Consumer Behaviour Analysis for Customer Segmentation and Purchase Prediction

Amritha Gopakumar
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Amritapuri, India
amrithagopakumarengoor@gmail.com

Aathira Shine
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Amritapuri, India
aathirashine1999@gmail.com

Amina Ajim
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Amritapuri, India
amina.ajimshah@gmail.com

Anjali T
Department of Computer Science and Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Amritapuri, India
anjalit@am.amrita.edu

2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS) | 979-8-3503-8436-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICACCS60874.2024.10717226

Abstract—**E-commerce, short for electronic commerce, is the practice of placing orders online or negotiating a sale's terms and price. Although e-commerce is often associated with selling products and services to consumers, business-to-business (B2B) transactions also make use of it. To market their goods, many smaller consumer enterprises rely only on online sales. Nowadays, people are choosing to make purchases online as online transaction mechanisms and shopping platforms develop. Nevertheless, given that both clients and Merchants are unable to interact with clients face-to-face, there is a limited understanding of consumer demands. Due to this, businesses may struggle to gauge nuanced cues such as body language and personal preferences. Limited direct engagement can hinder the ability to address specific customer needs and preferences, potentially resulting in misinterpretation or overlooking crucial details. In turn, businesses may face difficulties in tailoring their products or services to meet the nuanced expectations of their customers, emphasizing the importance of alternative strategies such as detailed customer feedback mechanisms and data analytics to bridge this gap. The online system logs user activity and gathers data on consumer behaviour, allowing for the prediction of consumer purchasing patterns. Therefore, to improve the shopping experience of people, our study aims to use historical sales data to analyze consumer behaviour and do customer segmentation using RFM (Recency, Frequency, and Monetary) based K-means clustering and purchase prediction using various ML models to improve sales, out of which the combination of Logistic Regression and XGBoost performed the best.**

Keywords—E-commerce, Behaviour Analysis, Customer Segmentation, Clustering, Machine learning, Purchase Prediction

I. INTRODUCTION

The scope of e-commerce transactions is vast, encompassing both business-to-consumer (B2C) and B2B interactions. In the realm of consumer transactions, e-commerce facilitates the online buying and selling of a diverse range of products and services, offering convenience and accessibility to a global audience. Additionally, e-commerce has revolutionized B2B transactions, streamlining supply chain processes, reducing costs, and enhancing efficiency for businesses by providing a digital platform for procurement, sales, and collaboration between enterprises. Many smaller consumer enterprises too increasingly rely on online sales as a primary avenue for marketing their goods due to the extensive reach and accessibility of the digital marketplace. The online platform allows these businesses to overcome geographical constraints, reaching a broader audience without the need for a physical presence in multiple locations. Moreover, the cost-effectiveness of online marketing and sales channels enables smaller enterprises to compete on a level playing field with larger counterparts, as they can leverage digital advertising, social media, and e- commerce platforms to showcase and sell their products to a global customer base. The digital landscape provides smaller businesses with a powerful tool to establish brand visibility, connect with diverse audiences, and compete in the modern marketplace. This digital paradigm shift has transformed traditional commerce, offering new opportunities and efficiencies across various sectors. The evolution of online purchases has witnessed significant advancements in transaction mechanisms and shopping platforms. Initially reliant on basic payment gateways, the landscape expanded with the introduction of secure and convenient methods such as credit cards and digital wallets. Over time, the rise of e-commerce giants and the advent of mobile technology led to the development of user-friendly shopping platforms, enhancing the overall online shopping experience. Subsequent innovations, including one- click purchasing, personalized recommendations, and seam- less integrations with social media, have further transformed online transactions, making them more efficient, secure, and tailored to individual consumer preferences. Today, diverse payment options, advanced security features, and immersive shopping interfaces characterize the dynamic evolution of online purchases. This evolution has also led to notable shifts in consumer behavior. With the advent of secure transaction mechanisms and user-friendly shopping platforms, consumers have become more inclined to make high-value transactions online, including purchasing electronics, luxury items, and even groceries. The convenience of mobile commerce has fueled impulsive

979-8-3503-8436-9/24/$31.00©2024 IEEE

buying behavior, with consumers increasingly relying on smartphones for on-the-go shopping. Moreover, the emphasis on personalized shopping experiences, driven by advanced algorithms and recommendations, has elevated consumer expectations, making them more discerning and seeking tailored products and services. Overall, these developments have contributed to a significant increase in online spending and a preference for seamless, personalized shopping journeys. This has also prominently featured considerations for scalability and adaptability, acknowledging the necessity of robust infrastructure to handle peak user traffic, ensuring a seamless shopping experience. Additionally, adaptability to diverse scenarios, such as fluctuations in demand, emerging market trends, and technological advancements, is crucial for staying competitive. E-commerce platforms that successfully incorporate scalable, adaptable technologies can effectively navigate the dynamic landscape, accommodating growth and evolving consumer preferences.

Customers are an organization’s most valuable asset, and they are crucial to increasing the company’s performance and competitiveness in the market [1]. According to studies, it generally costs more to acquire new customers than it does to keep an existing one. An organisation will make more money from its current clients if it keeps up a positive relationship with them over time. It is necessary that data on consumer consumption behaviours must be analyzed carefully because only then can reliable service, marketing, advice, and early warning of transaction hazards be provided. Hence, it is important to forecast the users who will participate in promotional activities and identify among them those loyal consumers who will make repeated purchases in order to reduce promotion expenses and maximise return on investment. In e-commerce sites, purchasing products is accompanied by a number of actions, such as clicking, gathering, and adding to shopping baskets [2]. After a campaign is over, the e-commerce platform retains extensive user log data, and deep mining of this data can indicate consumers’ preferences for a given product and their propensity to buy it. Optimizing a company’s inventory is also advantageous because it is a significant strategy to reduce overstocking and to increase sales and prevent under- stocking, which can cause sales to decline due to a lack of product availability. As a result, a new online retailer needs to develop and deploy a system for forecasting sales and product recommendations. In order to assist such retailers in undergoing business transformation, we decided to implement a customer segmentation model and purchase prediction model utilizing machine learning algorithms with good performance. In this paper, we have implemented customer segmentation using the RFM (Recency, Frequency, and Monetary) based K-means clustering method to determine the significance of each customer group. We also did purchase prediction using XGBoost [3], Random Forest Classifier and various other machine learning algorithms [4] to build high performance prediction models.

II. Related Work

Different studies have already been conducted in the past to help ecommerce businesses around the world. According to research by Wang XingFen et al., [5] Logistic regression using the XGBoost approach is possible, and the model’s evaluation index is superior to that of any method applied independently. Dirk Van den Poel and Wouter Buckinx [6] used Logit modeling. Accuracy and AUC are used to measure classification performance for predicting the online purchasing behaviour of various customers. Aidin Salamzadeh et al. use supervised learning and unsupervised machine learning techniques with Python to group customers and predict consumer behaviour. These techniques include the Gaussian mixture model and multi-layer perceptron. [7]. The most widely used grocery apps in Hungary and Iran, according to the findings, are Woltand and Snappfood. The MLP algorithm’s Mean Squared Error value is less than 0.1, which indicates a tolerable level of error. The results of overfitting show that the MLP model is correctly fitted.

III. Proposed Methodology

In this study, there are mainly 2 parts as shown in Fig. 1: Customer Segmentation and Purchase Prediction. The first part is achieved using RFM-based K-means clustering, to define the different kinds of customer groups and their corresponding value to a business. The second part is achieved by applying 7 different machine learning models and comparing their performance, to identify the best purchase prediction model. This will enable us to predict if a customer will purchase a product that is in their shopping cart or not.

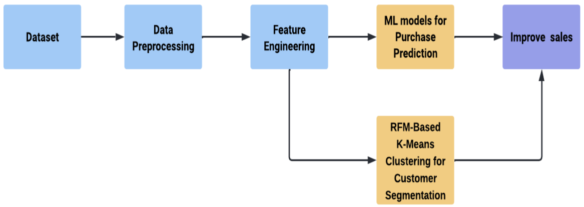
Dataset
Data Preprocessing
Feature Engineering
ML models for Purchase Prediction
Improve sales
RFM-Based K-Means Clustering for Customer Segmentation

Fig. 1. Proposed Model.

A. Dataset

For this study, we used the “eCommerce Events History in Cosmetics Shop” dataset from Kaggle, that includes behaviour information from a medium-sized online cosmetics store for 5 months (Oct 2019–Feb 2020) [8]. A different event is represented by each row in the file. Every event is associated with consumers and the items they buy. The “event time” attribute, for example, is an attribute to describe the time that an event occurred. “event type” is a categorical feature that indicates if a person has viewed, added to their shopping cart, removed from the cart or purchased an item. “product id” is to identify a unique product while “category id” specifies a unique category etc. Table I displays an overview of the characteristics of the dataset in detail.

The counts of various customer actions, such as viewing products, adding items to the cart, making purchases and removing items from the cart are shown in Fig. 2. The line plot in Fig. 3 shows that from October 2019 to November 2019, there was a spike in the total sales and then it sharply declines from November 2019 to December 2019. The amount of sales then rises again from December 2019 to January 2020. The top 10 brands that are purchased the most are shown in Fig. 4.

TABLE I. DESCRIPTION OF PROPERTIES

Attributes	Description
event_time	Time at which event occurred (in UTC)
event_type	Actions performed bythe user: view, cart, remove_from_cart, purchase
product_id	distinct product identifier
category_id	distinct category identifier
category_code	distinct category of anitem
brand	name of the brand
price	cost of an item
user_id	distinct user identifier
user_session	session id of the user

B. Data Preprocessing

To make the dataset appropriate for machine learning algorithms, numerous data processing operations are necessary in most cases. Additionally, this helps with minimizing execution time and enhancing the outcomes. In order to facilitate this, we first removed the duplicate rows. After analysing the data distribution, it was found that the columns brand and category code had a lot of missing values. Hence, both the columns were eliminated. We then dropped the rows that contained null values. While observing the data, it was found that a user session could have different user ids, therefore the “user session” column was dropped as the “user id” is a better feature for identifying a unique user. The most important part that we did comes after the data cleaning part, which is the feature engineering. The dataset was transformed into such a way that only one record is maintained for each product in the cart for a specific user. As shown in Table II, we introduced some new features to the dataset. These extracted features along with user id, product id, category code, price and month were merged in order to help determine whether consumers would buy a particular item, provided it was added to the cart.

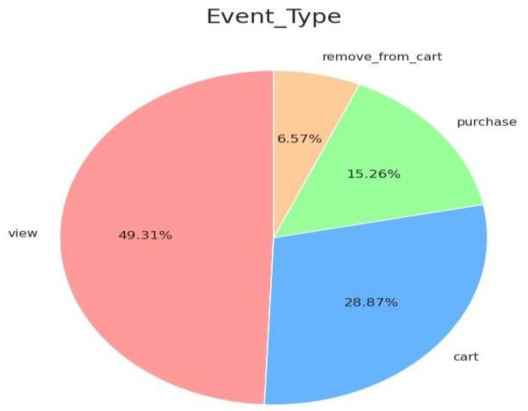

Fig. 2. Event Type Trend.

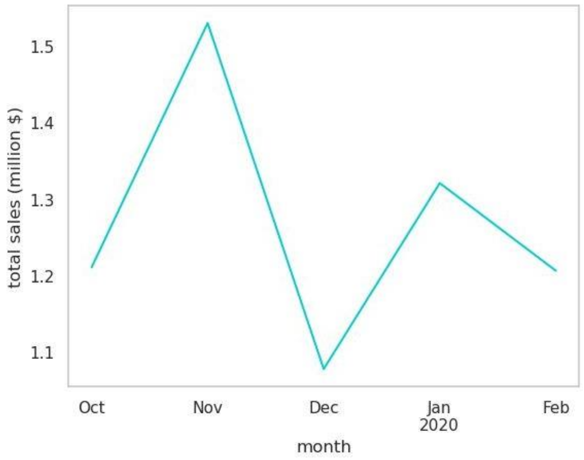

Fig. 3. Monthly Sales Trend.

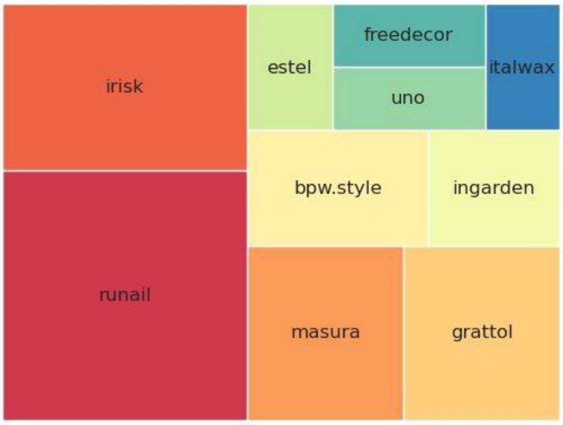

Fig. 4. Top 10 Most Purchased Brands.

TABLE II. THE NEWLY ADDED FEATURES AFTER FEATURE ENGINEERING

Features	Description
is_ purchased	0 if the item in thecart is not purchased; 1 if the item in thecart is purchased
event_ weekday	weekday based on thetime
view_ count	number of times theuser viewed the item
cart_ count	number of times the user has added the product to the cart
purchase_ count	number of times the item has been purchased

C. Customer Segmentation

Customer segmentation is the grouping of customers based on shared characteristics so that firms may promote to each group efficiently and properly. In business-to-business marketing, a corporation may categorise its customers. Here, customer segmentation is done by applying RFM-based K- means clustering.

As our goal was to segment customers to help with marketing strategies, we decided to start with the simple and effective RFM model. In short, RFM model is based on 3 factors - how recently (Recency), how often

(Frequency), and how much (Monetary Value) did the customer buy. We then sorted the data by user id and estimated the Recency (how long had it been since the client's most recent purchase), Frequency (how frequently had the consumer made a purchase from October 2019 to February 2020), and Monetary Value for each customer (How much did the client spend between October 2019 and February 2020) as shown in Fig. 5. Also, prior to grouping, we eliminated outliers.

In Fig. 6, it is seen that when the RFM data were first analyzed, it was discovered that: (1) customers were roughly uniformly distributed over the recency curve; (2) most customers made hardly ten purchases; and (3) most customers spent less than one hundred dollars.

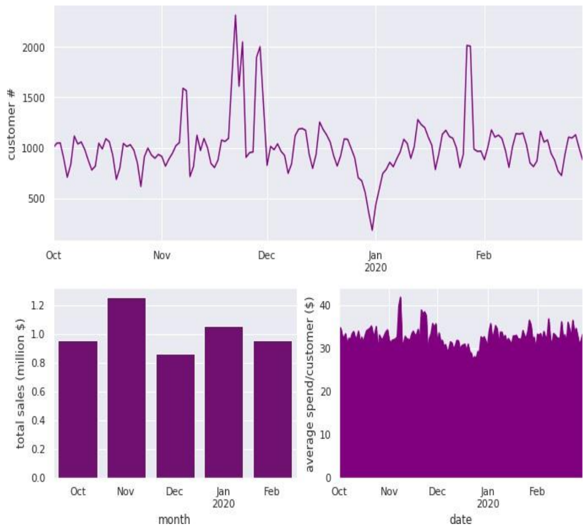

Fig. 5. Distribution of Customers.

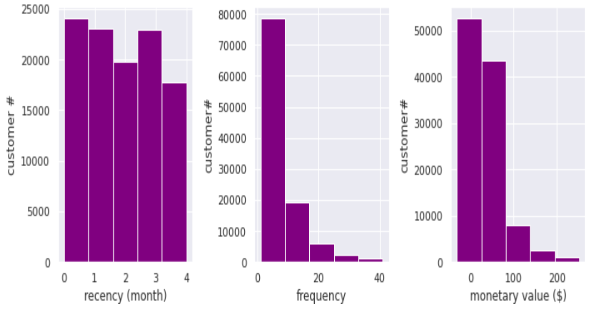

Fig. 6. Distribution of Customer on applying RFM.

D. Purchase Prediction

The scenario that we considered in our study is to determine whether a user would purchase a particular product, provided he/she had added it to the cart. A lot of research has been done by previous researchers regarding the best algorithms for purchase prediction. Due to the greater accuracy of XGBoost and Random Forest, they are extensively used. Apart from this, there are other models that combine algorithms like the LR+XGBoost model [5] and the Decision tree ensembles Bagging [9] that have resulted in high accuracies as well. In our paper, we compared 7 different machine learning algorithms: XGBoost, Random Forest, Logistic Regression, Support Vector Machine, Decision Tree, Decision Tree ensembles bagging and finally, LR+XGBoost.

IV. Experimental Results

A. Customer Segmentation

After applying RFM Model,each customer will be given a score for each RFM factor. These scores are merged and used for segmentation. As a result, we choose to perform RFM analysis using K-Means clustering. At first, we standardised the data before using the elbow approach as seen in Fig. 7 to get the ideal number of clusters. We decided to group our customers into 4 clusters by K-Means. After applying K- means clustering, the customers are segmented into 4 groups as shown in Fig. 8:

1) Loyal Customers: These clients made the most expensive purchases and made them frequently. They shopped from Oct 2019 to Feb 2020 with a median recency of 1 month.

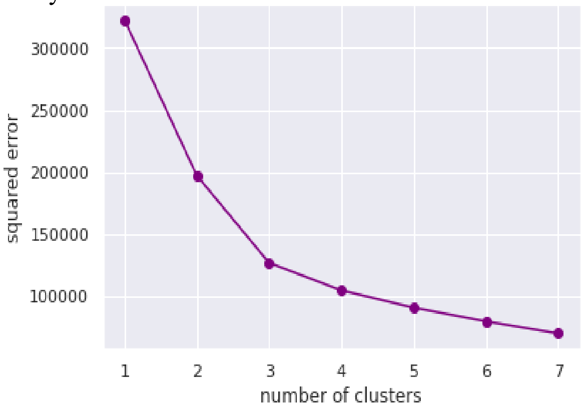

Fig. 7. Elbow Method.

2) Potential Loyalist: Although not as frequently as the regular consumers, this segment made frequent purchases and paid fair prices (though not as high as the loyal customers).

3) New Customers: They only recently began shopping, therefore they didn't buy frequently or spend a lot of money.

4) At-Risk: This category is the biggest in size. These clients are recent, infrequent, and have minimal spending power.

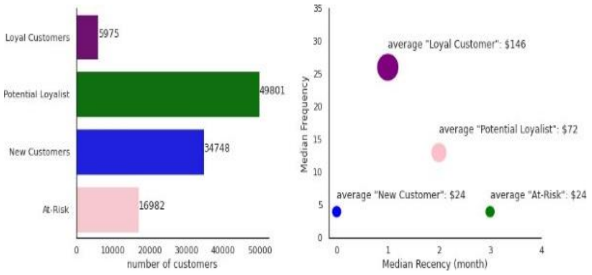

Fig. 8. The 4 groups of customers.

A. Purchase Prediction

In order to assess the efficiency of the different ML algorithms [10] and find out the best model for this problem, we applied them on our dataset that included the new features. We applied the K-fold cross validation technique for a better estimate of the model performance. We also did hyperparameter tuning using Grid Search for getting the optimal parameters which results in increased performance of the model and also helps to avoid overfitting. The results of this experiment are displayed in Table III.

TABLE III. Comparison Of Applied Ml Models

	Model						
	LR+XGB	*DT ensembles bagging*	*XGB*	*RF*	*DT*	*SVM*	*LR*
Accuracy	96.2	95.7	93.1	92.6	92.2	90.1	88.4

The model that performed the best was the LR+XGBoost which had the highest accuracy of 96.2%, closely followed by the DT ensembles Bagging model with an accuracy of 95.7%. The evaluation metrics for the best model is shown in Table IV.

TABLE IV. Evaluation Parameters Of Lr+Xgboost

Evaluation criteria of the best model	
Parameters	*Value*
Accuracy	96.2
Precision	93.4
Recall	98.2
F1-score	96.8

Due to its simplicity and efficiency, logistic regression has grown to be one of the most popular algorithms. The fact that it is only linear, however, prevents it from receiving complex information, and as a result, it requires a certain level of feature engineering to identify the necessary attributes. While LR excels at handling discrete data points, XGB excels at handling continuous data points, therefore combining it achieves the best results, no matter what evaluation criteria is used. Bagging uses parallel learning to increase the accuracy of the weak classifiers. The bias and variance is lowered by using bagging and it also boosts the accuracy. Bagging enhances the base model's accuracy, which is DT here which already performs well on its own. This is why it results in a good model as well.

V. Conclusion And Future Works

In this paper, we did customer segmentation and purchase prediction through consumer behaviour analysis. For ensuring the best performance of the models, we cleaned the data and extracted features via feature engineering. Using RFM-Based K-Means Clustering, we identified 4 distinct consumer groups, which is important for e-commerce businesses. We did a comparison study for finding the best purchase prediction using 7 different ML models, out of which the LR+XGBoost model performed the best. Different measures such as accuracy, precision, recall and F1-score [11] were used to evaluate the model. This study can help an online retailing handle its stock of supplies better and improve sales and thereby enhance its market presence. As part of future study, we intend to create a recommendation system that recommends products to consumers utilizing the WR-MF method, the Apriori algorithm, and collaborative filtering based approach. Five-fold cross validation will be used to assess the effectiveness of the suggested method. In subsequent research, the application of RFM-based K-means clustering and machine learning (ML) models could be extended by incorporating additional features or variables that capture a more comprehensive view of customer behavior. Integration of contextual data, such as customer demographics or psychographic information, could enhance the precision of segmentation and further refine targeting strategies. Additionally, exploring advanced ML algorithms beyond K-means, such as hierarchical clustering or ensemble methods, may improve the accuracy of customer segmentation and predictive modeling, providing more nuanced insights into customer preferences and behaviors for businesses.

References

[1] S Abhishek;Harsha Sathish;Arvind Kumar K;Anjali T.Aiding the Vi- sually Impaired using Artificial Intelligence and Speech Recognition Technology.2022.10.1109/ICIRCA54612.2022.9985659

[2] T Anjali;T R Krishnaprasad;P Jayakumar.A Novel Sentiment Classification of Product Reviews using Levenshtein Distance.2020.10.1109/ICCSP48568.2020.9182198

[3] Vipina Valsan;A.M. Abhishek Sai;Aryadevi Remanidevi Devidas;Maneesha Vinodini Ramesh.Regression based Prediction of Rainfall for Energy Management in a Rural Islanded Micro-Hydro Grid in Kerala.2022.10.1109/GlobConPT57482.2022.9938353

[4] T Anjali;K Chandini;K Anoop;V L La- jish.Temperature Prediction using Machine Learning Approaches.2019.10.1109/ICICICT46008.2019.8993316

[5] Wang XingFen;Yan Xiangbin;Ma Yangchun.Research on User Con- sumption Behavior Prediction Based on Improved XGBoost Algo- rithm.2018.10.1109/BigData.2018.8622235

[6] Dirk Van den Poel; Wouter Buckinx.Predicting online-purchasing be- haviour.2021.10.1016/j.ejor.2004.04.022

[7] Aidin Salamzadeh;Pejman Ebrahimi ;Maryam Soleimani;Maria Fekete- Farkas.Grocery Apps and Consumer Purchase Behavior: Applica- tion of Gaussian Mixture Model and Multi-Layer Perceptron Algo- rithm.2022.10.3390/jrfm15100424

[8] Kaggle "eCommerce Events History in Cosmetics Shop" https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history- in-cosmetics-shop; https://rees46.com/.

[9] Fatemeh Safara.A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic.2020.0.1007/s10614-020-10069-3

[10] Rahan Manoj;S Abhishek;Anjali T.A Strategy for Identification and Prevention of Crime using various Classifiers.2022.10.1109/ICCCNT54827.2022.9984364.

[11] S Abhishek;Harsha Sathish;Arvind Kumar;T Anjali.A Strategy for Detecting Malicious Spam Emails using various Classi- fiers.2022.10.1109/ICIRCA54612.2022.9985614