[1]Hossam H. Ahmed
[2]Mohamed H. Khafagy
[3]Mostafa R. Kaseb

Model for Partial and Total Churn Prediction in E-Commerce

Journal of Electrical Systems

Abstract: - E-commerce is booming market now because every company are launching in this sector to serve customers and reach out to the variety of product from sitting at home without any hassle. But we all know that thanks to the plethora of available ecommerce sites customer can take their purchase from one site to another or even split what they intended on 1st place. While this is a trend now, it poses significant challenges for companies as getting new customers are even more expensive than retaining existing ones. This uses a customer churn prediction model on the e-commerce market. When it comes to customer churn, we are talking about the percentage of them that at a certain moment stop using your product or service. The model was trained on a large dataset from an B2C multi-category e-commerce application for customer behavior and interactions. Patterns are defined in the model to predict what type of customer churn may take place i.e. either total (customer stops using e- commerce site wholly) or partial(customer cuts back on purchases or becomes less profitable). The dynamic churn definition step will allow the model to capture both kinds of churn. In the first phase, we clarify whether a customer has churned or not using LRFM model and K-means algorithm. In the second part of our study, Behavioral and interactional used to build the prediction model based on XGBoost. The proposed model achieved 98% in detection accuracy by Albough-Partial+Total churn detect algorithm. The individual accuracy rates for the partial and total churn algorithm stand at 98% and 99%, respectively.

Keywords: *E-commerce; Churn Prediction; Customer Churn; LRFM; K-means; Behavior Analysis; XGBoost.*

I. INTRODUCTION

The e-commerce sector has grown exponentially in the last ten years as additional suppliers compete for entire market. Businesses in e-commerce are not like the traditional businesses, customers here aren't tied up with contracts that make them stay on website/app [1]. But with this expansion came new obstacles that demand a focused, strategic response: namely attrition rates - aka customer churn [2]. Churn (or attrition) Churning of customers may happen in totality or T. F.Reich argues increasing customer retention by 5% can cause profits by over 25% [3]. In addition to this the fact that customers are still able to shop across multiple sites make it challenging for e-commerce companies trying to keep their client pool.

Customer Churn Impact on E-commerce Companies - A Key Feature to Be Considered Churn is a big problem for these companies, as it can cause revenue and profits to fall. A loss of current revenue from the churned customer and also a lost potential for future revenues [4, 5]. In addition, it is reported that the cost of customer acquisition could be up to 25 times more than the amount of money needed for keeping your existing one[6], proving churn prevention as an essential strategy for ecommerce companies. Additionally, the reality is that there are a number of things at play in churn in e-commerce. For instance, customers may be unhappy with the quality of services or products that a specific site offers as individual needs to buy something from Home Depot online; they might feel uncomfortable browsing through it concretes jungle.

Data analytics and machine learning are become very powerful technique for detecting customers likely to churn [7, 8]. After examining how customers interact with an ecommerce site or app, companies can learn what is causing churn and begin to fix the problem. This could be alleviated with site design [9], pricing strategies, customer service processes or even supply chain management by online retailers. For continuous tracking and analysis of customer behavior, companies can understand where they need to improve on and make changes accordingly which further helps in reducing churn and boosting retention.

[1] *Corresponding author: Hossam H. Hassan, Computer Science Department, Faculty of Computers and Artificial Inelegance, Fayoum University, Fayoum 63514, Egypt.

[2] Computer Science Department, Faculty of Computers and Artificial Inelegance, Fayoum University, Fayoum 63514, Egypt;

[3] Computer Science Department, Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum 63514, Egypt;

Emails: [1] hh1279@fayoum.edu.eg ; [2] mhk00@alexu.edu.eg ; [3] mrk00@fayoum.edu.eg

Take Amazon as an instance. They employ customer feedback to spot items often brought back, and then collaborate with the producers of those products to better quality or design. This action has contributed in lessening returns and, consequently, elevating client contentment. In this way, Amazon is able to avoid buyers who are not satisfied with the product by bringing another one. Also, they do not ask for the first product to be returned because of this strategy [10]. As difficult as it is to gain additional customers, the hard work of retaining your existing customer base can be even more challenging. There is an ideal way to deal with this problem, and they are long-term approaches and short-term approaches. The long-term strategies will aim to optimize the e-commerce site or application, service quality by collecting data as described above and an analysis of these activities with recommendations from experts on finances, both in system churn reduction (churn) and additional sales revenues per lead generated because consultants tend to see needs for evolving products' presentation leading marketing effort according website economically exploited. Short-term approaches, in the meantime, aim to detect customer segments which are at risk of discontinuing [11, 12] or attribute behaviors contributing these profiles for those targeted customers before they actually churned out by detecting behavior attributes that create such patterns and directly retarget them accordingly [13-15].

Previous research has taken the short-term timelines to be faster than long-terms one, a useful consideration if we are interested in saving customers from churning [16], data is collected during sessions through interactions of users with an application and its processing allows detecting the point at which they start terminating their engagement with e-commerce website or applications linking this moment for targeted intervention regarding retention strategies as well preventing defection [[12–17]. As far as we know, this is one of the first papers to model partial churn in a comprehensive view. But unlike previous approaches which rely on a specific time range (e.g. no purchase in the last X days), this new method factors in all of an individual user's unique engagement patterns.

This is also important because using only a fixed timeframe to calculate churn may not deliver an accurate impression of what your customers are doing, particularly if you're in a recurrent business (like e-commerce houses) or sells with longer purchase cycles. In addition, this new model uses session behavioral data to more accurately identify each type of churn. Our model solves these problems, thereby increasing churn prediction accuracy as well the user behavioral nuances.

The approach consists of two primary stages. Initially, we identify patterns for two categories of customers: those who have partially churned and those who have already churned. We monitor customers over a period of time to observe changes in their behavior by clustering them in two different groups using the LRFM model [18, 19] and the K-means algorithm. The integration between these two methods enhances the efficiency of customer clusterization through K-means running on LRFM parameters derived from 16 properties based on average business value— thereby optimized by the K-means clustering number.The model is composed of dual primary phases. Initially, we embark on delineating what we shall christen the churning definition. This entails setting up the templates for those customers who have partially churned as well as their already churned counterparts. We take a longitudinal approach in tracking our clientele and endeavor to unearth how their behavioral patterns metamorphose over time. This is achieved through a clustering exercise conducted in two different temporal epochs. The first involves the use of the LRFM model (which is a technique harnessed in customer value analysis) supplemented with inputs from previous studies [18, 19].

The second period employs the K-means algorithm whose primary task is grouping similar data points together based on pre-defined parameters like proximity or similarity. An intriguing facet that emanates from these disparate methodologies lies in their convergence point: an integration mechanism. K-means is made to run on LRFM parameters— a deliberate effort aimed at enhancing efficiency levels within our customer clusters. These are generated by the LRFM model and consist of 16 properties for each parameter (more or less than average value). Business derives this uniqueness from K-means optimization through the clusters count it should form, thus ensuring optimal resource allocation strategies are put into place. The next step would involve extraction of features that depict client behavior in the light of customer sessions for predictive analysis of churners. At this juncture, we make use of XGBoost which is a machine learning algorithm to predict the pattern of customers. The phase involves applying XGBoost as a predictive tool for identifying patterns among clients during the development lifecycle.

The model aim at answering a straightforward question: what is the estimation of customer status (non-churned, total churned, or partially churned)? The text that follows describes about the target.

A model like the one we have suggested is quite revolutionary for e-commerce firms. We define churn based on business data, which means our model can be applicable at any point of the business life cycle. The development of our model relies on customer behavioral data, typically sourced from user application interactions— this information allows us to identify two types of churns. All major contributions of our model are discussed here: Let us build a single model which can forecast both total and partial churn and thus assist the e-commerce firms in establishing an unwavering marketing plan. Let's focus on developing a scalable model that is adaptable to various types of ecommerce businesses; this is because we initiated a churn definition stage that can delineate the customer churn prediction process. Emphasize profitable customers during the formulation of the model, with an aim to ensure its effectiveness. We will evaluate the model on real-world multi-category e-commerce scenarios, allowing us to validate its practical applicability.

The architecture of this paper is described as follows: Section II introduces the Related work. Section III presents details of the construction data and model description, as well as test results for comparison with existing studies. Section IV summarizes the findings of the discussion. Section V draws conclusions from the findings. Section VI provides an appendix that includes the calculation of properties of attributes. Last Section provides references for further reading..

II. RELATED WORK

Customer churn has been a hot research topic for decades. Some work with full churn, while others operate with partial churn.

X. Li and Z. Li [20]proposed a hybrid estimation model for customer transactions in e-commerce X. Li, model uses logistic regression and Extreme Gradient Boosting (XGBoost) algorithms to detect customers who intentionally move through an e-commerce website or app. These models are trained using order data, customer profiles, preferences and post-sale events, and these measurements are the focus of this study. The disadvantage here is that the research only considers all losers and the context of loss is limited to a specific time and situation, customers who do not create orders for two months, regardless of the type of business. Logistic regression is a method for binary distribution problems and can be used to identify factors that cause customer churn. The hybrid model achieved an accuracy rate of 76.6%.

Xiahou and Y. Harada [21] proposed a B2C e-commerce customer churn prediction model based on the combination of k-means customer segmentation and support vector machine (SVM). The aim of the model is to predict which customers should not shop from the e-commerce site in the future and to identify the factors that cause customer churn. The main aim of this study is to add groups before predictions, thus improving long-term results. This indicates that the model takes into account not only the customer's current behavior but also their behavior over time and the many changes that may affect their behavior (night shopping, night shopping, night browsing). This approach is important because consumer behavior can be complex and affected by many factors. Further analysis of consumer behavior can lead to better predictions, which combine behavioral patterns over time. K-means is used to separate customers into various groups based on their purchasing behavior. This helps identify patterns and trends in customer behavior over time, as well as identify customers who are expected to leave. This is a high accuracy compared to the past because they are trying to solve the entire problem in e-commerce so that the model can predict customers' behavior well The concept variables applied in the model include various consumer behaviors such as shopping frequency. Purchase, purchase cost, purchase type and purchase time. These changes are vital because they help determine whether customers will leave by providing information about customer behavior and preferences. The model is limited to all churn issues and ignores behavioral data such as partial churn, session length, and time between orders, as well as additional changes to carts and purchases.

P.Berger and M. Kompan [22] proposed a model to predict user churn events based on user behavior of a web application using a vector machine (SVM) with a basis value cost (RBF) kernel. This model uses a combination of special techniques that allow users to interact with the web application during the session. In this case, the S VM algorithm is used to predict how the user will continue to use the website. The RBF kernel is a powerful dat

a transformation that makes algorithms more manageable, especially when dealing with nonlinear relationships. This process includes information about the user's interaction with the application, such as the number of sessi ons created, time spent in each session, and actions. Other features include measuring the profile. For examp le, if users spend too little time on an app, it may mean that they are not interested in the app and should stop usi ng it. On the other hand, if users spend more time in an app and visit more pages, it will show that they are intere sted in the app and will continue to use it. The model achieved an 84% accuracy rate. By using SVM with a com bination of RBF kernels and feature sets, the model can accurately predict whether the user will continue to use t he application. The main advantage of this work is that it is not used to predict the behavior, but it also has disa dvantages such as:(1) Since it is valid for all users who make more than 4 types of applications, what needs to be done is a lot of data, so a large amount of data needs to be done. (3) This study only focused on all driving f actors.

V. L. Migueis and D. Van den Poel [23]. A model is proposed to predict volatility in the grocery industry using logistic regression. The model is based on RFM (Recency, Frequency and Monetary) behavioral data, which is a way to analyze customer behavior based on their recent visits, frequency, purchases and cost per purchase. The highest accuracy achieved by this model is 87.42%.

This model developed a method to measure customer churn, identify different types of churn, segment customers into groups based on business and customer behavior, and investigate churn patterns by tracking customers on time. Second, the current approach focuses on all customers, regardless of their wealth. This is ineffective as it can be costly to make bulk products for unprofitable customers; we only focus on core and new customers

III. METHODOLOGY

In this study, a well-structured two-stage method was employed to tackle the issue of customer churn in the e-commerce sector. The first stage referred to churn definition. Specifically, this phase is about how to spare the data set to make it viable for further use, extract the most relevant customer attributes, divide the churns into classes of total, partial, and non-churn with the k-mean cluster, and then label the data that shown in Figure 1. The second stage referred to churn prediction. This stage concerns implementable outputs of the first one by preprocessing the data again, feature extraction and organization of features into a feature matrix, and employing the xyboost algorithm to produce a powerful forecaster capable of anticipating the customer churn status virtually that shown in Figure 1. It may better understand for e-commerce businesses and address customer attrition through this systematic and comprehensive approach, and the hardware specifications utilized are listed in Table 1.

Table 1: Hardware Specifications.

Table 1.	**Machine hardware specifications**
CPU	Intel® Core™ i7-6820HQ CPU @ 2.70GHz × 8
GPU	Intel® HD Graphics 530 (SKL GT2)
RAM	8 GB
CUDA Version	9.1
OS	Ubuntu 18.04

Fig. 1: The main steps of this research

A. Data Pre-processing

The research process began with a crucial data preprocessing phase which ultimately set the stage for its analysis. This first stage was to clean the data for further analysis. Including cleaning the data, and identifying and remove any outliers or missing values, as well few transforms into a neat format for the subsequent step which is feature extraction.

B. Feature Extraction

Feature Extraction Steps Involve the process of getting raw data ready for examination. In this scenario, the research depended on data collected from a variety of online shopping platforms, indicating a wide range of factors and data points to take into account. This data was utilized to identify statistical and behavioral characteristics, which might encompass aspects such as buying history and how customers browse.

Length-Recency-Frequency-Monetary (LRFM) Analysis LRFM is a technique for segmenting customers in marketing and customer relationship management (CRM) to sort them based on their buying patterns. The LRFM analysis looks at four main indicators to assess a customer's value:

- Length: the duration since a customer's initial purchase.
- Recency: the interval since a customer's last purchase.
- Frequency: the rate at which a customer makes purchases.
- Monetary: the total amount spent by the customer.

Each of these factors is given a score according to set standards (for example, purchase frequency limits, time since last purchase limits, etc.). Customers are then categorized into groups based on their scores across these four factors. Typical categorizations include:

- Loyal Customers: Have a long history, frequent purchases, and spend a lot.
- High-Potential Customers: Make purchases often and spend a lot, but are relatively new.
- Churn Risk Customers: Make purchases less often and spend less, with longer gaps between purchases.
- Value Customers: Spend a lot but make purchases less frequently.

C. Churn Definition

Typically, churns are categorized into two main types: total churn and partial churn. Most studies concentrate on total churn, which looks at the likelihood of a user returning to use a service. However, this research considers both total and partial churn. Rather than setting churn as a point in time when a customer stops ordering, the research acknowledges that the duration of not ordering can vary across different businesses. Consequently, the research aims to develop a model capable of identifying churners based on the nature of the business and the actions of other users. This approach is designed to pinpoint the most valuable customers who are central to the business and those who might be at risk of leaving. By achieving this, the research intends to assist businesses in keeping their valuable customers.

Keeping valuable customers is essential for companies as they are a major source of income. Losing these customers can adversely affect a company's earnings and expansion. Thus, by identifying churners and helping businesses keep them, the research seeks to support companies in sustaining or enhancing their income and securing their future.

This foundational stage, comprising the data preprocessing and feature extraction steps detailed in Sections III.A and III.B, laid the groundwork for the subsequent stages of the research process. Having established a clean and well-structured dataset with the relevant customer attributes, the study then moved on to the next phases, as elaborated upon in the following sections as shown in figure 2.

Fig. 2: Churn Definition

1) K-Means and Customer Clustering

This study aims to gain a better understanding of customer behavior and preferences by using the LRFM method. These parameters will then be combined with the k-means algorithm, a common clustering algorithm

used in machine learning. The goal is to replace static LRFM thresholds with the dynamic k-means algorithm, which can work on different commerce types and provide a dynamic churn definition for both partial and total churn types. By using these methods, patterns in customer behavior can be identified, and it may be possible to predict which customers are at risk of churning (i.e. discontinuing their relationship with the e-commerce platform). In addition to the LRFM parameters, the study also looks at behavioral parameters that are generated from customer session data. This refers to information on how customers interact with the e-commerce application during their online sessions.

This study notes that such data is commonly available in online merchants' reactive web interaction logs. This study extracted five sets of attributes from this data to gain a better understanding of customer behavior and engagement with the e-commerce platform, these attributes could include session data, purchasing history, customer behavior, customer interactions, and type of actions.

The k-means method combined with the LRFM model to segment the e-commerce customers into homogeneous clusters based on their length, recency, frequency, and monetary [25].

The Clustering technique K-means [26] is the most popular method of classifying n vectors into k partitions based on characteristics. The procedure begins by selecting k randomly chosen initial centroids, then uses Euclidean distance to assign vectors to the closest centroid, and finally recalculates the new centroids using supplied data vectors. Repeat this procedure until the vectors no longer change the clustering between iterations. The challenge here is how to select the correct number of clusters.

In this study, we used the sum of squared errors (SSE) method to optimize a criterion for the accuracy of the k-means when the data points were similar in magnitude. The lower the SSE, the more accurate the number of clusters (n).

The k-means technique is applied to different numbers of clusters (k), then the elbow is drowned with the SSE against the number of clusters, and the optimal number at which a knee exists in the elbow is selected [27].

In line with H. Chang and S. Tsay [25], we will utilize the Average LRFM metrics for each cluster as a basis for comparison against the overall Average LRFM metrics of all clusters. When the average (L, R, F, M) value of a cluster exceeds the total average, it will be denoted with an over bar. Conversely, if the average (L, R, F, M) value of a cluster falls below the total average, it will be denoted with an under bar. For instance, a higher R-value indicates that a customer has recently made a purchase, while a lower R-value suggests that the customer has not made a purchase on the online store for an extended period. The target of this research was to determine the main segments of the customer in e-commerce to enable the business to determine the best way to treat the segment and which one is more profitable and limit the number of segments.

H. Chang and S. Tsay [25], summarized the customer segments into sixteen combinations of LRFM, based on every attribute higher or lower than the LRFM total average, to five categories of customers: Core, Potential, Lost, New, and Resource consumption customers and described these categories with the patterns shown in Table 2.

Table 2: Category Definition based on LRFM

Category	Patterns
Core	**L R F M, L R F M, L R F M**
Potential	**L R F M, L R F M, L R F M**
Lost	**L R F M, L R F M, L R F M, L R F M**
New	**L R F M, L R F M, L R F M, LR F M**

Resource consumption	**L R F M, L R F M**

To detect churn types, customer clustering was performed on two different time frames (T and T+1), and churning in customer clusters was observed based on customer changes from one cluster in T to another cluster in T+1, if a customer moved from the core cluster in T to a potential customer or low resource consumption in T+1, then it's a partial churn; if the customer changed to a lost customer or high resource consumption, then it's a total churn. If customers change from a new cluster to a potential cluster or have low resource consumption, then it's a partial churn, but if they change to high resource consumption or lost clusters, then it's a total churn. This study focuses on core and new customers because those are the main profitable categories. Table 3 shows the churn status based on the changes in customer clusters between the two time frame clusters.

Table 3: Churn Status Based on Customer Behavior Changes

Churn Type	T Clusters	T+1 Clusters
Partial	**Core**	**Potential, Low resource consumption**
	New	
Total	**Core**	**Lost, High resource consumption**

1) Labeled Data

Following the data clustering stage, where customers were segmented into distinct groups employing techniques like k-means, the researchers moved on to label the data based on the clustering process's output. Specifically, customers were categorized as either partial churn, total churn, or non-churn, offering a clear and actionable classification of various types of customer attrition within the e-commerce environment.

The data after applying K-mean cluster algorithm are shown in figure 3.

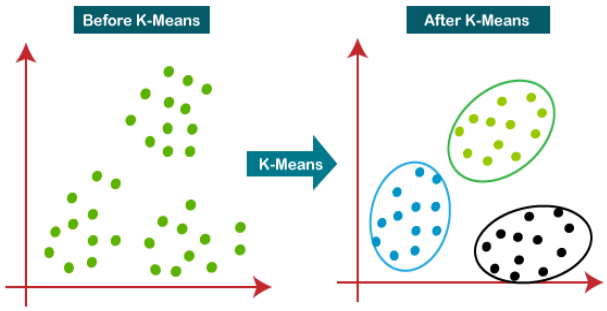

Fig. 3: The data before and after applying K-mean algorithm

D. Churn Prediction

The output of the churn definition stage is the input of the churn prediction where Multi-class prediction is used to detect the class of customers [non-churned, total churned, and partial churned], the preprocessing and Feature Extraction that explained in III.A. and III.B sections as shown in figure 4. The outputs of these stages act as an input to Extreme gradient boosting [XGBoost] classifier based on behavior data that was extracted from an online store after merging the result of the churn definition stage with session customer data. Session customer data gives insights into engagement levels and potential disengagement signals, purchasing data that are indicators of satisfaction, loyalty, and potential shifts in buying patterns, behavior changing data to give signals of evolving interests or dissatisfaction, and interaction with the application to measure engagement level and action made on the session to measure potential roadblocks. These attributes are chosen because they enable us to focus on churn behavior and have a comprehensive view of the customer interaction and its nuances. All

these behavior data in Table 4 extracted from customer sessions are the inputs to the XGBoost classifier. We display the equations for every attribute from Table 4 in the Appendix. Customer churn status can be identified after the end of the session. A higher number of sessions for the customer mean greater accuracy in the prediction.

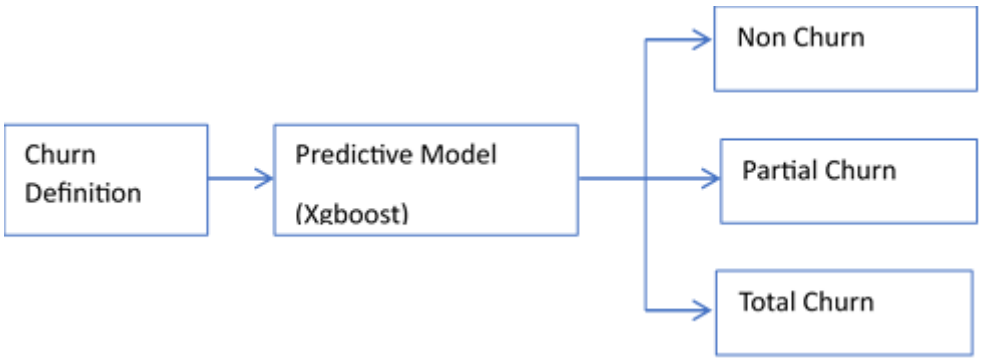

Figure 4: Churn Prediction

Table 4: Behavior data parameters used for prediction

Attribute	Description
Session Order	Number of sessions created by the customer
Last Session Time	Last session duration
No. of Actions Within Session	Number of actions taken on the last session
Average Time on Action	Average Time taken to do an action
Day of The Week	Last session day of the week
Day of The Month	Last session day of the month
Weekend	Is the last session done on the weekend?
Number of Purchases	How many sessions did the purchase event have?
Total Sum Paid	How much money was paid by the customer?
Last Sum Paid	How much money was paid on the last purchase event?
Session to Purchase Ratio	Percentage of number of sessions compared by the number of sessions that had purchase event
Add to Basket to Purchase Ratio	Percentage of number of sessions that had added to basket event compared by the number of sessions that had purchase event
Add to Basket Since the Last Purchase	Number of sessions that had added to the basket event since the last purchase event
Session Length Change	Difference between the last session duration and the average session duration
Number of Session Actions Change	Difference between the last session's number of actions and the average session number of actions
Session Gap Change	The difference between the last session duration and the user's average session duration
Number of Sessions Since the Last Purchase	Number of sessions created since the last session that had purchase event
Time Since the Last Purchase	The time in seconds between the user's first action of the actual session and the last action of his/her most recent session with the purchase event

Time Since the Last Visit	The time in seconds between the last action in the user's previous session and his/her first action in the actual session
Add to Basket	A binary flag referring to whether the-add-to-basket event is present in the actual session
Purchase	A binary flag referring to whether the purchase event is present in the actual session

IV. EXPERIMENTAL RESULTS

The present research provides findings for three cases of churn prediction: the initial case is Partial churn, the second case is Total churn, and the third case is Total and Partial churn. All of these cases use the k-mean for clustering and Xgboost for the prediction model.

A. Datasets

The model assessment involved the use of data from multi-category e-commerce platforms. Following this, we split the data using the train_test_split method, allocating 75% for model training and 25% for testing. The dataset comprises behavioral data for November 2018 and October 2019, covering a period of two months.

This compilation comprises 277,130 active users obtained from extensive multi-category online e-commerce [28]. For a comprehensive understanding of the data points incorporated in this research, please consult Table 5, where each attribute is clearly specified.

This dataset was used to extract behavioral attributes that were used in churn definition and prediction. We have relied on this dataset as it's organized and comes from real multi-category e-commerce with important info like the session of each user, we use it to extract the behavior parameters as described in Table 3. This data set is relatively large as it has 14.68 GB of transactions.

Table 5: Dataset attributes

Column	Type	Values
event_time	Date Time	Event Time in UTC
event_type	Selection	View, cart, purchase, remove_from_cart
product_id	Many2one	Product ID
category_id	Many2one	Product category ID
category_code	String	Meaningful name
Brand	String	Brand name
Price	Float	Product price
user_id	Many2one	Permanent user ID
user_session	String	User session ID

B. Comparative Analysis of Findings

1) Partial churn

We aimed to evaluate the performance of our outcome against other approaches that have been created to forecast partial churn. In order to accomplish this, we examined various cutting-edge research papers on the topic and contrasted our precision with theirs. The findings of this comparison can be seen in Figure 5.

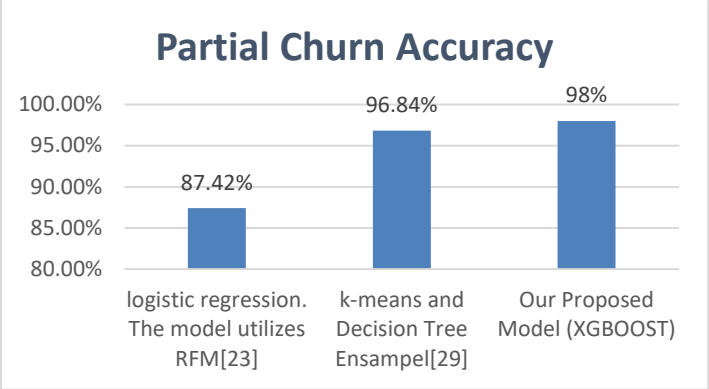

Figure 5: Partial Churn Accuracy Comparison

The model by V. L. Miguéis and D. Van den Poel [23] aims to predict partial churn in the grocery sector using logistic regression. The model utilizes RFM behavior data to analyze customer behavior based on recency, frequency, and monetary factors. Logistic regression is used to predict which customers are likely to partially churn and divide their purchases among different companies. The model achieved an accuracy of 87.42%.

The framework by, A.D. Rachid, A. Abdellah, B. Belaid, and L. Rachid [29], utilizes clustering algorithms and predictive models to identify patterns and predict customer defection. The study utilizes real-world data from an e-commerce platform to test the proposed framework. The authors use the k-means clustering algorithm to group customers based on their behavior and purchase history. A decision tree model is then used to predict the likelihood of customer defection. It's observed that identical parameters were employed for both churn definition and churn prediction phases. However, it is recommended that behavioral data be utilized during the churn prediction phase to identify churn patterns among customers.

Our model uses the LRFM and k-means clustering algorithm and machine learning to group customers based on their behavior and purchase history and machine learning to identify patterns and predict customer defection that achieves an accuracy of 98% for partial churn detection.

Our Model achieved better accuracy compared with the others in Figure 2, as we depend on a dynamic churn definition process that defines the churn based on the behavior of the customers not on a static parameter like V. L. Miguéis and D. Van den Poel [23] which depends on the static percentage of customer spent from one quarter to another if the customer reduces his spent less than 40%, he marked as partially churned, but on our model the customer marked as partially churned if he moved from higher cluster to lower cluster as mentioned on table 2. A.D. Rachid, A. Abdellah, B. Belaid, and L. Rachid [29] avoided the issue of static parameters but relied on a small number of behavior data. We avoided this by depending on more behavior parameters like session information, purchase information, behavior information, and interaction information as mentioned in Table 3. We used one data set in this experiment and achieved better accuracy. Theoretically, we think it will achieve the same on any e-commerce data set because the process of churn definition is dynamic and can detect the churn types based on the e-commerce status and is not related to static parameters.

2) Total Churn

We wanted to see how our result compares to other methods that have been developed to predict total churn. To do this, we looked at several state-of-the-art papers on the subject and compared our accuracy to theirs. The results of this comparison are shown in Figure 6.

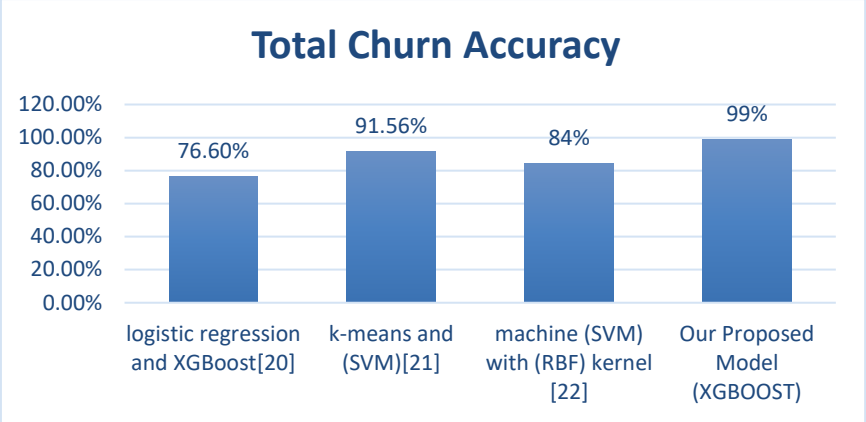

Fig. 6: Total Churn Accuracy Comparison

X. Li and Z. Li [20], proposed a hybrid prediction model for customer churn in the e-commerce industry. The model uses logistic regression and XGBoost algorithms to detect customers who are likely to churn. The model is trained using order information, customer profiles, preferences, and after-sales scenarios.

X. Xiahou and Y. Harada [21], proposed a loss prediction model for B2C e-commerce customers. The model uses k-means customer segmentation and support vector machines (SVM) to predict which customers are likely to stop shopping with the e-commerce site in the future. The model considers not only a customer's current behavior but also their behavior over time, as well as multiple variables that can influence their behavior.

P. Berger and M. Kompan [22], use a combination of feature sets to predict user churn. The feature sets include data on the user's interactions with the application, such as the number of sessions created, the time spent on each session, and the actions taken. The model also includes rating data. By combining these feature sets, the model can identify patterns in user behavior that indicate whether a user is likely to churn. The model uses a support vector machine (SVM) with a radial basis function (RBF) kernel.

Our model uses the LRFM and k-means clustering algorithm and machine learning to group customers based on their behavior and purchase history and machine learning to identify patterns and predict customer defection that achieves an accuracy of 99% for total churn detection.

In the case of total churn our model still gives better accuracy than X. Li and Z. Li [20], P. Berger and M. Kompan [22], which use a static number of months as a churn indicator; if the user hasn't created a purchase order within the defined period, he is marked as churned. In the case of X. Li and Z. Li [20], they use two months as the static period. Berger Patrik, and Michal Kompan [22] used one month. . This led us to the next note, namely, processing the whole data of the e-commerce users and not focus on the profitable customers. Our model solves these notes by measuring the churn definition based on the LRFM model and clustering technique which gives more flexibility in the process of detecting the real behavior of churners and focuses on profitable clusters like core, and new customers. X. Xiahou and Y. Harada [21], rely on a limited number of behavioral parameters, whereas our model considers session data and other relevant behavioral parameters, as described in Table 3. The data set used on total churn is the same data set used on partial churn.

3) Total and Partial churn

In this study, we proposed a new model for predicting customer churn in e-commerce. The model uses a combination of behavioral attributes and the XGBoost algorithm to identify customers who are likely to churn. This model is a multi-class prediction because it could detect total churned, partial churned, or non-churned customers unlike the previous two algorithms worked only on binary class prediction. The model was evaluated on a dataset of real-world user data and achieved an accuracy of 98%.

V. CONCLUSIONS

The prediction of customer churn in e-commerce has become a critical point with the growth of e-commerce market. Acquiring new customers is between 6 to 7 times more expensive than saving loyal ones, so e-

commerce companies begin to take preventive actions to save them and develop their customer relationship management to early detect customer intention of churning.

In this paper, the training and testing dataset comes from a B2C multi-category e-commerce application that describes customer behavior and interactions with the application. The proposed model can be used to predict customer intention to churn, with the type of churn being partial or total.

Our model analyzes the behavioral and interactional data for customers to detect different churn patterns based on session data and clicks in the sessions and the output of the churn prediction stage. Our model achieves 98% accuracy for partial churn and achieves 99% for total churn and 98% for both total and partial churn prediction based on the XGBoost on prediction state. Our model is better at predicting churn than state-of-the-art because it considers the unique needs of each business and the individual behavior of each customer. This makes our model more accurate and more useful for businesses.

Our dynamic churn definition method is a valuable innovation, as it allows businesses to get a more accurate picture of their churn rates and to identify the customers who are most at risk of partial and total churning. This information can then be used to develop targeted strategies to reduce churn and improve customer retention.

Our results are a significant contribution to the field of churn prediction, and our model is likely to be of great value to businesses that are looking to reduce churn and improve customer retention.

Future research will be on different types of e-commerce with varying data sizes to enhance the generalization of the model.

ACKNOWLEDGMENT

We want to express our sincere gratitude to Watan First Digital Company for their support and encouragement throughout this research project. Their dedication to advancing scientific knowledge is truly inspiring.

VI. APPENDIX: BEHAVIORAL DATA ATTRIBUTES

- Let the $Se_{u,n}$ be the nth session Se of a user u. Let $Ac_{u,n,m}$ be the sequence of m Actions made by a user u during his/her nth session then session order = $|Se_u|$
- Last session time = timestamp($Ac_{u,last,m}$) -timestamp($Ac_{u,last,1}$), m = $|Ac_{u,last}|$
- No. of actions within session = $|Ac_{u,last}|$
- Average time on action = $\frac{\text{session time}}{|Ac_{u,last}|}$
- Day of the week = get_day_of_the_week(timestamp($Ac_{u,last}$))
- Day of the month = get_day_of_the_month(timestamp($Ac_{u,last}$))
- weekend = 1, if the day of the weekend, 0 otherwise.
- Let $PurchAc_{u,n}$, PurchAc ∈ $Ac_{u,n}$ be the set of purchases made by the user u during the nth session.
- Similarly, let $AddAc_{u,n}$; AddAc ∈ $Ac_{u,n}$ be the set of add-to-basket actions made by the user u during the nth session.
- Number of purchases = $\sum_{i=1}^{last-1} |PurchAc_{u,i}|$
- Total sum paid = $\sum_{i=1}^{last-1}$ get_price($PurchAc_{u,i}$)
- Last sum paid = get_price($PurchAc_{u,last}$)
- Session to purchase ratio = $\frac{|Se_u|}{\text{number of purchases}}$
- Add to basket to purchase ratio = $\frac{\sum_{i=1}^{last} |AddAc_{u,i}|}{\text{number of purchases}}$
- Add to basket since the last purchase = $\sum_{i=last(PurchAc_u)}^{last} |AddAc_{u,i}|$
- Session length change = last session time $- \frac{\sum_{i=1}^{last-1} timestamp(Ac_{u,i,m}) - timestamp(Ac_{u,i,1})}{\text{session order}}$, m = $|Ac_{u,i}|$
- Number of session actions change = $|Ac_{u,last}| - \frac{\sum_{i=1}^{last-1} |Ac_{u,i}|}{\text{session order}}$

- Session gap change = $\left(\mathrm{Ac}_{u,last,1}\right) - \mathrm{timestamp}\left(\mathrm{Ac}_{u,last-1,m}\right) - \frac{\sum_{i=1}^{last-1} \mathrm{timestamp}\left(\mathrm{Ac}_{u,i,1}\right) - \mathrm{timestamp}\left(\mathrm{Ac}_{u,i-1,m}\right)}{\text{session order}}$,m = $\left|\mathrm{Ac}_{u,i}\right|$
- Number of sessions since last purchase = session order – get_order (last ($\mathrm{PurchAc}_{u,i}$))
- Time since last purchase = $\mathrm{timestamp}\left(\mathrm{Ac}_{u,last,1}\right)$ – timestamp (last ($\mathrm{Ac}_{u,i,m}$))
- Time since last visit = timestamp (last ($\mathrm{Ac}_{u,last,1}$)) $- \mathrm{timestamp}\left(\mathrm{Ac}_{u,last-1,m}\right)$, m = $\left|\mathrm{Ac}_{u,i}\right|$
- Add to basket = 1 if the add to basket event is present in the actual session, 0 otherwise
- Purchase = 1 if the purchase event is present in the actual session, 0 otherwise

REFERENCES

[1] A. S. Dick and K. Basu, "Customer loyalty: toward an integrated conceptual framework," Journal of the Academy of marketing science, vol. 22, pp. 99-113, 1994.

[2] Wikipedia. "Customer attrition." https://en.wikipedia.org/wiki/Customer_attrition (accessed Sep 28, 2023).

[3] F. Reichheld, "Prescription for cutting costs," Bain & Company. Harvard Business School Publishing, 2001.

[4] T. O. Jones and W. Sasser, "Why satisfied customers defect," IEEE Engineering Management Review, vol. 26, no. 3, pp. 16-26, 1998.

[5] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," Expert Systems with Applications, vol. 38, no. 12, pp. 15273-15285, 2011.

[6] J. H. Roberts, "Developing new rules for new markets," Journal of the Academy of Marketing Science, vol. 28, pp. 31-44, 2000.

[7] K. Matuszelański and K. Kopczewska, "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach," Journal of Theoretical and Applied Electronic Commerce Research, vol. 17, no. 1, pp. 165-198, 2022.

[8] Sadeghi, Maryam, Mohammad Naderi Dehkordi, Behrang Barekatain, and Naser Khani. "Improve customer churn prediction through the proposed PCA-PSO-K means algorithm in the communication industry." The Journal of Supercomputing, vol. 79, no. 6, pp.6871-6888, 2023.

[9] J. Lee, M. Podlaseck, E. Schonberg, and R. Hoch, "Visualization and analysis of clickstream data of online stores for understanding web merchandising," Data mining and knowledge discovery, vol. 5, pp. 59-84, 2001.

[10] Forbes. "Amazon Warns Customers About Frequently Returned Items." https://www.forbes.com/sites/walterloeb/2023/03/28/amazon-warns-customers-about-frequently-returned-items/?sh=5abb7a9a39d0 (accessed Sep 28, 2023).

[11] Ö. G. Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking," Expert Systems with Applications, vol. 41, no. 17, pp. 7889-7903, 2014.

[12] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," European Journal of Operational Research, vol. 164, no. 1, pp. 252-268, 2005.

[13] 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 Systems with Applications, vol. 32, no. 2, pp. 277-288, 2007.

[14] N. Gordini and V. Veglio, "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, vol. 62, pp. 100-107, 2017.

[15] A. T. Jahromi, S. Stakhovych, and M. Ewing, "Managing B2B customer churn, retention and profitability," Industrial Marketing Management, vol. 43, no. 7, pp. 1258-1268, 2014.

[16] V. L. Miguéis, D. Van den Poel, A. S. Camanho, and J. F. e Cunha, "Modeling partial customer churn: On the value of first product-category purchase sequences," Expert systems with applications, vol. 39, no. 12, pp. 11250-11256, 2012.

[17] M. Clemente-Císcar, S. San Matías, and V. Giner-Bosch, "A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings," European Journal of Operational Research, vol. 239, no. 1, pp. 276-285, 2014.

[18] H.- H. Wu, S.-Y. Lin, and C.-W. Liu, "Analyzing patients' values by applying cluster analysis and LRFM model in a pediatric dental clinic in Taiwan," The Scientific World Journal, vol. 2014, 2014.

[19] A. Amine, B. Bouikhalene, and R. Lbibb, "Customer segmentation model in e-commerce using clustering techniques and LRFM model: The case of online stores in Morocco," International Journal of Computer and Information Engineering, vol. 9, no. 8, pp. 1993-2003, 2015.

[20] X. Li and Z. Li, "A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm," Ingénierie des Systèmes d'Information, vol. 24, no. 5, 2019.

[21] X. Xiahou and Y. Harada, "B2C E-commerce customer churn prediction based on K-means and SVM," Journal of Theoretical and Applied Electronic Commerce Research, vol. 17, no. 2, pp. 458-475, 2022.

[22] P. Berger and M. Kompan, "User modeling for churn prediction in E-commerce," IEEE Intelligent Systems, vol. 34, no. 2, pp. 44-52, 2019.
[23] V. L. Miguéis, D. Van den Poel, A. S. Camanho, and J. Falcão e Cunha, "Predicting partial customer churn using Markov for discrimination for modeling first purchase sequences," Advances in Data Analysis and Classification, vol. 6, pp. 337-353, 2012.
[24] P. Huntington, D. Nicholas, and H. R. Jamali, "Website usage metrics: A re-assessment of session data," Information Processing & Management, vol. 44, no. 1, pp. 358-372, 2008.
[25] H. Chang and S. Tsay, "Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation," 2004.
[26] Wikipedia. "k-means clustering." https://en.wikipedia.org/wiki/K-means_clustering (accessed Sep 28, 2023).
[27] F. Berzal and N. Matín, "Data mining: concepts and techniques by Jiawei Han and Micheline Kamber," ACM Sigmod Record, vol. 31, no. 2, pp. 66-68, 2002.
[28] Kaggle. "eCommerce behavior data from the multi-category store." https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store (accessed Sep 28, 2023).
[29] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context," International Journal of Electrical and Computer Engineering, vol. 8, no. 4, p. 2367, 2018.

Disclaimer/Publisher's Note: The authors declare that we know of no conflicts of interest associated with this publication**.**

Data Availability :The dataset used in this study is eCommerce behavior data from a multi-category store, which is available on Kaggle at [https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store/data].

The dataset contains customer interaction from a large multi-category online store, such as event type, product ID, category ID, category code, brand, price, user ID, and user session.