TheAnalyzer: Clustering-Based System for Improving Business Productivity by Analyzing User Profiles to Enhance Human-Computer Interaction

D. S. A. Nanayakkara, K. J. P. G. Perera

***Abstract*—**E-commerce platforms have revolutionized the shopping experience, offering convenient ways for consumers to make purchases. To improve interactions with customers and optimize marketing strategies, it is essential for businesses to understand user behavior, preferences, and needs on these platforms. This paper focuses on recommending businesses to customize interactions with users based on their behavioral patterns, leveraging data-driven analysis and machine learning techniques. Businesses can improve engagement and boost the adoption of e-commerce platforms by aligning behavioral patterns with user goals of usability and satisfaction. We propose TheAnalyzer, a clustering-based system designed to enhance business productivity by analyzing user-profiles and improving human-computer interaction. TheAnalyzer seamlessly integrates with business applications, collecting relevant data points based on users' natural interactions without additional burdens such as questionnaires or surveys. It defines five key user analytics as features for its dataset, which are easily captured through users' interactions with e-commerce platforms. This research presents a study demonstrating the successful distinction of users into specific groups based on the five key analytics considered by TheAnalyzer. With the assistance of domain experts, customized business rules can be attached to each group, enabling TheAnalyzer to influence business applications and provide an enhanced personalized user experience. The outcomes are evaluated quantitatively and qualitatively, demonstrating that utilizing TheAnalyzer's capabilities can optimize business outcomes, enhance customer satisfaction, and drive sustainable growth. The findings of this research contribute to the advancement of personalized interactions in e-commerce platforms. By leveraging user behavioral patterns and analyzing both new and existing users, businesses can effectively tailor their interactions to improve customer satisfaction, loyalty and ultimately drive sales.

***Keywords*—**Data clustering, data standardization, dimensionality reduction, human-computer interaction, user profiling.

I. Introduction

E-COMMERCE platforms have experienced significant growth and have revolutionized the way people shop, providing convenient and efficient ways to purchase products and services. As the popularity of online shopping continues to rise, these platforms face a significant challenge that sets them apart from physical stores [1]. Unlike their brick-and-mortar counterparts, where sales people and social pressure can influence customer decisions to stay, e-commerce platforms face the reality that users have the power to initiate and terminate interactions with just one click [2]. To succeed in this dynamic landscape, understanding user behavior and preferences is crucial for businesses to enhance customer interactions, deliver personalized experiences, and optimize their marketing strategies.

By tailoring interactions with users based on their behavioral patterns, e-commerce platforms can increase sales, improve customer satisfaction, and foster long-term customer loyalty [3]. Thus, the ability to detect and analyze user behavioral patterns in e-commerce platforms has become increasingly important [4]. User behavior encompasses various actions performed by individuals while navigating the platform and these actions provide valuable insights into user preferences, decision-making processes, and overall satisfaction levels [5] as leveraging these behavioral patterns e-commerce platforms can gain a deeper understanding of their users, customize interactions, and provide personalized recommendations and offers [6].

Most of the existing research in this domain relies on manual methods, such as surveys which are commonly used to collect data on user behavior and preferences to gather data and insights. However, a potential drawback of these methods is their susceptibility to manipulation meaning that the respondents providing inaccurate or misleading information intentionally or unintentionally as social desirability bias can lead respondents to present themselves in a certain way, providing responses they believe are expected or socially acceptable, rather than reflecting their true behaviors and attitudes [7]. A more practical approach to gathering information for business purposes would involve collecting data from users while they interact with the platform, without their explicit knowledge. This method ensures that the user base remains relevant and the data collected accurately represent the actual usage of the platform for business purposes.

Other existing solutions comprise of models that only take purchase behaviors into consideration rather than overall cognitive behavior of the user [8]. One other limitation of existing research is their tendency to be domain-specific, focusing on particular industries such as agriculture [9], fashion e-commerce [10] etc. This narrow scope restricts the generalizability of findings and may not capture the full range of behavioral patterns and user experiences across different domains or industries.

Some other researches analyze customer behavior using web

D. S. A. Nanayakkara is a graduate of University of Westminster, London, W1B 2HW, UK (phone: 512-975-5113; e-mail: shaini.nanayakkara@gmail.com).

K. J. P. G. Perera is a graduate of University of Moratuwa, 10400, Sri Lanka (phone: 512-975-0005; e-mail: pravinda92@gmail.com).

logs [11] but stops in analysis state providing product recommendations to individual users [4] rather than offering a comprehensive understanding of customer behavior to business owners. This limitation hinders the ability to personalize marketing strategies [6] based on a holistic view of individual customers and their preferences.

In this research, we are going to bridge these gaps by leveraging data-driven analysis and machine learning techniques to identify behavioral patterns in e-commerce platforms. The study aims to recommend tailored interactions to users without relying on manual intervention. This approach ensures that recommendations that are tailored to individual users are given based on objective data so that businesses can refine their marketing strategies, optimize personalized recommendations, and enhance user engagement.

Rest of the paper is organized as follows. Section II introduces to e-commerce customer types, Z-Score normalization, Principal Component Analysis, elbow method, Silhouette Coefficient, K-means clustering, Human-Computer Interaction and the problem statement. Section III outlines the proposed technique. In Section IV, implementation details are provided, offering insights into how the proposed technique was implemented and executed. Section V focuses on the qualitative and quantitative evaluation of the proposed technique, discussing the results and findings obtained from the research. Section VI presents future work and concluding remarks, highlighting potential areas for further exploration and summarizing the key takeaways of the research.

II. Preliminaries

A. Impulsive Buyer

The key characteristic of this type is doing immediate purchases without a pre-plan to buy anything or having to fulfill a specific need [12]. This behavior involves quick decision-making and has a likelihood of the immediate product acquisition without forethought or deliberation [13]. Their buying patterns are spontaneous and arbitrary resulting them to be often unpredictable.

Impulsive buyers often succumb to the allure of immediate gratification and the excitement associated with making impulsive purchases. The sense of excitement or urgency during the buying process is the main motive of impulsive buyers.

B. Browsing Customer

This type of users can be described as web-window shoppers who engages in the activity of exploring and navigating through various products on the e-commerce platform without the specific intention of making an immediate purchase [14]. This type of customer usually spends more time researching and exploring different options, comparing prices, reading product descriptions, and assessing the available choices before making a final decision. They can be gathering information for a future purchase, seeking inspiration or simply exploring the platform as they are interested to know what it has to offer.

C. Loyal Customer

This is a type of user who makes repeated purchase over an extended period of time [15], which demonstrates a strong sense of allegiance, trust, and commitment to the brand or platform making consistent purchases showing brand support. Not only they as buyers produce consistent revenue to the business but also act as a strong marketing ally providing positive word of mouth about the business.

D. Bargain Hunter

This is a type of customer who is attracted to the platform because of deals, discounts or flash sales while prioritizing cost savings over product features. Their usual behavior comprises of browsing sale sections, using coupon codes, participating in loyalty programs, or waiting for seasonal promotions to maximize their savings. They also play a major role in putting pressure on pricing strategies and driving competition among businesses [16].

E. Z-Score Normalization (Standardization)

Z-score normalization (also known as standardization) standardizes the features by subtracting the mean and dividing by the standard deviation [17]. This results in features with zero mean and unit variance, centered around the mean of 0. It helps in achieving a standard distribution and maintains the relative relationships between the data points.

F. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used to simplify and understand complex datasets by reducing their dimensions [18]. It accomplishes this by transforming the original variables into a new set of uncorrelated variables called principal components. These components are ordered in terms of their importance, with the first component explaining the largest amount of variance in the data.

PCA works by identifying the directions in which the data vary the most and projecting the data onto these directions. This allows for a reduction in the number of variables [19] while preserving the most critical information. The resulting principal components are linear combinations of the original variables, and they capture the maximum amount of variability present in the data.

PCA is commonly employed in various fields, such as data analysis, pattern recognition, and image processing [20]. It is particularly useful when dealing with high-dimensional datasets, as it helps identify the underlying structure and relationships among the variables. By reducing the dimensionality of the data, PCA enables easier visualization, data exploration, and subsequent analysis.

G. Elbow Method

The Elbow Method is a graphical technique used to determine the optimal number of clusters in a dataset for clustering algorithms like k-means [21]. It is named after the shape of the plot, which resembles an elbow.

The Elbow Method works by calculating the sum of squared distances between data points and their cluster centroids for different values of k (the number of clusters). The idea is to find the value of k at which the reduction in the sum of squared distances significantly decreases, resulting in an elbow-like bend in the plot.

To apply the Elbow Method, the algorithm is executed multiple times for different values of k. For each iteration, the sum of squared distances is calculated and plotted against the corresponding value of k. The plot typically starts with a high reduction in the sum of squared distances as k increases, indicating better clustering. However, as k continues to increase, the reduction in the sum of squared distances becomes less significant.

The optimal number of clusters is determined by visually inspecting the plot and identifying the point where the decrease in the sum of squared distances starts to level off, forming an elbow shape. This point suggests that further increasing the number of clusters does not provide much improvement in clustering performance.

Although the Elbow Method is subjective to some extent, it provides a useful guideline for selecting the number of clusters in unsupervised learning tasks. It helps to strike a balance between capturing meaningful patterns in the data and avoid overfitting or excessive complexity in the clustering model.

H. Silhouette Coefficient (Silhouette Score)

The Silhouette method is a popular technique used for evaluating the quality and consistency of clustering results [22]. It provides a measure of how well each sample fits within its assigned cluster compared to other clusters. Research literature proves that Silhouette method is particularly useful for assessing the effectiveness of clustering algorithms, such as K-means, Hierarchical Clustering [23], DBSCAN [24] and the literature often supports the efficacy of the Silhouette method [22] in assessing the performance of these clustering algorithms.

To compute the Silhouette coefficient for a specific sample, two distances are considered: the average distance between the sample and all other samples within the same cluster (a), and the average distance between the sample and all samples in the nearest neighboring cluster (b). The Silhouette coefficient for the sample is then calculated as (b - a) divided by the maximum value between a and b. By analyzing the Silhouette coefficients of all samples and averaging them, an overall Silhouette score is obtained. This score serves as an indicator of the clustering quality, with higher scores representing more cohesive and distinct clusters.

I. K-means Clustering

The K-means clustering algorithm is a popular unsupervised machine learning technique used for partitioning a dataset into K distinct clusters [25]. It aims to group similar data points together while minimizing the intra-cluster distance and maximizing the inter-cluster distance.

The algorithm starts by randomly selecting K initial cluster centroids from the dataset. Each data point is then assigned to the nearest centroid based on its distance, typically using the Euclidean distance metric. This step forms the initial clusters.

In the next iteration, the algorithm calculates new centroids for each cluster by taking the mean of all the data points assigned to that cluster. Then, each data point is reassigned to the nearest centroid based on the updated centroids. These steps of centroid recalculation and data point reassignment are iteratively repeated until convergence.

Convergence occurs when the centroids no longer change significantly or when a predetermined number of iterations is reached. At this point, the algorithm has found stable clusters, and the process terminates.

The K-means algorithm aims to minimize a cost function called the within-cluster sum of squares [26], which represents the sum of squared distances between each data point and its assigned centroid. The algorithm iteratively updates the assignments and centroids to optimize this objective.

J. Human-Computer Interaction

HCI, or Human-Computer Interaction, refers to the study, design, and evaluation of the interaction between humans and computer systems [27]. It focuses on creating effective and user-friendly interfaces that facilitate seamless communication and interaction between humans and technology.

The goal of HCI is to enhance the usability and user experience of computer systems, software applications, websites, and other digital interfaces [28]. It takes into account various aspects of human cognition, behavior, and ergonomics to design interfaces that are intuitive, efficient, and satisfying for users [29].

HCI encompasses a multidisciplinary approach, drawing from fields such as computer science, psychology, design, sociology, and ergonomics. It involves understanding user needs, preferences, and behaviors, as well as considering the context in which the interaction takes place.

HCI has a significant impact on various domains, including software development, website design, mobile applications, virtual reality, and wearable technology [30]. By prioritizing user needs and preferences, HCI helps create technology that is more user-centered, efficient, and enjoyable to use.

K. Problem Statement

TheAnalyzer provides a clustering-based system for improving business productivity by analyzing user profiles to enhance HCI. Hence, the problem that this research attempts to address can be formulated as follows:

- How to improve business productivity, by enhancing human-computer interaction, through analyzing user profiles?

III. Proposed Technique

TheAnalyzer, comprises of six main stages as shown in Fig. 1.

A. Data Collection

TheAnalyzer begins its process by gathering user browsing data using web and mobile analytics. Presently, TheAnalyzer focuses on the following analytics to perform user profiling:

1) Number of times the user clicks to view reviews per visit.

2) Number of times the user clicks to filter by price per visit.
3) Number of times the user clicks to view new items per visit.
4) Number of times the user clicks to view promotions per visit.
5) Average time spent on each item from adding to cart to completing checkout.

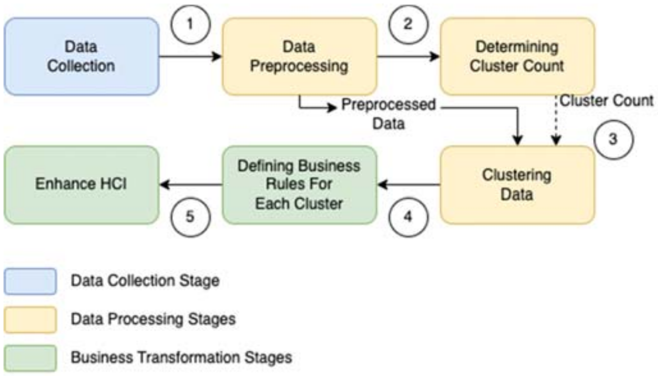

Fig. 1 TheAnalyzer: High-level Architecture

By utilizing the aforementioned analytics, TheAnalyzer can determine the user's behavior and preferences without requiring additional information or survey data. This enables TheAnalyzer to gather a substantial amount of data quickly and effortlessly, without imposing any extra burden on users or system administrators.

B. Data Preprocessing

Data preprocessing plays a crucial role within TheAnalyzer in preparing the dataset before performing clustering. It is a two-step process aiming at improving data quality and transforming features to facilitate effective clustering. Proper data preprocessing helps to ensure accurate and meaningful cluster analysis results.

Since the datasets provided to TheAnalyzer are generated using web and mobile analytics, they generally require minimal data cleaning. The focus at this stage primarily revolves around data standardization and dimensionality reduction. Standardizing the data is essential as the K-means clustering algorithm used by TheAnalyzer is sensitive to the scale and range of features. By bringing all features to a similar scale, the dominance of any single feature is avoided. TheAnalyzer uses Z-score normalization as the standardization technique.

To address dimensionality reduction, TheAnalyzer utilizes the PCA technique. By applying PCA, the dataset's dimensionality is reduced while preserving critical information and capturing the underlying data structure. This approach proves advantageous during clustering as it mitigates the curse of dimensionality and enhances computational efficiency.

Overall, the preprocessing stage for TheAnalyzer mainly involves standardizing the data and reducing its dimensionality through PCA. These steps ensure fair comparisons between features and facilitate efficient clustering analysis.

C. Determining the Cluster Count

During the third stage, TheAnalyzer will be determining the appropriate number of clusters in order to effectively perform k-means clustering. While there is no definitive method to determine the exact cluster count, there are several popular techniques which provide valuable insights for this decision-making process.

TheAnalyzer will be applying the elbow method on the preprocessed data to find the most optimum number of clusters that it needs to input to the k-means algorithm. Here, TheAnalyzer will be evaluating the within-cluster sum of squares, also known as the "inertia." The inertia measures the sum of squared distances between each data point and its assigned centroid within a cluster. As the number of clusters increases, the inertia typically decreases since smaller clusters tend to minimize the distances within each cluster.

D. Clustering Data

In this stage, TheAnalyzer will perform the actual clustering of the data in order to arrive at meaningful user profiling. TheAnalyzer will be using preprocessed data which are standardized as well as reduced in dimensionality. TheAnalyzer will also input an optimum cluster count which it identified during the previous stage using elbow method. TheAnalyzer will be using K-means clustering technique during this stage which is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct groups or clusters. It is a centroid-based algorithm that iteratively assigns data points to clusters based on their proximity to cluster centroids.

E. Defining Business Rules for Each Cluster

Defining business rules for each generated cluster is an essential step in utilizing the insights gained from cluster analysis to drive meaningful actions and decision-making in a business context. During this stage TheAnalyzer will define business rules for each system generated cluster with the help of domain specialists or business leadership. Business rules provide guidelines or conditions based on which specific actions or strategies can be implemented for each cluster. These rules help in understanding the characteristics, behaviors, or preferences of the users within each cluster, enabling targeted and personalized approaches to interact with them.

F. Enhancing HCI

In the final stage of TheAnalyzer, the focus is on improving the HCI based on user profiling and relevant business rules. This stage involves determining the most appropriate cluster association for each user and applying the corresponding business rules to enhance the user's experience with the business application. By assigning the user to the relevant cluster, TheAnalyzer establishes a personalized profile for the user.

Once the user's cluster association is determined, the defined business rules specific to that cluster are attached to the user's profile. These business rules serve as guidelines for customizing the user's interaction with the business application. By incorporating the updated business rules, the business application can tailor its interactions and services to meet the specific needs and preferences of each user. This could involve providing personalized recommendations, targeted promotions, optimized pricing strategies, or any other relevant enhancements based on the user's cluster association.

The ultimate goal is to create a more engaging and satisfying user experience by leveraging the insights gained from user profiling and aligning it with the business objectives. Through this process, TheAnalyzer enables the business application to deliver a more tailored and effective interaction that enhances user satisfaction and potentially improves business outcomes.

IV. Implementation

TheAnalyzer is designed and developed to receive a defined set of web and mobile-based analytics related to the business application that it is connected to, process the received data and then perform certain business transformations to enhance user interaction with the business application resulting higher business productivity. Once connected to the business application platform(s), TheAnalyzer begins collecting user data and once it reaches a user defined threshold in terms of collected data volume, it will initiate the data processing steps to cluster the associated user analytics.

This section will outline the step-by-step process that TheAnalyzer follows in terms of data processing operations. First TheAnalyzer will initiate the data preprocessing stage. Next it will then delve into determining the optimal cluster count and then proceed to the final stage of data processing operation by proceeding with clustering the preprocessed user profiling data. Additionally, this section will provide an overview of the implementation details of the business transformation stages, which are defining business rules for each cluster and enhancing the HCI aspect. As the specific transformations implemented by TheAnalyzer depend on the connected business, Section V will provide practical examples of the transformative changes conducted by TheAnalyzer in relation to two real-world businesses.

A. Preprocessing the Dataset

TheAnalyzer will first normalize the collected user analytics using the Z-score normalization. TheAnalyzer uses the standardization capability provided via scikit-learn library to perform this operation.

In completing this first step, TheAnalyzer achieves many benefits. It allows TheAnalyzer to create a comparable scale. All features within the collected data are now on a comparable scale. TheAnalyzer achieves this with Z-score normalization because it transforms the data to have a mean of zero and a standard deviation of one. Given TheAnalyzer will always process a numerical dataset which has multiple units, this is an extremely important step in the process. Performing standardization allows TheAnalyzer to eliminate outliers. Z-score normalization helps in identifying and handling outliers effectively. Since the standardization process is based on the mean and standard deviation, extreme values that deviate significantly from the mean will have larger Z-scores. It also improves algorithm performance. To perform effective clustering and derive meaningful insights, the data should be normally distributed and should have similar scales. The more this remains applicable in the dataset, the more accurate and improved results will be. TheAnalyzer uses Z-normalization to make sure the collected data become normally distributed, and all features share a similar scale. Normalization has a positive impact on reducing dimensionality. TheAnalyzer next performs PCA to reduce the dimensionality of the dataset and in doing so having normally distributed and features with similar scales improves the accuracy of the PCA outcome.

Next, TheAnalyzer will conduct PCA on the standardized dataset. Currently TheAnalyzer accepts five different features hence the dataset prior to performing PCA would be a five-dimensional (5D) dataset. TheAnalyzer uses the PCA capability provided via scikit-learn library to reduce the dimensionality of the standardized dataset. TheAnalyzer will reduce the 5D dataset to a two-dimensional (2D) dataset upon performing PCA. This completes the data preprocessing steps.

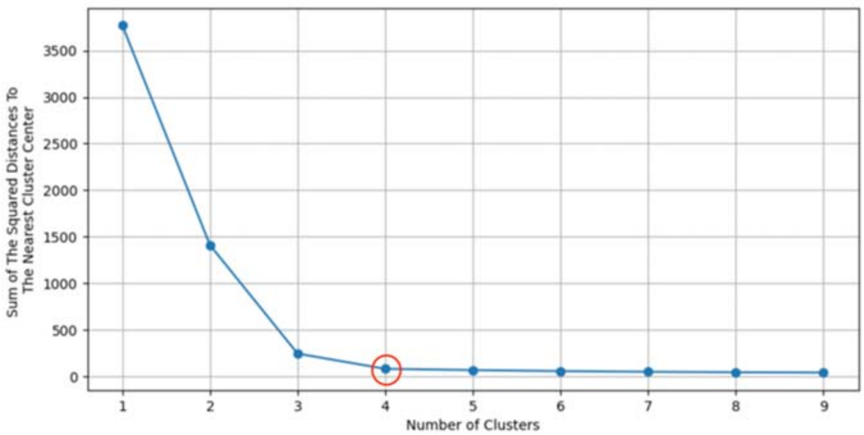

Fig. 2 Elbow graph

B. Determining the Optimum Cluster Count

TheAnalyzer will next work on determining the optimal number of clusters, which is a crucial step prior to performing the actual clustering using K-means algorithm. TheAnalyzer uses the capabilities of the K-means clustering technique provided via scikit-learn library to perform elbow method. The elbow method helps identify the number of clusters that best capture the underlying structure of the data without overfitting or underfitting.

TheAnalyzer will begin by applying the K-means clustering algorithm to the dataset. TheAnalyzer will start with two clusters and gradually increase the count up to a reasonably large value. This maximum value is a configurable value and provided as an input per each business application evaluation. During the process, TheAnalyzer will compute the Within-Cluster Sum of Squares (WCSS). For each cluster count within the specified range, we calculate the sum of squared distances between each data point and its cluster centroid. This is known as the WCSS. The WCSS measures the compactness or cohesion of the clusters. Next TheAnalyzer will plot the Elbow Curve. The graph contains the WCSS values against the corresponding cluster counts. Typically, the WCSS decreases as the number of clusters increases because smaller clusters tend to be more compact. However, as the number of clusters keeps increasing, the rate of decrease in WCSS diminishes. Finally, by examining the shape of the plotted curve, the elbow point will be derived. The elbow point is the cluster count at which the rate of decrease in WCSS significantly levels off. The curve resembles an arm, and the elbow point represents the optimal number of clusters. Fig. 2 shows an Elbow graph generated by TheAnalyzer for one of the real-world business applications. As per that application for the provided dataset it suggests that having four clusters (circled in red) best describes the underlying structure without overfitting or underfitting.

C. Performing K-means Based Clustering

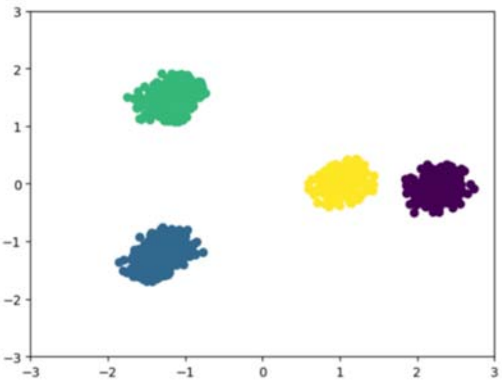

Fig. 3 K-means clustering

As the final step of the data processing stages, TheAnalyzer will perform the K-means clustering upon the preprocessed dataset providing the optimum number of clusters derived via the Elbow graph as the expected cluster count. Fig. 3 shows the output of the K-means clustering performed on the same real-world business dataset which was used to generate the Elbow graph.

D. Defining Business Rules to Enhance User Experience

The final step in the process involves defining business rules to enhance the user experience. Once the clustering process is completed by TheAnalyzer, each user profile is assigned a specific cluster label. The labeled dataset is then extracted, and with the guidance of domain experts, business rules are generated for each cluster. By associating user profiles with their respective clusters, the defined business rules are inherited by these profiles. As a result, the user experience and HCI aspects are improved, ultimately contributing to enhanced business productivity. This strategic approach ensures that the business application optimally caters to the needs and preferences of each user segment, thereby fostering increased productivity for the business. Furthermore, if a new user starts using the business application, as TheAnalyzer will not have enough information to determine which cluster the user will belong to, the user will experience default business rules and user experience but with time when TheAnalyzer collects enough analytics, the new user will also be categorized to an existing cluster relying on K-means predict functionality provided by the scikit-learn library.

Section V contains two real-world business applications transformed through the insights provided by TheAnalyzer which will contain more insights into what specific business rules are, how those get mapped to the identified clusters and how those resulted in business productivity improvements.

V. Evaluation

This section highlights the successful implementation of TheAnalyzer in two real-world business applications, showcasing its effectiveness in clustering users and analyzing user-driven data. It demonstrates how TheAnalyzer has facilitated the assignment of distinct business rules to different user groups, consequently enhancing user interactions and overall application experience. These improvements have led to a notable increase in business productivity.

The clustering process has allowed them to identify distinct user segments with similar characteristics, preferences, and behaviors. Subsequently, tailored business rules have been established for each user group, ensuring personalized experiences and targeted interactions. The impact of these personalized experiences on user engagement and satisfaction has been commendable. TheAnalyzer's implementation has resulted in improved user interactions, intuitive interfaces, and seamless navigational flows within the applications. By aligning the user experience with individual needs and preferences, the businesses have successfully fostered a higher level of engagement and satisfaction among their users.

A. Online Fashion Retail Platform

This online fashion retail platform offers a web and mobile application that specializes in fast fashion, providing trendy clothing, accessories, and personal care items. Customers can seamlessly navigate through the purchase cycle, filter items based on type, price etc., just like in any other retail e-commerce platform. Moreover, the platform ensures convenient communication channels through email and SMS services to foster enhanced customer engagement and send out marketing flyers to subscribers. Additionally, the brand is actively promoted on various social media platforms owned by the platform.

TheAnalyzer analyzed three months' worth of web and mobile user analytics data from the business's online platform. The dataset consisted of a total of 1,069 distinct user analytics. The results of the analysis are presented in Fig. 2, which displays the Elbow graph, and Fig. 3, which illustrates the clustering outcome. Based on the clustering results, four distinct user groups were identified.

After the completion of the clustering process, TheAnalyzer labeled each user profile in the analytics dataset and assigned them to one of the four clusters. This labeled dataset, which indicates the cluster assignment for each user based on their analytics, was then used by domain experts to interpret the clusters and define corresponding business rules.

The four clusters were interpreted to represent different user buying patterns, each with its own unique characteristics. The defined business rules were established to cater to the specific needs and preferences of each cluster. These rules serve as guidelines for enhancing the user experience, tailoring interactions, and optimizing business strategies for different user groups.

1) Cluster 01 - Representing a browsing customer type behavior: show more limited time deals in prominent sections within the application, send alerts on limited time deals, send alerts on items related to recent search history.
2) Cluster 02 - Representing a loyal customer type behavior: tailor the loyalty program benefits to be in line with customer's purchase preferences, early access to new collections, additional exclusive benefits: free shipping, free returns, priority customer service etc.
3) Cluster 03 - Representing a bargain hunter customer type behavior: clearance section items to be prominent within the application, send alerts on clearance items and bundle deals, promote free shipping items more prominently.
4) Cluster 04 - Representing an impulsive buyer type behavior: show limited edition items in prominent sections within the application, send alerts on limited edition items, rewarding for reaching purchase milestones.

Next, TheAnalyzer conducted a randomized experiment. A 50% sample of users from each cluster was selected, and the new business rules were applied to their profiles, providing them with a customized user experience. In contrast, the remaining 50% of users in each cluster continued to receive the default user experience.

Fig. 4 presents a comparison of monthly sales percentages between the two groups: those who experienced the customized user experience and those who were treated with the default user experience. Within each cluster, the users were randomly divided into two groups, ensuring a fair comparison.

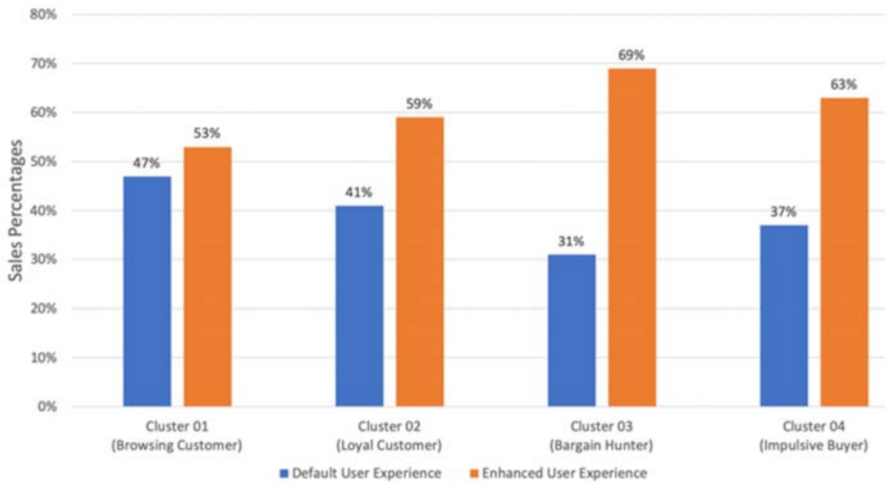

Fig. 4 Sales percentages – enhanced user experience vs. default user experience for “Online Fashion Retail” platform

The results clearly demonstrate the effectiveness of the customized user experience. Across all clusters and over the three-month period, the group that received the customized experience generated higher sales compared to the group that encountered the default behavior.

B. Online Retail E-commerce Platform

This application is a comprehensive e-commerce platform that offers a diverse selection of products spanning electronics, clothing, home goods, and books. The user-friendly web application enables seamless navigation through features like search, sorting, filtering, and the option to add products to wishlist and shopping cart. Furthermore, this platform leverages e-mail marketing to provide registered customers with promotional materials through e-mail flyers.

TheAnalyzer analyzed four months' worth of user analytics data from the business's web application. The dataset consisted of a total of 1,028 unique user analytics. Fig. 5 shows the Elbow graph which confirms that the optimum number of clusters that best reflects the 4 months' dataset underlying structure would be three clusters. Fig. 6 shows the clustering outcome when K-means based clustering is performed on the dataset using three clusters. Upon identifying the three clusters the same next steps were followed in performing the business transformation to this platform too and in that process following are the cluster identification and associated business rules the domain experts arrived at:

1) Cluster 01 - Representing a browsing customer type behavior: sending reminders on the items in the wishlist and cart, send alerts on items related to recent search history.
2) Cluster 02 - Representing a bargain hunter customer type behavior: default sort by option to be set as "Lowest Price First" in the product search/list page, show discounted items in prominent sections within the application, send alerts on discounted items.
3) Cluster 03 - Representing an impulsive buyer type behavior: upsell "bought together" items during the checkout process, default sort by option to be set as "Recommended" in the product search/list page.

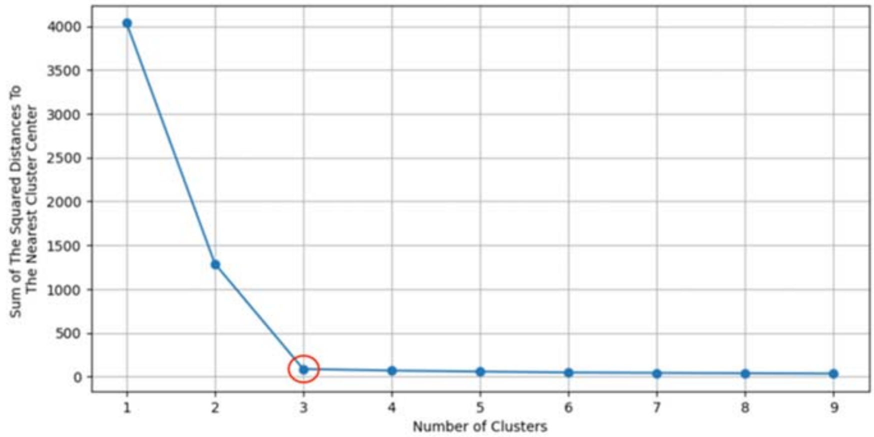

Fig. 5 Elbow graph – "Online Retail E-commerce" platform

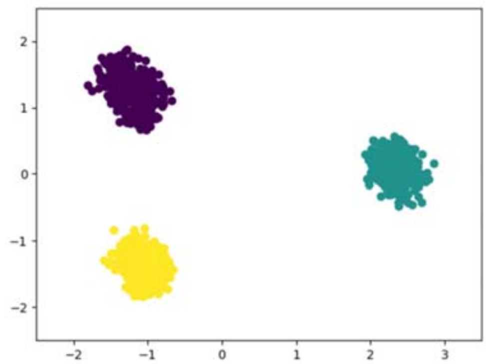

Fig. 6 K-means clustering

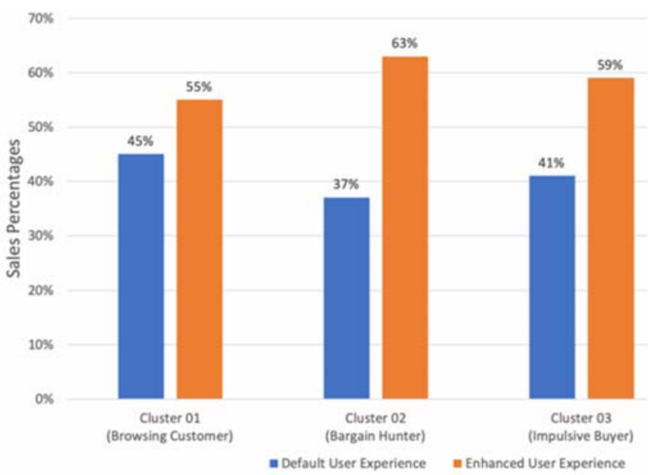

Fig. 7 Sales percentages – enhanced user experience vs default user experience for "Online Retail E-commerce" platform

Upon identifying the three clusters representing browsing, bargain hunter and impulsive buyer customer types as shown in Fig. 6, the identified business rules were implemented within the application to enhance the user experience. Next, the impact of the improved user experience was assessed in terms of the increase in sales percentage. The results in Fig. 7 show that over the three-month period, the group that received the enhanced customer experience resulted in generating higher sales percentage compared to the group that continued as it is with the default behavior.

The findings indicate that tailoring the user experience based on the defined business rules positively influenced user behavior and engagement with the application. TheAnalyzer clustered each application user to a specific cluster and randomly selected 50% of each cluster and attached the newly defined business rules to their profiles so that they get to experience the enhanced user experience and interaction with the business application while the other half in the cluster will feel no difference in terms of user experience when it comes to interacting with the applications. The cluster-wise sales percentage breakdowns shown in Figs. 4 and 7 show that for both applications, the group with enhanced customer experience and interaction with the application generated more sales compared to the other half. Some findings, observations and feedback received after with the experiments carried out with transforming real-world businesses with TheAnalyzer, are as follows:

1) The input dataset features have been carefully and effectively chosen so that it allows to derive meaningful clusters aligned with different user buying patterns.
2) TheAnalyzer has been designed in a manner where user analytics are obtained through natural user interactions

with the business application without requesting users to do any additional work such as: questioners, surveys etc. This has resulted in gathering an organic and faster dataset.

3) Cluster identifications have been accurate and able to represent associated business's user base without overfitting or underfitting.
4) The customized experience TheAnalyzer has produced, succeeded in capturing users' attention, satisfying their preferences, and ultimately driving more sales.

TheAnalyzer's randomized experiment provides compelling evidence of the value and impact of the customized user experience. These results serve as a basis for confident decision-making, encouraging businesses to adopt personalized strategies that align with user needs and preferences. By leveraging TheAnalyzer's capabilities, businesses can optimize their business outcomes, enhance customer satisfaction, and drive sustainable growth.

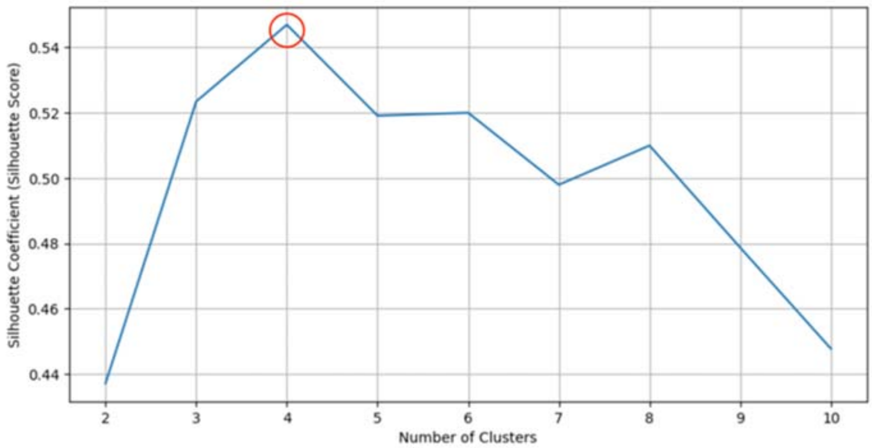

Fig. 8 Silhouette Coefficient – on “Online Fashion Retail” platform dataset

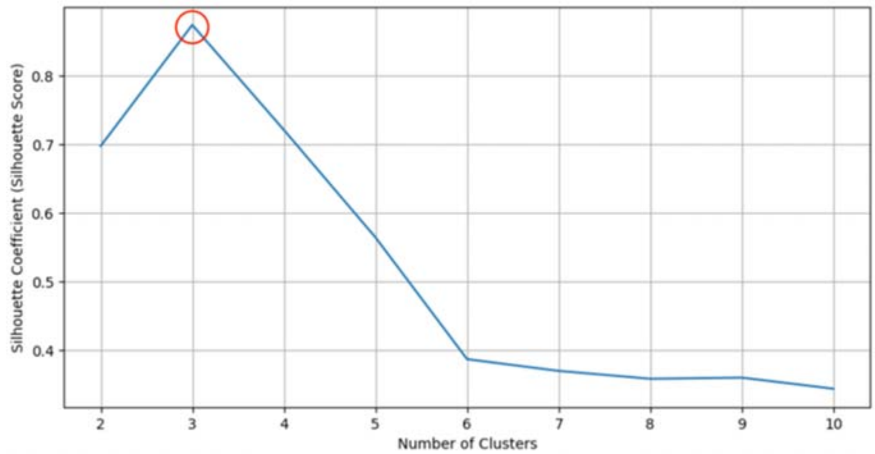

Fig. 9 Silhouette Coefficient – on “Online Retail E-commerce” platform dataset

C. Silhouette Coefficient (Silhouette Score)

This metric measures how well each sample fits within its assigned cluster compared to other clusters. It calculates the average Silhouette coefficient across all samples, where higher values indicate better clustering results. Fig. 8 displays the results of the Silhouette scores for different cluster counts. Notably, the highest Silhouette score is achieved when utilizing four clusters, coinciding with TheAnalyzer's decision to separate the dataset of the "Online Fashion Retail" platform into four distinct clusters. Similarly, Fig. 9 demonstrates that the highest Silhouette score is attained with three clusters, aligning with TheAnalyzer's choice to employ K-means clustering with three clusters for the "Online Retail E-commerce" platform. These findings indicate that the data processing and clustering approach of TheAnalyzer yield favorable results, as evidenced by the positive evaluation provided by the Silhouette coefficients.

VI. Conclusion

We present TheAnalyzer, a solution aimed at improving business productivity by enhancing HCI through user profile analysis. TheAnalyzer is designed to seamlessly integrate with business applications, allowing for the collection of relevant data points based on users' natural interactions without imposing any additional burdens such as questionnaires or surveys.

The effectiveness of TheAnalyzer has been evaluated quantitatively using both research techniques, such as assessing the Silhouette coefficient, and business statistics, such as comparing sales percentages with and without the influence of TheAnalyzer. Furthermore, qualitative evaluation has been conducted, incorporating feedback from domain experts and businesses. This feedback highlights the quality of feature selection, interpretability, and domain relevance of the data clustering and user profiling process facilitated by TheAnalyzer.

As next steps, our plan is to expand the feature set to capture a broader range of user analytics. Additionally, we aim to enhance TheAnalyzer by implementing a knowledge base that tracks the performance of each business rule applied to different clusters. This knowledge base will enable us to provide feedback to domain experts and business owners regarding the efficacy of specific business rules attached to clusters, facilitating continuous improvement and optimization of the system.

Acknowledgment

Authors would like to thank the anonymous reviewers for their valuable insights and comments about the paper.

References

[1] Y. Bakos, H. C. Lucas, Jr., W. Oh, G. Simon, S. Viswanathan and B. Weber, "The Impact of E-Commerce on Competition in the Retail Brokerage Industry," 2005.
[2] M. G. Helander and H. M. Khalid, "Modeling the customer in electronic commerce," in Applied Ergonomics 31 (2000) 609}619, 2000.
[3] P.-M. Lee, "Behavioral Model of Online Purchasers in E-Commerce Environment," Electronic Commerce Research, vol. 2, 2002.
[4] J. Bang, Y. Cho and M. S. Kim, "Getting Business Insights through Clustering Online Behaviors," Modelling and Simulation in Engineering, vol. 2015, 2014.
[5] S.-T. Li, L.-Y. Shue and S.-F. Lee, "Business intelligence approach to supporting strategy-making of ISP service management," in Expert Systems with Applications, 2008.
[6] W.-P. Lee, C.-H. Liu and C.-C. Lu, "Intelligent agent-based systems for personalized recommendations in Internet commerce," Expert Systems with Applications, vol. 22, no. 14, pp. 275-284, 2002.
[7] J. J. Vaske, "Advantages and Disadvantages of Internet Surveys: Introduction to the Special Issue," 2011.
[8] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 1785-1792, 2022.
[9] D. S. Schwering, W. I. Sonntag and S. Kühl, "Agricultural E-commerce: Attitude segmentation of farmers," in Computers and Electronics in Agriculture, 2022.
[10] A. Tamhane, S. Arora and D. Warrier, "Modeling Contextual Changes in User Behaviour in Fashion e-Commerce," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2017.
[11] S. Hernandez, J. Fabra and J. Ezpeleta, "By standardizing the data, all features are brought to a similar scale, ensuring that no single feature dominates the clustering algorithm.," 2017.
[12] S. E. Beatty and M. E. Rerrell, "Impulse Buying: Modeling Its Precursors," 1998.
[13] G. Muruganantham1 and R. S. Bhakat, "A Review of Impulse Buying Behavior," International Journal of Marketing Studies, vol. 5, no. 3, 2013.
[14] Y. Chen, Y. Lu, S. Gupta and Z. Pan, "Understanding "window" shopping and browsing experience on social shopping website: An empirical investigation," in Information Technology & People, 2020.
[15] V. Leninkumar, "The Relationship between Customer Satisfaction and Customer Trust on Customer Loyalty," International Journal of Academic Research in Business and Social Sciences, vol. 7, no. 4, 2017.
[16] N. Sudo, K. Ueda, K. Watanabe and T. Watanabe, "Working Less and Bargain Hunting More: Macroimplications of Sales during Japan's Lost Decades," 2016.
[17] D. Sing and B. Singh, "Investigating the impact of data normalization on classification performance," Applied Soft Computing, vol. 97, 2020.
[18] S. Deng, B. Li and K. Wu, "Analysing the impact of high-tech industry on regional competitiveness with principal component analysis method based on the new development concept," Emerald Group Publishing Limited, 2022.
[19] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," Journal of Soft Computing and Data Mining, vol. 2, no. 1, 2021.
[20] E. O. Omuya, M. W. Kimwele and G. O. Okeyo, "Feature Selection for Classification using Principal Component Analysis and Information Gain," Expert Systems with Applications, vol. 174, 2021.
[21] F. Liu and Y. Deng, "Determine the number of unknown targets in Open World based on Elbow method," in IEEE Transactions on Fuzzy Systems, 2020.
[22] M. Saputhra, D. Saputhra and L. D. OSWARI, "Effect of Distance Metrics in Determining K-Value in KMeans Clustering Using Elbow and Silhouette Method".
[23] W. Yang, X. Wang, J. Lu, W. Dou and S. Liu, "Interactive Steering of Hierarchical Clustering," 2020.
[24] Y. Chen, N. Bouguila, L. Zhou, C. Wang, Y. Chen and J. Du, "BLOCK-DBSCAN: Fast clustering for large scale data," 2021.
[25] A. Ashabi, S. B. Sahibuddin and M. S. Haghighi, "The Systematic Review of K-Means Clustering Algorithm," in International Conference on Networks, Communication and Computing, 2020.
[26] D. Pollard, "Strong consistency of k-means clustering," 1981.
[27] C. Frauenberger, "Entanglement HCI The Next Wave?," 2019.
[28] T. Issa and P. Isaias, "Usability and Human–Computer Interaction (HCI)," in Sustainable Design, SpringerLink, 2022, pp. 23-40.
[29] S. S. Feger, S. Dallmeier-Tiessen, P. W. Woźniak and A. Schmidt, "The Role of HCI in Reproducible Science: Understanding, Supporting and Motivating Core Practices," in CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.
[30] N. Dell and N. Kumar, "The Ins and Outs of HCI for Development," 2016.