IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 May 2024, accepted 19 June 2024, date of publication 27 June 2024, date of current version 22 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420221

RESEARCH ARTICLE

LRFS: Online Shoppers' Behavior-Based Efficient Customer Segmentation Model

RIYO HAYAT KHAN[1], DIBYO FABIAN DOFADAR[1], MD GOLAM RABIUL ALAM[1], (Member, IEEE), MOHAMMAD SIRAJ[2], (Senior Member, IEEE), MD RAFIUL HASSAN[3], AND MOHAMMAD MEHEDI HASSAN[4], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh
[2]Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh 11543, Saudi Arabia
[3]Department of Computer Science, Central Connecticut State University, New Britain, CT 06050, USA
[4]Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Md Golam Rabiul Alam (rabiul.alam@bracu.ac.bd)

This work was supported in part by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSP2024R118; and in part by the Computer Science Department of Central Connecticut State University, USA.

ABSTRACT In the realm of digital commerce, online shopping has witnessed unprecedented growth globally, becoming a cornerstone of modern consumer behavior. This research introduces an advanced customer segmentation model, named LRFS, which builds upon the traditional LRF framework (Length of Relationship, Recency of Purchase, and Frequency of Purchase), specifically tailored for the e-commerce sector. The innovation of the LRFS model lies in the integration of a novel component, ''S'', which quantifies the Staying Rate relative to the revenue generated by customers on a specific website. This addition aims to enhance the granularity and efficacy of customer segmentation by leveraging data extracted from Google Analytics. To operationalize the LRFS model, this study employs two renowned clustering algorithms, K-Means and K-Medoids, analyzing the dataset through the lens of three distinct dimensionality reduction techniques: PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), and Autoencoder. This methodological approach facilitates a robust comparative analysis between the LRFS model and its predecessors — LR, LF, and LRF — utilizing K-Means clustering to evaluate the precision of customer cluster assignments. The empirical findings of this research underscore the superiority of the LRFS model in achieving more accurate and insightful customer segmentation. Additionally, a composite Customer Classification and Customer Relationship Matrix was deployed to discern the nuanced traits of clustered groups, identifying the fusion of K-Medoids and t-SNE as the most effective strategy for capturing the full spectrum of customer dynamics. The research further elucidates several test cases and use case scenarios, demonstrating the practical applications of the LRFS model in conjunction with K-Means and PCA to refine marketing strategies and foster a deeper understanding of the online customer base. Through the development and application of the LRFS model, this study contributes significantly to the field of e-commerce by providing a more nuanced tool for businesses to tailor their marketing initiatives, ensuring alignment with the evolving preferences and behaviors of their online clientele.

INDEX TERMS Customer segmentation, unsupervised machine learning, K-means, K-medoids, RFM analysis, LRFM analysis, dimensionality reduction, PCA, t-SNE, autoencoder, deep learning, Google analytics.

I. INTRODUCTION

E-commerce, synonymous with online shopping, has become an integral component of contemporary digital lifestyles, enabling transactions of goods and services across global

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

boundaries with unprecedented ease [1]. This digital marketplace serves as a vital platform for third-party transactions over the internet, where consumers can effortlessly explore merchandise, place orders, and make payments online, culminating in the delivery of products directly to their doorsteps via internet courier services [2]. The advent of the Covid-19 pandemic further underscored the convenience and necessity of online shopping, propelling it to become the preferred method for procuring goods amidst government restrictions and health concerns, thereby cementing e-commerce's role as a pivotal aspect of modern consumer behavior, supported by continual advancements in internet technology [3].

In response to the burgeoning trend of online shopping, businesses have increasingly turned their attention to analyzing consumer purchasing behaviors with the aim of enhancing service offerings, boosting customer satisfaction, fostering repeat purchases, and ultimately driving profitability. At the heart of these efforts lies the practice of customer segmentation, a strategic process that enables businesses to dissect their customer base into distinct groups based on shared characteristics such as demographics, purchasing patterns, and brand preferences, thereby allowing for the delivery of targeted marketing messages and personalized shopping experiences [1].

The complexity and volume of data inherent in e-commerce necessitate sophisticated approaches to customer segmentation. The RFM (Recency, Frequency, Monetary value) model and its variants have been widely adopted for their efficacy in identifying valuable customer traits. Moreover, the integration of data mining and machine learning techniques has revolutionized the ability to emulate the personalized interactions characteristic of small business-customer relationships [4]. Clustering, a form of unsupervised machine learning, plays a crucial role in this context, enabling the identification of customer groups with similar purchasing behaviors, which can then be targeted with tailored marketing strategies, thereby enhancing customer retention and attracting new clients [5].

Recognizing the competitive landscape of the e-commerce sector, this research endeavors to develop a sophisticated model dedicated to the segmentation of online shoppers. By integrating a suite of unsupervised algorithms and dimensionality reduction techniques, the study aims to dissect and understand customer behavior in the online shopping milieu more deeply.

At the core of marketing strategy lies the imperative to augment sales and ensure the sustained profitability of businesses. This objective demands a granular understanding of customer behavior during online interactions and transactions [6]. Service providers and manufacturers are tasked with not only forecasting future purchasing trends but also identifying and mitigating any deterrents to purchase decisions. The proliferation of irrelevant information can overwhelm consumers, potentially deterring purchases and leading to missed opportunities [7]. Thus, businesses are compelled to prioritize the enhancement of the online user experience and the alignment of product offerings with customer interests, thereby bolstering consumer retention and trust [8].

This study introduces the LRFS (Length, Recency, Frequency, Staying Rate for Revenue) model, a novel approach to customer segmentation that incorporates a "Staying Rate for Revenue" metric, designed to provide a deeper understanding of the online shopping community. By leveraging advanced clustering algorithms and dimensionality reduction techniques, the LRFS model aims to bridge the gap in existing segmentation models, offering a more nuanced reflection of customer interests derived from website-specific features.

This research marks a significant advancement in the field of e-commerce customer segmentation through the introduction of the LRFS model. This model uniquely captures customer behaviors and preferences by analyzing data derived from online shopping interactions. The key contributions of this study include:

- Introduction of the "S" component within the LRFS model, demonstrating its potential to significantly enhance revenue generation by improving the precision of customer segmentation [6].
- A detailed exploration of the LRFS model, emphasizing the importance of preprocessing, feature investigation, and the creation of new metrics tailored to the objectives of the model [7].
- A comprehensive comparative analysis among various unsupervised machine learning algorithms and dimensionality reduction techniques, identifying the most effective strategies for implementing the LRFS model [4].
- Evidence of the LRFS model's superiority over traditional segmentation models, highlighting its capacity for more accurate customer clustering [5].
- The application of a combined customer classification and relationship matrix to analyze clustered groups, yielding insights into customer behaviors and preferences, and illustrating potential use cases [8].

The subsequent sections will delve into a literature review, detailing studies relevant to the development of the LRFS model and its components. Following this, the methodology section describes the data analysis process, including the selection of clustering algorithms and dimensionality reduction techniques, and the rationale behind their use. The results section presents the findings from the application of the LRFS model, comparing its performance with traditional segmentation models and highlighting its effectiveness in identifying distinct customer segments. A discussion on the implications of these findings for e-commerce marketing strategies is also provided. Finally, the conclusion summarizes the key points of the research, acknowledges its limitations, and suggests directions for future studies in the realm of e-commerce customer segmentation.

II. RELATED WORK

In today's fiercely competitive marketplace, business proprietors are keen on attracting new clientele and nurturing enduring relationships with their most profitable customers [9], [10]. Recognizing the distinctive attributes of each client is pivotal for cultivating mutual trust [11]. Consequently, customer segmentation has long been a staple in practice, with its application spanning various sectors including healthcare [12], e-commerce [5], telecommunications [10], [13], [14], [15], retail [16], [17], dining [18], insurance [10], energy [19], travel and hospitality [20], [21], [22], financial technology [23], [24], and recommendation systems [25], among others. Various clustering methodologies, including K-means [26], hierarchical clustering [24], DBSCAN [19], [27], Mean-Shift [5], and Fuzzy C-Means, alongside heuristic techniques like Gaussian Peak Heuristic Clustering (GPHC) [28], have proven to be effective in segmenting customer bases. The utility of unsupervised machine learning algorithms lies in their capacity to deliver data-driven segmentation outcomes, regardless of data precision or the features present [22].

The advancement of information technology has significantly broadened access to consumer behavior and purchase intent data, facilitating the development of increasingly sophisticated segmentation strategies. Despite this complexity, the value of simplicity in segmentation cannot be overstated [29]. RFM analysis, focusing on recency, frequency, and monetary value, has long been a focal point for researchers, serving as a behavioral-based method to profile clients and evaluate the consistency of their purchasing behaviors over time [14], [30]. This analysis not only aids in marketing decision-making but also in tailoring strategies to engage the right consumers effectively [31]. Notably, some researchers have proposed a two-phased model technique as an innovative segmentation solution, leveraging the RFM model [32], with comparative studies also exploring the juxtaposition of AI methods and RFM analysis [27].

Furthermore, empirical research [33] employing a decision tree methodology for segmenting the electronic toll collection (ETC) customer base revealed its proficiency in analyzing travel behaviors, valuation, and potential appreciation for ETC services. This approach, integrating decision tree-based explainability with clustering via the ExKMC algorithm [34], enhances the interpretability of cluster assignments. The RFM model, alongside data modeling techniques, has been applied to identify customer behavior patterns, validated through various classification methods including multi-layer perceptron (MLP), support vector machine (SVM), and decision tree classification (DTC) [35].

Beyond traditional RFM models, modifications incorporating new weights or dimensions have been explored to address inherent limitations and enhance segmentation outcomes. Such adaptations recognize that customers with higher RFM scores are generally more profitable and responsive, yet traditional RFM analyses do not account for changes in customer behavior over time or the varying emphasis on RFM components across different industries [36], [37]. Since the importance of the three variables in the RFM model can vary based on product features and industry characteristics, researchers have experimented with the order of importance of each variable [43], [44], [45], while creating formulas to calculate the RFM score [46]. Analytic Hierarchy Process (AHP) was implemented to determine the various weights of RFM variables [47]. WRFM or the Weighted RFM has also been used by [48] along with a range of data mining approaches such as K-means, ARM, and neural networks. Their proposed model demonstrates how WRFM paired with clustering methods exceeds traditional RFM and enhances business strategy, resulting in higher company profits.

Innovations like the inclusion of "Length" (L) [38], [39] as a new dimension has brought up a new scope of improvement to the conventional RFM model. In the authors' opinion, by only using RFM models, firms are often unable to clearly differentiate between short-term and long-term customers. To overcome this shortcoming, they have extended the RFM model by adding the attribute Length (L) and used an unsupervised neural network clustering method SOM, with which they succeeded in for dental services marketing more effectively. The application of SOM was also considered in the work of [56] to determine the optimal number of clusters, for their proposed LRFM model to identify customers with profit potential. The development of models such as RFMTC [36] and LRFMV [40] underscore efforts to refine customer value analysis and segmentation strategies, leveraging advanced clustering techniques and deep learning approaches for improved accuracy and insight into consumer behavior. In RFMTC model [36], additional components refer to the "Time since first purchase" (T) and "Churn probability" (C) respectively. In another work [49], the RFM model was expanded by including two more components, duration (L) and cost (C). The authors in [12], have taken three years of patient behavior data to figure out who has the potential for loyalty towards the clinic. Generally speaking, the longer the relationship between the customer, the more they are likely to be loyal. Hence, they first calculated the R, F, M, and L separately, then eventually calculated the weighted Life Time Value (LTV) for each cluster. They have described their model as Customer Life Time Value or CLV, where the greater the value of CLV is, the more loyal the customer base is. The concept of customer value analysis was reviewed in [50] as well. Customers have been divided based on k means clustering using the weighted RFM model for calculating customer lifetime value (CLV). The relative weights of the RFM model were established using the Analytical Hierarchy Process (AHP). According to the findings, CLV analysis using RFM measures can assist the company in better understanding its customers and locating the most profitable customers. Enhanced RFM scores integrated with CLV matrix were proposed by [37], where they also applied to enhance normal distribution formulas

to eliminate outliers before analyzing customer purchasing patterns using hard clustering, Expectation Maximization (EM) and soft clustering, Fuzzy C-Means methods. With its capacity to assign excellent initial points in the smaller dataset, the EM method has scaled far better than the Fuzzy C-Means approach.

In another study, the LRFMP model [51], where the 'P' feature denotes the periodicity of the customer return, was applied to access the consumers in the grocery retail industry. The periodicity of customer returns is a crucial consideration in various organizations, such as grocery stores, when analyzing the behavior of customers. LRFMP was also implemented in [52] where the authors have used two-stage clustering along with the LRFMP model for segmenting customers and analyzing their characteristics across Iranian Fintech companies. Following the initial clustering, more clustering experiments were conducted in segments that required further variable analysis for a better understanding of the important clients. This study extends and applies both the RFM and LRFMP models in the B2B scenario, and proposes a new approach strategy based on a two-stage clustering method that contributes to a deeper understanding of customer behavior. In another recent research, the authors have developed a web content retrieving system [53] to fetch customers' data from a website, and using the data, they have made an extension to the existing RFM model. They have modified the conventional RFM model by introducing a new dimension "T" which refers to "Interpurchase Time", the interval between a customer's two subsequent purchases. The concept of T itself was not new; researchers had previously employed this attribute to increase the effectiveness of product recommendations [54], introducing a loyalty program [55].

The dataset used in this research, namely online purchase intention dataset [41], was applied by several researchers to predict customers' likely purchase behavior. A comparative analysis of deep neural network has been made along with conventional machine learning methods XGBoost, Random Forest, CatBoost, AdaBoost, Support Vector Machine, and Decision Tree [2]. Deep Neural Network Algorithm outperformed them all by producing higher accuracy, precision, recall, f1 score, as well as AUC compared to other algorithms. In another work [4], a prediction algorithm was created to identify whether or not a consumer would make a purchase before leaving the website and allow customers to instantly act on their needs. It also demonstrates how the various components have the potential to improve the platform's effectiveness. The authors have stated their prediction module can be applied to a variety of aspects of a website, not just predicting a user's future purchase. In both of these papers, the researchers have provided a comprehensive overview of the different elements that a platform can utilize to strengthen its buyer decision journey. However, they have used this dataset to treat it as a binary classification problem, by predicting the target class Revenue. Hence, it can be said that this work would be the first attempt to apply an unsupervised technique, or segmentation based on LRF framework, using this dataset so far.

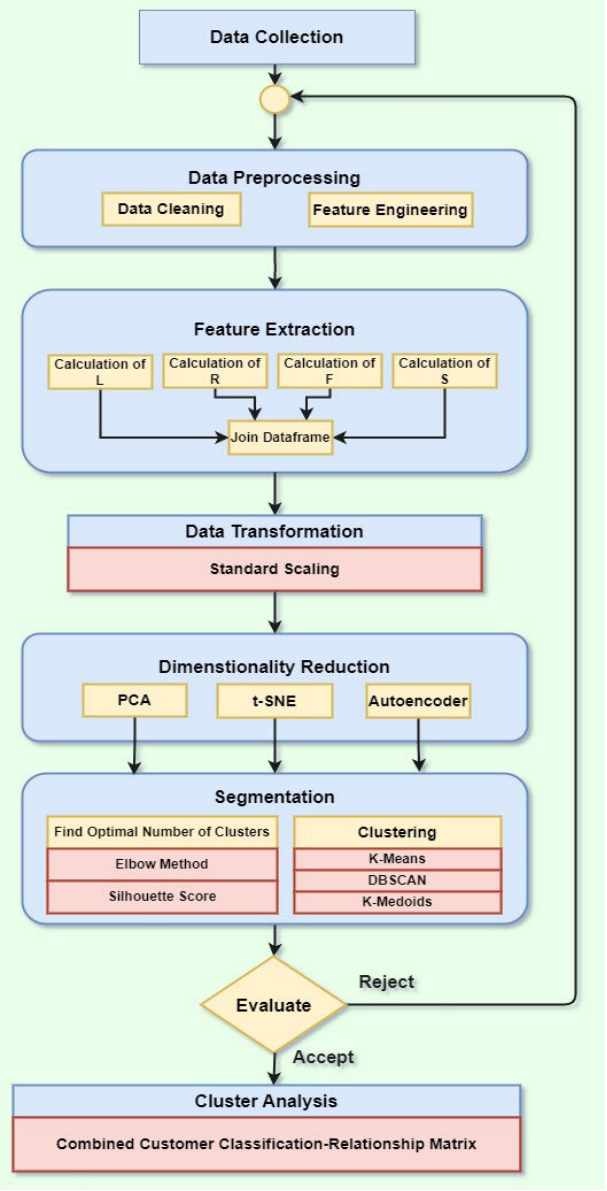

FIGURE 1. Top level layout of the proposed LRFS model.

III. METHODOLOGY

A. PROPOSED LRFS MODEL

The first and foremost step was to find an appropriate e-commerce dataset [41] with Google Analytics features from the UCI Machine Learning Repository website. After that, preprocessing steps such as Data Cleaning and Feature Engineering have been performed on the dataset. Exploratory data analysis has been conducted where the correlation of the variables was checked to avoid any kind of dependencies between the existing features. The necessary categorical features were encoded in the following step. Then, the components L, R, F, S were calculated from the existing

features, for establishing the LRFS model. After separately computing the required features, all four components were merged into one dataframe. Next, the data were transformed using Standard Scalar to avoid bias. The following step was dimensionality reduction, which was one of the most crucial parts of this research. The dataset was shrunk separately using three different dimensionality reduction techniques, namely PCA, t-SNE, and Autoencoder. From this point, the analysis was done individually for these three newly created reduced dataframes. For each dataframe, the optimal number of clusters was computed using Elbow Method and Silhouette Coefficient, and subsequently, two clustering algorithms namely K-Means and K-Medoids were applied on each of the compressed dataframe. The outputs were visualized using different colors for different clusters. From the generated figures, cluster analysis was demonstrated accordingly. Each of the clusters was again made into a separate dataframe along with the "Revenue" column to understand the relationship between each component and the Revenue. Afterwards, the segmented data were evaluated and labeled correspondingly using a combined customer classification and relationship matrix. Fig.1 visualizes the bird's eye overview of the proposed model.

B. DATA PREPROCESSING

Each entry in the applied dataset [41] is made up of a feature vector that has information on a user's visit or "session" on an e-commerce website. The dataset comprises information acquired from 12,330 sessions, each associated with a unique user to prevent any sort of user impact on the model. There are a total of 18 features in this dataset. 10 of them are numerical and the other 8 are categorical features. The features mainly describe the information of customers who are visiting the e-commerce website. For example, the number of visits and time spent on pages like Administrative, Informational, and Product Related Pages, and some Google Analytics features such as Bounce Rates, Exit Rates, and Page Values are also included in the dataset.

The dataset was organized neatly and did not contain any null values. It did have 125 duplicate values, and they were removed at the beginning of the preprocessing steps. Overall the dataset was good in terms of quality except for some minor issues with outliers. The outliers were not handled as important information would have been erased if they were removed.

In the dataset, "Revenue" denotes binary values, in other words, whether the user ended up in a transaction or not. In the whole year, only 15.6% of the sessions made by the visitors did end up in purchasing something, while the other 84.4% resulted in leaving the website without any transaction; making the dataset imbalanced.

Fig.2 shows the volume of traffic across different traffic types on special days. The special day represents the closeness to any significant events or holidays. It is commonly assumed that the purchase rate will be higher whenever there is a special occasion approaching. The image shows that

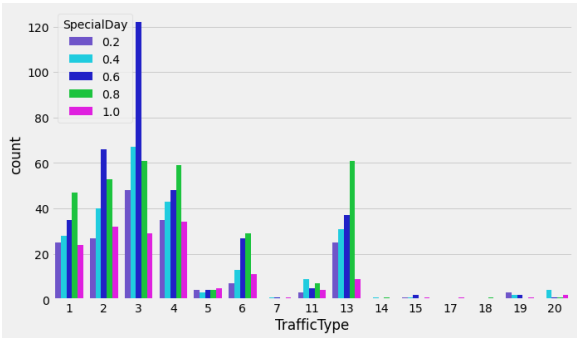

FIGURE 2. Traffic volume around special day per traffic type.

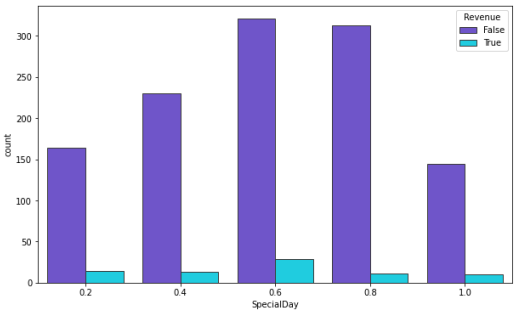

FIGURE 3. Revenue before SpecialDay.

most traffic was on Traffic Type 3 when there were four days remaining for a special day. It is also noticeable that regardless of the traffic type, the users are browsing the website when the special day is getting closer. Fig.3 shows that even though the users mostly browsed the platform when the day was near to a special occasion, it did not necessarily end up in much revenue generation. More traffic near special occasions did not result in more revenue generation, so it can be stated that for this particular dataset, the feature "SpecialDay" is not affecting the Revenue.

The purpose of feature engineering is to aggregate common features and record feature interactions. As features like "Administrative", "Informational", and "ProductRelated" were all representing the number of page visits by the user, they were aggregated into one, named "total_page_view". The following equation (1) shows the general calculation for creating the feature "total_page_view". Here, m represents the number of different kinds of pages, the value of m would be 3 for three different pages.

$$total\_page\_view = \sum_{m=1}^{n} visit_m \tag{1}$$

Similarly, duration-related features for different pages in the website such as "Administrative_Duration", "Informational_Duration", and "ProductRelated_Duration" were summed into a new column "total_page_duration". Equation (2) shows the generalized formula, where m stands for various kinds of pages a user visits, and for each of the pages,

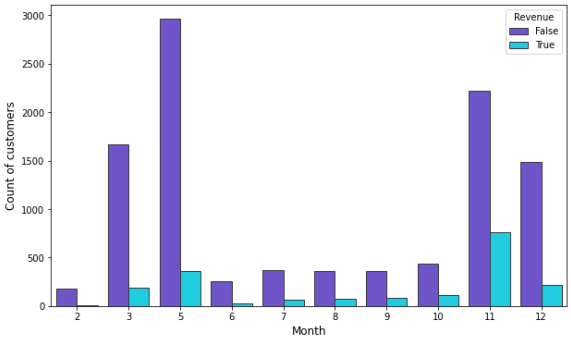

FIGURE 4. Revenue and month analysis.

the amount of spent time would be added.

$$total\_page\_duration = \sum_{m=1}^{n} duration_m \tag{2}$$

After creating these two new features, the individual six features were dropped from the dataframe. After some thorough investigation, it was found that features such as "SpecialDay", "OperatingSystems", "Browser", "Region", "TrafficType", and "Weekend" were not related to the proposed LRFS analysis. Hence, those six columns were removed as well. After that, the remaining categorical values "Month", "VisitorType", and "Revenue" was replaced by numerical values.

The feature "Month" is important for any e-commerce analysis, as it shows the time of the year in which the site might be crowded. In the data, it is observable that May has the highest percentage of sessions (27.3%) and the least sessions were held in February (1.5%). It should be noted that regardless of whether they made a purchase, visitors have browsed the platform all year. So, it would be better to analyze the distribution of customers throughout the year and check whether the visit has contributed to the Revenue or not. Interestingly, Fig.4 visualizes that even though most sessions were held in May, they did not necessarily end up with the highest revenue contribution. The month of May has the highest number of "False" Revenue, which means, the highest number of visitors have left the website in May without buying anything. On the other hand, November had the second most frequency of sessions (24.45%), and some of the population in this month have led to transaction completion.

One of the most prominent categorical features in this dataset was the Visitor Type, which tells whether a customer came back to the platform or not. The three types of visitors were observed in the dataset, where 85.5% of the visitors are repeating customers (labeled as 3) and 13.9% of them are new (labeled as 1). The "Other" visitor type (labeled as 2) is holding a very negligible percentage, so for the sake of easier calculation, it was considered with the "New_Visitor" type in this research. It is often assumed that the old customers would tend to purchase more compared to the new customers, but in this dataset, it seemed to be otherwise. Fig.5 shows that

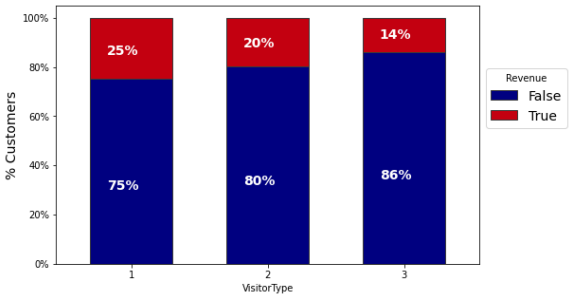

FIGURE 5. Revenue by visitor type.

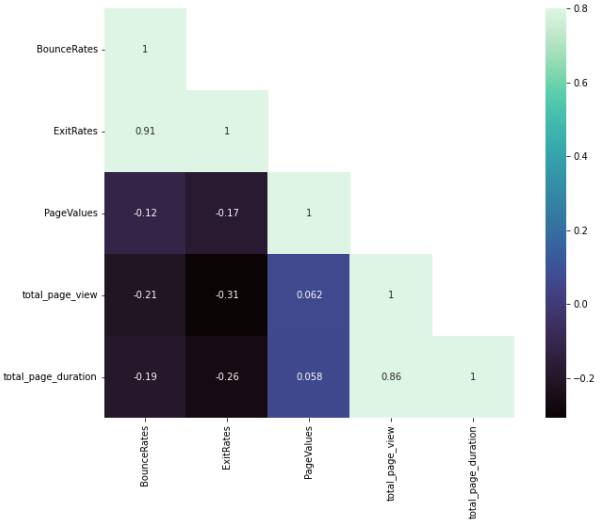

FIGURE 6. Correlation matrix.

the most revenue is coming from the new visitors (25%) and the old customers are contributing lesser in revenue (14%) compared to the new ones. Considering the fact that most customers were the repeated ones, 14% is quite low. But, they cannot be discarded as potential customers since they might prove their loyalty in a later period of time. On the other hand, percentage of the new visitor to this site was less than 15%, but still, they had 25% revenue contributions. At this point, the company could think of a strategy to handle them more carefully so that if they browse the website again, they do not have to repeat similar behavior as the existing returning customers.

Using the above mentioned features, correlation matrix was illustrated to understand how they are connected to each other. Fig.6 shows the correlation among the extracted features. In the figure, it is visible that a strongly positive correlation (0.91) between Exit Rates and Bounce Rates is still there. For both of these features, there is one common factor and that is, the customer is leaving the website. While Bounce Rate considers the exit from the very first page the customer is visiting, Exit Rate considers all exits. So, all Bounce Rates are considered to be the Exit Rates. So, to avoid redundancy, in this research, only the Exit Rates were taken for analysis purposes. A high correlation (0.86) was also found between Total Page Duration and the Total Page Views,

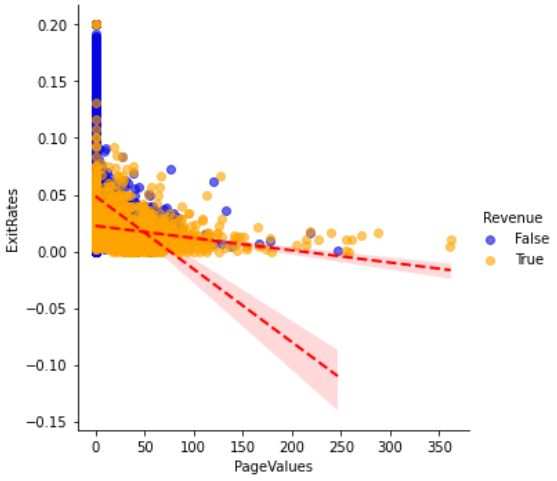

FIGURE 7. Relation between page values and exit rates.

which makes sense as the more the pages are being accessed by the users, the more time they are going to spend. So, the Total Page Duration was removed this time and only the Total Page Views were taken for further calculation for building the LRFS model.

In Fig.7 another interesting fact can be visualized which is, when the values of Exit Rates are decreasing, Page Values are increasing. Generally, if customers keep exiting the website frequently, it does not lead to better revenue generation. On the other hand, if the user spends more time on the website, it could increase the possibility of more purchases from the website. Hence, the Staying Rate, which is the opposite of Exit Rates was given a vital role to introduce the new dimension of the LRFS model.

C. FEATURE EXTRACTION

Each session of this dataset corresponds to a unique customer to avoid any sort of influence on the model, which means that there is no information about multiple visits of a single customer. Hence, it is not possible to directly calculate each customer's recency or frequency. However, there are some very important features in this dataset which has the potential to make better customer segmentation. Keeping that in mind, new features were created by combining or tweaking the existing ones and were mapped as components namely "L", "R", "F", and "S".

1) CALCULATION OF COMPONENT L

Length indicates the interval between two particular visits. So in general equation of Length would be,

$$L = v_l - v_i \tag{3}$$

Here, v_l stands for the latest visit and v_i denotes the initial visit. However, things are a little different in this dataset. As the dataset was created in a way that each session would belong to an individual customer for one year. So technically, the overall Length should be 12 months whereas, for each

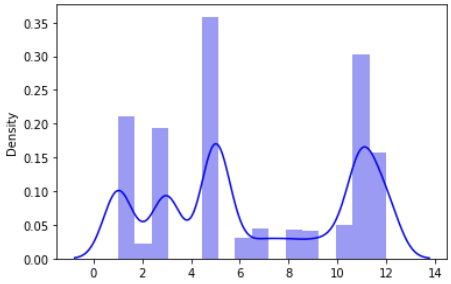

FIGURE 8. Displot of length.

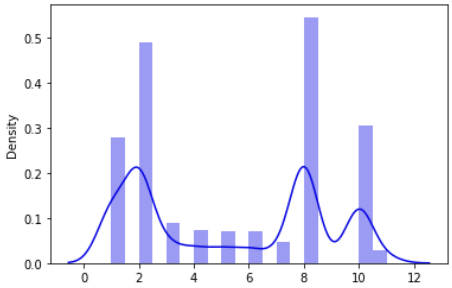

FIGURE 9. Displot of recency.

customer, the calculation will differ. Moreover, the dataset did not have any DateTime feature to track, so a few tricks have been used to calculate the value of L. There were two features (which were originally categorical but encoded in the data preprocessing step) that came in handy, there was a "Month" feature, and the other feature was the "VisitorType". If the visitor type of a customer is returning, it would mean that the customer has been associated with the website before, and since the earliest month of this dataset is January, the difference from January to the current month has been considered for returning customers. On the other hand, the default number of month 1 was taken for the new customers, since it was the first time the customer is visiting the website. Fig.8 shows the displot of length, where the data is distributed in a multimodal manner.

2) CALCULATION OF COMPONENT R

Recency tells how recently the customer has made a visit. As mentioned in the previous subsection, there is no specific time interval or DateTime feature to calculate the most recent visits of customers. So the highest value for the month of the whole dataset, meaning the most recent month for that particular website (December in this case) was taken and subtracted from the current month for each customer. So the Recency for each customer i was calculated as,

$$R_i = month_{max} - month_c + 1 \tag{4}$$

Here, $month_c$ stands for the current purchased month for that particular customer. 1 was added because if someone purchases something in December, it should also be considered. Fig.9 shows the displot of the feature Recency. The data

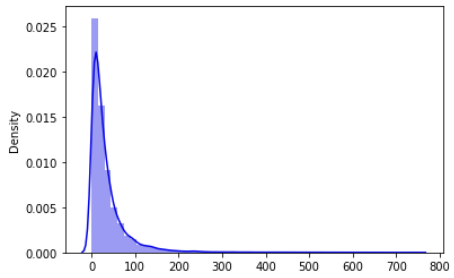

FIGURE 10. Displot of frequency.

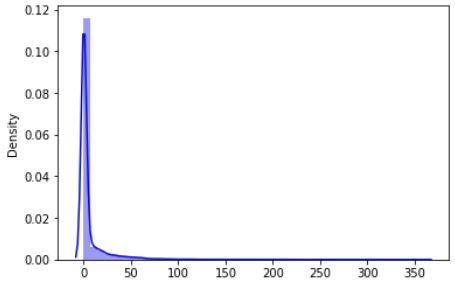

FIGURE 11. Displot of staying rate for revenue.

is shown to have more than two peaks, so it is a bimodal distribution.

3) CALCULATION OF COMPONENT F

Frequency tells the number of times a customer has visited a website, so the general equation of Frequency would be pretty straightforward,

$$F = count(v_n) \tag{5}$$

where v_n represents the total number of visits made by a particular customer.

The number of total page visits made by the customer, namely column ''total_page_view'', the summation of Administrative, Informational, and ProductRelated sessions, were taken as the component Frequency. Fig.10 shows the displot of the feature Frequency, where it is showing the data distribution is right skewed.

4) CALCULATION OF COMPONENT S

In this research, the main purpose was to introduce a new dimension in terms of customer segmentation analysis for online shoppers, and the component ''S'' plays a big role in this. ''S'' is short for ''Staying Rate for Revenue'', which basically symbolizes the essence of a particular customer staying on the website, and eventually contributing to the revenue. From the perspective of the company that owns the e-commerce website, this attribute would play a vital role to understand customer purchase intention. To understand the importance of this variable, one must dig deep into two of the most essential features in Google Analytics, ''Page Values'' and ''Exit Rates''. Page Values refer to the average value of each successful transaction completed by that customer, which can be defined as follows,

$$PageValues = \frac{WebsiteRevenue + TotalGoalValue}{\#UniquePageviewsForAGivenPage} \tag{6}$$

According to Google Analytics [57], Page Value is the average value of each successful transaction completed by the user. In the numerator, the summation of Website Revenue and Total Goal Value represents the successful transaction. The definition of a successful transaction could differ depending on the website owners. For some website owners, a successful transaction could mean when a user is just adding items to a cart, saving any content for the future, simply interacting with videos and visuals, or any other user engagement with the website set by the website owners. In this paper, in the definition of ''S'', the term ''Revenue'' refers to both Website Revenue and Goal Value.

Contrarily, Exit Rates reflect the points at which website users chose to leave. For example, suppose a user goes to the landing page of the website, then moves to a product related page, but after that, without adding the product to the cart or let alone completing the transaction, the visitor simply leaves the site. It would mean that the specific product-related page was the last one in the session, and the exit rate will be calculated based on that. The following formula is used to determine the exit rate.

$$ExitRates = \frac{\#ExitsFromAGivenPage}{\#PageviewsForAGivenPage} \tag{7}$$

In Fig.7, it was visible that the ExitRates have a negative relationship with the feature PageValues. Since it is generally considered a negative metric, a high exit rate of a page would mean that page is not intriguing to the customer. The reason behind it could be anything, page speed, bad design, high expenses for the desired product, unorganized page content, and so on. So the goal should be to minimize the exit rates as much as possible by consistently checking on it and updating the websites accordingly. That being said, (1 - Exit Rates) is the exact opposite of the Exit Rates, referring to the value with which the user stayed on the page, which is why it was named Staying Rates in this research.

Now, multiplying it with Page Value gives us a ratio of Revenue per Staying Rate, so looking at the component ''S'' would tell us how much revenue a certain customer is contributing on the basis of how long the customer has been staying on that website. The final equation would be:

$$S = PageValues * (1 - ExitRates) \tag{8}$$

Fig.11 displays the displot of Staying Rate for Revenue, where the data distribution is seen to be heavily right skewed. After computing all the necessary features L, R, F, and S, they were joined into a single dataframe as LRFS. Next, feature scaling was performed with the help of a data standardization technique, standard scaling; the retrieved features are rescaled to a specific range, where the attribute's mean value is 0 and the distribution's standard deviation is 1.

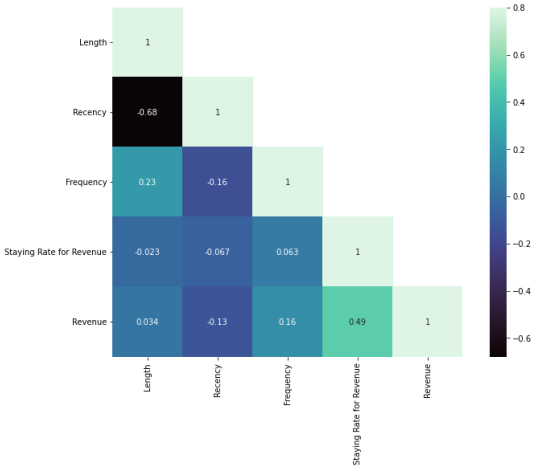

FIGURE 12. Correlation matrix of LRFS.

5) CORRELATION OF COMPONENTS WITH REVENUE

For identifying the customer segments, all the components of the LRFS model also the Revenue column have to be taken into account. Every variable must have unique features that do not coincide with those of other properties. The updated correlation matrix shows the relationship between each parameter. It can be observed in Fig.12 that the newly introduced parameter S has a weak association with each of the other parameters. With L it has -0.023, with R -0.067, with F 0.063 which makes S a unique trait that cannot be replaced by other features. Also, it has a moderate correlation of 0.49 with the Revenue. As a result, it can be said that the LRFS model could be used as a special model for customer segmentation in online shopping.

D. MODEL SPECIFICATION

Several well-known techniques, including Elbow Method, Silhouette Coefficient, the Cumulative Explained Variance Ratio, dimensionality reduction techniques (PCA, t-SNE, Autoencoder), and two of the most widely used clustering algorithms (K-Means and K-Medoids) were used for the model specification. Each of these methods has assisted in the discovery of the independent attributes of the proposed model.

IV. RESULT ANALYSIS AND DISCUSSION

In the used dataset, the Revenue column contains binary values, so from this feature, it can only be found whether a particular customer has purchased anything from the online store or not. In a given cluster, the value of Revenue would represent the percentage of customers in that cluster who had contributed to Revenue generation. Three of the dimensionality techniques PCA, t-SNE, and Autoencoder have been separately used for each clustering algorithm to closely examine which combination works better. The output clusters from K-Means and K-Medoids were visualized, and the four components (L, R, F, S) along with the feature Revenue were plotted for further cluster analysis, from which one can observe how this model is going to serve the firm with a proper description of each of the clusters generated from each algorithm.

A. K-MEANS CLUSTERING VISUALIZATION AND ANALYSIS

To implement K-Means clustering, the first step was to figure out the optimal number of clusters. WSS curve and Silhouette Scores were used to find the number of k. After that, K-Means clustering was implemented and each of the clusters was analyzed. For each of the dimensionality reduction methods, similar steps were followed.

1) K-MEANS ANALYSIS USING PCA

The optimal number of clusters was determined to be 4 and the four different colors are showing the four different customer groups. In Fig. 13, Cluster 0 shows the generated revenue is 10%, and the other components are also showing lower values for features L (3.6), F (22.8) and S (2.3), except for the R (7.6). The Cluster 1 also has 10% of revenue, but the combinations of the features are different. L (10.8) yields the highest value, and the remaining ones are R (2.2), F (30.9), and S (30.9). It is observed that Cluster 2 has the second highest Length of 9.5, however, it has much bigger Frequency which is 182.8, and in the end, the value of Revenue is yet quite low (30%). Here, the most revenue generation can be seen in Cluster 3, and it has the highest value of S (72.7). It is noticeable that in this cluster the other components such as L (5.6), R (4.5), and F (35.5) are not that high compared to the S. However, Cluster 2 has the highest value of F (182.8), and the second highest value of L (9.5) yet the Revenue is showing to be 30%. This shows how important feature S is to determine whether the customer would purchase anything from the store or not.

2) K-MEANS ANALYSIS USING T-SNE

Similar to the previous one, for t-SNE, the optimal number of clusters came out to be 4 as well. In Fig.14, it is displayed that the overall Revenue came out to be quite low. The most revenue was generated in Cluster 3 with the value of 30%, and it had the most value of S (16.2), and greatest F (56.9), while moderate L (6.9) and R (5.3). Cluster 1 showed close to 0% of Revenue generation, even though it had an L of 7.2, R of 5.8, and F of 28.7. This implies that even though a customer is associated with the website for more than seven months, it does not mean that the customer would purchase products from them. Both clusters 0 and 2 had 10% of Revenue although the value of other components varied from each other. Cluster 0 had a moderate value for L (4.9) and R (5.8), low F (17.1) and very low S(0.5). Cluster 2 had comparatively higher L (7.1), same value of R (5.8) as cluster 0, slightly larger F (23.5) and almost negligible amount of S (0.2).

3) K-MEANS ANALYSIS USING AUTOENCODER

After performing the Elbow method and generating Silhouette scores, it was found that, in the case of Autoencoder, the

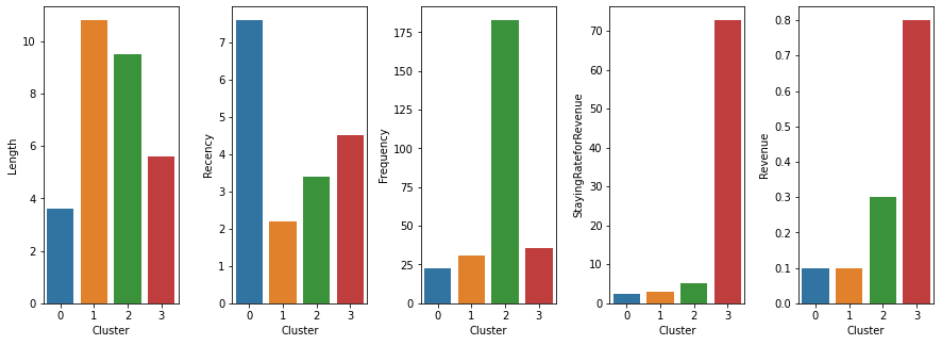

FIGURE 13. K-means clustering analysis of LRFS model (PCA).

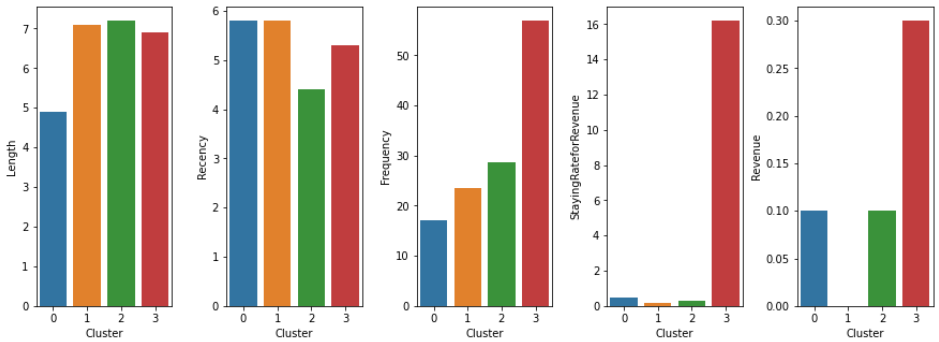

FIGURE 14. K-Means clustering analysis of LRFS model (t-SNE).

optimal number of clusters was also 4. Looking further into the cluster analysis in Fig.15, Cluster 0 holds only 3% of the revenue, where the S (0.4) is trivial. The other features are L (4.2), R (8.2), and F (19.7). Cluster 1 has been grouped with the customer base in which 30% of them contributed to revenue. This cluster has the highest value of F (180.3), high L (8.3), moderate R (4.7), and S (9.6). Cluster 2 comprises the highest value of Revenue (50%), and the value of S (29.4) was also the highest among all clusters, whereas L, R, and F had a value of 3.1, 3.8, and 25.0 respectively. It means even though the customers spend only three to four months on this website, still, 50% of customers in this group purchased product(s) from the website. On the other hand, although Cluster 1 had the biggest value of F (180.3), a decent L of more than eight months, and R of almost five months, in the end in terms of Revenue it could not beat Cluster 2.

B. K-MEDIODS CLUSTERING VISUALIZATION AND ANALYSIS

Similar to the K-Means, the initial step for K-Medoids was also to find the number of k. Again, the WSS curve and Silhouette Scores were implemented to get the optimal number of clusters. For each dimensionality reduction method, the Elbow method and Silhouette Score were generated and then according to the value, K-Medoids clustering was implemented.

1) K-MEDOIDS ANALYSIS USING PCA

Unlike the K-Means, the optimal number of clusters came out to be 7 for the reduced dataset by PCA. Looking into Fig. 16, Cluster 0 and Cluster 4, seemed to have 80% of Revenue. Cluster 4 had the highest S here, which was 110.6, and the other values were moderate such as L (7), R (3.2), and F (39.7). Interestingly, although the amount of Revenue generation was very closer, Cluster 0 had only S (43.4), L (4.7), R (5.8), and F (32.3). Cluster 3 has a Revenue of 30% and cluster 6 has 20 %, with L (9.3), R (3.6), F (200.8), S (4.6), and L (10.8), R (2.2), F (72.1), S (5.6) respectively. Clusters 1, 2, and 5 had a value of 0.1 (10%) in terms of Revenue, but these were separated mainly because they had a very different amount of the variable L. Cluster 1 had a large L (11.1), small R (1.9), F (16.9), and very low (1.6). Cluster 2 had a small L (2.5) but a big R (9.8), a small F (19.8), and a very small S (1.2). For Cluster 5, the S (1.3) was still very minimal, L (4.5) and R (6.4) were moderate, and small F (23.4). This again implies that the Length does not always ensure the certainty of purchase, and sometimes the customer may not stay for a lot of sessions to make a purchase.

2) K-MEDOIDS ANALYSIS USING T-SNE

According to the Elbow Method and Silhouette Score, this time the value of k was 6. Later in this section, Cluster Analysis using Customer Classification Matrix and Customer

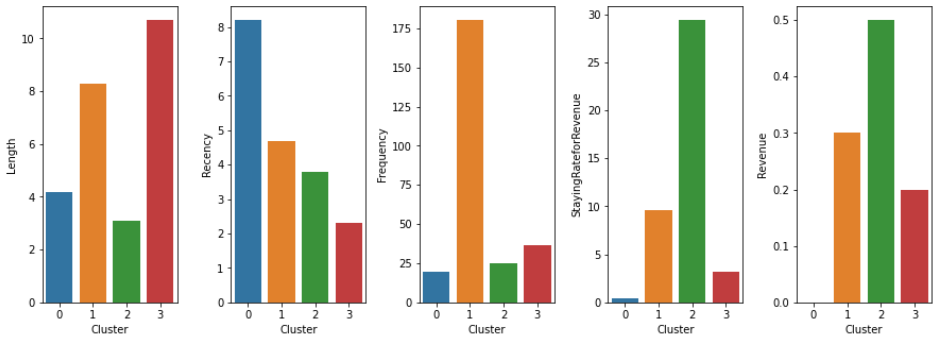

FIGURE 15. K-medoids clustering analysis of LRFS model (autoencoder).

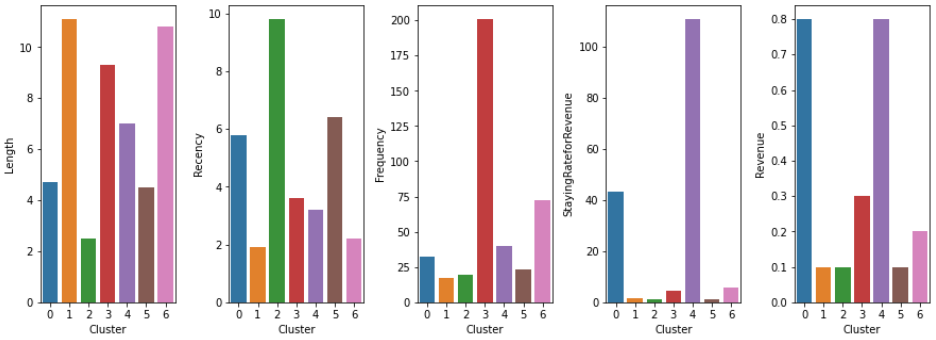

FIGURE 16. K-medoids clustering analysis of LRFS model (PCA).

Relationship Matrix will be conducted using these six clusters.

Fig.17 shows the most value for Revenue was 50%, in Cluster 1. It had a moderate L (5.6), R (4.9), F (32.0), and S (24.5). The other clusters did not have much Revenue generation, Cluster 5 had a Revenue of 20% with low S (7.5), even though it had higher L (7.2), moderate R (5.5), and quite large F (74.1). Clusters 0 and 3 had a Revenue of 10%, and both of them had a very low value of S, even though their values of L were not that small. Cluster 0 had a high L (7.7), moderate R (4.5), and F (26.3) with very small S (0.3), and Cluster 3 had a moderate L (7.0), R (5.8), and F (37.0) but again, a very minimal amount of S (0.4). Clusters 2 and 4 had almost no Revenue, to be precise they were 4% and 3% respectively. Both of these clusters had one thing in common and that is, the value of S was negligibly small as well. For Cluster 2, S was 0.012156 and for Cluster 4, it was 0.018869. The other features for Cluster 2 were L (5.1), R (5.4), and F (16.4), and for Cluster 4 were L (6.7), R (6.2), and F (18.5). From these two clusters, it can be said that, even though the customers had a relationship with the website for six to seven months, it does not guarantee their purchase from the platform.

3) K-MEDOIDS ANALYSIS USING AUTOENCODER

For the autoencoder, the optimal number of clusters was found to be 4, similar to the K-Means and PCA combinations.

Fig.18 shows Cluster 2 had the most Revenue of 60%, with the highest value of S (43.3) and the second highest value of F (40.6). In this cluster, L was close to 9 months and R was 3 months. The revenue of Cluster 1 was 20% with a low S (2.7), even though it had the highest F (144.4), second highest L, which was more than 9.2 months, and moderate R, which was close to 4 months. The other two clusters, Cluster 0 and Cluster 3 had only 10% Revenue. Interestingly, Cluster 3 had almost 11 months of Length but less than 2 months of Recency, whereas Cluster 0 had only 3.6 months of Length but a Recency was 7.6. The remaining features for Cluster 0 were F (20.7), S (2.7), and for Cluster 3, they were and F (24.7), S (0.3) respectively.

C. COMPARATIVE ANALYSIS AMONG LR, LF, LRF, AND LRFS

To verify the efficiency of the proposed LRFS model, it must be compared to other RFM models to demonstrate whether it outperforms them or not. Since the used dataset did not have monetary features (the product price, total generated revenue amount, etc.), the monetary portion in RFM and modified RFM models have been omitted. Keeping all these in mind, a comparative analysis was conducted between the proposed LRFS and the variations of the RFM model excluding the feature M, such as, LRF (Length Recency Frequency), LR (Length Recency), and LF (Length Frequency). In the case of LR and LF models, dimensionality

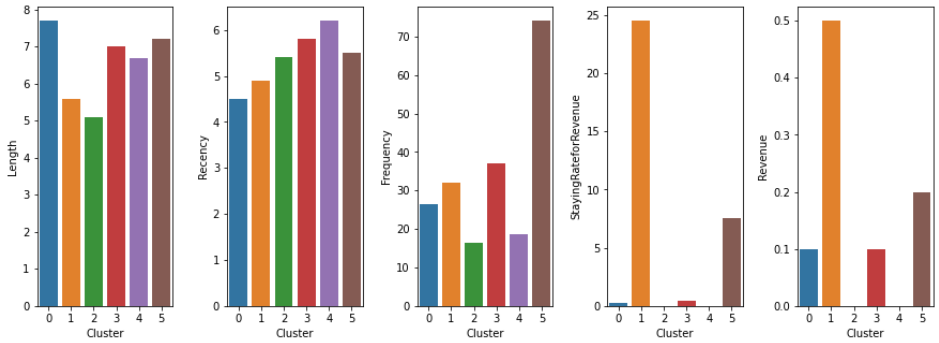

FIGURE 17. K-medoids clustering analysis of LRFS model (t-SNE).

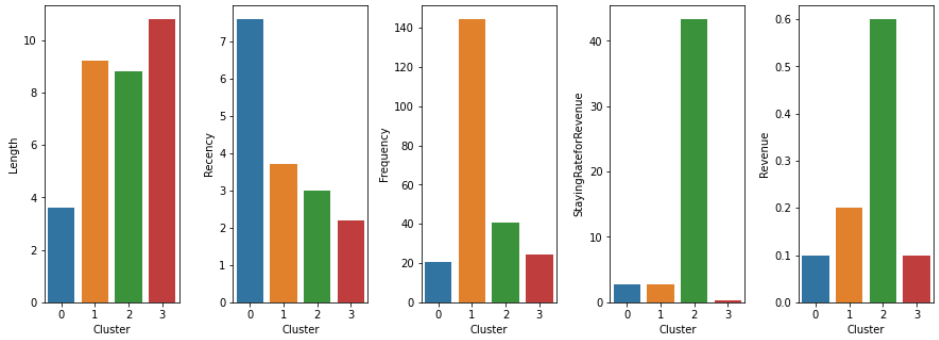

FIGURE 18. K-medoids clustering analysis of LRFS model (Autoencoder).

reduction techniques were not applied to them as they have only two features. To compare several models against each other, a common algorithm should be chosen so that the differences can be distinguished clearly. In the previous section of clustering analysis, it was seen that the K-Means algorithm performed better overall for the LRFS model. Hence, K-Means has been used for comparison purposes.

Figure 19 shows the differences among these 4 models in terms of their generated portion of Revenue for each cluster. The x-axis shows each cluster and to differentiate the models, four separate colors have been used to plot the bars. The y-axis shows the percentage of Revenue generated from each clustered group. The LR model groups the customer base into three clusters where their Revenue is 26%, 19%, and 10% respectively. The LF shows a little better cluster assignment by segmenting the data in such a way that the Revenue generation is separated into 16%, 26%, 13%, and 35%. The LRF model has shown its Revenue percentage to be 13%, 17%, and 34%. It has been seen that the other three models LR, LF, and LRF have grouped customers with Revenue of not more than 35% and not less than 10%. On the other hand, in LRFS, cluster generation makes more sense. In Cluster 0, the least amount, only 1% of the customer base has purchased something from the website, then Cluster 1 shows the revenue has increased to 14%, Cluster 2 has a customer of 33% and the fourth one, Cluster 3 groups the customer base from which 81% has completed transaction. The main difference among

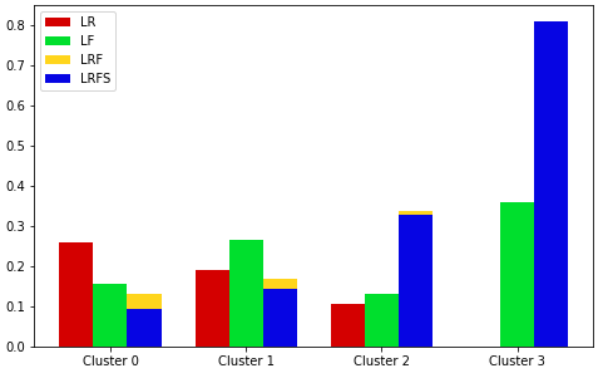

FIGURE 19. Differences of revenue among different clusters generated by LR, LF, LRF, and LRFS.

these models is the Revenue generation of the fourth cluster of the LRFS model. The reason behind this could be that the used dataset overall did not contain a lot of clients who have contributed to more Revenue generation. As the number of customers who ended up purchasing was small, the other three models may have overlooked this information while the LRFS model was successful in distinguishing the range of customers with low to high Revenue generation.

Table 1 summarizes the means of each cluster for each model, and the additional Revenue column has been added to examine which cluster led to more percentage of Revenue generation. Looking at the percentages separately, these four

TABLE 1. Comparison among LR, LF, LRF, and LRFS.

	Cluster	L	R	F	S	% of Revenue
LR	0	1.000000	2.139229	-	-	26%
	1	10.828845	2.171155	-	-	19%
	2	4.120000	8.520486	-	-	10%
LF	0	10.855578	-	24.366285	-	16%
	1	9.198840	-	111.946147	-	26%
	2	3.543425	-	20.801671	-	13%
	3	9.554878	-	296.365854	-	35%
LRF	0	3.597152	7.527638	23.044947	-	13%
	1	10.838451	2.161549	31.008233	-	17%
	2	9.508929	3.421131	180.891369	-	34%
LRFS	0	3.649919	7.609199	22.775743	2.263332	1%
	1	10.841568	2.158432	30.895384	2.999515	14%
	2	9.527132	3.431008	182.803101	5.114335	33%
	3	5.606947	4.524680	35.495430	72.663165	81%

models can be ranked from lowest to highest as LR < LRF < LF < LRFS. From the analysis, it can be stated that only knowing the total duration of the customer's association with the website and the recent visit is not enough to distinguish groups of customers. Rather, how frequent a customer is, can play a vital role to detect a potential customer. LF showing better cluster generation than LRF proves that increasing the dimension of the model will not necessarily lead to the improvement of its performance, but the significance of the individual feature should have more focus. LRFS has made a notable improvement in this regard by introducing the new dimension "S". It was the reason why the LRFS model was able to group customers into the fourth cluster.

D. CUSTOMER CLASSIFICATION AND RELATIONSHIP MATRIX

After segmenting the customers, firms often use customer analysis methods to understand the segmented customer base in a better way. So far several analysis techniques have been used by the researchers for market segmentation purposes. Factors like motives, attitudes, usage patterns, and user preferences have been taken into account, which is known as database marketing [29]. One of those methods is called Customer Profit Analysis or CPA. According to the studies by Raaij, Vernooij, and Triest [42], it is the process of allocating revenues and expenses to a particular customer or groups of customers in order to determine their profits. The ultimate goal of any business owner would be to ensure more profits, and CPA can help in this regard by demonstrating which group of customers is yielding more profits and how much cost it requires to serve them. Since the used dataset in this research does not contain the "Profit" or any attribute directly related to the profit made by each customer, the feature "Revenue" were be used instead. There are four types of customer groups in CPA, namely "Passive", "Carriage Trade", "Bargain Basement", and "Aggressive". The passive customer generates high revenues without requiring many costs. They can be considered to be the most loyal clients and the firm has to show appreciation to them and take special care of them so that they retain their loyalty. The Carriage Trade customers also spend money by purchasing products, but the firm has to spend for them as well. If only their revenue exceeds the cost to serve them, the business will get profits. Another customer type is Bargain Basement, who do not spend much using the website, but at the same time, they do not need high costs. The remaining group is named Aggressive since they require high costs but they do not generate much revenue. So after classifying the customer groups, the main target for a company would be to increase the number of Passive customers, that way the business will flourish more.

Another method to investigate customer relationships is known as Customer Relationship Matrix, or CRA, created by Wu, Lin and Liu [39] which was inspired by Marcus's Customer Value Matrix [29], or CVA. In this approach, the authors have taken the suggestion from Marcus that the Length and Recency of customer relationships with company should be amalgamated. They have omitted the feature "Monetary" from their proposed model since their monetary value was fixed. They have demonstrated a 2×2 matrix by combining the two attributes "Length" and "Recency". The customer Relationship Matrix has four customer types, such as "Loyal", "Potential", "New", and "Uncertain". If a customer group has increasing L but decreasing R, they are considered to be Potential Customers. On the other hand, if the opposite happens, they are considered to be New customers. If a customer has been spending a long time with the firm and also made a visit recently, they are Loyal customers. And lastly, if a customer is not active, both L and R would be decreased and they are marked as uncertain customers.

1) CLUSTER ANALYSIS USING COMBINED CUSTOMER CLASSIFICATION AND RELATIONSHIP MATRIX

In previous sections, it was visualized that in terms of clustering, overall K-Means has exhibited better clusters. The K-Medians and t-SNE combination worked pretty good

TABLE 2. Average of L, R, F, S for the dataset.

Feature	Average of Components
L	6.558637
R	5.348013
F	34.550203
S	5.793039
Revenue	0.154745

as well. In terms of dimensionality reduction, Autoencoder showed promising results, even though the input dimension were not that high. The performance of PCA was more compatible with K-Means than K-Medians. Among all clusters, K-Medoids with t-SNE combination was chosen for the Customer Classification Matrix, since it generated the most number of distinct clusters and resulted in all four customer traits in terms of CPA and also CRM. Based on this, the firms will be able to classify the type of customers and can plan their marketing strategies accordingly.

The average values for each feature of the LRFS model for the entire dataset are calculated and shown in Table 2. Feature L shows that for six to seven months on average the customers were associated with the website and R shows the average recent time the customers have visited the website is more than five months. F shows that the average visit to the pages of the website was a little more than 34 and S shows the average Staying Rate for Revenue was close to 6. The overall revenue of the customers was only 0.15, which means almost 15% of all the customers ended up purchasing something. The Customer Classification Matrix and the Customer Relationship Matrix are combined in Table 3. For each of the clusters, the values of L, R, F, and S were also computed. The mean values of features of the whole dataset and the mean values of features from each cluster were compared against each other. The arrow defines the status of whether each component is greater or less than the average value or not. If the average of each component L, R, F, S of each cluster were found to be greater than the average of L, R, F, S of the complete dataset, an up arrow (↑) was drawn and if it was the opposite, a down arrow (↓) was added in the matrix. The same method was followed for the Revenue column as well. In the last column, the customer type (Aggressive, Bargain Basement, Carriage Trade, Passive) is allocated to each cluster.

Cluster 0 shows L(↓) R(↑) F(↓) S(↓), which means this group of customers visited the website a long time ago but have not stayed on the website that much. Also, they have not visited many pages and have not spent much either. So they are considered to be **Bargain Basement, Potential Customers**. Some customers may buy products occasionally, but whenever they are buying them, they would buy a large amount. Cluster 1 shows them as **Passive Customers**, as do not need much maintenance, and they do not visit the website frequently, but whenever they do, they end up purchasing products. However, by the defining Customer Relationship Matrix, Cluster 1 falls into the category of **Uncertain Customers**, since they have a low number of L and R. Cluster 2 has L(↑) R(↓) F(↑) S(↓), although the customers in this group have recently visited the website, they still did not contribute that much, so they are assumed to be the group of **New Customers**. Possibly, they are the people who came to visit the website after hearing some reviews from their friends. They are new to the website and have not purchased anything yet, so they are labeled as **Cluster 2 (Bargain Basement, New Customer)**. Still, there is a possibility to turn them into potential, or even loyal customers if proper strategies are taken. Some free points or discount coupons could be granted to them for the first time registering on the website.

In Cluster 4, L(↑) R(↑) F(↓) S(↓), although the customers have a long term relationship with the website, and have recently visited the website as well, they have less number of visits. These three clusters also generated low Revenue, hence, they fall into **Bargain Basement**. An example of these type of customers could be household owners who prefer to buy regular products on a routine basis. It could be daily groceries such as bread-butter, dairy items, or any other household goods like toothpaste, dishwasher, etc. They are not necessarily spending a lot of money, since the types of products they are buying are not that expensive. However, they have been visiting the website for a long time, and the last visit is also recent. It means they are loyal, and it does not cost much to retain them. They do not need to interact with the website that much, as they already know what they are looking for. This is the reason why their page visits are low. These types of habitual customers have to be maintained regularly by giving them rewards for being consistent in their shopping. For example, for each purchase, they could be offered loyalty points, when they reach a certain limit of points, they can use them to buy something. Or, they could be presented a card on which they could get a stamp each time they are completing a transaction, and after they reach a certain number of stamps, they could get a free present or discount on their purchase. This type of strategy could encourage them to buy other products as well. Also, if there are any supplementary products, those should be introduced while they make their regular purchase. Special messages could be sent to them from time to time as well, to show appreciation for their loyalty.

In Cluster 3, although the LRF is (↑), S indicates (↓), which means the customers are visiting the website frequently for a long time, but not spending time or money on the purchase. Some of the customers could be more into trying out different products. However, they do not want to take risks by buying products without giving them much thought. Since the field of E-commerce or online shops is vast, they have a lot of options to look into. Hence, it could be difficult for them to make a concrete decision without taking a longer period of time. So, they could be mapped in **Aggressive, Loyal Customer**. On the other hand, there are people who do not want to take a lot of time to purchase something. They often act on impulse and are risk-takers. They would engage with the website at a

TABLE 3. Comparison amongst each feature and cluster using customer classification matrix.

Cluster	L	R	F	S	Revenue	Customer Type
0	↑	↓	↓	↓	↓	Bargain Basement, Potential Customer
1	↓	↓	↓	↑	↑	Passive, Uncertain Customer
2	↓	↑	↓	↓	↓	Bargain Basement, New Customer
3	↑	↑	↑	↓	↓	Aggressive, Loyal Customer
4	↑	↑	↓	↓	↓	Bargain Basement, Loyal Customer
5	↑	↑	↑	↑	↑	Carriage Trade, Loyal Customer

Notes: Here, up-arrow(↑)defines the status of whether each component is greater than the average value, and down-arrow(↓)defines the status of whether each component is less than the average value.

moderate amount. They would not purchase all the time, but they are willing to purchase if they are given something extra in return. For example, if there is a sale going on, they would prefer to take that opportunity. This kind of customer group falls into **Cluster 5 (Carriage Trade, Loyal Customer)**. It shows all of the features facing upwards, which implies it is the Carriage Trade category. Although the customers are generating Revenue, they are always visiting the websites, which requires the staff of the firm to take care of them whenever they need any assistance, also running a website has its own cost. One important detail is that for both of these cases, the customers are spending their time on the specific website. Even though the cost spent on them is higher than the revenue they are generating, they are still loyal to the shop. They have a very high number of page views, which means they are moving back and forth to various pages, most likely product related ones.

2) CUSTOMER SEGMENTATION USE CASE SCENARIOS WITH TEST CASE VISUALIZATION

From the perspective of the sellers, more frequent purchases by customers and gaining more profit would be the main target. In order to get an increasing number of frequent buyers they need to make sure that the users of their websites are satisfied. By getting better knowledge about the customer base, the company could offer them personalized rewards, advertisements, and recommendations. Also, according to the loyalty of the customers, the business owners could give them some kind of special benefits so that they keep coming back. The users who are not engaging that much could be offered lucrative deals to grab their attention. Depending on different scenarios, the grouping of customers may differ. So knowing which segments behave in what manner could allow one to effectively allocate budgets to them. In this section, some probable scenarios would be discussed and visualized using the LRFS model with K-Means Clustering and PCA implementation. Three datapoints were taken to visualize where they are going to be assigned in the generated cluster groups. Table 4 contains three test cases along with their mapping to the customer type.

a: TEST CASE 1

Suppose, Customer X is one of some university students planning to go on a study tour. She and her friends need to buy clothes, backpacks, water bottles, umbrellas, and some other relevant things. Sometimes they check various products on the website but they cannot make a decision immediately. So instead, they keep the relevant pages on their bookmark list and decide to check them again later. They simultaneously look into different websites to check whether they find any cheaper option. These types of customers perfectly fit in **"Bargain Basement, Potential Customer"** category.

TABLE 4. Some test cases using the LRFS model with K-means clustering and PCA.

Test Case	L	R	F	S	Customer Type
Customer X:	↑	↓	↓	↓	Bargain Basement, Potential Customer
Customer Y:	↓	↓	↑	↑	Carriage Trade, Uncertain Customer
Customer Z:	↓	↓	↓	↑	Passive, Uncertain Customer

Currently, they are not spending a lot of money, but they have been associated with the website for a long time. Also, they are not just sticking to one website, rather, they are trying out different shops. As they are often conscious about the pricing of the products, they go for more budget friendly options available. It is important to retain these types of customers because they have the potential to become loyal customers later on. Otherwise, they will move on to rival online shops and might not come back again. Appropriate strategies have to be taken to make sure that they do not lose interest. Some strategies could be to offer a student friendly package such as buying 3 products at the same time that could cost them less than buying individually. Seasonal buy one get one offer could be another good option. Even on their desired product related pages, some other products that are related to traveling such as sunglasses, tracking shoes, etc. could be shown as recommended products. Here, the Length is high (↑), where as other features R (↓), F (↓), and S (↓) are lower compared to the average values of these features. The Fig.20 shows that the datapoint of Customer X falls into the **"Bargain Basement and Potential Customer"** category.

b: TEST CASE 2

Customer Y came to know about the website after hearing positive reviews from a friend and decided to give it a go. He noticed a sale has been going on the website, and even though he was not planning to buy right away, he browsed a lot of pages and finally ended up in ordering some products.

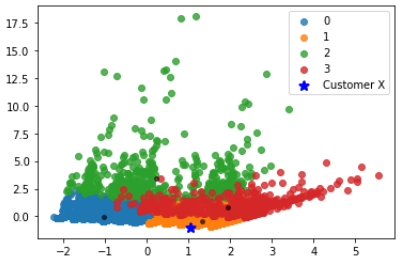

FIGURE 20. Customer X test case.

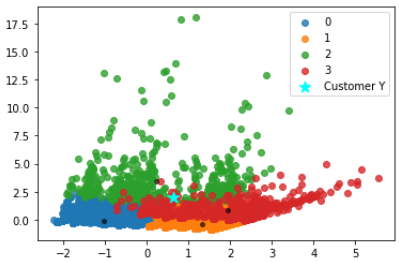

FIGURE 21. Customer Y test case.

This type of customers falls into **"Bargain Basement and Potential Customer"** category, shown in Fig.21.

Looking into each of the components, the Length and Recency values are lower, which means whether the customer is going to visit the website again is uncertain. However, both Frequency and Staying Rate for Revenue had higher values. These customers are mainly interested in the offers or discounts. So the expense to serve them could cost more to the company compared to the revenue generated by them. Although it would cost more to retain these customers, they are still contributing to the revenue, so appropriate measures should be taken so that they turn into Loyal customer group from the Uncertain group. As they are likely to be quite engaged in the shopping process, but struggle to choose the product they desire, they often could seek confirmation that their decision was the right one. In these cases, They should be reminded of the benefits of the items they are considering to buy. Other alternative options or related products could be shown to them as recommendations as well. For the users who are attracted to lucrative offers, some approaches could be taken such as free giveaways, points for daily login, free presents for birthdays, and so on.

c: TEST CASE 3

Customer Z works at a corporate job and wants to buy something for their family whenever a holiday is approaching. The user is not short in terms of budget, however, is incapable of spending too much time browsing various pages on the website. So, he made the decision quickly without checking out many products and spent a decent amount of money purchasing gifts. These types of clients are the perfect example of **Passive, Uncertain Customer**.

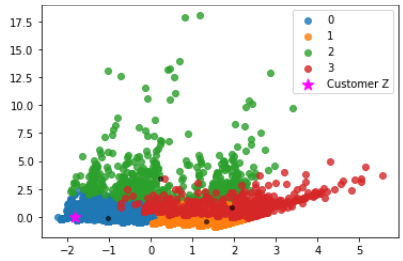

FIGURE 22. Customer Z test case.

Here, only feature S is high (↑), and the other three features Length, Recency, and Frequency are low (↓). As their purchase behavior is occasional, their total number of page visits to the website is not high and they have not visited the website recently either. Although their interaction with the website is not much, still they are contributing to a significant amount of revenue generation. Therefore, this group of people plays an important role in this aspect. In such cases, they could be given coupons for buying large amounts of products. Also, personalized messages or emails could be sent to them on special occasions so that the chance that they might visit the website at a later time would increase. Fig.22 shows that the datapoint of Customer Z falls into a separate cluster, mapped as **Passive, Uncertain Customer** category.

V. CONCLUSION

The objective of this paper was to present a detailed methodology for carrying out a new LRFS model for online shoppers by taking additional variable S into account. In the vast field of customer segmentation research, this was probably the first work in which a link has been formed between Google Analytics features like Exit Rates or Page Values and the conventional LRF model. After analyzing each of the components of LRFS against Revenue, a feature from the original dataset, some interesting insights have been found about the relationship between each of the components of the model. For example, if a customer has a long association with the website, it does not mean that the customer is going to be still interested and spend more money on the website. Applying different dimensionality reduction approaches along with three different clustering algorithms had worked very well for some combinations, while not suitable for others. On the other hand, the K-Means algorithm has performed well with all dimensionality reduction techniques. K-Median with t-SNE and Autoencoder showed decent performance whereas K-Medoids with PCA together did have more overlappings comparatively.

While conducting the research, there were some constraints associated with the used dataset. The main drawback of this research was the dataset did not contain some key features required for LRFS analysis. For instance, the dataset did not have any DateTime feature. Calculating Length and Recency would have been more precise if there were information about the customer's access date to the website.

The number of products as well as the amount of profit were also missing from the dataset. If these two features existed, profit analysis could have been carried out. The relationship between purchased quantity, profit, and the proposed LRFS model would have led to some better insights into the business. Another limitation was Revenue column was binary, hence it was not possible to use this feature to calculate how much a particular customer has spent.

For future works, experimenting with the proposed LRFS model on different datasets would be a good idea. The dataset needs to contain Google analytics features and also include some other features about the purchase like DateTime, product quantity, price, and profit. In the case of dimensionality reduction approaches, the Autoencoder model could be improved by tweaking its hyperparameters. Clustering algorithms other than the ones that have been implemented in this work such as Hierarchical Agglomerative Clustering, Mini-batch K-Means Clustering, Mean Shift Algorithm, OPTICS Algorithm, etc. could be explored as well. Moreover, recently researchers have already proposed a new clustering method based on deep learning namely DEC (Deep Embedded Clustering), which could be carried out by integrating it with the LRFS model. And last but not least, the hybridization of algorithms could open new doors of possibilities.

REFERENCES

[1] I. K. Rachmawati, M. Bukhori, F. Nuryanti, and S. Hidayatullah, "Collaboration technology acceptance model, subjective norms and personal innovations on buying interest online," *Int. J. Innov. Sci. Res. Technol.*, vol. 5, no. 11, pp. 115–122, 2020.

[2] C. I. Agustyaningrum, M. Haris, R. Aryanti, and T. Misriati, "Online shopper intention analysis using conventional machine learning and deep neural network classification algorithm," *Jurnal Penelitian Pos dan Informatika*, vol. 11, no. 1, pp. 89–100, Nov. 2021.

[3] J. A. Al-Gasawneh, M. H. Al-Wadi, B. M. Al-Wadi, B. E. Alown, and N. M. Nuseirat, "The interaction effect of comprehensiveness between social media and online purchasing intention in Jordanian pharmacies," *Int. J. Interact. Mobile Technol.*, vol. 14, no. 15, p. 208, Sep. 2020.

[4] S. Ahsain and M. A. Kbir, "Predicting the client's purchasing intention using machine learning models," in *Proc. E3S Web Conf.*, vol. 351. Les Ulis, France: EDP Sciences, 2022, p. 01070.

[5] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022.

[6] A. Balkaya, E. Tüzünkan, K. Ayaz, Y. Akçay, F. Abut, M. F. Akay, S. Erdem, and A. Alsaç, "Developing customer segmentation models for digital marketing campaigns using machine learning," in *Proc. 6th Int. Medit. Sci. Eng. Congr. (IMSEC)*, Alanya, Türkiye, 2022, pp. 697–701, Paper 318.

[7] B. Han, M. Kim, and J. Lee, "Exploring consumer attitudes and purchasing intentions of cross-border online shopping in Korea," *J. Korea Trade*, 2018, doi: 10.1108/JKT-10-2017-0093.

[8] M. A. Khan and S. Khan, "Service convenience and post-purchase behaviour of online buyers: An empirical study," *J. Service Sci. Res.*, vol. 10, no. 2, pp. 167–188, Dec. 2018.

[9] K. Fang, Y. Jiang, and M. Song, "Customer profitability forecasting using big data analytics: A case study of the insurance industry," *Comput. Ind. Eng.*, vol. 101, pp. 554–564, Nov. 2016.

[10] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries," *Appl. Soft Comput.*, vol. 73, pp. 816–828, Dec. 2018.

[11] S. Ahsain and M. A. Kbir, "Data mining and machine learning techniques applied to digital marketing domain needs," in *Proc. 3rd Int. Conf. Smart City Appl.* Cham, Switzerland: Springer, 2021, pp. 730–740.

[12] M. Mohammadzadeh, Z. Z. Hoseini, and H. Derafshi, "A data mining approach for modeling churn behavior via RFM model in specialized clinics case study: A public sector hospital in Tehran," *Proc. Comput. Sci.*, vol. 120, pp. 23–30, Jan. 2017.

[13] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA," *J. Big Data*, vol. 7, no. 1, pp. 1–23, Dec. 2020.

[14] E. H. S. Addin, N. Admodisastro, S. N. S. M. Ashri, A. Kamaruddin, and Y. C. Chong, "Customer mobile behavioral segmentation and analysis in telecom using machine learning," *Appl. Artif. Intell.*, vol. 36, no. 1, pp. 1–21, Dec. 2022.

[15] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, "Integrated churn prediction and customer segmentation framework for Telco business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021.

[16] S. P. Nguyen, "Deep customer segmentation with applications to a Vietnamese supermarkets' data," *Soft Comput.*, vol. 25, no. 12, pp. 7785–7793, Jun. 2021.

[17] R. Shirole, L. Salokhe, and S. Jadhav, "Customer segmentation using RFM model and K-means clustering," *Int. J. Sci. Res. Sci. Technol.*, vol. 8, pp. 591–597, Jun. 2021.

[18] A. Alghamdi, "A hybrid method for customer segmentation in Saudi Arabia restaurants using clustering, neural networks and optimization learning techniques," *Arabian J. Sci. Eng.*, vol. 48, no. 2, pp. 2021–2039, Feb. 2023.

[19] X. Wang, C. Zhou, Y. Yang, Y. Yang, T. Ji, J. Wang, J. Chen, and Y. Zheng, "Electricity market customer segmentation based on DBSCAN and k-means: A case on Yunnan electricity market," in *Proc. Asia Energy Electr. Eng. Symp. (AEEES)*, May 2020, pp. 869–874.

[20] M. Nilashi, S. Samad, B. Minaei-Bidgoli, F. Ghabban, and E. Supriyanto, "Online reviews analysis for customer segmentation through dimensionality reduction and deep learning techniques," *Arabian J. Sci. Eng.*, vol. 46, no. 9, pp. 8697–8709, Sep. 2021.

[21] E. Yadegaridehkordi, M. Nilashi, M. H. N. B. M. Nasir, S. Momtazi, S. Samad, E. Supriyanto, and F. Ghabban, "Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques," *Technol. Soc.*, vol. 65, May 2021, Art. no. 101528.

[22] R. van Leeuwen and G. Koole, "Data-driven market segmentation in hospitality using unsupervised machine learning," *Mach. Learn. Appl.*, vol. 10, Dec. 2022, Art. no. 100414.

[23] E. Umuhoza, D. Ntirushwamaboko, J. Awuah, and B. Birir, "Using unsupervised machine learning techniques for behavioral-based credit card users segmentation in Africa," *SAIEE Afr. Res. J.*, vol. 111, no. 3, pp. 95–101, Sep. 2020.

[24] A. Abdulhafedh, "Incorporating k-means, hierarchical clustering and PCA in customer segmentation," *J. City Develop.*, vol. 3, no. 1, pp. 12–30, 2021.

[25] I. Pawełoszek, "Customer segmentation based on activity monitoring applications for the recommendation system," *Proc. Comput. Sci.*, vol. 192, pp. 4751–4761, Jan. 2021.

[26] K. M. Manero, R. Rimiru, and C. Otieno, "Customer behaviour segmentation among mobile service providers in Kenya using k-means algorithm," *Int. J. Comput. Sci. Issues*, vol. 15, no. 5, pp. 67–76, 2018.

[27] B. Turkmen, "Customer segmentation with machine learning for online retail industry," *Eur. J. Social Behav. Sci.*, vol. 31, no. 2, pp. 111–136, Apr. 2022.

[28] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Inf. Process. Manage.*, vol. 59, no. 6, Nov. 2022, Art. no. 103085.

[29] C. Marcus, "A practical yet meaningful approach to customer segmentation," *J. Consum. Marketing*, vol. 15, no. 5, pp. 494–504, Oct. 1998.

[30] M. Tavakoli, M. Molavi, V. Masoumi, M. Mobini, S. Etemad, and R. Rahmani, "Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: A case study," in *Proc. IEEE 15th Int. Conf. e-Bus. Eng. (ICEBE)*, Oct. 2018, pp. 119–126.

[31] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking—An effective approach to customer segmentation," *J. King Saud Univ., Comput. Inf. Sci.*, vol. 33, no. 10, pp. 1251–1257, Dec. 2021.

[32] A. Sheshasaayee and L. Logeshwari, "An efficiency analysis on the TPA clustering methods for intelligent customer segmentation," in *Proc. Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Feb. 2017, pp. 784–788.

[33] C. Qian, M. Yang, P. Li, and S. Li, "Application of customer segmentation for electronic toll collection: A case study," *J. Adv. Transp.*, vol. 2018, pp. 1–9, Aug. 2018.

[34] R. H. Khan, D. F. Dofadar, and M. G. R. Alam, "Explainable customer segmentation using K-means clustering," in *Proc. IEEE 12th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Dec. 2021, pp. 639–643.

[35] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," *J. Retailing Consum. Services*, vol. 61, Jul. 2021, Art. no. 102566.

[36] I.-C. Yeh, K.-J. Yang, and T.-M. Ting, "Knowledge discovery on RFM model using Bernoulli sequence," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5866–5871, Apr. 2009.

[37] F. Yoseph and M. AlMalaily, "New market segmentation methods using enhanced (RFM), CLV, modified regression and clustering methods," *Int. J. Comput. Sci. Inf. Technol.*, vol. 11, no. 1, pp. 43–60, Feb. 2019.

[38] J.-T. Wei, S.-Y. Lin, C.-C. Weng, and H.-H. Wu, "A case study of applying LRFM model in market segmentation of a children's dental clinic," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5529–5533, Apr. 2012, doi: 10.1016/j.eswa.2011.11.066.

[39] 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," *Sci. World J.*, vol. 2014, Jun. 2014, Art. no. 685495, doi: 10.1155/2014/685495.

[40] R. Mahfuza, N. Islam, M. Toyeb, M. A. F. Emon, S. A. Chowdhury, and M. G. R. Alam, "LRFMV: An efficient customer segmentation model for superstores," *PLoS ONE*, vol. 17, no. 12, Dec. 2022, Art. no. e0279262.

[41] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Comput. Appl.*, vol. 31, no. 10, pp. 6893–6908, Oct. 2019.

[42] E. M. van Raaij, M. J. A. Vernooij, and S. van Triest, "The implementation of customer profitability analysis: A case study," *Ind. Marketing Manage.*, vol. 32, no. 7, pp. 573–583, Oct. 2003.

[43] B. Stone, *Successful Direct Marketing Methods*. Lincolnwood, IL, USA: NTC Bus. Books, 1995, pp. 37–57.

[44] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Proc. Comput. Sci.*, vol. 3, pp. 57–63, Jan. 2011.

[45] C.-C. Shen and H.-M. Chuang, "A study on the applications of data mining techniques to enhance customer lifetime value," *WSEAS Trans. Inf. Sci. Appl.*, vol. 6, no. 2, pp. 319–328, 2009.

[46] J. R. Miglautsch, "Thoughts on RFM scoring," *J. Database Marketing Customer Strategy Manage.*, vol. 8, no. 1, pp. 67–72, Aug. 2000.

[47] Y.-Y. Shih and C.-Y. Liu, "A method for customer lifetime value ranking—Combining the analytic hierarchy process and clustering analysis," *J. Database Marketing Customer Strategy Manage.*, vol. 11, no. 2, pp. 159–172, Dec. 2003.

[48] P. A. Sarvari, A. Ustundag, and H. Takci, "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis," *Kybernetes*, vol. 45, no. 7, pp. 1129–1157, Aug. 2016.

[49] R. A. Soeini and E. Fathalizade, "Customer segmentation based on modified RFM model in the insurance industry," in *Proc. 4th Int. Conf. Mach. Learn. Comput. (IPCSIT)*, vol. 25. Singapore: IACSIT Press, 2012, pp. 104–108.

[50] T. Belhadj, "Customer value analysis using weighted RFM model: Empirical case study," *Al Bashaer Econ. J.*, vol. 7, no. 3, p. 932, 2021.

[51] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: A case study," *Marketing Intell. Planning*, vol. 35, no. 4, pp. 544–559, May 2017.

[52] A. Sheikh, T. Ghanbarpour, and D. Gholamiangonabadi, "A preliminary study of fintech industry: A two-stage clustering analysis for customer segmentation in the B2B setting," *J. Bus.-to-Bus. Marketing*, vol. 26, no. 2, pp. 197–207, Apr. 2019.

[53] J. Zhou, J. Wei, and B. Xu, "Customer segmentation by web content mining," *J. Retailing Consum. Services*, vol. 61, Jul. 2021, Art. no. 102588, doi: 10.1016/j.jretconser.2021.102588.

[54] J. Guo, Z. Gao, N. Liu, and Y. Wu, "Recommend products with consideration of multi-category inter-purchase time and price," *Future Gener. Comput. Syst.*, vol. 78, pp. 451–461, Jan. 2018.

[55] L. Meyer-Waarden, "The influence of loyalty programme membership on customer purchase behaviour," *Eur. J. Marketing*, vol. 42, no. 1/2, pp. 87–114, Feb. 2008.

[56] Y.-T. Kao, H.-H. Wu, H.-K. Chen, and E.-C. Chang, "A case study of applying LRFM model and clustering techniques to evaluate customer values," *J. Statist. Manage. Syst.*, vol. 14, no. 2, pp. 267–276, Mar. 2011, doi: 10.1080/09720510.2011.10701555.

[57] Google. *How Page Value is Calculated—Analytics Help*. Accessed: May 25, 2024. [Online]. Available: https://support.google.com/analytics/answer/2695658?hl=en#:~:text=Page%20Value%20is%20the%20average,more%20to%20your%20site's%20revenue

RIYO HAYAT KHAN received the B.Sc. and M.Sc. degrees in computer science and engineering from BRAC University, Bangladesh, in 2021 and 2023, respectively. During her undergraduate years, she was a Student Tutor, and after her graduation, she was a Contractual Lecturer with the Department of Computer Science and Engineering, BRAC University. She has experience in writing scientific research articles and implementing various machine learning algorithms using Python. She has several publications at international conferences and also presented some of those works. Her research interests include artificial intelligence, machine learning, and deep learning domains.

DIBYO FABIAN DOFADAR is currently pursuing the master's degree with BRAC University. He is also a full-time Lecturer with the Department of Computer Science and Engineering, BRAC University. Previously, he was a Contractual Lecturer and a Student Tutor with the Department of Computer Science and Engineering. Out of many, one of his favorite activities is to solve and optimize various problems through coding. His research interests include the artificial intelligence, machine learning, and deep learning domains and has a few publications in these fields.

MD GOLAM RABIUL ALAM (Member, IEEE) received the B.S. degree in computer science and engineering and the M.S. degree in information technology and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2017. He was a Postdoctoral Researcher with the Computer Science and Engineering Department, Kyung Hee University, from March 2017 to February 2018. He is currently a Professor with the Computer Science and Engineering Department, BRAC University, Bangladesh. His research interests include healthcare informatics, mobile cloud and Edge computing, ambient intelligence, and persuasive technology. He is a member of IEEE IES, CES, CS, SPS, CIS, and ComSoc. He received several best paper awards from prestigious conferences.

MOHAMMAD SIRAJ (Senior Member, IEEE) received the Bachelor of Engineering degree in electronics and communication engineering from Jamia Millia Islamia, New Delhi, the Master of Engineering degree in computer technology and applications from Delhi College of Engineering, Delhi, India, and the Ph.D. degree from Universiti Technologi Malaysia. He was a Scientist C with the Defense Research and Development Organization, India. Currently, he is an Assistant Professor in electrical engineering with King Saud University. His research interests include cognitive wireless networks, wireless mesh networks, sensor networks, the Internet of Things, cloud computing, and telecom optical networks. He has numerous peered publications in well-known international journals and conferences. He is a reviewer of many well-known international journals and conferences.

MD RAFIUL HASSAN received the Ph.D. degree in computer science and software engineering from the University of Melbourne, Australia, in 2007. He is currently an Associate Professor in computer science with Central Connecticut State University, USA. His research interests include artificial intelligence, machine learning, and computational intelligence, with a focus on developing new data mining and machine learning techniques for big data analysis, cyber-security, and the IoT. He is the author of around 50 papers published in recognized international journals and conference proceedings. He is a member of Australian Society of Operations Research (ASOR) and IEEE Computer Society; and is involved in several Program Committees of international conferences. He also serves as a Reviewer of a number for renowned journals, such as *BMC Cancer*, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Neurocomputing*, *Applied Soft Computing*, *Knowledge and Information Systems*, *Current Bioinformatics*, *Information Sciences*, *Digital Signal Processing*, and IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS AND COMPUTER COMMUNICATIONS.

MOHAMMAD MEHEDI HASSAN (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in February 2011. He is currently a Professor with the Information Systems Department, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia. He has authored or co-authored more than 365 publications including refereed journals (333 SCI/ISI-indexed journal articles, 42 conference papers, one book, and two book chapters. His research interests include cloud computing, edge computing, the Internet of Things, body sensor networks, big data, deep learning, mobile cloud, smart computing, wireless sensor networks, 5G networks, and social networks. He has served as the Chair and a Technical Program Committee Member for numerous reputed international conferences/workshops, such as IEEE CCNC, ACM BodyNets, and IEEE HPCC. He is one of the top 2% Scientists in the world in the networking and telecommunication field.