true negative rate by measuring the proportion of non-churners that are correctly classified [50]. The f-measure is calculated with the harmonic mean of recall and precision [15], [26], [45], [69], [72], [76], [96], [105], [107], [111], [112], [114], [121].

The lift allows us to calculate the churn rate by group, and the top decile focuses on the measurement of the 10% of cases that are more likely to churn [8], [95], [123], [125], [127], while the profit approach in the uplift expands this approach to achieve the maximization amount of profit [16], [33], [50], thus providing a cost benefit perspective. The top-decile lift supports the development of more proactive actions to retain customers at risk of churn [95], [123].

The problem was formalized by Devriendt *et al.* [33] using the following four categories regarding churn: (1) customers

who would never churn; (2) customers who churn independently of the retention actions; (3) customers who do not churn because of a retention campaign; and (4) customers who will churn if exposed to a retention campaign.

It is considered positive if approaches are developed to retain the customers, but the following problems should be considered: (1) customers who have a greater risk of dropout should be targeted to provide a base for a better ROI in the retention strategies [95], [123] and (2) retention strategies should be developed focusing on customers with higher satisfaction, or its inclusion could be a reminder of the contractual agreement nearing an end and leading to churn [33]. Although using uplift models seems to be a good strategy, they should also consider factors other than risk and customer satisfaction, as not in consideration they could be counterproductive, which means that they should be removed from retention strategies. Devriendt *et al.* [33] suggested that the use of uplift models could outperform predictive models and contribute to a greater profitability in the development of retention campaigns to reduce dropout. These approaches provide important benchmarking metrics by considering not only the model accuracy but also the financial performance to maximize the retention strategies. Better approaches should be provided to increase the return of investment in marketing campaigns [33], where the business objective is to reduce customer churn and customers who are about to churn but cannot be retained should be excluded from the campaign, as targeting them will be a waste of scarce resources. However, customers with a higher risk of churning may not be the best targets, as suggested by Ascarza [18]. Investments in retention strategies are investments that should be developed to distinguish churners susceptible to marketing actions from those that will leave anyway [28].

IV. OPEN QUESTIONS

The systematic literature review identified open questions about customer dropout. From RQ1, some underresearched business areas were identified, such as the energy sector, education, logistics and hospitality. Compared with other business areas, such as telecom or the financial sector, research on the energy or water supply sectors is lacking, considering the contractual settings that are assumed to provide such types of services. Considering the software as a service SaaS business model of many software companies, the number of studies in this area is also surprisingly low.

RQ2 provided an overall perspective related to the algorithms that are being utilized to predict customer dropout. The first viewpoint could be the importance and wider adoption of decision trees and random forests [49], [74], [89] and logistic regression [91], which could be attributed to its higher interpretability and flexibility [45]. Interpretability is an important aspect for the marketing department in the extraction of valuable information from the model to develop effective retention strategies [50]. The problem arises in the balancing between interpretability and the higher performance of the algorithms inspired by nature (such as neural networks).

From a business perspective, dropout prediction should also be considered as business objective, which requires more than predicting whether the customer will churn [33], where higher interpretability provides better support in the development of retention strategies. The developed SLR also raises the possibility of integrating different algorithms using ensemble methods or integrating several models using a hybrid approach. None of the studies integrated the survival approach to predict customer dropout, for example, using a hybrid approach.

It is considered positive if actions are developed to retain customers, but the following problems should be considered, such: (1) customers who have a greater risk of dropout should be targeted to gain a better ROI from the retention strategies [95], [123] and (2) the retention strategies should be developed by focusing on customers with higher satisfaction, or its inclusion could be a reminder of the contractual agreement nearing an end and could lead to churn [33].

From RQ3, several types of features were identified, such as demographic, behavioral, and economic indicators, pictorial data, network relationships or high cardinality features. The problem that arises is that certain studies employed used data and features that were not described, and this creates a major issue. How can reproducibility be developed in a study without the availability of data or the identification of the features used? Considering that science is driven by data, with the development of new technologies, the increasing complexity of research and the amount of data collected, the challenge is to ensure that research is available to all [132]. This inclusivity requires the availability of both the data and algorithms so that they can be explored by other researchers. The features are selected mainly to verify the performance of the models and are essential to performance prediction, accuracy, and the steps for processing the data, which are fundamental to improve the model accuracy [77].

There are several challenges pertaining to the timing of dropout and the dynamic behavior of a customer with the intent to dropout [15]. The importance of understanding when dropout will occur and the risk when discarding the temporal perspective of the problem seems to be an element that should be addressed. Few studies have considered this issue [86], [92]. This gap in the research represents an opportunity to address the importance of the timeframe and its influence on the efficiency of the model.

According to each business model, the timeframe could be addressed considering the survival probability according to the customer relationship age. Dropout predictions could be developed according to these survival probabilities, as suggested by Esteves and Mendes-Moreira [106], to investigate which data timeframe produces the best results and how the efficiency of the models is influenced by this timeframe. Exploring the duration of the relation and the understanding of the features that increase or decrease this duration seems to be an important approach that could complement the existing approaches to predicting dropout.

From RQ5, the literature analysis showed that different types of questions arise. Which approaches are best for analysing the performance in predicting dropout? Several metrics, such as AUC, sensitivity, specificity, recall, precision, and F-score are employed. However, the goal of customer dropout is to improve the performance of organizations in retaining customers, which is a management problem for which data mining is adopted [50]. The goals of the model should be formulated considering the context of the problem that is being addressed; in marketing retention strategies, the uplift supports the development of proactive actions to minimize the investment in retention strategies [95]. Some assumptions that underlie the adoption of uplift metrics consider that customers with a higher risk of churning may not be the best targets, as suggested by Ascarza [18]. Other researchers have addressed the problem of using the top-decile lift to develop more proactive actions to retain customers at risk of churning [95], [123]. This approach considers that the top the 10% of customers with greater risk, and investments in retention strategies that distinguish churners susceptible to marketing actions from those who will leave anyway should be developed [28]. Although uplift models seem to be good strategies, they should also consider factors other than risk and customer satisfaction, as not considering other factors could be counterproductive, in which case the model should be removed from the retention strategy.

The true business objective is to reduce customer churn. Customers who are about to churn but cannot be retained should be excluded from the campaign, as targeting them would be a waste of scarce resources [33]. Using these models seems to be a good strategy, as they can outperform predictive models that consider only accuracy from a profitability business perspective. It should be considered that customers with a higher risk of churning may not be the best targets to develop retention strategies toward. These perspectives entail that a business context, or the clarification of a business objective underlying the prediction of customer dropout, should be developed to clarify which objectives should be achieved before employing machine learning algorithms. Surprisingly, the analyzed studies did not address the customer lifetime value as an objective to optimize considering the profitability of reducing dropout.

V. VALIDITY THREATS

This section describes the validity threats of this study. The common validity threats are listed as follows [133]: (1) construct validity; (2) internal validity; (3) external validity and (4) conclusion validity.

Construct validity: This threat evaluates whether the measures being adopted correctly support the concepts that are being employed. To reduce the threat to the search process, we selected several databases (SpringerLink, Scopus, Science@Direct, ISI Web of Science, IEEE Digital Library and ACM Digital Library) and employed the PICOC method to define the search string. The extracted measures were

developed using bibliometric analysis and overall trends using language processing applied to the extracted article PDFs; additionally, the data extraction was developed to support our research questions. The sample size of our retrieved studies (420 articles) lowers this threat, supporting the generalization of the results.

Internal validity: This threat validates if a causal relationship, can be identified, i.e., if existing conditions could lead to other conditions, obtaining valid outcomes and evidence using a rigorous process. This category evaluates the implementation of the SLR process. To improve the internal validity, we followed the guidelines of Kitchenham and Charters [55] to develop a rigorous process. To reduce errors our process was developed integrating the results from multiple databases (SpringerLink, Scopus, Science@Direct, ISI Web of Science, IEEE Digital Library and ACM Digital Library), where article screening was developed using machine learning and validated by the researchers, which allowed us to automatically rank the documents reducing human error identifying relevant articles [39]. The adoption of contractual or membership tries to reflect the contractual relationship between a customer and an organization where the moment of dropout is known.

External validity: External validity addresses whether the findings of the study can be generalized and applied to another study. We selected 87 articles between 2000 and 2020 that addressed the prediction of customer dropout using machine learning with contractual settings in multiple databases. Scopus and ISI Web of Knowledge allowed us to reduce the risk of bias during the search, considering that they are the two main bibliographic databases with more than 77.8 and 79 million records [134]. Furthermore, the sample size gives some confidence in the generalization of the results.

Conclusion validity: The conclusion validity addresses issues that affect the ability to draw correct conclusions. The analysis in the selected articles was carried out using two approaches, a bibliometric and a qualitative analysis. The data extraction was developed considering each research question to increase the study reproducibility supported in the adoption of ASReview in the selection of eligible studies. The conclusions in our study could have been threatened by two main factors: the first one is related to the set of identified articles that address the subject of interest and how the analysis developed provides new knowledge. Nevertheless, several strategies have been employed to mitigate this threat.

VI. CONCLUSION AND FUTURE WORK

This study proposed a systematic literature review to identify different critical areas of research for customer dropout using machine learning. We developed a systematic literature review of relevant articles in the last 20 years. The studies explored different business areas, and it seems that there are research gaps that need to be filled and improved. Some business areas are underresearched, such as the energy sector, where studies targeting the dropout problem in a contractual setting are lacking. However, in recent years, we have seen

studies in new areas such as the logistics, hospitality, and education sectors.

The variety of algorithms is impressive, and approximately 60 different types of algorithms were identified. Where a strategy in the data preparation should be developed is a key aspect that is probably more important than which algorithms to use, where interpretability is fundamental, considering the possibility of rule extraction.

The usage of algorithms depends on not only their performance but also their interpretability and simplicity, which could be related to the higher adoption of decision trees and logistics regression. The adoption of hybrid approaches to improve the performance of the predictions where the output of a model is the input of another model has been noted also.

The features used to predict dropout are dependent on the availability of the data and are related to a business area. The studies applied demographic and behavioral data, complemented with other features, such as social relationships with other customers. The process related to the preparation of the dataset is fundamental to the adoption of machine learning algorithms to prepare the data before developing the prediction model.

The performance of the algorithms should be measured in a business context where dropout prediction is addressed, whether a goal is to maximize profit using customer retention or whether the number of customers is considered an objective. The metrics predicting dropout should consider performance indicators such as accuracy but also metrics such as uplift, allowing us to support a profitability perspective in the employed approaches. Investment in counteractions to reduce customer dropout also needs to consider customer responsiveness, distinguishing customers who are more responsive to marketing actions from those who will churn independently of organizational effort. The studies addressed the problem with a static perspective considering the risk of dropout at a specific moment or as a dynamic problem that is considered a temporal perspective, where the risk of dropout varies according to the customer relationship age. Several authors present approaches that are more defined than predicting only dropout or non dropout. The performance of the models should consider a business objective that should aim to increase future earnings from the customer, which should be articulated with the interpretability of the algorithms to support the development of retention actions. Dropout timing is an important element to complement existing approaches, providing a temporal perspective for when retention actions should be developed.

REFERENCES

- [1] J. N. Sheth, B. Mittal, and B. Newman, *Customer Behavior: Consumer Behavior and Beyond*, 1st ed. Fort Worth, TX, USA: South-Western College, Aug. 1998.
- [2] M. J. A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd ed. Indianapolis, IN, USA: Wiley, 2004.
- [3] J. Han and M. Kamber, "Data mining: Concepts and techniques," in *The Morgan Kaufmann Series in Data Management Systems*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2006.

- [4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, Sep. 1996.
- [5] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Syst. Appl.*, vol. 31, no. 3, pp. 515–524, Oct. 2006.
- [6] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and techniques," in *The Morgan Kaufmann Series in Data Management Systems*, 3rd ed. Amsterdam, The Netherlands: Morgan Kaufmann, 2012.
- [7] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
- [8] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 15273–15285, Nov. 2011.
- [9] D. L. García, A. Nebot, and A. Vellido, "Intelligent data analysis approaches to churn as a business problem: A survey," *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 719–774, Jun. 2017.
- [10] A. D. Athanassopoulos, "Customer satisfaction cues to support market segmentation and explain switching behavior," *J. Bus. Res.*, vol. 47, no. 3, pp. 191–207, Mar. 2000.
- [11] M. A. Jones, D. L. Mothersbaugh, and S. E. Beatty, "Switching barriers and repurchase intentions in services," *J. Retailing*, vol. 76, no. 2, pp. 259–274, Jun. 2000.
- [12] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," *Eur. J. Oper. Res.*, vol. 197, no. 1, pp. 402–411, Aug. 2009.
- [13] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, Apr. 2012.
- [14] M. Edward and S. Sahadev, "Role of switching costs in the service quality, perceived value, customer satisfaction and customer retention linkage," *Asia Pacific J. Marketing Logistics*, vol. 23, no. 3, pp. 327–345, Jun. 2011.
- [15] N. Alboukaey, A. Joukhadar, and N. Ghneim, "Dynamic behavior based churn prediction in mobile telecom," *Expert Syst. Appl.*, vol. 162, Dec. 2020, Art. no. 113779.
- [16] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *J. Marketing Res.*, vol. 43, no. 2, pp. 204–211, May 2006.
- [17] S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker, and S. Sriram, "Modeling customer lifetime value," *J. Service Res.*, vol. 9, no. 2, pp. 139–155, 11 2006.
- [18] E. Ascarza, "Retention futility: Targeting high-risk customers might be ineffective," *J. Marketing Res.*, vol. 55, no. 1, pp. 80–98, Feb. 2018.
- [19] E. Ascarza and B. G. S. Hardie, "A joint model of usage and churn in contractual settings," *Marketing Sci.*, vol. 32, no. 4, pp. 570–590, Jul. 2013.
- [20] P. S. Fader and B. G. S. Hardie, "How to project customer retention," *J. Interact. Marketing*, vol. 21, no. 1, pp. 76–90, Jan. 2007.
- [21] W. Xue, Y. Sun, S. Bandyopadhyay, and D. Cheng, "Measuring customer equity in noncontractual settings using a diffusion model: An empirical study of mobile payments aggregator," *J. Theor. Appl. Electron. Commerce Res.*, vol. 16, no. 3, pp. 409–431, Nov. 2020.
- [22] N. Prasasti and H. Ohwada, "Applicability of machine-learning techniques in predicting customer defection," in *Proc. Int. Symp. Technol. Manage. Emerg. Technol. (ISTMET)*, May 2014, pp. 157–162.
- [23] J. Hopmann and A. Thede, "Applicability of customer churn forecasts in a non-contractual setting," in *Innovations in Classification, Data Science, and Information Systems* (Studies in Classification, Data Analysis, and Knowledge Organization), D. Baier and K.-D. Wernecke, Eds. Cham, Switzerland: Springer, 2005, pp. 330–337.
- [24] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," *Appl. Soft Comput.*, vol. 14, pp. 431–446, Jan. 2014.
- [25] E. Sivasankar and J. Vijaya, "Hybrid PPFCM-ANN model: An efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network," *Neural Comput. Appl.*, vol. 31, no. 11, pp. 7181–7200, Nov. 2019.
- [26] A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, and K. Huang, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, May 2017.
- [27] F. Shirazi and M. Mohammadi, "A big data analytics model for customer churn prediction in the retiree segment," *Int. J. Inf. Manage.*, vol. 48, pp. 238–253, Oct. 2019.
- [28] K. Coussement, S. Lessmann, and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry," *Decis. Support Syst.*, vol. 95, pp. 27–36, Mar. 2017.
- [29] H. Risselada, P. C. Verhoef, and T. H. A. Bijmolt, "Staying power of churn prediction models," *J. Interact. Marketing*, vol. 24, no. 3, pp. 198–208, Aug. 2010.
- [30] F. F. Reichheld, "Learning from customer defections," *Harvard Bus. Rev.*, vol. 74, no. 2, pp. 56–67, Mar./Apr. 1996. [Online]. Available: <https://hbr.org/1996/03/learning-from-customer-defections>
- [31] C. B. Bhattacharya, "When customers are members: Customer retention in paid membership contexts," *J. Acad. Marketing Sci.*, vol. 26, no. 1, p. 31, Dec. 1998.
- [32] K. Coussement and D. V. D. Poel, "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6127–6134, Apr. 2009.
- [33] F. Devriendt, J. Berrevoets, and W. Verbeke, "Why you should stop predicting customer churn and start using uplift models," *Inf. Sci.*, vol. 548, pp. 497–515, Feb. 2021.
- [34] Y. Ekinci, N. Uray, and F. Ülengin, "A customer lifetime value model for the banking industry: A guide to marketing actions," *Eur. J. Marketing*, vol. 48, no. 3/4, pp. 761–784, Apr. 2014.
- [35] J. H. Friedman, "An overview of predictive learning and function approximation," in *From Statistics to Neural Networks* (NATO ASI Series), V. Cherkassky, J. H. Friedman, and H. Wechsler, Eds. Cham, Switzerland: Springer, 1994, pp. 1–61.
- [36] P. Ongsulee, V. Chotchaung, E. Bamrunsi, and T. Rodcheewit, "Big data, predictive analytics and machine learning," in *Proc. 16th Int. Conf. ICT Knowl. Eng. (ICT KE)*, Nov. 2018, pp. 1–6.
- [37] J. D. Kelleher, B. M. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 1st ed. Cambridge, Massachusetts: The MIT Press, 7 2015.
- [38] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.
- [39] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: A new paradigm to machine learning," *Arch. Comput. Methods Eng.*, vol. 27, no. 4, pp. 1071–1092, Sep. 2020.
- [40] S. Agrawal, A. Das, A. Gaikwad, and S. Dhage, "Customer churn prediction modelling based on behavioural patterns analysis using deep learning," in *Proc. Int. Conf. Smart Comput. Electron. Enterprise (ICSCEE)*, Jul. 2018, pp. 1–6.
- [41] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2354–2364, Mar. 2011.
- [42] K. W. De Bock and D. V. D. Poel, "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12293–12301, Sep. 2011.
- [43] W.-H. Au, K. C. C. Chan, and X. Yao, "A novel evolutionary data mining algorithm with applications to churn prediction," *IEEE Trans. Evol. Comput.*, vol. 7, no. 6, pp. 532–545, Dec. 2003.
- [44] K. Ng and H. Liu, "Customer retention via data mining," *Artif. Intell. Rev.*, vol. 14, no. 6, pp. 569–590, Dec. 2000.
- [45] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," *Appl. Soft Comput.*, vol. 24, pp. 994–1012, Nov. 2014.
- [46] K. W. D. Bock and D. Van den Poel, "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6816–6826, Jun. 2012.
- [47] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to churn prediction: A data mining approach," *Expert Syst. Appl.*, vol. 23, no. 2, pp. 103–112, Aug. 2002.
- [48] A. T. Jahromi, S. Stakhovych, and M. Ewing, "Customer churn models: A comparison of probability and data mining approaches," in *Looking Forward, Looking Back: Drawing on the Past to Shape the Future of Marketing*, C. Campbell and J. Ma, Eds. Cham, Switzerland: Springer, 2016, pp. 144–148.

- [49] D. F. Benoit and D. Van den Poel, "Improving customer retention in financial services using kinship network information," *Expert Syst. Appl.*, vol. 39, no. 13, pp. 11435–11442, Oct. 2012.
- [50] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, Apr. 2012.
- [51] P. Schober and T. R. Vetter, "Survival analysis and interpretation of time-to-event data: The tortoise and the hare," *Anesthesia Analgesia*, vol. 127, no. 3, pp. 792–798, Sep. 2018.
- [52] D. Van den Poel and B. Larivière, "Customer attrition analysis for financial services using proportional hazard models," *Eur. J. Oper. Res.*, vol. 157, no. 1, pp. 196–217, Aug. 2004.
- [53] J. Ehrlinger, "GgRandomForests: Exploring random forest survival," 2016, *arXiv:1612.08974*.
- [54] Ö. G. Ali and U. Arntürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking," *Expert Syst. Appl.*, vol. 41, no. 17, pp. 7889–7903, Dec. 2014.
- [55] B. Kitchenham and S. Charters, "Guidelines for performing structural literature reviews in software engineering," Keele Univ., Keele, U.K., Tech. Rep. EBSE-2007-01, 2007.
- [56] A. Fink, *Conducting Research Literature Reviews: From Internet to Paper*. Newbury Park, CA, USA: Sage, 2010.
- [57] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Malden, MA, USA: Oxford: Blackwell Pu, 2006.
- [58] C. S. Williams, "Surviving survival analysis. An applied introduction," in *Proc. South East SAS Users Group*, 2008, pp. 1–17.
- [59] G. Rätsch, "A brief introduction into machine learning," in *Proc. 21st Chaos Commun. Congr.*, Tübingen, Germany: Friedrich Miescher Laboratory of the Max Planck Society, 2004, pp. 1–6. [Online]. Available: <https://events.ccc.de/congress/2004/fahrplan/files/105-machine-learning-paper.pdf>
- [60] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [61] P. Ottolinger, T. Leeper, M. Salmon, P. Egeler, and E. X. Esposito, (May 2019). *Bib2Df: Parse a Bibtex File to a Data Frame*. [Online]. Available: <https://CRAN.R-project.org/package=bib2df>
- [62] I. E. Allen, "Estimating time to conduct a meta-analysis from number of citations retrieved," *JAMA, J. Amer. Med. Assoc.*, vol. 282, no. 7, pp. 634–635, Aug. 1999.
- [63] I. J. Marshall and B. C. Wallace, "Toward systematic review automation: A practical guide to using machine learning tools in research synthesis," *Systematic Rev.*, vol. 8, no. 1, p. 163, Jul. 2019.
- [64] R. van de Schoot, J. de Bruin, R. Schram, P. Zahedi, J. de Boer, F. Weijdemans, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands, A. Harkema, J. Willemsen, Y. Ma, Q. Fang, S. Hindriks, L. Tummers, and D. L. Oberski, "An open source machine learning framework for efficient and transparent systematic reviews," *Nature Mach. Intell.*, vol. 3, pp. 1–9, Feb. 2021.
- [65] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
- [66] J. Jesson, *Doing Your Literature Review: Traditional and Systematic Techniques*, 1st ed. Newbury Park, CA, USA: Sage, Mar. 2011.
- [67] S. Khodabandehlou and M. Zivari Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," *J. Syst. Inf. Technol.*, vol. 19, nos. 1–2, pp. 65–93, Mar. 2017.
- [68] N. R. Al-Molhem, Y. Rahal, and M. Dakkak, "Social network analysis in telecom data," *J. Big Data*, vol. 6, no. 1, p. 99, Dec. 2019.
- [69] K. Kianmehr and R. Alhaji, "Calling communities analysis and identification using machine learning techniques," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6218–6226, 2009.
- [70] K. Kianmehr and R. Alhaji, "A fuzzy prediction model for calling communities," *Int. J. Netw. Virtual Organisations*, vol. 8, nos. 1–2, p. 75, 2011.
- [71] K. C. Lee and N. Y. Jo, "Bayesian network approach to predict mobile churn motivations: Emphasis on general Bayesian network, Markov blanket, and what-if simulation," in *Future Generation Information Technology*, T.-H. Kim, Y.-H. Lee, B.-H. Kang, and D. ŚleTzak, Eds. Berlin, Germany: Springer, 2010, vol. 6485, pp. 304–313.
- [72] P. Amornvetchayakul and N. Phumchusri, "Customer churn prediction for a software-as-a-service inventory management software company: A case study in Thailand," in *Proc. IEEE 7th Int. Conf. Ind. Eng. Appl. (ICIEA)*, Apr. 2020, pp. 514–518.
- [73] K.-H. Liao and H.-E. Chueh, "Applying fuzzy data mining to telecom churn management," in *Intelligent Computing and Information Science*, vol. 134, R. Chen, Ed. Berlin, Germany: Springer, 2011, pp. 259–264.
- [74] E. Antipov and E. Pokryshevskaya, "Applying CHAID for logistic regression diagnostics and classification accuracy improvement," *J. Targeting, Meas. Anal. Marketing*, vol. 18, no. 2, pp. 109–117, Jun. 2010.
- [75] X. Liu, M. Xie, X. Wen, R. Chen, Y. Ge, N. Duffield, and N. Wang, "Micro- and macro-level churn analysis of large-scale mobile games," *Knowl. Inf. Syst.*, vol. 62, no. 4, pp. 1465–1496, Apr. 2020.
- [76] N. P. Martono, K. Kanamori, and H. Ohwada, "Utilizing customers' purchase and contract renewal details to predict defection in the cloud software industry," in *Knowledge Management and Acquisition for Smart Systems and Services*, vol. 8863, Y. S. Kim, B. H. Kang, and D. Richards, Eds. Cham, Switzerland: Springer, 2014, pp. 138–149.
- [77] M. Azeem, M. Usman, and A. C. M. Fong, "A churn prediction model for prepaid customers in telecom using fuzzy classifiers," *Telecommun. Syst.*, vol. 66, no. 4, pp. 603–614, Dec. 2017.
- [78] S. Mitrović, B. Baesens, W. Lemahieu, and J. D. Weerd, "On the operational efficiency of different feature types for telco churn prediction," *Eur. J. Oper. Res.*, vol. 267, no. 3, pp. 1141–1155, Jun. 2018.
- [79] M. Ballings and D. Van den Poel, "Customer event history for churn prediction: How long is long enough?" *Expert Syst. Appl.*, vol. 39, no. 18, pp. 13517–13522, Dec. 2012.
- [80] S. Mitrović, G. Singh, B. Baesens, W. Lemahieu, and J. de Weerd, "Scalable RFM-enriched representation learning for churn prediction," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2017, pp. 79–88.
- [81] M. Ballings, D. Van den Poel, and E. Verhagen, "Improving customer churn prediction by data augmentation using pictorial stimulus-choice data," in *Management Intelligent Systems*, vol. 171, J. Casillas, F. J. Martínez-López, and J. M. Corchado Rodríguez, Eds. Berlin, Germany: Springer, 2012, pp. 217–226.
- [82] J. Moeyersoms and D. Martens, "Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector," *Decis. Support Syst.*, vol. 72, pp. 72–81, Apr. 2015.
- [83] G. Benedek, Á. Lublóy, and G. Vastag, "The importance of social embeddedness: Churn models at mobile providers: Social embeddedness and churn," *Decis. Sci.*, vol. 45, no. 1, pp. 175–201, Feb. 2014.
- [84] R. Mohanty and C. N. R. Sree, "Churn and non-churn of customers in banking sector using extreme learning machine," in *Proc. 2nd Int. Conf. Comput. Intell. Inform.*, vol. 712, V. Bhateja, J. M. R. Tavares, B. P. Rani, V. K. Prasad, and K. S. Raju, Eds. Singapore: Springer, 2018, pp. 51–58.
- [85] R. Mohanty and K. J. Rani, "Application of computational intelligence to predict churn and non-churn of customers in Indian telecommunication," in *Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Dec. 2015, pp. 598–603.
- [86] J. Burez and D. Vandenpoel, "Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department," *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 497–514, Jul. 2008.
- [87] M. V. Naik and S. S. Reddy, "An innovative optimized model to anticipate clients about immigration in telecom industry," in *Proc. 3rd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Dec. 2017, pp. 232–236.
- [88] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, Apr. 2009.
- [89] J. Burez and D. Van den Poel, "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services," *Expert Syst. Appl.*, vol. 32, no. 2, pp. 277–288, Feb. 2007.
- [90] M. Óskarsdóttir, B. baesens, and J. vanthienen, "Profit-based model selection for customer retention using individual customer lifetime values," *Big Data*, vol. 6, no. 1, pp. 53–65, Mar. 2018.
- [91] K. Coussement, D. F. Benoit, and D. Van den Poel, "Improved marketing decision making in a customer churn prediction context using generalized additive models," *Expert Syst. Appl.*, vol. 37, no. 3, pp. 2132–2143, Mar. 2010.

- [92] A. Perianez, A. Saas, A. Guitart, and C. Magne, "Churn prediction in mobile social games: Towards a complete assessment using survival ensembles," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2016, pp. 564–573.
- [93] K. Coussement and K. W. De Bock, "Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning," *J. Bus. Res.*, vol. 66, no. 9, pp. 1629–1636, Sep. 2013.
- [94] D. Radosavljevik and P. van der Putten, "Preventing churn in telecommunications: The forgotten network," in *Advances in Intelligent Data Analysis XII*, vol. 8207, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Tucker, F. Höppner, A. Siebes, and S. Swift, Eds. Berlin, Germany: Springer, 2013, pp. 357–368.
- [95] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 313–327, Jan. 2008.
- [96] P. Routh, A. Roy, and J. Meyer, "Estimating customer churn under competing risks," *J. Oper. Res. Soc.*, vol. 72, no. 5, pp. 1138–1155, Aug. 2020.
- [97] K. W. D. Bock and D. Van den Poel, "Ensembles of probability estimation trees for customer churn prediction," in *Trends in Applied Intelligent Systems*, vol. 6097, N. García-Pedrajas, F. Herrera, C. Fyfe, J. M. Benítez, and M. Ali, Eds. Berlin, Germany: Springer, 2010, pp. 57–66.
- [98] M. Saravanan and G. S. V. Raajaa, "A graph-based churn prediction model for mobile telecom networks," in *Advanced Data Mining and Applications*, vol. 7713, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S. Zhou, S. Zhang, and G. Karypis, Eds. Berlin, Germany: Springer, 2012, pp. 367–382.
- [99] S. E. Schaeffer and S. V. Rodriguez Sanchez, "Forecasting client retention—A machine-learning approach," *J. Retailing Consum. Services*, vol. 52, Jan. 2020, Art. no. 101918.
- [100] J. Semrl and A. Matei, "Churn prediction model for effective gym customer retention," in *Proc. Int. Conf. Behav. Econ., Sociocultural Comput. (BESC)*, Oct. 2017, pp. 1–3.
- [101] D. Delen, K. Topuz, and E. Eryarsoy, "Development of a Bayesian belief network-based DSS for predicting and understanding freshmen Student attrition," *Eur. J. Oper. Res.*, vol. 281, no. 3, pp. 575–587, Mar. 2020.
- [102] H. Shao, G. Zheng, and F. An, "Construction of Bayesian classifiers with GA for predicting customer retention," in *Proc. 4th Int. Conf. Natural Comput.*, 2008, pp. 181–185.
- [103] T. Dierkes, M. Bichler, and R. Krishnan, "Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks," *Decis. Support Syst.*, vol. 51, no. 3, pp. 361–371, Jun. 2011.
- [104] M. A. H. Farquard, V. Ravi, and S. B. Raju, "Data mining using rules extracted from SVM: An application to churn prediction in bank credit cards," in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, vol. 5908, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. Sakai, M. K. Chakraborty, A. E. Hassanien, D. Ślęzak, and W. Zhu, Eds. Berlin, Germany: Springer, 2009, pp. 390–397.
- [105] C.-F. Tsai and M.-Y. Chen, "Variable selection by association rules for customer churn prediction of multimedia on demand," *Expert Syst. Appl.*, vol. 37, no. 3, pp. 2006–2015, Mar. 2010.
- [106] G. Esteves and J. Mendes-Moreira, "Churn prediction in the telecom business," in *Proc. 11th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2016, pp. 254–259.
- [107] I. Ullah, H. Hussain, I. Ali, and A. Liaquat, "Churn prediction in banking system using K-means, LOF, and CBLOF," in *Proc. Int. Conf. Electr. Commun., Comput. Eng. (ICECCE)*, Jul. 2019, pp. 1–6.
- [108] M. A. H. Farquard, V. Ravi, and S. B. Raju, "Churn prediction using comprehensible support vector machine: An analytical CRM application," *Appl. Soft Comput.*, vol. 19, pp. 31–40, Jun. 2014.
- [109] M. A. H. Farquard, V. Ravi, and S. B. Raju, "Analytical CRM in banking and finance using SVM: A modified active learning-based rule extraction approach," *Int. J. Electron. Customer Relationship Manage.*, vol. 6, no. 1, p. 48, 2012.
- [110] T. Verbraken, W. Verbeke, and B. Baesens, "Profit optimizing customer churn prediction with Bayesian network classifiers," *Intell. Data Anal.*, vol. 18, no. 1, pp. 3–24, Jan. 2014.
- [111] M. Gök, T. Özyer, and J. Jida, "A case study for the churn prediction in turksat internet service subscription," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2015, pp. 1220–1224.
- [112] J. Vijaya and E. Sivasankar, "An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing," *Cluster Comput.*, vol. 22, no. S5, pp. 10757–10768, Sep. 2019.
- [113] J. Vijaya, E. Sivasankar, and S. Gayathri, "Fuzzy clustering with ensemble classification techniques to improve the customer churn prediction in telecommunication sector," in *Recent Developments in Machine Learning and Data Analytics*, vol. 740, J. Kalita, V. E. Balas, S. Borah, and R. Pradhan, Eds. Singapore: Springer, 2019, pp. 261–274.
- [114] B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on SVM model," *Proc. Comput. Sci.*, vol. 31, pp. 423–430, Jan. 2014.
- [115] B. Q. Huang, M.-T. Kechadi, and B. Buckley, "Customer churn prediction for broadband internet services," in *Data Warehousing and Knowledge Discovery*, vol. 5691, T. B. Pedersen, M. K. Mohania, and A. M. Tjoa, Eds. Berlin, Germany: Springer, 2009, pp. 229–243.
- [116] Y. Wang and J. Xiao, "Transfer ensemble model for customer churn prediction with imbalanced class distribution," in *Proc. Int. Conf. Inf. Technol., Comput. Eng. Manage. Sci.*, Sep. 2011, pp. 177–181.
- [117] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. A. H. Farquard, V. Ravi, and S. B. Raju, "Rule extraction from support vector machine using modified active learning based approach: An application to CRM," in *Knowledge-Based and Intelligent Information and Engineering Systems*, vol. 6276, R. Setchi, I. Jordanov, R. J. Howlett, and L. C. Jain, Eds. Berlin, Germany: Springer, 2010, pp. 461–470.
- [118] G.-E. Xia and W.-D. Jin, "Model of customer churn prediction on support vector machine," *Syst. Eng. Theory Pract.*, vol. 28, no. 1, pp. 71–77, Jan. 2008.
- [119] A. Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification," *Int. J. Speech Technol.*, vol. 39, no. 3, pp. 659–672, Oct. 2013.
- [120] J. Xiao, Y. Xiao, A. Huang, D. Liu, and S. Wang, "Feature-selection-based dynamic transfer ensemble model for customer churn prediction," *Knowl. Inf. Syst.*, vol. 43, no. 1, pp. 29–51, Apr. 2015.
- [121] R. Jafari-Marandi, J. Denton, A. Idris, B. K. Smith, and A. Keramati, "Optimum profit-driven churn decision making: Innovative artificial neural networks in telecom industry," *Neural Comput. Appl.*, vol. 32, no. 18, pp. 14929–14962, Sep. 2020.
- [122] J. Xiao, L. Xie, C. He, and X. Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3668–3675, Feb. 2012.
- [123] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5445–5449, Apr. 2009.
- [124] M. Jiang, N. Chu, and X. M. Bi, "Research on customers churn prediction model based on logistic," *Adv. Mater. Res.*, vols. 989–994, pp. 1517–1521, Jul. 2014.
- [125] X. Zhang, J. Zhu, S. Xu, and Y. Wan, "Predicting customer churn through interpersonal influence," *Knowl.-Based Syst.*, vol. 28, pp. 97–104, Apr. 2012.
- [126] E. Kaya, X. Dong, Y. Suhara, S. Balcisoy, B. Bozkaya, and A. S. Pentland, "Behavioral attributes and financial churn prediction," *EPJ Data Sci.*, vol. 7, no. 1, p. 41, Dec. 2018.
- [127] B. Zhu, B. Baesens, A. Backiel, and S. K. L. M. van den Broucke, "Benchmarking sampling techniques for imbalance learning in churn prediction," *J. Oper. Res. Soc.*, vol. 69, no. 1, pp. 49–65, Jan. 2018.
- [128] J. Forsman, M. van den Bogaard, C. Linder, and D. Fraser, "Considering Student retention as a complex system: A possible way forward for enhancing student retention," *Eur. J. Eng. Educ.*, vol. 40, no. 3, pp. 235–255, May 2015.
- [129] W. Fang, X. Li, M. Zhang, and M. Hu, "Nature-inspired algorithms for real-world optimization problems," *J. Appl. Math.*, vol. 2015, pp. 1–2, Oct. 2015.
- [130] K. C. B. Steer, A. Wirth, and S. K. Halgamuge, "The rationale behind seeking inspiration from nature," in *Nature-Inspired Algorithms for Optimisation*, (Studies in Computational Intelligence), R. Chiong, Ed. Berlin, Germany: Springer, 2009, pp. 51–76.

[131] E. Junqué de Fortuny, D. Martens, and F. Provost, "Predictive modeling with big data: Is bigger really better?" *Big Data*, vol. 1, no. 4, pp. 215–226, Dec. 2013.

[132] B. Hanson, A. Sugden, and B. Alberts, "Making data maximally available," *Science (New York, N.Y.)*, vol. 331, no. 6018, p. 649, Feb. 2011.

[133] X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang, "A map of threats to validity of systematic literature reviews in software engineering," in *Proc. 23rd Asia-Pacific Softw. Eng. Conf. (APSEC)*, Dec. 2016, pp. 153–160.

[134] R. Pranckutė, "Web of science (WoS) and Scopus: The titans of bibliographic information in today's academic world," *Publications*, vol. 9, no. 1, p. 12, Mar. 2021.



DOMINGOS DOS SANTOS MARTINHO received the master's and Ph.D. degrees from the University of Lisbon, Portugal. He is currently a Coordinating Professor with ISLA Santarém, Portugal. His research interests include online learning systems, information systems and technologies, and decision support systems.



JOSÉ G. ALONSO (Member, IEEE) received the Ph.D. degree (with European Mention), in 2014. He is currently an Associate Professor with the Department of Informatics and Telematics System Engineering, University of Extremadura, Spain, and a Co-Founder of Startups Gloin and Viable. His research interests include quantum software engineering, eHealthcare, eldercare, mobile computing, context-awareness, and pervasive systems.



PEDRO SOBREIRO received the B.S. degree in computer science from ISLA Santarém, Portugal, and the M.S. degree in computer science from Portucalense University, Portugal. He is currently pursuing the Ph.D. degree with the Universidad de Extremadura, Spain. He is also an Associate Professor with the Sport Sciences School of Rio Maior, Polytechnic Institute of Santarém, Portugal. His interests include customer dropout, information systems, and machine learning.



JAVIER BERROCAL (Member, IEEE) received the Ph.D. degree (with European Mention), in 2014. He is currently an Associate Professor with the Department of Informatics and Telematics System Engineering, University of Extremadura, Spain, and a Co-Founder of Startup Gloin. His research interests include mobile computing, context awareness, pervasive systems, the Internet of Things, and fog computing.

...