I.J. Information Engineering and Electronic Business, 2020, 4, 1-14
Published Online August 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2020.04.01

Designing a Real-Time Data-Driven Customer Churn Risk Indicator for Subscription Commerce

Alexandros Deligiannis, Charalampos Argyriou
Research & Development Department, Apifon S.A., Thessaloniki, 570 01, Greece
Email: {a.deligiannis, c.argyriou}@apifon.com

Received: 27 March 2020; Accepted: 08 May 2020; Published: 08 August 2020

Abstract: One of the main goals of customer relationship management is to reduce or eliminate "customer churn", i.e. loss of existing customers. This paper introduces a prototype algorithm to estimate a continuously updated indicator of the probability of an existing customer to cease purchasing from a subscription commerce business. The investigation is focused on the case of repeat consumers of subscription commerce products which require regular replacement or replenishment. The motivation is to help marketers to make targeted proactive retention actions by categorizing regular customers into groups of similar estimated churn risk. The proposed algorithm re-computes the probability of churn for each customer at regular intervals using past purchase transaction data and incorporating subscription-based business logic. We describe the detailed process from data collection and feature engineering to the algorithm's design. We also present evaluation results of the algorithm's performance based on a pilot test that took place on a consumables e-commerce business. The results suggest a significant capability of the proposed algorithm in capturing the purchasing intentions of repeat customers, regardless of the risk group they belong to.

Index Terms: Churn prediction, Customer relationship management, Prototype algorithm, Purchase transaction data, Conversion rate.

1. Introduction

Customer churn is defined as the tendency of customers to stop purchasing services from a company [1] and is usually referred to as churn rate or rate of attrition. From a mathematical perspective, it is defined as the percentage of users who stop using a service within a given period of time.

High customer defection represents a significant challenge for companies in several conventional service industries, such as financial, utility, healthcare, or telecommunications services [2]. This is also true for companies adopting new forms of service-based business models, such as subscription commerce businesses which sell regularly replenished consumer goods.

It has been suggested that acquiring a new customer is 5 to 25 times more expensive than retaining one [3], while reducing churn by just 5% could boost profitability by 75%. Van den Poel and Larivière showed that improving retention has up to 4 times greater impact on growth than acquisition [4].

For those reasons, marketers in service industries find great value in technological tools that can help them understand what causes customer churn [5]. Churn risk prediction modelling is not a simple process, since a wide variety of factors may come into play and conditions can differ significantly between businesses and customers. Reasons for churn may include high service cost, poor user experience, lack of desirable features, superior competitor products, evolving customer needs or other factors resulting in low market fit [6].

Much of the research published on the topic of customer churn estimation modelling for service industries focuses on quantifying churn as a lagging indicator. That means that the loss of customers is captured and interpreted after it actually happened. From a business-wise perspective, however, a proactive solution that gives marketers the opportunity to change the mind of a dissatisfied customer before the customer actually leaves, is much more powerful. Relatively less research has been published in this direction and has mainly focused on indicators that can quantitatively predict the churn of known customers before they cease purchasing, acting like "red flags" to the marketer's attention (i.e., leading indicators) [7].

Many state-of-the-art predictive models have been developed for customer churn prediction. That models mostly include purely statistical methods, neural networks, as well as sophisticated machine learning algorithms [8]. However, most of them suffer from 2 limitations: (a) they cannot effectively capture the business needs and constraints from the each business sector; and (b) they are not interpretable by non-experts in order to both reveal the reasons of customer churn and get further optimized.

The research presented in this paper is part of project PRIME (Predictive Personalization of Conversational Customer Communications with Data Protection by Design). This is an internal R&D program carried out at Apifon (www.apifon.com) with the goal of developing a next-generation business-to-consumer messaging platform with advanced personalization capabilities [9, 10]. One envisaged capability of the PRIME platform is to automate the delivery of a personalized message to any customer whose estimated risk of churn exceeds a certain threshold. We aim to build this capability leveraging data from past customer purchases and past exchanges of messages between the business and the customer, while ensuring data protection through a GDPR-compliant architecture [11].

The current study aims: (a) to introduce a custom algorithm that can estimate the repeat customers' churn risk for a subscription commerce business; (b) to incorporate the aspect of transactional sequence over time per customer utilizing real-time transactional data (i.e., churn risk based on the status at the given time); and (c) to effectively categorize repeat customers into distinct segments based on their churn risk score in order to help marketers contact them in a personalized manner.

2. Related work

Customer churn prediction involves building a prediction model that ranks customers based on their likelihood to stop consuming the services offered by a company. An accurate and effective churn prediction model assigns high churn probabilities to customers who actually stop purchasing, and lower ones otherwise, based on past customers' behavioral data [12]. Retaining measures for customers who tend to churn is generally referred to as customer retention.

Historically, the study of customer churn prediction has gone through the following evolution steps:

- The first stage was based on statistical predictions, including Bayesian classification, clustering, logistic regression and decision trees. A comparative study of the application of such methods in telecommunications industry is presented by Vafeiadis et al. [13]. They applied 5 different classification methods in order to predict customer churn. Based on their investigation's results, boosted Support Vector Machine classification achieved the highest accuracy.
- The second stage relied on more complex artificial intelligence algorithms. The main methods were artificial neural network and self-organizing mapping. In order to predict the customer churn, Ahmad et al. applied 4 advanced machine learning algorithms on big telecommunication data [14]: (i) Decision Tree; (ii) Random Forest; (iii) Gradient Boosted Machine Tree (GBM); and (iv) Extreme Gradient Boosting (XGBoost). The results revealed that XGBoost algorithm outperformed the rest ones, since it relies on the results of more than one machine learning models.
- The third stage incorporated integrated business rules into predictive algorithms. To that direction, Wu and Meng introduced a hybrid predictive approach, since they firstly used clustering to identify high-value customers, and then applied the Synthetic Minority Over-sampling Technique (SMOTE) to process e-commerce customer imbalanced data [15]. They eventually used the AdaBoost algorithm to perform churn prediction, concluding that the model got better results after customer segmentation.

As the market is getting more and more competitive over the years, the cost of acquiring new customers is continuously getting increased. The cost of maintaining old customers is proved to be more affordable than the cost of acquiring new customers [16]. Based on that assumption, more and more companies try to apply efficient customer segmentation aiming to personalize their customer retention strategies and reduce the rate of their customers' churn. Zhuang built a machine learning model to predict the churn of customers before and after their segmentation [17]. The research showed that the prediction accuracy was higher after customer segregation, although the results were mainly based on customer activity in social networks.

Customer segmentation refers to the classification of customers according to their attributes, behaviors, needs, preferences and values, in a clear strategic business model, in a specific market. Dividing customers into different groups, according to certain criteria, and marketing them with targeted manners can improve customer loyalty, and reduce customer churn. In this scope, Ascarza [18] defined a customer segmentation algorithm that classifies customers based on a combination of customer value, needs and interests. However, the evaluation of this method applied to a single retention campaign, whereas most companies typically run multiple campaigns as part of their retention efforts.

Many researchers have identified multiple attributes from accessible data that reveal the implicit behavioral patterns of customers. Some of the behavioral features that are proven to be essential predictors of the churning customers, stem from the Recency, Frequency and Monetary (RFM) value of the customers [19]. This rationale is usually used in combination with features calculated based on the location and the time (i.e., spatio-temporal features), the customer segmentation and the customer lifetime value. Regarding the three notions that constitute the RFM analysis, recency is the time interval since the last purchase or transaction is made, frequency is the number of purchases made in a specified time window and monetary value is the amount spent during a specified time window [20]. These attributes help organizations define custom thresholds of customer retention rates.

On top of those attributes, Kaya et al. used purchase transaction data to determine core behavioral traits of customers which led to better churn prediction [21]. They defined even more attributes, like diversity, loyalty, regularity, and tracked individual customer choices for making financial transactions. That kind of features is characterized as spatio-temporal features, since they are based on the location and timestamp of transactions. They particularly aim to measure and analyze the behavior of a customer who tends to churn and this is the core objective of the current research.

Keramati and Ardabili defined customer satisfaction as an experience-based assessment that derives from the degree that the customer service meets the customer expectations [22]. Their customer churn analysis utilized telecommunication services' data, such as demographic data, call detail records, length of duration since a customer is paired to a service provider, and count of logged complaints taken from call center data. This study showed that data from CRM tools help to get information about customer satisfaction level, which can be used along with other attributes for better analysis and prediction of churning customers.

The advent of machine learning brought many data mining approaches that have been developed trying to analyze common customer usage patterns in order to predict their churn [23]. Euler developed a decision tree algorithm to indicate the telecommunications where the customers are most likely to churn [24]. He used the data pre-processing capabilities of the behavioral mining techniques to derive predictive features that were not present in the original data. The extracted findings can effectively extend business intelligence focused on the problems of customer retention, targeted campaign management and churn prediction.

Coussement and Van den Poel applied support vector machines to improve the performance of predicting churn for a newspaper subscription service [25]. The results revealed that interaction between the customers and the businesses is an important predictor of churn. They also extended their study of customer-to-business interaction by adding emotions stemming from customer emails to their model [5]. Towards the same direction, Hadden et al identified predictive features into the customer complaints, and found that decision trees outperform neural networks and regression in terms of overall model accuracy [26].

The problem of understanding the user intent in an online environment has been heavily studied by applying several machine learning modelling approaches. Romov and Sokolov were the winners of a relevant competition by applying a Gradient Boosting Machine (GBM) model with extensive feature engineering [27]. Their recommender system performed especially well in handling a huge amount of user interaction data, in tracking real-time changes, as well as in tackling the cold-start problem.

Sheil et al. leveraged multi-layer recurrent neural networks to model semi-structured input data of categorical, quantified and unique instances [28]. The results showed classification accuracy of 98%, without the need for any domain-specific feature engineering on the purchasing events. However, their model lags behind in the cognitive capturing of the connection between unrelated instances. To predict the shopping behavior, many published solutions use a mixture of Recurrent Neural Networks (RNNs), effectively combining a predictive with a recommendation model. Thus, they aim to utilize temporal features from user preferences, in order to improve recommendations, and predict future interests in the e-commerce sector [29, 30]. However, most of that research is still difficult to adapt to the need of personalization in production level and often acts as a black box.

Association rules for customer churn estimation, is also a widely adopted strategy that could fit into approaches like Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Data Mining by Evolutionary Learning (DMEL), and Bayesian networks [31]. Based on that, Chiang et al. introduced a goal-oriented algorithm for identifying patterns in potential churners using association rules that identify relationships amongst variables [32]. They defined a two-step process for finding out association rules. In the first step, they defined the large item set that requires compliance with certain minimum conditions of support and a minimum confidence. In the second phase, they used the Apriori algorithm to explore the rules of association. However, the proposed solution still have some caveats because of the specific nature of the churn prediction problem, such as the handling of imbalanced and noisy data and the ranking of subscribers according to their likelihood to churn.

Verbeke et al. pointed out that data mining methods for business applications should not only consist of poorly interpretable machine learning models of high accuracy, but they should also provide comprehensible models for non-experts to understand [33]. Towards this direction, their proposed method of rule extraction incorporates personalized customer parameters giving the change to adequately justify a customer's churn. Huang et al. also proposed an interpretable rule-based classification algorithm that applies several methods to learn a set of rules and then uses these rules to classify customers in the telecommunications sector and predict their behavior, with promising results [34].

3. Research Scope and Goals

A. Targeted proactive measures for customer retention

From a marketing perspective, the ultimate goal of the timely estimation of the risk of a customer's attrition at a given period of time, is to take the necessary actions to retain them. The type and the level of intensity of those actions is determined by the risk level that each particular customer belongs to [35].

Towards this direction, user experience designers employ a range of practices to improve customer engagement in subscription commerce applications. The following practices are some of the most common [36]:

- *Push notifications:* Send automated push notifications to customers who are about to churn in order to encourage repeat visits, engagement, and purchases. With a personalized approach, these notifications can reactivate consumers who are at risk of churning.
- *Personalized content:* With personalized interactions and relevant message content, customers feel like marketers are actually speaking to them using data like first name, behaviors and preferences to customize interactions.
- *Deep linking:* Deep links (i.e., direct URI links) can take users right to a particular screen inside a website or an application. This results in launching the web page or the application from exactly where the user left off, or take him/her to a specific product page immediately. This can be an additional approach to increase the conversion rate of inactive customers.
- *In-app messaging:* In-app messaging shows users the right message at the right time. Thus, a website can welcome new users coming for the first time and help them discover new features. Alternatively, a personalized promo could be displayed when they view a particular product.

B. Previous PRIME research

As stated in the introduction, the customer churn risk indicator presented in this paper is part of a set of services included in the PRIME platform [10]. These services were built to offer personalized marketing tools pertaining to subscription commerce businesses.

After analyzing the needs and the way that the aforementioned business sector operates, the PRIME platform was designed to provide ecommerce marketers with the following suite of services:

- *Date & Time Optimization:* The goal of any business that promotes its own products through direct marketing campaigns is to find the best time to send out the campaign content, so as to achieve the highest possible conversion rate. A corresponding use case that the PRIME platform supports, is to automatically determine the right date and time to send out a specific direct marketing campaign message to each customer individually, based on his/her unique profile [37].
- *Segment Recommendation:* A great challenge faced by many marketing professionals in their daily work has to do with how to choose the recipient list for a particular campaign message. Marketers often lack the tools to be able to specify a highly relevant target audience for a campaign. Messages are often sent out to large lists of non-relevant customers, resulting in low engagement rates and poor customer experience. Through automated segmentation, the PRIME platform enables marketers to reach those subscribers who are most likely to find the content of a specific campaign relevant and compelling enough to engage.
- *Keyword Suggestion:* One of the most important factors that define a campaign's success is its text content. Personalization means using the most appropriate vocabulary, in terms of keyword efficiency, to speak to each individual customer. PRIME offers personalized content enrichment for the message of an upcoming campaign by automatically suggesting effective words to be added to the campaign text.
- *CTR Estimation:* One of the most important challenges in everyday marketing practice is the lack of foresight on the effectiveness of a messaging campaign during planning. Marketers often have no way of knowing how effective a campaign will be before executing it. The PRIME platform aims to produce a reliable estimation of a messaging campaign's CTR before it is actually sent [11]. This is tied to predicting who of the recipients will successfully receive the message and will subsequently engage with the content (e.g., by opening a link in the message).

In addition to the above tools, the PRIME platform provides a prototype customer churn risk estimation algorithm tailored to subscription commerce businesses. This service aims to facilitate marketers to timely anticipate when a customer is highly likely to cease purchasing in order to take preventive measures, like sending out special offers. After computing a reliable estimator of the customer's risk of leaving at any given time, the platform can dynamically categorize customers into groups, based on their projected risk score.

C. Research objective

The goal of the research presented in this paper is to develop a prototype algorithm that can estimate the probability of a customer to cease purchasing from a business at the given period of time. Based on real-time purchase transaction data, the introduced algorithm is built on top of subscription-based business logic stemming from each customer's purchasing behavior. The ultimate goal is to maintain a dynamically updated categorization of customers based on their projected churn risk score. This would allow marketers to communicate with buyers who belong to different risk categories in a targeted manner in the context of proactive customer retention.

As discussed in the related work section, the majority of measures against customer churn are based on a personalized way to communicate with the customer or website visitor. This fact reveals that personalization plays a crucial role in every proactive strategy for customer retention. Significant progress has already been made in developing algorithms capable of assessing customers' intention to stop buying.

Towards this direction, the current research aims to contribute in two ways: a) customer churn prediction for the business sector of subscription commerce; b) system to be used as part of customer churn management which combines rule-based customer classification and churn risk scoring based on customer purchasing behavior.

Our approach draws inspiration from three areas of related research: (a) rule-based classification of customers [34]; (b) segmentation of customers based on purchasing behavior data [17, 18]; and (c) grouping of repeat customers based on RFM analysis principles [19, 20].

4. Data Collection & Processing

A. Integration of purchase transaction data

Purchase transaction data refers to records of customers' previous purchases of products sold by a subscription commerce business. This data is owned and controlled by the said business type and needs to be continuously imported, in streaming format, into the PRIME platform to be analyzed in real-time.

Table 1 provides an example list of key features found in the purchase transaction data.

Table 1. Typical incoming purchase transaction entry example.

Feature	Explanation	Type
Account Id	Unique identifier of the company managing the transaction	Numerical
Transaction Date	Date of the transaction	Date
Customer Code	Unique identifier of the customer who made the transaction	Numerical
Transaction Code	Unique identifier of the transaction	Numerical
Order Source	Way that the customer made the transaction	Categorical
Products	List of products purchased in the transaction	List of Categorical
Total Price	Total amount of money paid for the transaction	List of Categorical
Payment Method	Way that the customer paid for the transaction	Categorical
Order Status	Status of the payment	Categorical

Data integration between the PRIME platform and the data owners is achieved through a "publish/subscribe" messaging infrastructure. The infrastructure developed in the scope of this research is built on top of the RabbitMQ open source message broker software [38]. The underlying Advanced Message Queuing Protocol (AMQP) provides a channel to send and receive information that remains secure end to end. The protocol's most important features are message addressing, routing, reliability and security [39].

B. Data processing & feature engineering

"Feature engineering" refers to the process of extracting features from raw data through data mining or domain-specific rules. For instance, "Transaction Date" (shown in Table 1) is a feature already present in the incoming data stream, whereas "Minimum Consumption Duration" (shown in Table 2) needs to be extracted from the data source through a computation step.

The process of feature engineering needs to be repeated on the fly, such that the data features are continuously being updated in real-time with new purchase transaction entries. Going one step before feature engineering, the incoming data must go through a predefined series of logical transformations and enrichments (i.e., pre-processing stage). This a necessary stage that must be executed each time along with the main churn risk estimation algorithm.

To gain insights into the purchasing behavior of subscription commerce customers, we focused on the time aspect of the purchase transactions. Based on that, we processed the aforementioned type of data and extracted features that characterize a typical customer of the above businesses. For instance, we take into account the number of days lapsed from the last purchase of a specific product by a known customer. The full list of the engineered features that are utilized to feed the churn risk estimation algorithm are presented in Table 2.

Table 2. Features to feed the churn risk estimation algorithm.

Feature	Explanation	Type
Last Order	Date of the most recent transaction of each customer	Date
Minimum Consumption Duration	Minimum days among two consecutive transactions per product unit	Numerical
Average Consumption Duration	Average days among two consecutive transactions per product unit	Numerical
Maximum Consumption Duration	Maximum days among two consecutive transactions per product unit	Numerical
Average Fluctuation	Standard deviation of the days among two consecutive transactions per product unit	Numerical

C. Preserving data privacy

No processing of consumers' personal data is happening on the platform without the data subject's consent. The way that this consent is obtained depends on the data privacy and marketing communication policies of the data owner, i.e. the subscription commerce business which maintains the direct relationship with the consumer. Therefore, consent is managed outside the whole PRIME platform.

Inside the platform, there is a number of data management services which are built to support GDPR-specific provisions [40], such as the right of access (i.e., data portability) and the right to erasure of the data subject. These terms describe the right of a customer, as a data subject, to receive a copy of all the information held by the PRIME platform regarding his/her own activity, as well as to request the permanent removal of such information.

Additionally, the platform encrypts sensitive data and protects personal identifiers through mappings to second-level identifiers which are physically segregated. No single processing node in the platform architecture can ever have read access to the full set of data on an individual. Personal identity information and customer profiles are kept separate. For example, customer's name and address are kept separate from that customer's purchasing time and purchased product information.

5. Customer Churn Risk Estimation

A. Definition of repeat customers

Accurate churn prediction is only feasible for customers whose transaction history provides a sufficient amount of data points, i.e. repeat customers who make regular and consistent purchases. Customers with a high degree of consistency in their purchasing behavior allow for the highest accuracy in behavior prediction. Consistency of the time that lapses between consecutive purchases is a very important factor. The respective metric is described as "Average Fluctuation" in Table 2.

To answer the question of what makes a repeat customer, we utilized anonymized customer purchase transaction data collected from the internal information systems of the investigated business and focused on a specific product. To quantify the definition we built a logistic regression model to predict the next purchase transaction date and captured its accuracy against six sets of customers of the pilot company. The customers were segmented with respect to different combinations of values for: (a) total number of product purchases; (ii) number of last months to take into account; and (c) time consistency between the customer's repurchases. The values for these three metrics can be seen in Table 3 under "Transactions", "Last Months" and "Average Fluctuation", respectively.

Table 3. Performance investigation of several repeat customers' definitions.

Transactions	Last Months	Average Fluctuation	MAE	R^2
≥ 2	2	< 15	4.48	0.37
≥ 2	2	< 20	4.41	0.45
≥ 3	2	< 15	5.16	0.39
≥ 3	2	< 20	4.88	0.43
≥ 3	3	< 15	5.05	0.4
≥ 3	3	< 20	4.91	0.42

The definition of repeat customer that we ended up with, for the business domain of the particular pilot company, refers to customers who: (a) made at least 2 purchases of a specific product in the last 2 months; and (b) had less than 20 days of average fluctuation, ever since the first time they bought the product.

As shown in Table 3, we measured the predictive accuracy of the utilized model on the aforementioned business scenario, based on the below metrics [41]:

- *Mean Absolute Error (MAE):* The average of the absolute errors (1). The MAE units are the same as the predicted target, which is useful for understanding whether the size of the error is of concern or not being robust to outliers. The smaller the MAE is, the better the algorithm's performance [42]. In (1), N is the total number of errors and $|x_i - x|$ equals the absolute errors.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - x|. \quad (1)$$

- R^2 Score: The percentage of variation in the dependent variable, which is predictable by the independent variable. The R^2 value varies between 0 and 1, where 0 represents no correlation between the predicted and the actual value and 1 represents complete correlation [43].

As it can be seen from Table 3, the definition of the smallest MAE and, at the same time, the highest R^2 score refers to the customers with consistency days between their purchases of less than 20 (i.e., days of affinity among purchases), and who have made at least 2 purchases in the last 2 months.

B. Algorithm design & deployment

The customer risk factor is a percentage that determines how likely a repeat customer is to churn during a certain period of time. This is a dynamically updated value as long as new transactions are continuously being made, affecting the parameters of the algorithm over time.

Purchase transaction data from subscription commerce businesses is usually of large scale, since that kind of businesses are often acquiring more and more detailed data per customer over time, building a continuously growing profile. Thus, processes like conditional filtering on big data structures is a costly computation process that involves a large amount of data [44]. To address this challenge we applied parallel data processing using the Apache Spark framework [45].

Based on the customer's purchasing profile, the current study introduces relevant benchmark features that pertain to the duration between important moments of the customer's purchasing behavior. For example, the research takes into account the average number of days between a customer's purchases and the maximum number of days elapsed between two consecutive purchases of a given customer.

Table 2 contains the complete list of the features that are utilized by the prototype algorithm in order to capture the risk of churn of each regular customer in specific periods of time. According to the business rationale that is followed by this paper, the extracted churn risk score is affected by a different combination of features that is unique for each time interval after the date of a customer's last order (i.e., scenarios).

As stated, the aforementioned purchasing behavior features help the proposed algorithm to adapt to each one of the intended scenarios. The mathematical formula that quantifies the risk of churn, is able to adapt to the special conditions of each scenario. The description of each possible situation, along with the associated risk estimation formulas, are organized as follows:

1. Zero risk: The period elapsed between the time of the customer's latest order and the current time is shorter than the customer's shortest previously recorded period between reorders (Fig. 1). For instance, let us assume that the customer's last order for a certain product was placed 15 days ago, while the shortest period of time which was ever recorded between two purchases of the same product by this customer was 20 days. In this case the customer is considered to be of zero risk (2).

$$Risk = 0 \quad (2)$$

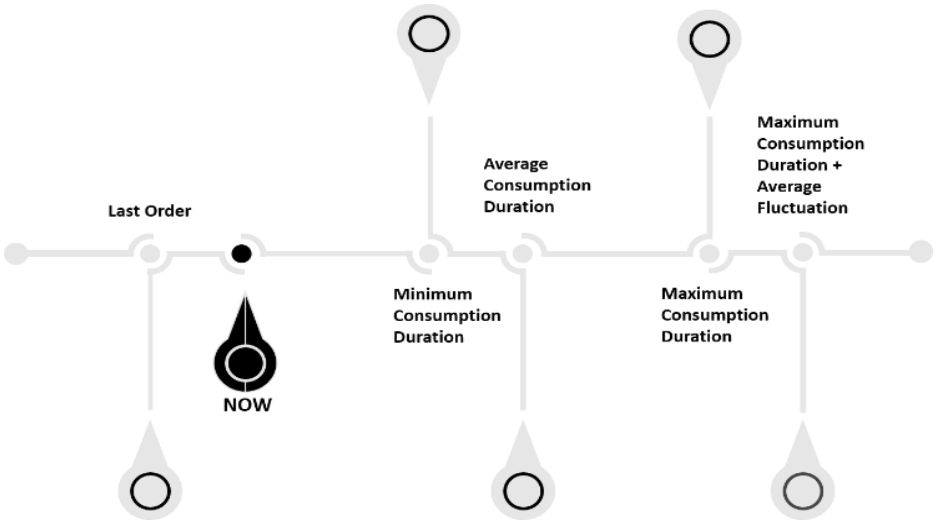

Fig. 1. Depiction of current situation in the first scenario.

2. Low risk: The period elapsed between the time of the customer's latest order and the current time is shorter than the customer's longest previously recorded period between reorders (Fig. 2). For instance, let us assume that the customer's last order for a certain product was placed 22 days ago, while the longest period of time which was ever recorded between two purchases of that same product by this customer was 25 days. In this scenario, the client has a minimal risk given by formula (3).

$$Risk = \frac{Max.consumptio\ n\ duration - Days\ from\ last\ order}{Max.consumptio\ n\ duration} \tag{3}$$

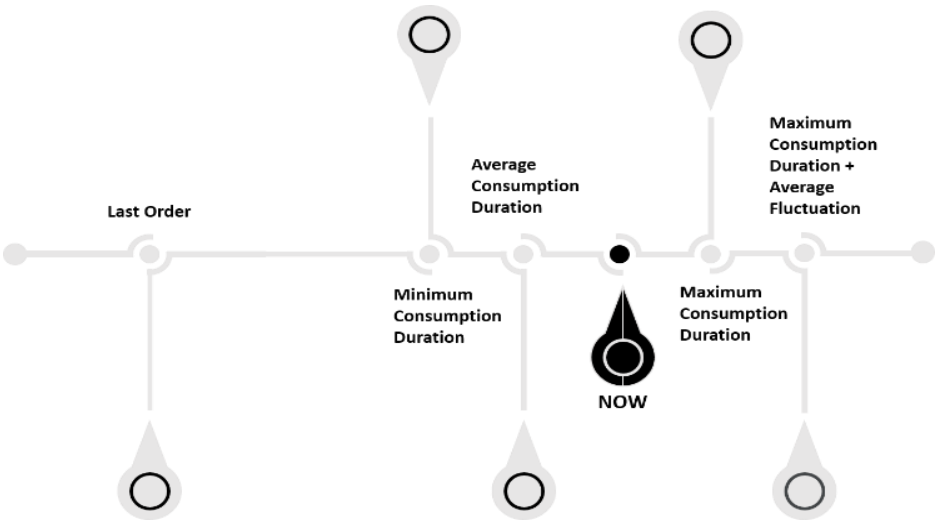

Fig. 2. Depiction of current situation in the second scenario.

3. Moderate risk: The days since the customer's last order are more than the number of days that have elapsed to complete its most delayed one. However, these days are less than the sum of the average deviation between the customer's purchases recorded so far and the maximum number of days elapsed between two consecutive purchases of the same customer (Fig. 3) For instance, let us assume that the customer's last order for a certain product was placed 28 days ago, while the longest period of time which was ever recorded between two purchases of that same product by this customer was 25 days and the aforementioned average deviation of that customer is 18 days. In this scenario, the customer has a measurable risk given by the formula (4).

$$Risk = \frac{Max.consumptio\ n\ duration + Avg.Fluctuatio\ n - Days\ from\ last\ order}{Avg.Fluctuatio\ n} \tag{4}$$

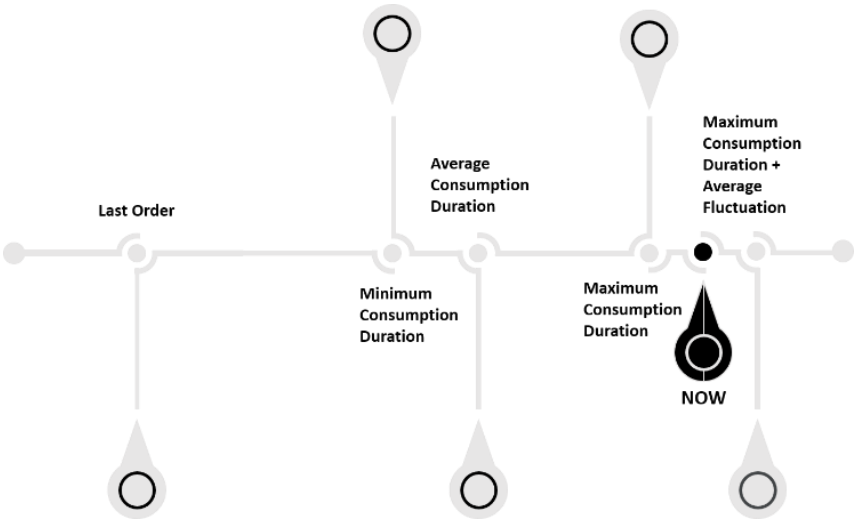

Fig. 3. Depiction of current situation in the third scenario.

4. High risk: The days since the last order of the customer are more than if we added the days that have elapsed to make his most delayed order to the respective average fluctuation (Fig. 4). For instance, let us assume that the customer's last order for a certain product was placed 45 days ago, while the longest period of time which was ever recorded between two purchases of that same product by this customer was 25 days and the aforementioned average deviation of that customer is 18 days. In this scenario, the customer has a significant risk given by the formula (5).

$$Risk = \frac{Days\ from\ last\ order - (Avg.Fluctuation + Max.consumption\ duration)}{Avg.Fluctuation} \quad (5)$$

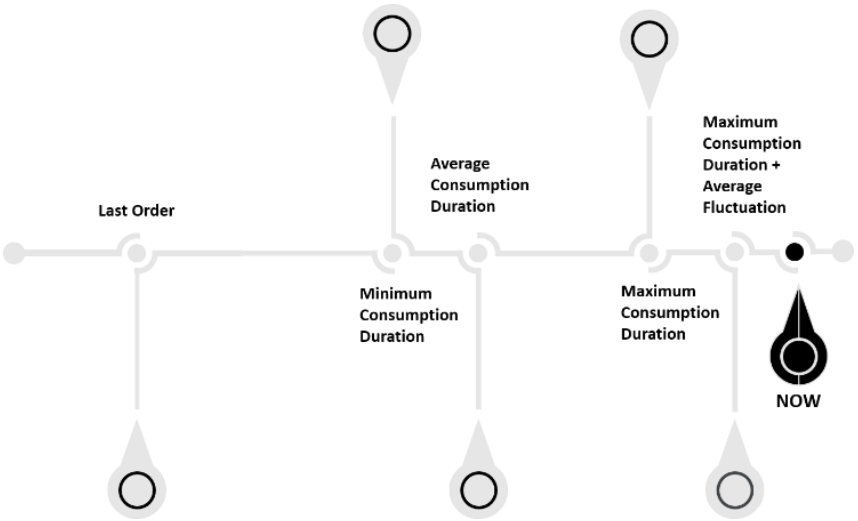

Fig. 4. Depiction of current situation in the fourth scenario.

5. Critical risk: In any of the aforementioned scenarios, when the churn risk score is (close to) 1, the customer is very likely to have already ceased purchasing from the business and is considered as "lost" (6):

$$Risk = 1 \quad (6)$$

A typical example of calculating a customer's churn risk score is the following. Let us assume that according to the given customer's profile, 26 days have passed since the customer's last transaction and the associated average consumption duration is 23.3 days. Moreover, the average fluctuation between two consecutive purchases of the same customer (i.e., consistency between purchases) is 10.8 days and the maximum consumption duration ever recorded between two of the customer's orders is 37 days. Based on the described algorithm's design, the above customer falls under the second category, and, therefore, the churn risk of the given customer is almost 0.3 or 30% for the given period of time.

Technically, the described estimation algorithm is being implemented as a scheduled Python Jupyter Notebook with an Apache Spark configuration (i.e., cluster of computers), storing the extracted values in a cloud-based database utilizing the IBM cloud infrastructure [46, 47, 48]. Figure 5 depicts an overall view of the way that the described components interact with each other during the algorithm's deployment to the cloud.

Fig. 5. Overview of the algorithm deployment architecture.

The aforementioned code is therefore regularly updating the churn risk score of each customer. The scheduling strategy is unique for each one of the businesses (i.e., customers of Apifon), since it is affected by the number and the frequency of purchase transactions that are being recorded for the particular client. Based on this score, any of the personalization services in the PRIME platform, which has been authorized to access the risk indicator for a particular customer, can use it to perform targeted retention actions.

6. Evaluation

A. Pilot test scope

As highlighted in the sections above, the research presented in this paper addresses the challenge of maintaining a dynamically updated customer churn risk indicator. Especially for subscription commerce businesses, gaining knowledge about the probability of a customer to churn at the given time, helps marketers to proactively plan effective retention actions.

The objective is to efficiently categorize the regular customers of an e-commerce business into churn risk groups based on their projected level of uncertainty - from the least possible to cease purchasing to the most possible to do so. The ultimate goal is to optimize the way the marketer contacts each group of customers through targeted messages in terms of both timing and content, as well as to gain insights on the overall satisfaction level of their business at a given time.

To evaluate the effectiveness of our approach we conducted a pilot test with one of Apifon's clients, a European subscription commerce business in the market of baby products. This is a company that sells direct to consumers. Its revenues and profitability depend on its customers making regular product repurchases. The company sends regular reminders to its subscribed customers through text messaging to let them know that their next shipment of products is approaching.

The way that the company has been using reminder messages is as follows. Once a customer orders a product of a specific category a reminder is automatically scheduled for delivery to the customer after a fixed number of days. The number of days between a purchase and sending a reminder message is determined by the product type and is the same for all customers. The message is scheduled for delivery after taking the customer's explicit consent to receive such kind of messages.

B. Experiment design

Based on the test scope of the current research, we designed our evaluation approach so as to capture the purchase transaction behavior of repeat customers during the evaluation period after predicting the respective churn risk scores. More specifically, we defined fixed churn risk probability bins, dynamically allocating to them customers, taking into account the daily estimations of the introduced algorithm. The evaluation metric that was utilized to quantify the effectiveness of the algorithm, was the percentage of the repeat customers who eventually made a purchase during the

evaluation period per risk category (i.e., conversion rate). Good results were expected to show a decreasing trend in conversion rates for risk factor bins of greater scores.

To select the duration of the evaluation period, we measured the average time between 2 consecutive purchases per regular customer (i.e., average consumption duration) of the pilot subscription commerce business. The statistical analysis of transactional data of almost 2 years long revealed that the average consumption duration of the repeat customers – as defined in previous section – is 22 days.

Based on the available purchase transaction data of the same business, we analyzed the percentage of the customers who used to place an order within the average consumption duration period (i.e., conversion rate). The analysis revealed that almost 40% of the regular customers made at least one repurchase within 22 days after their last purchase. We also noticed that the average daily number of customers who fall under the given definition of repeat customers, is 20 customers.

Taking into account the above statistics, we conducted a two-tailed sequential likelihood ratio test in order to decide on the acceptable sample size of our experiment. In particular, given the baseline conversion rate of 40% and setting the minimum detectable relative change in conversion rate to be 15%, with statistical significance of 95% (p-value 0.05), the sample size must consist of 650 repeat customers.

Based on the daily number of regular customers and the required sample size for the test to be statistically significant, our evaluation followed a backward measurement approach for a period of one month during August 2019. According to that approach, we applied the estimation algorithm based on continuously updating purchase transaction data (day by day), simulating the process of a real-time data acquisition infrastructure. More specifically, we tracked the purchase transactions of totally 615 repeat customers meeting the following criteria: (a) they made at least 2 purchases of a specific product in the last 2 months; and (b) they also had less than 20 days of variance between their purchases of that product, ever since the first time they bought the product.

After the backward measurement was completed, we classified the participated customers based on their churn risk probability into 4 major classes. Each class corresponds to a fixed probability range. We considered scores between 0 and 0.5 as negligible and a score of 1 as lost customer. To define the bin lengths for the risk values between 0.5 and 1, we adopted a 2:3 rule (i.e., split into ranges of 0.5 to 0.7 and 0.7 to 1, respectively). We preferred to allocate more customers to the most dangerous zone in order the indicator to act as a proactive alert to the marketer's attention. Based on the above, our investigation works like an exploratory research imitating the rationale of unsupervised learning algorithms.

7. Results and Discussion

Based on the above evaluation approach, the aggregated results per churn risk range are shown in Table 4.

Table 4. Pilot test result per risk factor range.

Risk factor bin	Allocated customers	Customers purchased	Conversion rate
[0, 0.5]	596	438	73%
(0.5, 0.7]	12	8	67%
(0.7, 1)	3	1	33%
1	4	1	25%

The results suggest that the score that the proposed algorithm extracted for the regular customers who were tracked for the test, proved to indeed follow their actual purchasing intention. This is evident on each particular risk range estimation, since the values of the conversion rate are inversely proportional to the respective values of the churn risk probabilities.

More specifically, almost 75% of the regular customers with less than 0.5 probability of churn, made at least one additional purchase, while only one customer considered as lost actually moved to a next transaction. Also, repeat customers with projected churn risk indicator between 0.5 and 0.7, proved to be hesitant to continue their orders since 67% out of them actually purchased. Finally, customers with estimated churn risk factor between 0.7 and 1 were very reluctant to buy more products since only 33% out of them did so.

Based on the detailed results, the algorithmic approach of the proactive estimation of customer churn introduced by the current research, allows businesses that offer consumer products and services to apply more efficient customer retention strategies. Thus, marketers can have a continuously updated churn risk score per customer, having the chance to perform the most appropriate targeted actions to retain several customer segments.

8. Conclusion

Modern marketing automation strategies are mostly focused on the efficient segmentation of repeat customers in order to contact them in a personalized manner. Subscription-based commerce businesses that are based on regular repurchasing model are proved to be profitable when they aim to retain their existing customers. Towards this direction, the most recent studies attempt to find leading indicators of customer churn applying several state-of-the-art machine learning models based on large-scale data.

We have found limited existing research on churn risk estimation approaches covering the following topics: (a) indicative properties of repeat buyers of e-commerce subscription-based products; (b) incorporation of common business rationale into the churn risk score prediction process in order to move from complex algorithms' black boxes to more interpretable solutions; (c) the aspect of purchase transaction sequence over time utilizing real-time data in order to take into account the status of a customer at the given time; and (d) effective categorization of customers into groups based on their estimated churn risk score in order to allow marketers to apply targeted proactive retention strategies.

The research analyzed in this paper focuses on the intersection of the aforementioned areas. More specifically, this paper presents a prototype algorithm that can estimate the probability of a known customer to cease purchasing from a subscription commerce business. This is a solution to the problem of planning consent-based, targeted marketing approaches to groups of repeat buyers who belong to a specific level of churn risk. On top of that, a notable contribution of the current paper is the investigation method followed to identify the repeat customers in the subscription commerce sector. Another one contribution is the engineering of features that come from real-time purchase transaction data helping to keep customers' estimations continuously updated. Moreover, the design of the introduced algorithm is based on human interpretable features. It is therefore an understandable process that can easily extend to even more retail sectors and is open to new optimizations.

The ultimate goal is to improve the way the marketer approaches customers in order to better reflect their satisfaction at the given period of time, leading to a better customer experience and, eventually, to increased conversion rate.

The evaluation of our proposed algorithm took place utilizing a backward measurement approach with actual customers of a European retail brand in the market of baby products. The results revealed that 75% of the low-risk customers actually bought again, while 25% of the customers who considered as lost, made an additional purchase. Moreover, 66% of the customers with considerable churn risk and 33% of the high-risk customers proceeded to a next purchase transaction. The above statistics suggest that our algorithm can capture the actual customers' intent to churn to a significant degree.

The results of the pilot test are encouraging although the experiment is limited in two ways. Firstly, the total sample size is not large enough in order to safely generalize our solution. Secondly, we used fixed risk segmentation ranges rather than picking out the most efficient one after experimental trials. It is understood that these factors could potentially be contributors to measurable errors. However, the research team has tried to mitigate them by careful sample design.

As part of future work, we plan to design a larger-scale experiment and apply several definitions of customer segmentation in order to find the most insightful one. Additionally, we aim to incorporate data from even more sources (e.g., customer relationship management tools) into our algorithm containing information related to customers' level of satisfaction, such as complaints, social comments and emotion icons. It would also be meaningful to apply association rules for new customers without adequate purchase transaction data to feed the algorithm (i.e., cold start problem).

Acknowledgment

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation under the project code *T1EDK-04550*.

References

[1] Huang, B., Kechadi, M. and Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), pp.1414-1425.

[2] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of marketing research, 43(2), 204-211.

[3] Bischof, S. F., Boettger, T. M., & Rudolph, T. (2019). Curated subscription commerce: A theoretical conceptualization. Journal of Retailing and Consumer Services, 101822.

[4] Van den Poel, D. and Larivi ère, B. (2004). Customer attrition analysis for financial servicesusing proportional hazard models. European Journal of Operational Research, 157(1),pp.196-217.

[5] Coussement, K. and Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers.Expert Systems with Applications, 36(3), pp. 6127-6134.
[6] Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (pp. 1-4). IEEE.
[7] Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), 2902-2917.
[8] Baumann, A., Lessmann, S., Coussement, K., & De Bock, K. W. (2015). Maximize What Matters: Predicting Customer Churn With Decision-Centric Ensemble Selection. In ECIS.
[9] Siber, R. (1997). Combating the churn phenomenon. Telecommunications, 31(10), 77–81.
[10] Deligiannis, A., Argyriou, C. & Kourtesis, D. (2019). Predictive personalization of conversational customer communications with data protection by design. IEEE/WIC/ACM International Conference on Web Intelligence on - WI '19 Companion.
[11] Deligiannis, A., Argyriou, C., & Kourtesis, D. (2020). Building a Cloud-based Regression Model to Predict Click-through Rate in Business Messaging Campaigns. International Journal of Modeling and Optimization, 10(1), 26-31. doi:10.7763/IJMO.2020.V10.742
[12] Gordini, N. and Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Industrial Marketing Management, 62, pp. 100-107.
[13] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.
[14] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 28.
[15] Wu, X., & Meng, S. (2016). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In 2016 13th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-5). IEEE.
[16] Berman, B. (2016). Referral marketing: Harnessing the power of your customers. Business Horizons, 59(1), pp. 19-28.
[17] Zhuang, Y. Y. (2018). Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm. Management Science and Engineering, 12 (3), 51-56.
[18] Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80-98.
[19] Cao, L. (2010). In-depth behavior understanding and use: The behavior informatics approach. Information Sciences, 180(17), pp.3067-3085.
[20] Wang, C. (2010). Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. Expert Systems with Applications, 37(12), pp.8395-8400.
[21] Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B. and Pentland, A. (2018). Behavioral attributes and financial churn prediction. EPJ Data Science, 7(1).
[22] Keramati, A. and Ardabili, S. (2011). Churn analysis for an Iranian mobile operator. Telecommunications Policy, 35(4), pp.344-356.
[23] Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. Journal of Organizational Computing and Electronic Commerce, 19(2), 133-151.
[24] Euler, T. (2005). Churn prediction in telecommunications using mining mart. Proceedings of the Workshop on Data Mining and Business (DMBiz) at the 9th European Conference on Principles and Practice in Knowledge Discovery in Databases (PKDD).
[25] Coussement, K. and Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), pp.313-327.
[26] J. Hadden, A. Tiwari, R. Roy, and D. Ruta. (2006). Churn prediction using complaints data. Proceedings of world academy of science, engineering, and technology, 13:158–163.
[27] Romov, P. and Sokolov, E. (2015). RecSys Challenge 2015: Ensemble Learning with Categorical Features. In Proceedings of the 2015 International ACM Recommender Systems Challenge (RecSys '15 Challenge). ACM, New York, NY, USA, Article 1, 4 pages.
[28] Sheil, H., Rana, O., & Reilly, R.G. (2018). Predicting Purchasing Intent: Automatic Feature Learning using Recurrent Neural Networks. ArXiv, abs/1807.08207.
[29] Toth, A., Tan L., Fabbrizio G., and Datta, A.. (2017). Predicting Shopping Behavior with Mixture of RNNs. In ACM SIGIR Forum. ACM.
[30] Wu C. Y., Ahmed, A., Beutel, A., Smola, A. J., and HowJing. (2017). Recurrent Recommender Networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17). ACM, New York, NY, USA, 495–503.
[31] D'Angelo, G., Rampone, S., & Palmieri, F. (2017). Developing a trust model for pervasive computing based on Apriori association rules learning and Bayesian classification. Soft Computing, 21(21), 6297-6315.
[32] Chiang D, Wang Y, Lee S, Lin C (2003) Goal-oriented sequential pattern for network banking and churn analysis. Expert Systems with Applications 25(3): 293–302.
[33] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert systems with applications, 38(3), 2354-2364.
[34] Huang Y, Huang B, Kechadi MT. (2011). A rule-based method for customer churn prediction in telecommunication services. Advances in knowledge discovery and data mining. Berlin Heidelberg: Springer. pp. 411–22.
[35] Zhang, Y., Wang, Y., He, C. and Yang, T. (2014). Modeling and Application Research on Customer Churn Warning System Based in Big Data Era. International Journal of Multimedia and Ubiquitous Engineering, 9(9), pp. 281-298.

[36] Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. Expert Systems with Applications, 83, 326-332.
[37] Deligiannis, A., Argyriou, C. and Kourtesis, D. (2020). Predicting the Optimal Date and Time to Send Personalized Marketing Messages to Repeat Buyers. *International Journal of Advanced Computer Science and Applications*, 11(4).
[38] Naik, N. (2017). Choice of effective messaging protocols for IoT systems: MQTT, CoAP, AMQP and HTTP. 2017 IEEE International Systems Engineering Symposium (ISSE).
[39] Dobbelaere, Philippe & Sheykh Esmaili, Kyumars. (2017). Kafka versus RabbitMQ.
[40] Lachaud, E. (2019). Adhering to GDPR Codes of Conduct: A Possible Option for SMEs to GDPR Certification. SSRN Electronic Journal.
[41] Botchkarev, A. (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, pp.045-076.
[42] Willmott, C. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, pp.79-82.
[43] Alexander, D., Tropsha, A. and Winkler, D. (2015). Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. Journal of Chemical Information and Modeling, 55(7), pp.1316-1322.
[44] Singh, P., Anand, S. and B., S. (2017). Big Data Analysis with Apache Spark. International Journal of Computer Applications, 175(5), pp.6-8.
[45] Olasehinde, O., Johnson, O. and Fakoya, J. (2018). Computational Efficiency Analysis of Customer Churn Prediction Using Spark and Caret Random Forest Classifier. Information and Knowledge Management. Department of Computer Science, The Federal Polytechnic, Ile-Oluji, Ondo State, Nigeria, 8(2), pp.8-16.
[46] Bozhinov, I. (2019). AI and big data on IBM Power Systems servers. U.S.A.: IBM Corporation.
[47] Palmer, T. (2019). Predict and Optimize Business Outcomes with IBM Decision Optimization for Watson Studio and IBM Cloud Pak for Data. The Enterprise Strategy Group, Inc., pp.3-14.
[48] Dhoolia, P., Chugh, P., Costa, P., Gantayat, N., Gupta, M., Kambhatla, N., Kumar, R., Mani,S., Mitra, P., Rogerson, C. and Saxena, M. (2017). A cognitive system for business andtechnical support: A case study. IBM Journal of Research and Development, 61(1), pp.7:74-7:85.

Authors' Profiles

Alexandros Deligiannis obtained his MSc. in Big Data Science from the Queen Mary University of London, UK and his BSc. in Mathematics from the Aristotle University of Thessaloniki, Greece. Currently, he is working as data engineer in the Research and Development department of Apifon, Greece in the field of personalized communication in order to increase sales and improve customer experience. His main research interests include model-driven systems engineering and ad-hoc process optimization.

Charalampos Argyriou obtained his BSc. in Applied Informatics from the University of Macedonia, Thessaloniki, Greece. Currently, he is working as a Research and Development engineer in Apifon, Greece. His main research interests include, amongst others, machine learning, data mining, natural language comprehension and information retrieval, especially in the field of customer experience management.

How to cite this paper: Alexandros Deligiannis, Charalampos Argyriou, " Designing a Real-Time Data-Driven Customer Churn Risk Indicator for Subscription Commerce", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.12, No.4, pp. 1-14, 2020. DOI: 10.5815/ijieeb.2020.04.01