**University of North Texas**  
 **G. Brint Ryan College of Business**



**DSCI 5260: Section 002 | Business Process Analytics**

**Professor: Dr. Javier Rubio-Herrero**

**Project Name:** Customer Behavior Analysis and Personalized Recommendations on JD.com

***Submitted by:*** Group 6

**Group Members:**

Susmitha Gutta

Srivalli Obulasetty

Sindhuja Pulyala

Shobhana Singh

Sai Sri Madhuri Yatam

Table of Contents

[List of Figures 3](#_Toc165409037)

[List of Tables 4](#_Toc165409038)

[Acknowledgement 5](#_Toc165409039)

[1. Introduction 6](#_Toc165409040)

[2. Literature Review 8](#_Toc165409041)

[2.1 Customer Behavior Analysis: 8](#_Toc165409042)

[2.2 Recommendation Systems: 9](#_Toc165409043)

[3. Data Description and Analysis 12](#_Toc165409044)

[3.1 Data Description: 12](#_Toc165409045)

[3.1.1 User Table 12](#_Toc165409046)

[3.1.2 Delivery Table 13](#_Toc165409047)

[3.1.3 Order Table 13](#_Toc165409048)

[3.1.4 Click Table 14](#_Toc165409049)

[3.1.5 SKU Table 14](#_Toc165409050)

[3.1.6 Inventory Table 15](#_Toc165409051)

[3.2 Data Analysis 15](#_Toc165409052)

[3.2.1 Explorative Analysis 15](#_Toc165409053)

[3.2.2 Descriptive Analysis 18](#_Toc165409054)

[4. Methodology 25](#_Toc165409055)

[4.1 Recommender Systems 25](#_Toc165409056)

[4.2 Experimental Design 26](#_Toc165409057)

[5. Findings 29](#_Toc165409058)

[6. Conclusion 34](#_Toc165409059)

[References 35](#_Toc165409060)

# **List of Figures**

[Figure 1‑1 The U.S ecommerce Sales from year 1999-2023 5](#_Toc165408754)

[Figure 2‑1 Flowchart Of Recommendation Systems [24] 9](#_Toc165408755)

[Figure 3‑1 Distribution of Attribute-2 across all SKU 15](#_Toc165408756)

[Figure 3‑2 Distribution of Attribute-1 across all SKU 15](#_Toc165408757)

[Figure 3‑3 Distribution of Male(M) and Female(F) Across All Users 16](#_Toc165408758)

[Figure 3‑4 Distribution of Age Groups Across All Users 16](#_Toc165408759)

[Figure 3‑5 Distribution of Education Level Across All Users 16](#_Toc165408760)

[Figure 3‑6 Distribution of Marital Status Across All Users (M- Married, S- Single) 16](#_Toc165408761)

[Figure 3‑7 Distribution of Plus Members Across All Users (0 -Non Plus Members, 1-Plus Members) 17](#_Toc165408762)

[Figure 3‑8 Distribution of City Level Across All Users 17](#_Toc165408763)

[Figure 3‑9 Number of products sold in a day 18](#_Toc165408764)

[Figure 3‑10 Number of products sold throughout March 2018 18](#_Toc165408765)

[Figure 3‑11 Sales revenue Across Different Channels 19](#_Toc165408766)

[Figure 3‑12 Analyze Sales Trend over Year and Month-Wise 20](#_Toc165408767)

[Figure3**‑**13Top10UsersBasedonTotalAmountPurchasedandNumberofOrders 22](#_Toc165408768)

[Figure 3‑14 Top 10 Products Customers Ordered 22](#_Toc165408769)

[Figure 3‑15 Top 5 Customer Purchases vs Inventory Stock 23](#_Toc165408770)

# **List of Tables**

[TABLE 1 User Table Attributes and Their Description 12](#_Toc165408993)

[TABLE 2 Delivery Table Attributes and Their Description 13](#_Toc165408994)

[TABLE 3 Order Table Attributes and Their Description 13](#_Toc165408995)

[TABLE 4 Click Table Attributes and Their Description 14](#_Toc165408996)

[TABLE 5 SKU Table Attributes and Their Description 14](#_Toc165408997)

[TABLE 6 Inventory Table Attributes and Their Description 15](#_Toc165408998)

[TABLE 7 Users Engagement Through Different Channels 18](#_Toc165408999)

[TABLE 8 Revenue before Discounts Deductions 20](#_Toc165409000)

[TABLE 9 Revenue after Discount Deductions 20](#_Toc165409001)

[TABLE 10 Distribution of Packages Across Different Delivery Centers Origin and Destination 21](#_Toc165409002)

[TABLE 11 Purchases made Based on the Educational Level and Age 22](#_Toc165409003)

[TABLE 12 Purchases made Based on the Gender and Age 22](#_Toc165409004)

[TABLE 13 Findings based on similar items 32](#_Toc165409005)

[TABLE 14 Findings based on similar users 33](#_Toc165409006)

# **Acknowledgement**

We want to thank everyone who provided assistance to us in completing this project. Primarily, we would like to express our profound gratitude to every member of our team, whose dedication, expertise, and cooperation made this project possible. Each member enhanced our discussions and advanced our cause by bringing unique skills and perspectives to the table. We express our gratitude to each member of the team for your continuous commitment and diligence during the whole project. Their drive, expertise, and desire to go beyond were helpful to us in overcoming obstacles and accomplishing our objectives.

We also thank our lecturers and fellow students for their insightful criticism, direction, and encouragement. Their helpful criticism shaped our project and inspired us to pursue perfection.

Lastly, we would like to thank our friends and family for their steadfast understanding and support during these busy times.

This initiative would not have been possible without the united efforts of all those involved. We are grateful that you joined us on this journey and contributed to making it a really rewarding experience.

# **Abstract**

Understanding e-commerce consumer behavior and user (customer) engagement based on recommendation systems is vital for ecommerce profitability. This study aims to create a recommendation engine specifically for JD.com. We used predictive modeling, by finding specifically identified elements from JD.com’s purchase dataset, to predict future purchases for users. The five machine learning models used are: Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting. A range of validation methods were then applied to evaluate each model’s efficacy. Furthermore, a collaborative filtering recommendation system based on item SKU was implemented to find products and users that are similar, allowing for customized user interfaces and focused product recommendations. While this study had drawbacks, including hardcoded unique records in the dataset and a lack of categorical data, we highlighted how important customized recommendation systems and consumer behavior analysis are to raising user engagement and profitability on e-commerce platforms.

# **Introduction**

People do more online shopping today than they have before. When the first online shopping concept started it was not very popular because deliveries took longer to arrive. In today’s world, deliveries are fast and most of the household items can be delivered the same day or the next day. Advances in ecommerce technology help to make last-mile deliveries possible. For example, Walmart is planning to start drone deliveries in a few cities in the USA [1]. Drone deliveries would make last-minute deliveries even faster and will be potentially revolutionary. One of the main reasons ecommerce is better for consumers is because consumers can compare a variety of items just by sitting at home and comparing quality and price across online stores.

The ecommerce business is very important for a country’s economy, as it reduces the maintenance cost and provides job opportunities in the engineering sector and expands the delivery area. In addition, through ecommerce websites such as Amazon, people can start their own business which eliminates the hurdle of leasing and maintaining space for a brick-and-mortar retail shop. Moreover, throughout the COVID pandemic, when retails shops were closed across the world, online grocery shopping helped to provide staple items and helped to reduce the spread of covid.

In 2022, worldwide, retail ecommerce sales accounted for $5.72 trillion (about $18,000 per person in the US) [2]. This was a 9.7% increase in ecommerce sales from 2021. By 2025, ecommerce sales are expected to reach $7.53 trillion (about $23,000 per person in the US). In 2022, China was among the top three largest ecommerce market, followed by the UK and South Korea. Figure 1 (below) shows the growth of ecommerce over the last 24 years [3]. It is important to notice the exponential increase in ecommerce sales during the COVID pandemic (2019 Q4 to 2020 Q4). We can see ecommerce sales consistently increase year by year, attracting more and more customers.

A graph showing a line

Description automatically generated with medium confidence

Figure 1‑1 The U.S ecommerce Sales from year 1999-2023.

Despite the persistent increase, the ecommerce business also faces many challenges [4]. The ecommerce business setup cost is high and startup ecommerce businesses face a shortage of skilled workers. Moreover, ecommerce is a technology-driven industry, so there are technical demands to stay up to date with the current technology. Language can also be a barrier in the ecommerce business. Social challenges relating to people who do not understand ecommerce’s primary language, English. To attract non-English-speaking customers, ecommerce stores must advertise to them in the language customers can understand. Finally, there are legal challenges when following a country’s laws to make sure all operations are legal.

In the highly competitive ecommerce market, online sellers must stand out to reach the right customers. To identify the right customers and their needs, it is important to study consumer behavior. Consumer behavior analysis is the study which can help us to forecast what customer will need /buy in future from the product or services provided to them. This analysis will help the provider to enhance the product or service which will meet the customer’s need and thus build a strong relationship with the customer and will eventually help to retain the customer base. Consumer behavior can be divided based on multiple factors such as age, gender, and region. Social media marketing is one of the ways to reach out to targeted customers for specific products. Online shopping companies use various technologies such as Google analytics, YouTube analytics, and other social media analytics to identify their customers. Big companies like Amazon and Walmart have their own data analysis programs to analyze customer behaviors.

In the current market, customers have too many choices in crowded marketplaces and a short time to find a product which suits their needs. With new products constantly entering the market it is difficult to understand product specification easily. The ecommerce and brick-and-mortar retail sectors both face challenges in effectively engaging customers and increasing sales. Product recommendation engines generate personalized recommendations and predictive offers tailored to specific audiences. To address this problem, we aim to provide an algorithm using personalized information provided by users. This algorithm will recommend and offer products or suggestions that suit customer interests and behaviors.

To study consumer behavior and make personalized recommendation, we gathered data from China’s biggest ecommerce site, JD.com. JD.com was founded by Richard Liu Qiangdong in 2004. Unlike his competitors, Alibaba, JD.com focused on providing a platform for a third-party merchant, emphasized direct sales, vertical integration, and extensive supply chain management [5]. The data set includes a wide range of information. It captures a “full customer experience cycle” that begins when a customer browses through a product available on the platform, before placing his or her order, and ends when customers receive the product at his or her location. The data sets provide information on 2.5 million customers with over 30,000 SKUs in one specific product category during March in 2018.

We have developed various research questions to understand customer/consumer behavior. We wanted to study if customers’ behavior follows any pattern or trend in the kind of products they are more likely to buy. If there is any pattern shown between customer and product, then we can help customers find products they might be interested in. We would also like to develop a personalized product recommendation system that recommends products based on an individual customer’s browsing history, region, and purchase history. Then, we want to conduct customer segmentation analysis considering factors such as gender, age, and JD plus membership. We will create personalize product recommendation, here we will set up customer profile, find products and categories that drive sales, find comparable items and similar customers based on their behavior, perfect recommendations, and display them strategically to attract customers.

Our methodology will start by constructing a database which will consist of data derived from JD.com. This database includes seven different tables which consists of information like delivery of each product, inventory table will provide information about the availability of each SKU (SKU\_ID) at each warehouse(dc\_ID), network table provides information about the assignment if different warehouses located in different districts(dc\_ID) to different geographical regions(Region\_ID), order table consists of 486,928 unique customer orders, SKU, user table describes the characteristics of each of the 457,298 users table describes the characteristics of each the 457,298 users who purchased at least one of the SKUs and click table establishes the linkage between users and SKUs through their browsing history. Once we have all the tables ready in our database. Then we will connect to python and perform descriptive analysis. We will also use algorithms like neural networks and decision trees for predictive analysis.

We have found various factors during data exploration which contribute to JD.com’s success. By building a predictive model, we have discovered future customers for JD.com. This will be helpful in targeting certain groups of customers and providing more customized and personalized experiences. We have built a recommendation system for finding out similar products and similar users. This can be helpful in grouping similar users under one group to help personalize the user experience within each group and products can be recommended based on other users in same group.

The paper is organized as follows: section 2 consists of related research work, section 3 consists of data pre-processing, section 4 describes the explanation of various algorithms used in the project, results / findings will be elaborated in section 5 of the paper and section 6 will consists of conclusion.

# **Literature Review**

Our literature review will consist of two sections. In the first section, we will describe customer behavior analysis, the factors which affect customer behavior, and various techniques which help us identify the pattern recognition associated with customer behavior. In the second section, we will discuss the personalized recommendation system. Here, our focus is to study why personalized recommendation system is important, the evaluation of recommendation system and process involved in making a recommendation system and lastly, we will talk about some of the limitations of recommendation system.

## **Customer Behavior Analysis:**

Our goal is to understand the factors that impact consumers online purchasing decisions and evaluate the effectiveness of recommendation systems. Consumer behavioris the study of an individual or group and the processes they use to select or discard, secure, use, and dispose of products, services, experiences, and the impacts that these processes have on the consumer and society [6]. Consumer Behavior analysis can help us to forecast what a customer will need /buy in future from the products or services provided to them. This analysis will help the provider to enhance the product or service which will meet the customer’s need and thus build a strong relationship with the customer and will eventually help to retain a valuable customer.

Customer buying behavior is driven by many factors such as the word of mouth, Gender, Culture and Economic Structure [7]. The word-of-mouth influences internet purchasing decisions [8]. With the advent of the digital era, communication has discovered a new medium for exchange- the internet [7]. Information is no longer a commodity owned by a few people or organizations and the study also reveals that with the ease of expressing opinions on the internet, people can quickly share their dissatisfaction about a product they purchased online by posting negative reviews or comments which can significantly influence purchasing decisions [7].

In addition, there are some differences in how men and women approach online shopping [9]. It is essential to recognize that individual preferences and behaviors can vary greatly. For men, their decision-making process is mainly driven by practical considerations and functional factors [10]. Women, on the other hand, value the emotional [10] and social aspects of online shopping. Aspects like packaging aesthetics or product appearance can also appeal more to women's emotional engagement with shopping [11].

Online purchase behavior also varies by country and culture [12]. Western customers are more inclined to openly express rage and dissatisfaction, whereas Eastern customers tend to accumulate anger slowly and may not immediately express negative emotions [12].People in the countries with low GDP, focus on basic needs like food and safety. They tolerate lower-quality services, have less access to technology, and prefer human interaction over high-tech solutions [13].Whereas people in the countries with higher GDP, have more money and higher expectations. They expect top-notch service, have less tolerance for inefficiency, and are more tech-savvy, prioritizing high-tech solutions over personal interaction [13].

When customers are trying to buy or browse through the website to buy a particular product, they will follow some pattern to find those products [14]. These patterns define how the customer purchases or discards the product in relation to the quality, place of purchase, competing with other brands, frequency, and other factors. Sellers want to know consumer behavior so that buying and selling activities can be carried out productively [14]. Ensuring customer satisfaction and fostering trust in online shopping are crucial aspects of e-commerce, particularly for attracting interest and encouraging repeat purchases from the same platform. Customers can read reviews to gauge a company's reputation and service quality. Online shopping is perceived as a time and cost-saving option [7], providing customers with recommendations for related items based on their preferences. Additionally, customers can compare prices and discounts between online and offline products, and analyze various online payment options, including cashback rewards on credit cards. Companies are increasingly leveraging data mining techniques [9] to understand consumer behavior. Data mining characteristics such as pattern identification and information generation from past datasets are invaluable for understanding customer behavior.

There are various data modelling techniques which can be used to find these patterns. When a customer visits the online shopping sites, they leave some valuable information when they log on to the server. This information can be used to figure out the customer's buying habits and then to identify the business performance [15]. We can predict the future customer buying behavior by analyzing the historical data of earlier customers. We can also improve in areas where a company is not making profits. There has been various work done to study the technique involved in analyzing pattern recognition. One such study has used K-Means and cluster analysis to study pattern recognition [16]. In this paper, we want to start by exploring our data with K-means, cluster analysis and explore other techniques. Based on the conceptual framework, we will summarize the predictive models in three ways. We will first use the Probabilistic classifier followed by classical Data Mining Classifier, then Deep learning. These predictive models will help us to predict the buying behavior of customers, their intent and interest [16].

## **Recommendation Systems:**

Generally, recommender systems are used as information filtering tools to offer personalized content for users. Its main aim is to reduce the user’s effort and time required to search for similar information. They are widely used for different applications such as web, books, YouTube shots, Instagram reels, tourism, retail, movies, music, research resources. So, it is important to have a high-quality and personalized recommender system for providing recommendations to the users in different applications [17].

E-commerce removes the need for conventional “middlemen” like distributors, wholesalers, and retailers by enabling direct purchases between consumer and manufacturers [18]. However, given how quickly the world is changing, it might be difficult to offer customers everything that is reasonably priced. Recommendation algorithms were created at that time to fill this gap, and they are currently present on every e-commerce website [18].

Using Recommendation systems (RS) became essential to make decisions for almost every individual [19]. Daily, a lot of people struggle to choose how to spend their time and money at different stages of life to make their future better [19]. Normally, people seek opinions from others, use third party software, utilize online resources like the internet, or just follow the crowd [19]. Since these recommendations are generalized and not specific to a person, they are not effective and not much of some help [19]. To address this challenge, we use a recommendation system. RSs are software tools and techniques that enable the recommendation of “objects” that may be of most interest to targeted/designated users [20]. RSs aid customers to choose products when there is wide range of products [21].

Over the past 25 years the recommender system has evolved into a contemporary tool. Usenet is the global discussion/exchange system and became the origin for recommender system [22]. Between the late 1970s and1992, this system was used for very little research, primarily on content filtering. The initial literature expansion started towards the close of the 1990s [23]. The number of current research studies in recommender systems has expanded due to the enhanced advancement in data mining techniques and infrastructure developments that make big data analysis easy. When searching academic databases for “recommender system,” 30,945 documents retrieved after 2005, as opposed to 1,531 records between 1969 and 2005 shows the increase in research scope on recommender system. [24]

A basic flow chart of RS is shown in Figure 2 below [25]. Here, the User is a representation of multiple consumers with their own requirements and choices. There are different ways like numerical evaluation, verbal evaluation, and binary evaluation which can be used to fill the user item matrix. This matrix reflects the consumer views about the content in the system. There are different ways like numerical evaluation, verbal evaluation, and binary evaluation which can be used to fill the user item matrix [25]. Item information will have information about the product contents, year of manufacture, and typers.

Recommender systems deal with users and items, with each user giving a rating to an item [17]. User ratings are collected using implicit or explicit methods. Implicit ratings are the ratings of the item provided by user indirectly from the interaction with the item. Explicit ratings are provided directly by the user by selecting a value like rating the item within the scale provided [17]. Implicit ratings for different items are captured by clickstream data or the time spent by the user on the webpage. Most recommender systems use both explicit and implicit methods to get user ratings [17]. These feedback or ratings given by the user are arranged in a user-item matrix which is also called utility matrix.

The utility matrix may contain missing values because not all users provide product feedback. Finding these missing values in the utility matrix is the major problem [17]. This is a difficult task as the initial matrix is usually very sparse because user’s rate few items. We are highly interested in the high user ratings because we can suggest them back to the users. The efficiency of a recommender system generally depends on the type of algorithm used to create the user item matrix and the nature of the data source. [17]



**INPUT**

User Profile

Demographic Information

Trend, Interests

User – Item matrix

Product Information Database

**SYSTEM**

Recommendation System

**OUTPUT**

Recommendation List



**USER**

**USER**

Figure 2‑1 Flowchart Of Recommendation Systems [24]

Through their recommendation system, YouTube was able to increase the view count of a related video by 30% of total views of that video [26]. Furthermore, it helped to increase the diversity, which helped viewers to discover more videos of their interest instead of popular videos only [26]. So, when we develop a recommendation system for JD.com, customers will be able to explore the products based on their past purchase or preferences instead of seeing only popular products.

Although providing personalized recommendation services to customers contributes to the success of e-commerce, there are also various limitations. Some deployment related challenges like cold start problem, synonymy, scalability, and data sparsity [27]. Cold start problem is a condition where the system is not able to produce efficient recommendations for the new users either because the user has not rated any item or has rated very few items. It occurs when a novel item is added, or new users enter the system [27]. Synonyms occur when related items have different names, or when the same item is represented by many other names in the system. In this case, the recommender systems cannot make a distinction between two or more related items, such as babywear or baby cloth [28]. Recommended systems which use collaborative filtering techniques face the scalability issue. In this technique, much training data is needed, which causes scalability. This problem occurs when the amount of data used as input to a recommender system increases very fast. [28]

The data sparsity problem occurs when active users rate very few items. This leads to a sparse user-item matrix, not being able to find neighbors successfully, and producing a weak recommendation system [27]. Some other challenges are difficulty in handling real-time user feedback, difficulty in gathering implicit user data, difficulty in choosing the correct implementation techniques, and difficulty in measuring system performance [27].

# **Data Description and Analysis**

The JD.com datasets give a thorough understanding of the operations related to every single SKU during March 2018. There are six different tables. 1. User, 2. Click, 3. Order, 4. Delivery, 5. SKU, and 6. Inventory. Since we have six tables and over 20 million data points stored on these tables, we created a database for easy processing and data analysis.

We started by creating a database in MS SQL Server Management Studio. We created all six tables in the database by defining datatypes for each attribute and assigning a Primary Key (PK) to each table. Next, we imported data from Excel file into tables (created in the database). After successfully importing the data, we carefully studied all the tables and noticed that there were some duplicate records in the Orders table. In total, we dropped 8,890 duplicate records from the Orders table. Next, we used Pandas Library from Python to import data from SQL to Python (pd.read\_sql\_query) for further data analysis.

The rest of this section is divided into two parts. The first part is data description, in which a brief introduction about the tables and their attributes is provided. This introduction is followed by pre-processing findings, which detail the null values, duplicate values, and how they were handled. The second part is data analysis, where data is explored, and descriptive questions of this research are answered.

## **Data Description:**

### **User Table**

The User table holds the information on each customer who purchased at least one product from JD.com. The table has 457,298 unique users and ten columns capturing their demographic data. Every row is unique as it holds a unique UserID corresponding to each user. The details of each column are specified below.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attribute Name** | **Attribute Description** |
| 1 | UserID | The Unique code to identify each customer. |
| 2 | UserLevel | This attribute represents a value from 0 to any number. The higher the purchases, the higher the number assigned to the customer. |
| 3 | FirstOrderMonth | The attribute represents the month in which a customer made their first order in yyyy-mm format. |
| 4 | Plus | This attribute represents 0 for regular customers, 1 for plus members. |
| 5 | Gender | The attribute represents the gender of the user. |
| 6 | Age | This attribute represents the age of the user. |
| 7 | Marital status | This attribute represents the marital status of the user. M- married and S- single |
| 8 | Education | This attribute represents customer education level: 4 for postgraduate, 3 for a bachelor’s degree, 2 for high school, 1 for less than high school, and -1 for unknown). |
| 9 | PurchasePower | This attribute represents the user’s purchasing power (1 for highest purchasing power where 5 represents lowest purchasing power). |
|  |  |  |
| 10 | CityLevel | This attribute represents the level of industrialization of a city based on values between 1 to 5 (1 for highly industrialized areas and 5 for smaller cities). |

TABLE 1 User Table Attributes and Their Description

There were no duplicates present in the User table. Unique values for each column were explored to find if any unknown values are present. If any unknown values are present, we programmed a code to replace it with the value from a similar group. For example, if the marital status of a user is unknown, the program finds a similar user with same age, education, gender, plus membership, and purchase power. It replaces the unknown value of marital status with the marital status of a similar user. The Gender attribute had unknown values which were handled by the programmed code. The categorical values “F” and “M” are replaced with numerical values 1 and 0 respectively for analysis purpose with 1 given to the most occurring value.

The various age groups in Age column are “<=15,” “16-25,” “26-35,” “36-45,” “46-55,” “>=56,” and “U.” The unknowns in the Age column represented by “U” were replaced with value from a similar group. The string values representing the different age groups are converted to numeric values according to their order. 0 is assigned to the youngest age group and 5 to the oldest user age group.

In the Marital Status, we have three unique values:-, married represented by “M,” singles represented by “S,” and unknowns represented by “U.” The unknowns “U” are replaced with value from a similar group, and the values “S” and “M” were then converted to 1 and 0, where 1 was assigned to the most frequent occurrence, “M.” In the UserLevel column, the values were arranged in categorial order from 0 to 6. The value 0 represents the first-time buyer, and the number increases as the purchases increase up to level 5. Numeric value 6 represents vendors.

In the Education column, the unknowns represented by -1 are replaced with value from a similar group and then the values were assigned from level 0 to level 3 in order. The value 0 represents the lowest degree of education, and 3 represents highest education level. In the CityLevel column the unknown values represented by -1 are replaced with value from a similar group and then the values are assigned from value 0 to value 4. The value 0 represents the smallest level in cities and 4 represents highest industrialized cities like Shanghai and Beijing. In the PurchasePower column the unknown values represented by -1 were replaced with value from a similar group and then the values were assigned from 0 to 4, where 0 represents users having lowest purchase power, and 4 represents users with highest purchase power.

### **Delivery Table**

The Delivery table built the connection between each order (OrderID) and multiple shipping packages (multiple PackageIDs) as there are scenarios where an order is shipped in multiple deliveries. It consists of 6 columns and 293,229 rows which represent the number of packages delivered by JD Logistics at a particular event. The description of the attributes is given below.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attribute Name** | **Attribute Description** |
| 1 | PackageID | This attribute represents a unique code for package identification. |
| 2 | OrderID | This attribute represents unique code assigned on order. |
| 3 | Type | This attribute represents 3 types based on the party owned. 1 represents first party owned products, likewise for 2 and 3. |
| 4 | ShipOutTime | This attribute represents time of shipping from the warehouse (distribution center) |
| 5 | ArrStationTime | This attribute represents the time of arrival of the package at the station. |
| 6 | ArrTime | This attribute represents the time at which the customer received the package. |

TABLE 2 Delivery Table Attributes and Their Description

We discovered that there were no null values and no duplicates in the Delivery table. We ignored the columns with ID details and time details as their unique value count is extreme. We have also found that in the Type column, there are two types of values 1 and 0. Here, 1 describes the product sold by JD.com, and 0 describes a product sold by third party vendors.

### **Order Table**

The Order table gives information on all orders during March 2018. There are 541,098 unique orders captured in rows and 17 columns in this table. The same customer might make multiple purchases, so in that case, the same OrderID is assigned for all unique SKU\_IDs. The attributes of the table are listed below.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attribute Name** | **Attribute Description** |
| 1 | OrderID | This attribute represents unique code assigned on order. |
| 2 | UserID | This attribute represents a unique code to identify customers. |
| 3 | SKU\_ID | This attribute represents unique code for product. |
| 4 | OrderDate | This attribute represents the date on which the order is placed. |
| 5 | OrderTime | This attribute represents the time at which the order is placed. |
| 6 | Quantity | This attribute represents number of units placed |
| 7 | Type | This attribute represents 3 types based on the party owned. 1 represents first party owned products, likewise for 2,3. |
| 8 | Promise | This attribute represents how many days are expected to deliver the order. |
| 9 | OrginalUnitPrice | This attribute represents original unit price |
| 10 | FinalUnitPrice | This attribute represents purchasing price |
| 11 | DirectDiscountPerUnit | This attribute represents direct discounts on the products. |
| 12 | QuantityDiscountPerUnit | This attribute represents a discount on purchasing quantity. |
| 13 | BundleDiscountPerUnit | This attribute represents a discount on bundle purchase. |
| 14 | CouponDiscountPerUnit | This attribute represents a discount due to coupon. |
| 15 | GiftItem | This attribute represents gift promotion. 1 represents yes, 0 represents no |
| 16 | DcOr | This attribute represents the identity of the distribution center from where the order is shipped. |
| 17 | DcDes | This attribute represents the destination address for shipment. |

TABLE 3 Order Table Attributes and Their Description

There were no null values in the Order table of JD.com. We removed 8,890 duplicate records from the Order table in the initial phase when the records were loaded into the database. We also converted the datatype for the OrderDate attribute to datetime for future analysis.

### **Click Table**

Each record in this table represents the browsing history of the customer on specific SKUs. This table has over 20 million browsing history records of 2.5 million users (about twice the population of Hawaii) and four fields. However, the users on this table may or may not end up purchasing.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attribute Name** | **Attribute Description** |
| 1 | SKU\_ID | This attribute represents a unique identification code for a product. |
| 2 | UserID | This attribute represents a unique identification code for the user. |
| 3 | Request\_Time | This attribute represents the time of requesting SKU page by user or accessing time. |
| 4 | Channel | This attribute represents platform of purchase. |

TABLE 4 Click Table Attributes and Their Description

There were no null values found in the Click table. We have found 487,267 duplicates in the Click table, which we dropped. The users who haven’t made a purchase but contributed to click history are represented by a hyphen “-” in the dataset. We converted the Request\_Time attribute datatype to datetime and separated it into two attributes, Request\_Time and Request\_Date, for further processing.

### **SKU Table**

Inthe SKU (stock keeping unit**)** table**,** each record is distinguished by SKU\_ID which is specific to the product. However, there are products that are the same but sold by different parties (such as different brands) which are considered as different products. There are 31,868 products and 7 attributes in the table.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attribute Name** | **Attribute Description** |
| 1 | SKU\_ID | This attribute represents a unique identification code for a product. |
| 2 | Type | This attribute represents a party or brand. |
| 3 | BrandID | This attribute represents a unique identification code for brand. |
| 4 | Attribute1 | This attribute represents values from 1 to 4, the higher the value, the higher the performance of product. |
| 5 | Attribute2 | This attribute represents a value from 30 to 100, higher the value higher the performance of product. |
| 6 | Active\_Date | This attribute represents the date on which SKU was first introduced. |
| 7 | Deactive\_Date | This attribute represents the date on which SKU is terminated. |
| 8 | Rating | This attribute represents satisfaction rate of user with the product |

TABLE 5 SKU Table Attributes and Their Description

The data for SKU table is available only for March 2018. The products that are introduced or terminated out of the March have no data available under Active\_Date and Deactive\_Date. We have dropped these columnsas they are not useful for our analysis and converted attribute1 and attribute2 datatypes from object to float.

### **Inventory Table**

The Inventory table maintains the availability of the products at each distribution center or warehouse at the end of each day. We assumed that, if a SKU is not found in the center that means there were no records of the SKU shown in that center for that day.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attribute Name** | **Attribute Description** |
| 1 | DcID | This attribute represents the identity of distribution center. |
| 2 | SKU\_ID | This attribute represents a unique identification code for a product. |
| 3 | Date | This attribute represents the date of availability recorded. |
| 4 | RegionID | This attribute represents the region where the distribution center is located. |

TABLE 6 Inventory Table Attributes and Their Description

There are no null values found in the table. This table didn’t require any pre-processing for the further exploratory analysis done in the later part of this research.

## **Data Analysis**

This section is divided into 2 parts. The first part is explorative analysis. In this section, we will present findings discovered when the tabular data was analyzed. The second part of this section is a descriptive analysis where descriptive questions of this research are answered.

### **Explorative Analysis**

From the SKU table, we found that most of the products sold by JD.com have Attribute1 value as 3 and Attribute2 value as 100. Attribute1 and Attribute2 specify the performance and quality of products. Through this we can say that most of the products sold by JD.Com are of better quality as both their attributes of performance are above average level.

A graph with blue bars

Description automatically generated

Figure 3‑1 Distribution of Attribute-2 across all SKU

A graph with blue bars

Description automatically generated

Figure 3‑2 Distribution of Attribute-1 across all SKU

In the User table, the age, education level, gender, marital status, whether a user is a Plus member or not and city level of users are explored. In Figure 5, we can see that the value 1, which represents female customers, distinctly exceeds value 0, which represents male customers. Figure 6 shows the distribution of different age groups. The highest number of customers are represented by value 2. These customers are in the age group 26-35. Figure 7 depicts the education level. Education level-3 is the majority, which means most users have bachelor’s degree. Figure 8 shows married users tend to buy more than unmarried users. Figure 9 shows that very few users are plus members of JD.Com. Figure 10 shows city level 3 has the most users.

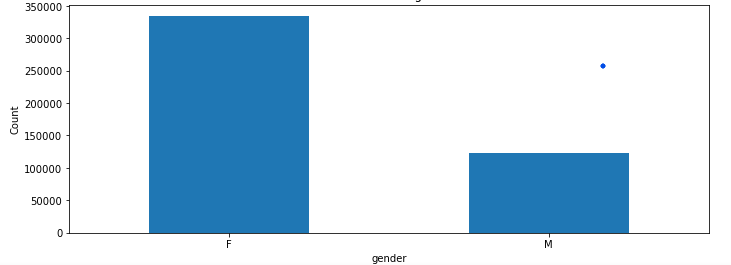


Figure 3‑3 Distribution of Male(M) and Female(F) Across All Users

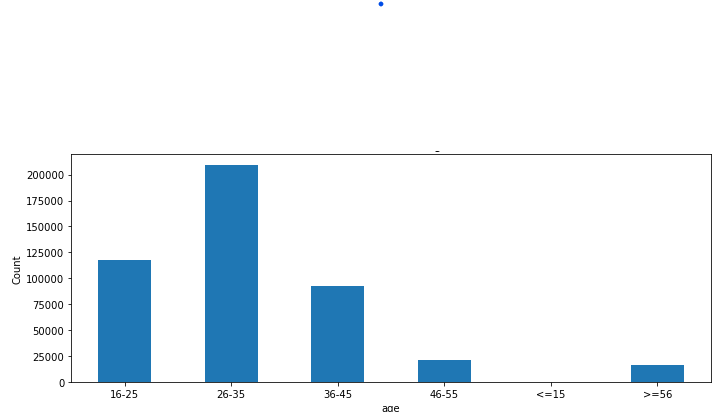


Figure 3‑4 Distribution of Age Groups Across All Users

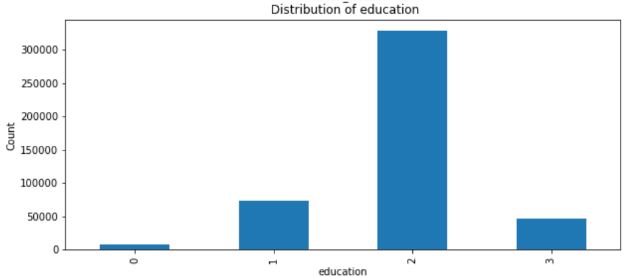


Figure 3‑5 Distribution of Education Level Across All Users

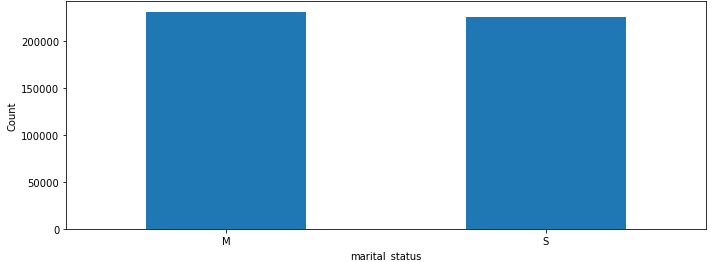


Figure 3‑6 Distribution of Marital Status Across All Users (M- Married, S- Single)

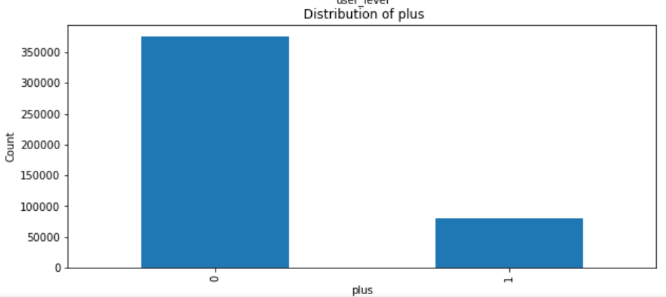


Figure 3‑7 Distribution of Plus Members Across All Users (0 -Non Plus Members, 1-Plus Members)

A graph of a number of blue squares

Description automatically generated with medium confidence

Figure 3‑8 Distribution of City Level Across All Users

Table 7 (below) illustrates the contribution of different channels in terms of user preference and engagement for JD.com. The 'app' channel represents JD.com's mobile application, which is experiencing significant user preference. Following closely is the 'WeChat' channel, which refers to a mini program operating within the WeChat social media platform. Channels 'PC' and 'Mobile' represent web browsers accessed via computers and mobile devices, respectively.

|  |  |
| --- | --- |
| **Channel** | **Number of Users for Each Channel** |
| App | 14,702,575 |
| WeChat | 2,591,321 |
| PC | 1,021,665 |
| Mobile | 958,385 |
| Others | 453,302 |

TABLE 7 Users Engagement Through Different Channels

### **Descriptive Analysis**

To analyze our descriptive questions, we have merged Order table and Delivery table on OrderID through inner join. This resulted in a dataset containing orders which are delivered. We have grouped OrderHour based on quantity; through this we have obtained hourly quantity sold in a day. Figure 11 (below) shows that the peak time for orders during a day is at 10 a.m. with a total quantity of 29,954. The second peak time is observed at 10 p.m. with a total quantity of 28,575.

A graph with orange dots

Description automatically generated

Figure 3‑9 Number of products sold in a day

We have grouped OrderDay based on quantity; through this we have obtained the day of the month with the highest number of orders. Figure 12(below) shows that the highest orders were placed on the first of the month with a total quantity of 37,803 followed by eighth of the month with a total quantity of 23,988. In March 2018 Chinese people celebrated Arbor Day on March 12, which is a public holiday to commemorate the passing of Dr. Sun Yat-sen, known as the father of modern China. The Zhonghe festival which is a traditional Chinese festival held on second day of the second month of the Chinese calendar was on March 18. These two festivals might be the reason for more orders on the first of the month and we can see a decrease in orders in between the holiday period i.e March-10 to March-20.

A graph with a line graph

Description automatically generated

Figure 3‑10 Number of products sold throughout March 2018

To find out the actual orders made through each channel we merged click table with order table on SKU\_ID. We found that sales were made through each channel. There are five different channels in total, named app (JD.com mobile app), WeChat, mobile, PC, and others. However, from combining orders and click table, the sales were mostly made on JD.com mobile app, mobile, WeChat (social media). Figure 13 (below) shows that almost 580k sales happened through WeChat followed by app and mobile.

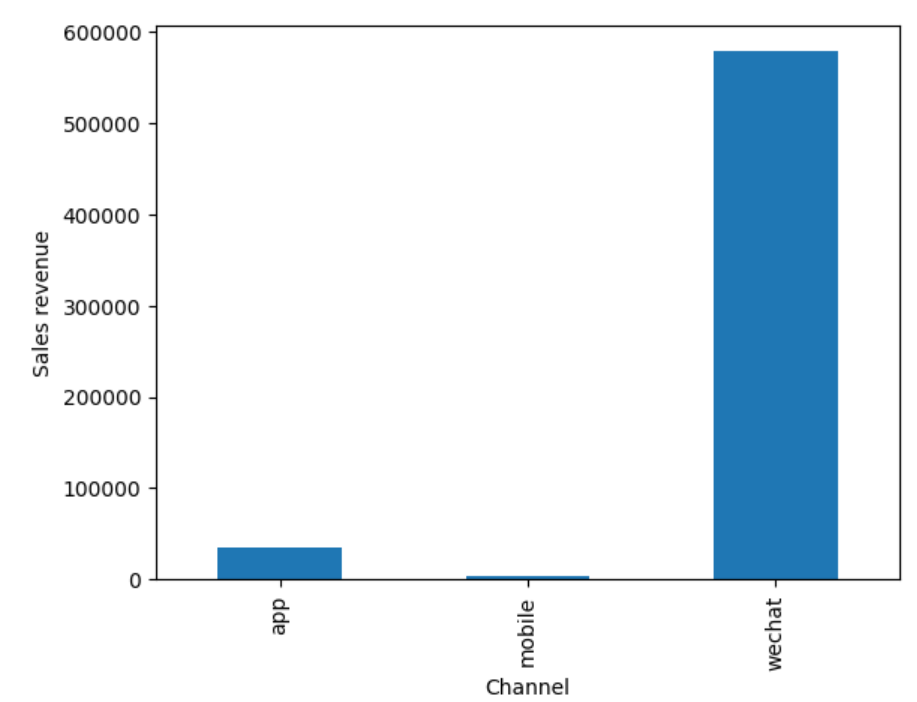


Figure 3‑11 Sales revenue Across Different Channels

Similarly, there are discounts available on products based on the channel type used for buying. The revenues on each channel, with and without discounts, are also calculated for each platform and the results are shown in Table 8 and Table 9. WeChat is a major platform for revenue for JD.com.

|  |  |
| --- | --- |
| **Channel** | **Revenue in RMB (With discounts)** |
| WeChat | 578,295 |
| JD.com mobile app | 34,980 |
| Mobile | 4,323 |

TABLE 8 Revenue before Discounts Deductions

|  |  |
| --- | --- |
| **Channel** | **Revenue in RMB (After discount deductions)** |
| WeChat | 352,787 |
| JD.com mobile app | 27,656 |
| Mobile | 3,055 |

TABLE 9 Revenue after Discount Deductions

Analyzing the sales trend for March in 2018 as we have Order table, SKU table with data available only for March. It reveals a consistent increase from March 19th to the end of the month. However, there is a notable spike on March 8th, followed by a sudden drop thereafter. Further observation highlights a pattern where sales decrease during weekends, suggesting that people tend to prefer outdoor activities over online shopping during those times.

A graph showing a line

Description automatically generated

Figure 3‑12 Analyze Sales Trend over Year and Month-Wise

The sales quantity by date indicates notable spikes on March 1st, 8th, and 28th, with each of these dates recording the highest number of sales within the observed period.

Table 10 (below) shows that top five delivery centers with the highest packages delivered. It also explains that more packages can be delivered if the delivery center origin and delivery center destination are same. It is observed that very few packages were delivered if the origin and destination number were not the same.

|  |  |  |
| --- | --- | --- |
| **Delivery Center Origin** | **Delivery Center Destination** | **Count of Packages Delivered** |
| 5.0 | 5.0 | 31,658 |
| 2.0 | 2.0 | 23,431 |
| 9.0 | 9.0 | 23,061 |
| 4.0 | 4.0 | 20,423 |
| 24.0 | 24.0 | 12,631 |

TABLE 10 Distribution of Packages Across Different Delivery Centers Origin and Destination

A graph of blue rectangular bars

Description automatically generated with medium confidence

Figure. 15. Distribution of Packages Across Different Delivery Centers Origin

It is observed that the age group of 26-35 (all genders included) with education level of bachelor’s degree are making more purchases. Orders on gender level for this age group is presented in the below table, indicates that females are placing more orders than males.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Education** | | | |
| Less than High School | High School Diploma or Equivalent | Bachelor’s Degree | Post-Graduate Degree |
| <=15 | 2 | 5 | 16 | 3 |
| 16-25 | 3,284 | 33,232 | 74,224 | 7,230 |
| 26-35 | 4,250 | 39,254 | 136,293 | 29,186 |
| 36-45 | 2,278 | 17,242 | 53,736 | 19,181 |
| 46-55 | 581 | 4,529 | 12,756 | 3,502 |
| >=56 | 436 | 3,441 | 10,245 | 2,492 |

TABLE 11 Purchases made Based on the Educational Level and Age

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Gender** | **Age** | | | | | |
| <=15 | 16-25 | 26-35 | 36-45 | 46-55 | >=56 |
| Female | 17 | 88,244 | 152,294 | 66,490 | 15,621 | 12,109 |
| Male | 9 | 29,726 | 56,689 | 25,947 | 5,647 | 4,505 |

TABLE 12 Purchases made Based on the Gender and Age

Figure 16 (below) shows the top 10 customers who have made a substantial contribution to JD.com's revenue. For security and privacy concerns, only IDs of the users and products are displayed, not the actual names. For example, customer with UserID “de8763924d” stands out by purchasing 1162 products in the month of March generating revenue of 21,2088.24 RMB (Chinese Yuan). It is 0.45% of total revenue generated in March. This customer is not a JD plus member, and we can recommend this user as a potential JD plus member to retain the valuable customer.

A graph with red and blue lines

Description automatically generated

Figure 3‑13 Top 10 Users Based on Total Amount Purchased and Number of Orders

Out of the top 10 most frequently ordered products, 5 products are highly stocked in inventory which shows it is highly available with effective inventory management to meet customer requirements. Figure 17 (below) has the top 5 products IDs that are most frequently ordered by the users and highly available items in the inventory. Only the IDs of the products are displayed for security and privacy concerns. For example, '8dc4a01dec' is ordered 9891 times with considerable remaining of 1305 units post the purchases.

A graph of blue bars

Description automatically generated

Figure 3‑14 Top 10 Products Customers Ordered

A graph of sales and inventory

Description automatically generated

Figure 3‑15 Top 5 Customer Purchases vs Inventory Stock

# **Methodology**

This section firstly introduces the recommender systems briefly, and explains the methods followed for the research in detail. The four distinct strategies of recommender systems are explained in the first part of this section. The second part explains the predictive models and the two recommender systems used in our research.

## **Recommender Systems**

Recommendation systems (RS) are built on several algorithms but mainly focused on four distinct strategies: (1) Content-Based filtering (CBF), (2) Collaborative Filtering (CF), (3) Knowledge- based filtering (KBF), and (4) Hybrid Methods (HF). The two primary filtering techniques used for the Recommendation system are content based and collaborative methods [29]. Information processing begins with a content-based approach [30]. CBF, which depends on correlation between contents, is one of the finest techniques for creating suggestions. However, CBF is not an optimal solution when feature values are entered manually. This can be manageable for small datasets, but thousands of new products are being added to JD.com daily, so this task is impossible [31]. Goldberg et al. [32] introduced a new technique called collaborative approach as an alternative to content-based approach. The purpose of CF was to clearly give users item-specific information [33]. This approach is one of the most widely used approaches in literature and became popular in product recommendation systems. It also achieved great success in the market [34]. Various studies use the CF algorithm with different techniques to create recommendation list, based mainly on a user’s past evolution. The CF algorithm works on aggregating the preferences of similar users assuming they have similar interests. How similar the two customers are can be determined by looking at how much customer 1’s past evaluations of the same product match those of customer 2. Users who closely resemble the active user take a crucial role in determining the active user’s interest in the product being evaluated. For example, Amazon uses CF methods to provide recommendations based on historical purchases made by a consumer as well as the customer who has bought the same or comparable products. E-commerce platforms employ this technique to divide the market according to consumer behavior in relation to psychographic and demographic criteria. This method’s primary goal is to compile user preferences into a database and analyze it as efficiently as possible to generate recommendations [35]. CBF recommendations are based on the content or descriptive attributes of the item [36]. Unlike CF this technique does not require an active database for purchase history [36]. For instance, consider a customer named user1 who has purchased a laptop through Amazon and rated the product with comments, but we do not have access to the other users who have purchased a laptop and their ratings, comments. In this case using CF is not useful. However, the laptop has a brand name, and its specifications will help to sort specific laptop accessories like laptop bags, and compatible drivers, which can be recommended to user1.

Even though, as being widely used algorithms, they both have their own limitations. Clear recommendations are often provided by CBF because of the use of contents or keywords; however, products with certain keywords will not be recommended if the user1 has never used those words. This problem of not being able to recommend is referred to as “cold start” [29]. This process tends to lessen the diversity of the RS. Whereas with CF, we face sparsity in cold start as the main problem [37]. Sparsity issues arise when many users have not liked or disliked any relevant item.

The challenges of CF and CBF are addressed by knowledge-based filtering and Hybrid methods. Solving the cold start problem is one of the main strengths of KB approach. CBF or CF methods alone cannot be used to suggest products because of the volume of data that has been gathered in recent years and the constantly shifting preference of customers [38]. For this reason, most customers favor browsing e-commerce sites that incorporate both approaches [39]. The solution to avoid the problems in CF and CBF is implementing both filtering techniques and aggregating the results. Knowledge-based RS uses explicit knowledge about products and users to create a knowledge-based criterion to generate recommendations [31]. But in the KB method, there is the requirement of creating a database to maintain product related information or area of use. For example, books stores, which store features like author, genre, and year in a database, provide offers to users based on these features. To overcome the limitations of these three methods, sometimes we combine two or more recommendation techniques, which can be termed as a hybrid approach [40]. For example, the KB approach combined with CF typically yields better recommendations. As a result, Tran [41] presented a hybrid RS architecture which integrates CF and KBF methods. Hybrid solutions are sometimes used to improve algorithm accuracy rather than to solve a specific issue. In fact, all techniques have issues like cold start, insufficient data, non-rational behavior. Numerous scholars have presented hybrid methodologies to tackle these issues. As such, one of the hybrid approach’s drawbacks is the requirement for knowledge-based engineering, which can be challenging.

## **Experimental Design**

The User table and the Order table are merged through left join, giving purchase details for users who purchased in March and null values for Order table columns for users who haven’t purchased any product in March. A new column named “Purchase” was added to the dataset which captures if user made a purchase in March or not. If the user made a purchase it is represented by 1. If a purchase is not made, it is represented by 0. Overall, we had 540,018 users who made a purchase in March and 3322 users who did not purchase in March. So, the random over\_sampling method is applied, and we created a dataset with an equal number of users who made a purchase and who haven’t made a purchase.

We divided the dataset into train dataset and test dataset and trained five models with “UserLevel”, “Plus”, “Gender”, “Age”, “MaritalStatus”, “Education”, “CityLevel” and “PurchasePower” as features and “MarchPurchase” as the target column. The modeling methods used for this project for building predictive models are Logistic Regression, Decision Tree, Random Forest, Naive Bayes and Gradient Boosting. Features importance is observed, and dimensionality reduction is done to tune the models.

Logistic regression is very common and a useful statistical method. The name of this technique is derived from the log-it transformation with the dependent variable [42]. This model was used to predict whether a customer will make a purchase or not based on various factors such as “UserLevel”, “Plus”, “Gender”, “Age”, “MaritalStatus”, “Education”, “CityLevel” and “PurchasePower.” The major advantage of using regression analysis is because it is used to predict the outcome of a dependent variable that is binary or categorical on nature. In logistic regression, a log-it transformation is applied on the odds, i.e., the probability of success divided by the probability of failure. This is known as log odds or natural logarithm of odds, and this is represented as below equation [43]

Where:

p = the probability that event Y occurs

= odds that event Y occurs

= represents the coefficients used in the model

= represents error

ln () = Natural log of the odds (log odds), or logit

In this project the dependent variable is the “Purchased” column. The column indicates whether the customer made a purchase in March or not. The model was trained on the training datasets, which was split into 80% training data and 20% test data using the scikit-learn library.

A Decision Tree is a set of nested tests which we use to “divide and conquer” (also known as recursive manner) a prediction problem [44]. The tree metaphor is useful for organizing or representing data. For instance, it can represent reporting the hierarchy in an organization. The advantage of using decision tree is how branching maps the notion of categorical outcomes. Decision trees let us view our data as filtering down through several decisions until finally coming to rest on a leaf. Each branch of a tree represents a test, each node represents the results of a test, and each leaf represents a class assignment. The instance is classified by navigating them from the root of the tree to a leaf, according to the results of the test along the path.

The Decision Tree algorithm, first proposed by Quinlan, stated that the algorithm operates on the principle of recursively partitioning a dataset and incremental tree construction [45]. The algorithm for decision tree consists of using training data then selecting the best attribute to split on. The algorithm identifies all the possible values for that attribute and for each value, create a new child node, and for each child node, if the node is pure, the algorithm stops. If the node is not pure, it recursively calls the algorithm to split again.

The Random Forest is a widely used machine learning algorithm, which was developed by Leo Breiman and Adele Cutler [46]. This algorithm is based on the output of multiple decision trees to a single output/result. The Random Forest algorithm is made up from a collection of Decision Trees. Each Decision Tree has a sample of data drawn from a training set with replacement (called Bootstrap sample). From the training sample, one-third of data is set aside, which is called as out of bag (OOB) sample. This is done to reduce the correlation between Decision Trees. The determination of prediction will vary depending on the type of problem.

For example, for regression tasks, the average of individual trees will be calculated, and for classification problems, a majority vote will be taken. Finally, the OOB sample will be used for cross-validation for finalizing the prediction. The key benefits of using random forest are it reduces the risk of overfitting, providing flexibility as it can handle regression and classification problems.

The Naive Bayes algorithm (Bayesian classifier) is a classification approach based on the probability of an event occurring, given that some of the events have already occurred [47]. The Navie Bayes classifier simplifies learning by assuming features are independent for a given class. The algorithm is called “naive” because it assumes that predictors are statistically independent. Even though this belief is never true, the results of Naive Bayes algorithm are more accurate than most of the other predictive algorithms and hence are used widely. Naive Bayes is based on Baye’s theorem, which states that the probability of y, given that x has already occurred is equal to the probability of X, given that Y occurred times the probability of Y divided by the probability of X [48].

Bayes theorem:

Where:

Y and x are events

P(y) is a class prior probability: the probability of event y

P(x) is a predictor prior probability: the probability of event x

P() is a likelihood probability: the probability of event x, given that y is true

P() is a posterior probability: the probability of event y, given that x is true

The advantage of using Naive bayes for predicting modelling is that this algorithm is based on mathematics, so it is relatively easy to understand and build. Also, it works very fast with large datasets. The major drawback is that it assumes independence of features [49].

The highly customizable Gradient Boosting Machine (GBM) learning algorithm is based on constructive strategy of ensemble formation [50]. The core idea of boosting is to add new models to ensemble sequentially. The algorithm consecutively fits the new model to ensemble to provide more accurate estimate of the target variable. In other words, the ensemble technique combines the prediction of various weak learners, typically decision tree. The main objective of using this algorithm is to improve the predictive performance by optimizing the model’s weight based on the error of the previous iteration and gradually reducing the prediction error and thus enhancing the model’s accuracy [51]. This algorithm can be used in two ways: if the problem is a classification problem, then we use Gradient Boosting Regressor and if the problem is classification problem, then we use Gradient Boosting Classifier. Since in this project we have classification problems, we are using Gradient Boosting Classifier. For Gradient Boosting Classifier we will use “loos function” as log likelihood.

The major advantage of Gradient Boosting is that they are more accurate than other models and train faster on large datasets. In this project we have large datasets, so it will be beneficial to use this model over other models. One of the major drawbacks is that they are computationally expensive as they take a long time to train the model [51].

In User-based Collaborative Filtering, predicting a user's interest in an item is done by analyzing ratings from similar users. Similarity between users is determined by using cosine similarity [52]. The predicted rating for an item by a test user is computed by considering ratings from similar users, with more similar users given higher weight. One user might rate items mostly high, while another might rate them lower overall. Adjusted cosine similarity adjusts these differences, so we can better compare their preferences [53]. It is particularly useful for item-based recommendations and performs equally well for user-based collaborative filtering. User-based Collaborative filtering method only considers a small portion of the user-item matrix for recommendation, specifically the known ratings from similar users, which are referred to as “similar user ratings.”

The Item-based Collaborative Filtering applies a similar concept to user-based approaches, but instead of comparing users, it compares items. Each item is sorted and re-indexed based on its dissimilarity to the test item in the user-item matrix, with ratings from more similar items weighted stronger [52]. The prediction formula involves summing the ratings of similar items weighted by their similarity. Similarity between items was computed using cosine similarity. Like top-5 similar items, a set of top-5 similar users can also be generated. Both the Collaborative Filtering approaches are widely used for product recommendations because they are effective. User-based filtering performs better with larger datasets, while item-based filtering performs better with smaller data [53]. Each method has its own strengths and weaknesses, but when used together, they create the core of collaborative filtering, a strong tool for customized recommendations.

# **Findings**

## **Predicting the Customer Purchases:**

We created a binary 'purchased' column based on the presence of 'order\_ID'. If a user made a purchase, we assigned 1 to indicate purchase, otherwise, it is assigned 0. Next, we selected relevant features and a target variable. We then checked for missing values in the features and the target variable. Finally, we visualized the correlation between the selected features using a heatmap.

A screenshot of a computer screen

Description automatically generated

Figure 5‑1 Top 5 Correlation Matrix

After preprocessing and checking for correlation, we divided our data into a 70% training and 30% testing split. Then, we trained four different classification models: Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting. After training, we evaluated each model's performance based on Accuracy and ROC AUC.

LogisticRegression()

Cross-Validation Scores: [0.99394856 0.99393752 0.9939]

Mean CV Score: 0.9939

Logistic Regression Accuracy: 0.9938

Logistic Regression Classification Report:

Logistic Regression ROC AUC: 0.5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logistic Regression | Precision | Recall | F1 score | Support |
| 0 | 0.00 | 0.00 | 0.00 | 1677 |
| 1 | 0.99 | 1.00 | 1.00 | 269993 |

TABLE 13 Logistic Regression Confusion Matrix

DecisionTreeClassifier()

Cross-Validation Scores: [0.99374979 0.99370562 0.99366138]

Mean CV Score: 0.9937055985415476

Decision Tree Accuracy: 0.9936503846578569

Decision Tree Classification Report:

Decision Tree ROC AUC: 0.5013926066119151

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decision Tree | Precision | Recall | F1 score | Support |
| 0 | 0.09 | 0.00 | 0.01 | 1677 |
| 1 | 0.99 | 1.00 | 1.00 | 269993 |

TABLE 14 Decision Tree Confusion Matrix

RandomForestClassifier()

Cross-Validation Scores: [0.99389335 0.99386022 0.99389328]

Mean CV Score: 0.9938822836934436

Random Forest Accuracy: 0.9937865793057754

Random Forest Classification Report:

Random Forest ROC AUC: 0.49997962910149524

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | Precision | Recall | F1 score | Support |
| 0 | 0.00 | 0.00 | 0.00 | 1677 |
| 1 | 0.99 | 1.00 | 1.00 | 269993 |

TABLE 15 random Forest Confusion Matrix

GradientBoostingClassifier()

Cross-Validation Scores: [0.99393752 0.99393752 0.99393745]

Mean CV Score: 0.9939374976992504

Gradient Boosting Accuracy: 0.9938307505429381

Gradient Boosting Classification Report:

Gradient Boosting ROC AUC: 0.5005944510220203

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gradient Boosting | Precision | Recall | F1 score | Support |
| 0 | 0.67 | 0.00 | 0.00 | 1677 |
| 1 | 0.99 | 1.00 | 1.00 | 269993 |

TABLE 16 Gradient Boosting Confusion Matrix

All models demonstrated extremely high accuracy of 99%, which means they were able to correctly predict the outcome for most cases. However, when we look at the ROC AUC (Receiver Operating Characteristic - Area Under the Curve) score, we see a value of 0.5 for all models, ROC AUC is a measure of how well a model can identify between classes. A score of 0.5 shows that none of the models can distinguish between the two classes any better than random chance, which would represent a bad model. ROC AUC suggests that the models do not effectively distinguish between the positive and negative classes. The models are essentially predicting all cases as the majority class (class 1) and not providing any useful information.

The precision, recall, and f1-scores for the minority class (Class 0) were low for all models, indicating an inability to correctly classify the minority class. This suggests that despite using different techniques, the imbalance in the dataset severely hampers the effectiveness of traditional models and evaluation metrics.

When we checked for class imbalance, it gave us this result.

|  |  |
| --- | --- |
| 1 | 540018 |
| 0 | 3322 |

TABLE 17 Class Distribution

The class distribution is highly skewed towards class 1, with 540,018 instances when compared to 3,322 instances of class 0. This represents a significant imbalance that could bias a machine learning model towards predicting the majority class. To address this issue, SMOTE is applied to the dataset. SMOTE works by synthesizing new examples for the minority class 0 based on the existing examples. It does this by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. Synthetic points are added between the chosen point and its neighbors. After applying SMOTE, the class distribution becomes balanced by having an equal number of instances 540,018 each. This balanced distribution is important for training machine learning models that are sensitive to class imbalance, as it helps improve the generalizability of the model by ensuring it is not biased towards the majority class.

|  |  |
| --- | --- |
| 1 | 540018 |
| 0 | 540018 |

TABLE 18 After Over-Sampling Class Distribution

We performed classification techniques again to predict customer purchases after balancing the classes in the data.

**Logistic Regression:** This model gave an accuracy of around 55%, meaning the model is 55% accurate in predicting whether a user will make a purchase or not. Precision for predicting purchases (class 1) is about 55%, and recall (the ability to find all positive instances) is around 0.54%.

LogisticRegression()

Cross-Validation Scores: [0.5467 0.5450 0.5437]

Mean CV Score: 0.5451

Logistic Regression Accuracy: 0.5462

Logistic Regression Classification Report:

Logistic Regression ROC AUC: 0.5462

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logistic Regression | Precision | Recall | F1 score | Support |
| 0 | 0.54 | 0.55 | 0.55 | 269783 |
| 1 | 0.55 | 0.54 | 0.54 | 270235 |

TABLE 19 Logistic Regression

**Decision Tree:** This model gave an accuracy of approximately 68%. Precision for predicting purchases is around 74%, and recall is around 54%. ROC AUC here it is about 0.68.

DecisionTreeClassifier()

Cross-Validation Scores: [0.6760 0.6754 0.6747]

Mean CV Score: 0.6754

Decision Tree Accuracy: 0.6749

Decision Tree Classification Report:

Decision Tree ROC AUC: 0.6751

|  |  |  |  |
| --- | --- | --- | --- |
| Decision Tree | Precision | Recall | F1 score |
| 0 | 0.64 | 0.81 | 0.71 |
| 1 | 0.74 | 0.54 | 0.64 |

TABLE 20 Decision Tree

In the **Naive Bayes** model, we achieved an accuracy of 56%. The precision scores, which gives the proportion of correctly identified items for each class, were 57% for the non-purchased class and 56% for the purchased class. The recall was 52% for the non-purchased class and 61% for the purchased class. This means the model performs moderately well in identifying customers who are likely to make a purchase.

GaussianNB()

Cross-Validation Scores: [0.5620 0.5651 0.5650]

Mean CV Score: 0.5640

Naive Bayes Accuracy: 0.5642

Naive Bayes Classification Report:

Naive Bayes ROC AUC: 0.5641

|  |  |  |  |
| --- | --- | --- | --- |
| Gaussian Naive Bayes | Precision | Recall | F1 score |
| 0 | 0.57 | 0.52 | 0.54 |
| 1 | 0.56 | 0.61 | 0.58 |

TABLE 21 Naive Bayes

**Random Forest:** This model gave an accuracy of 68%. The random forest model’s performance is like the decision tree. Precision for predicting purchases (class 1) is about 74%, and recall is around 55%.

Random Forest Classifier()

Cross-Validation Scores: [0.6764 0.6758 0.6751]

Mean CV Score: 0.6757

Random Forest Accuracy: 0.6751

Random Forest Classification Report:

Random Forest ROC AUC: 0.6753

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest | Precision | Recall | F1 score |
| 0 | 0.64 | 0.81 | 0.71 |
| 1 | 0.74 | 0.55 | 0.64 |

TABLE 22 Random Forest

|  |  |
| --- | --- |
| user\_level | 0. 250 |
| age | 0. 171 |
| education | 0. 122 |
| purchase\_power | 0. 114 |
| gender | 0. 049 |
| Marital\_status | 0. 0419 |
| plus | 0. 0327 |

TABLE 23 Feature Importance for Random Forest:

**Gradient Boosting:** This model achieved an accuracy of 59%, and precision for predicting purchases (class 1) is about 61%. Recall is around 53%. Overall, decision trees and random forest models gave the best accuracy and overall performance among the classifiers tested.

GradientBoostingClassifier()

Cross-Validation Scores: [0.5922 0.5947 0.5927]

Mean CV Score: 0.5932

Gradient Boosting Accuracy: 0.5929

Gradient Boosting Classification Report:

Gradient Boosting ROC AUC: 0.5930

|  |  |  |  |
| --- | --- | --- | --- |
| Gradient Boosting | Precision | Recall | F1 score |
| 0 | 0.58 | 0.66 | 0.61 |
| 1 | 0.61 | 0.53 | 0.57 |

TABLE 24 Gradient Boosting

|  |  |
| --- | --- |
| user\_level | 0.256 |
| age | 0.203 |
| plus | 0.047 |
| education | 0. 199 |
| purchase\_power | 0. 166 |
| Marital\_status | 0. 050 |
| gender | 0. 030 |

TABLE 25 Feature Importance for Gradient Boosting:

These models are Analyzed based on their ability to correctly predict whether a user will make a purchase or not. The decision tree and random forest models perform slightly better than logistic regression and Naive Bayes and gradient boosting. Among the models, logistics and gradient boosting have the lowest accuracy compared to the others. Feature importance indicates the contribution of each feature to the model's decision-making process. For example, in random forest, user\_level and city\_level are the most important features, while in gradient boosting, user\_level and plus (a feature indicating premium subscription) are the most influential. This information can be useful for analysis and decision-making in marketing or user-targeting strategies.

## **Product Similarity Analysis - Collaborative Filtering (KNN model)**

We have built a recommendation system in two ways. In the first method, we used an item based collaborative recommendations system. This method recommends new products based on previous purchases of customers. We started building a recommendations system by creating a user-item matrix with userID and sku\_ID attributes from the order table. This matrix was created by using scipy csr matrix library in python. After building a matrix we executed the K nearest neighbor (KNN) algorithm to find similar neighbors or users. KNN is a supervised machine learning algorithm which uses distance to find similar users [54]. For measuring the distance, we used the cosine similarity measure. This method is a matrix that calculates the cosine of an angle between two or more data points to determine if they are pointing in the same direction. The cosine similarity ranges between 0 and 1. A value of 1 indicates the highest level of similarity, whereas a value of 0 indicates the two data points are not similar at all [55]. For example, we provided the model sku\_ID as 01c99f44fd and found 3d7b8701a5 sku\_ID as the similar product with a similarity score of 0.975(Table 13). This means that if a customer brought a product with sku\_ID as 01c99f44fd, then the customer might also like the product with sku\_ID 3d7b8701a5.

|  |  |
| --- | --- |
| **Sku\_ID** | **Similarity score** |
| 3c24a96dfd | 0.991 |
| cb61667e38 | 0.989 |
| a7fd50fbf9 | 0.987 |
| 15c43e97a3 | 0.985 |
| 853a3eedc7 | 0.982 |
| d8038520e7 | 0.982 |
| c351610d10 | 0.982 |
| 48331e69ab | 0.982 |
| 2be549e2f1 | 0.982 |
| 3d7b8701a5 | 0.975 |

TABLE 26 Findings based on similar items.

In the second method we have given userID as input to find a similar user. This method will be helpful in grouping the similar user into one category. We can improve user engagement by targeting similar users in a group. This can be done by creating a specific promotion or discount which resonates with each user in a group, leading to higher engagement rates. By using the similarity between the user, we can create more customized targeted marketed campaigns. We can segment users based on their interests and characteristics, which can increase the effectiveness of marketing campaigns. Also, we can create a more personalized experience for each user by grouping them using their similar characteristics or behavior. For example, when the user with similar browsing or purchasing history is grouped together, we can modify recommendations with more effectiveness.

In our model for finding similar user, we have used KNN algorithm and cosine similarity measure. For example, we have provided userID as e49625c3ee and found that userID as 8d5f73c188 is like input userID with a similarity score of 0.951. Table 13 represents the results for top 10 similar user for input.

|  |  |
| --- | --- |
| **UserID** | **Similarity score** |
| c4c37a4b57 | 0.994 |
| 411ddd08e5 | 0.994 |
| 0757d730f6 | 0.994 |
| 904fbf8b97 | 0.993 |
| 95c58a09a4 | 0.991 |
| 09fca54905 | 0.987 |
| ec4b9ba058 | 0.984 |
| ad71b48a82 | 0.966 |
| aa0fcf0fb6 | 0.965 |
| 8d5f73c188 | 0.951 |

TABLE 27 Findings based on similar users.

# **Conclusion**

The aim of this paper is to study customer behavior regarding e-commerce and understanding and developing a recommendation system for JD.com, one of China’s largest e-commerce platforms. By exploring vast datasets that captures a complete customer experience cycle. We have found significant insights into customer purchasing patterns and developed predictive models to forecast future customers behavior. We identified some of the major factors and hidden insights during the data exploration which were crucial for generating revenue for JD.com.

Our data exploratory analysis revealed some important trends within data. We have observed a higher number of purchases among certain demographic groups, particularly among those aged between 25 to 35. Female customer predominated over male customers and customers holding bachelor's degree emerged as most frequent purchases on JD.com. We can provide targeted marketing strategies that can be more effective by focusing on these segments. Also, we have identified the most popular products and customers which are crucial for inventory and marketing strategies.

The project involved building five different machine learning algorithms to find future purchases for customers. We used models like Logistic Regression, Decision Trees, Random Forest and Gradient Boosting to predict future buying behavior of users based on factors such as demographics and past purchase activity. Then we checked how well predictions based on the cross-validation score, mean cv score, accuracy, classification report, and ROC AUC for evaluating these models. The accuracy for Logistic Regression is 0.55, for Decision Tree it's 0.67, for Random Forest it's around 0.68, and for Gradient Boosting and Naive Bayes it's 0.59 and 0.56 respectively. From the results we ca say that Decision Tree and Random Forest are performing well compared to the remaining models and the ROC AUC for Random Forest and Decision Tree is high, approximately 0.67, indicating that these models are performing well in predicting the unseen data.

Moreover, we have successfully built a recommendation system using an item-based collaborative system, in which we found similar users. This can be helpful in grouping similar users under one group to help personalize the user experience within each group, and products can be recommended based on other users in same group. Also, with the help of this recommendation model, we have grouped similar products under one category. This will be helpful in enhancing the shopping experience for customers by making it easier to find related products.

Despite the success, there are some limitations. The major limitation of our project is that we don't have any categorical data and challenges associated with cold start and data sparsity. Due to this limitation, we were not able to explore other methods of recommendation systems like content-based recommendation systems. Future work could explore hybrid models that involve various data sources and techniques, and every unique record is hardcoded in the JD.com datasets. Because of this, we don't have actual details of records like the product name or brand name.

In conclusion, our project has built a robust model for advancing JD.com’s capabilities in predicting analytics and personalized recommendations. Customer behavior analysis and personalized recommendation systems are important tools for businesses who want to thrive in today's competitive, online market. By understanding the needs of every customer, JD.com can enhance customer engagement, satisfaction, sales and thus generate more revenue.

# **References**

|  |  |
| --- | --- |
| [1] | "Sky High Ambitions: Walmart To Make Largest Drone Delivery Expansion of Any U.S. Retailer," 9 january 2024. [Online]. Available: https://corporate.walmart.com/news/2024/01/09/sky-high-ambitions-walmart-to-make-largest-drone-delivery-expansion-of-any-us-retailer. |
| [2] | M. Osman, *Is Ecommerce profitable? A study of the current landscape plus tips on how to thrive as an online business,* Nexcess, 2023. |
| [3] | "U.S. E-Commerce Sales (adjusted for seasonal variation)," [Online]. Available: marketplacepulse.com. |
| [4] | C. Cano-Espinoza, "Moving Beyond Start-Ups: Challenges & Opportunities For Ecommerce Businesses," 2019. |
| [5] | P. Westberg, "The Story of Richard Liu Qiangdong and JD.com," 2023. |
| [6] | J. P. Peter, "Consumer behavior and marketing strategy," 1987. |
| [7] | J. Arndt, "Role of product-related conversations in the diffusion of a new product," *Journal of Marketing Research,* vol. 4, no. 3, pp. 291-295, 1967. |
| [8] | S. J and P. S, "Exploring the effects of consumers′ dissatisfaction level on complaint behaviours," *European Journal of Marketing,* pp. 7-21, 1991. |
| [9] | G. Hofstede, Culture's Consequences: International Differences in Work-Related Values, Newbury Park, CA: Sage, 1984. |
| [10] | B. Hasan, "Exploring gender differences in online shopping attitude," *Computers in Human Behavior,* vol. 26, no. 4, pp. 597-601, 2010. |
| [11] | K. Soen and Bo Yin, "Customer Behaviour Analysis of E- Commerce," 2019. |
| [12] | G. Hofstede, Cultures and Organizations: Software of the Mind, London: McGraw-Hill Book Company, 1991. |
| [13] | N. Malhotra, F. Ulgado, J. Agarwal and I. Baalbaki, "A comparative evaluation of the dimensions of service quality between developed and developing countries," *International Marketing Review,* vol. 11, no. 2, pp. 5-15, 1994. |
| [14] | S. A. Kalaivani, *FACTORS INFLUENCING BUYERS BEHAVIOUR WHILE PURCHASING,* Madurai, 2016. |
| [15] | P. P. K. Maheswari, *Predicting customer behavior in online shopping using SVM classifier,* Tamil Nadu, India: Department of Computer Applications, 2017. |
| [16] | M. F. Mahendra Pratap Yadav, *Mining the customer behavior using web usage mining in e-commerce,* Coimbatore: IEEE, 2012. |
| [17] | D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data,* 2022. |
| [18] | D. A. B. Jadhav, "A Study of E-Commerce and Online Shopping," *Excel Journal of Engineering Technology and Management Science,* vol. 1, no. 20, 2021. |
| [19] | D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, "Recommender Systems: An Introduction," *Cambridge University Press,* 2010. |
| [20] | B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *Association for Computing Machinery: New York,* pp. 285-295, 2001. |
| [21] | P. Marx, T. Hennig-Thurau and A. Marchand, "Increasing Consumers’ Understanding of Recommender Results: A Preference-Based Hybrid Algorithm with Strong Explanatory Power," *Association for Computing Machinery: New York,* pp. 297-300, 2010. |
| [22] | V. Bhatnagar, "Collaborative Filtering Using Data Mining and Analysis," *IGI Global: Hershey,* 2016. |
| [23] | M. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," *Mach. Learn,* pp. 313-331, 1997. |
| [24] | G. Ş. C. I. E. Yıldız E, "A Hyper-Personalized Product Recommendation System Focused on Customer Segmentation: An Application in the Fashion Retail Industry.," *Journal of Theoretical and Applied Electronic Commerce Research,* vol. 18, no. 1, pp. 571-596, 2023. |
| [25] | F. Ricci, L. Rokach and B. Shapira, "Recommender Systems Handbook," *Springer: New York,* no. 2nd ed, 2015. |
| [26] | R. S. K. a. L. G. Zhou, "The impact of YouTube recommendation system on video views.," *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement,* 2010. |
| [27] | D. R. Mala Dutta, "A systematic review and research perspective on recommender systems," *Journal of Big Data ,* 2022. |
| [28] | Y. F. B. O. F.O. Isinkaye, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal,* vol. 16, no. 3, pp. 261-273, 2015. |
| [29] | R. Mahajan, "Collaborative Filtering Using Data Mining and Analysis," in *Review of Data Mining Techniques and Parameters for Recommendation of Effective Adaptive E-Learning System.*, IGI Global, 2017, pp. 1-23. |
| [30] | G. Akrivas, M. Wallace, G. Andreou, G. Stamou and S. Kollias, "Context—Sensitive Semantic Query Expansion," *Proceedings 2002 IEEE International Conference on Artificial Intelligence Systems,* pp. 109-114, 2002. |
| [31] | Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim and R. Kashef, "Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities," *Appl. Sci,* p. 7748, 2020. |
| [32] | D. Goldberg, D. Nichols, B. Oki and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM,* pp. 61-70, 1992. |
| [33] | J. F. D. H. J. S. S. Schafer, "Collaborative Filtering Recommender Systems," *Lecture Notes in Computer Science, Springer,* vol. 4321, 2007. |
| [34] | J. Schafer, J. Konstan and J. Riedl, "Recommender Systems in E-Commerce," *Association for Computing Machinery: New York,* pp. 158-166, 1999. |
| [35] | X. K. X. B. W. W. T. M. B. F. X. H. Liu, "Context-based collaborative filtering for citation recommendation," *IEEE Access,* vol. 3, pp. 1695-1703, 2015. |
| [36] | C. Aggarwal, "An Introduction to Recommender Systems," in *Recommender Systems*, Springer, 2016, pp. 1-28. |
| [37] | P. P. W. C. H. G. C. S., "The development of an Ontology-Based Adaptive Personalized Recommender System," in *Proceedings of the2010 International Conference OnElectronics and Information Engineering*, 2010. |
| [38] | I. K. M. R.-T. L. Hao Ma, "Mining Web Graphs for Recommendations," *IEEE Transactions on Knowledge & Data Engineering,* vol. 24, pp. 1051-1064, 2012. |
| [39] | P. M. a. N. N. J. a. H. M. a. S. A. A. a. D. A. Alamdari, "A Systematic Study on the Recommender Systems in the E-Commerce," *IEEE Access,* vol. 8, pp. 115694-115716, 2020. |
| [40] | R. Burke, "Hybrid Recommender Systems: Survey and Experiments," *User Model. User-Adapted Interact,* no. 331-370, 2002. |
| [41] | T. Tran, "Designing recommender systems for e-commerce: An integration approach," in *Proc. 8th Int. Conf. Electron. Commerce New E-Commerce Innov. Conquering Current Barriers Obstacles Limitations Conducting Successful Bus. Internet*, 2006. |
| [42] | Adarsh Anand, "Predicting Customer’s Satisfaction (Dissatisfaction) Using Logistic Regression," *Research Gate,* 2016. |
| [43] | S. Gupta, "What is Logistic Regression? A Guide to the Formula & Equation," 2021. [Online]. Available: https://www.springboard.com/blog/data-science/what-is-logistic-regression/. |
| [44] | O. M. Lior Rokach, "Decision Trees," *Research gate,* no. The Data Mining and Knowledge Discovery Handbook (pp.165-192), 2005. |
| [45] | A. S. T. S. N. S. G. Hrystyna Lipyanina, "Decision Tree Based Targeting Model of Customer". |
| [46] | [Online]. Available: https://www.ibm.com/topics/random-forest. |
| [47] | S. H. R. H. X. Z. Hong Chen, "Improved naive Bayes classification algorithm for traffic risk management," *Springer Open,* 2021. |
| [48] | "wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Bayes%27\_theorem. |
| [49] | "MLNerds," 30 july 2021. [Online]. Available: https://machinelearninginterview.com/topics/machine-learning/naive-bayes-classifier-advantages-and-disadvantages/. |
| [50] | A. K. Alexey Natekin, "Gradient boosting machines, a tutorial," *Front Neurorobot,* 2013. |
| [51] | A. Saini, "Gradient Boosting Algorithm: A Complete Guide for Beginners," *Analytics vidhya,* 2024. |
| [52] | J. Wang, D. V. A. P and R. Marcel JT, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 501-508, 2006. |
| [53] | P. Yu, "Collaborative filtering recommendation algorithm based on both user and item," *2015 4th International Conference on Computer Science and Network Technology (ICCSNT),* vol. 1, pp. 239--243, 2015. |
| [54] | H. W. D. B. Y. B. &. K. G. Gongde Guo, "KNN Model-Based Approach in Classification," *SpringerLink.* |
| [55] | J. P. Jiawei Han, "Cosine similarity," *Science direct,* 2012. |
| [56] | L. Z. W. W. X. &. L. S. Peng, "Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: Considering the impact of product involvement.," *Information & Management,* pp. 317-328, 2019. |