Consumer Behavior Analysis and Purchase Prediction in E-Commerce Using Spark-Based Deep Learning

author: Yasser Nael Fayeq Kuhail (ykuhail@students.iugaza.edu.ps)

Supervised by: Dr. Rebhi S. Baraka and is a requirement of the course Big Data (ICTS 6339), Faculty of Information Technology, The Islamic University of Gaza (IUG).

Abstract With the rapid increasing of e-commerce platforms, massive volumes of user behavior data are generated daily, including product views, cart additions, favorites, and purchases. Extracting valuable insights from this data and predicting actual purchasing decisions remain critical challenges. For businesses, each decision involves substantial financial and human resource investments, making accurate prediction of customer behavior essential for effective sales strategies. This report proposes a big data–driven framework for consumer behavior analysis and purchase prediction using sequential deep learning models—Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). The study consumes the large-scale User Behavior Data from Taobao (Alibaba), a benchmark dataset that fulfills big data requirements with millions of sequential user interactions, making it ideal for modeling large-scale behavioral patterns from other datasets. The pipeline begins with data preprocessing and exploratory analysis, followed by customer segmentation using RFM metrics and K-means clustering. Sequential modeling is then applied to capture temporal dependencies in user actions, which are essential for understanding behavior progression and predicting purchases in time-ordered sequences. RNN and LSTM are chosen because they are better at handling temporal dependencies in behavioral data, while traditional models ignore these sequences and process events independently, where LSTM further overcomes vanishing gradient issues to capture long-term dependencies. Experimental evaluations compare their performance using accuracy. The findings highlight clear patterns in customer activity, low conversion rates even among top-selling categories, and a large share of churn-risk users. Sequential modeling further confirms the advantage of deep learning, with LSTM outperforming RNN. Overall, the study demonstrates the potential of integrating big data and deep learning to improve recommendations, increase conversions, and support data-driven decisions in e-commerce.

***Index Terms*—Apache Spark, big data analytics, consumer behavior prediction, e-commerce, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), user behavior analysis**

I. Introduction

With the rapid increase of the e-commerce industry, customers can browse, compare, and purchase products from a wide range of categories with unprecedented convenience and flexibility [7]. Modern e-commerce platforms, such as Taobao and Alibaba, record vast volumes of user interaction data, including product views, cart additions, favorites, and purchases, within their web server logs [8]. This data represents a valuable resource for understanding consumer preferences, improving recommendation systems, and optimizing marketing strategies [5].

However, transforming raw behavioral data into actionable insights remains a challenging task. The continuous and automated collection of user actions produces massive datasets that may contain redundant, noisy, or incomplete information, which complicates the analysis process [9]. Additionally, traditional analytical methods often fail to capture the temporal dependencies inherent in sequential user actions, leading to suboptimal prediction of actual purchase decisions [10][11].

In an increasingly competitive market, where each business decision carries substantial financial and human resource implications [12], the ability to predict customer behavior accurately is critical. This includes identifying peak purchasing times, understanding the relationship between views and purchases, detecting early signs of customer churn, and predicting the next interaction in a user’s behavioral sequence [13].

To address these challenges, this research proposes a big data–driven pipeline that integrates distributed data processing using **Apache Spark**, which has been shown to be highly effective for large-scale data analytics due to its in-memory computation and scalability in big data environments [14], with sequential deep learning models—Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)—which are chosen not only for their ability to capture temporal dependencies but also for their

.

robustness in modeling user behavior as ordered sequences of actions. Unlike traditional machine learning models such as logistic regression or XGBoost, which treat user interactions as independent observations, RNN and LSTM can preserve the contextual relationship between actions (e.g., view → add-to-cart → purchase). This is particularly important in e-commerce, where the meaning of one action often depends on the actions that precede it. Moreover, while vanilla RNNs suffer from vanishing and exploding gradients, LSTM mitigates this issue through gating mechanisms, making it capable of learning long-term dependencies such as delayed purchases or recurring shopping patterns. These characteristics make RNN and LSTM more suitable for purchase prediction tasks in e-commerce than non-sequential models like CNN, which excel in spatial pattern recognition but are less effective in sequential behavioral modeling [15]. the study consumers the User Behavior Data from Taobao (Alibaba), which qualifies as big data under the 5Vs framework: it contains a massive volume of over one billion interaction logs from millions of users, collected at high velocity during continuous online activity; it offers variety by recording multiple event types such as view, cart, favorite, and purchase; it provides strong veracity as an official benchmark dataset released by Alibaba; and it delivers significant value for real-world e-commerce analysis and applications [1]. The pipeline involves data preprocessing, exploratory data analysis, customer segmentation, and sequential modeling to predict purchase behavior. The ultimate goal is to enhance recommendation accuracy, improve conversion rates, and support data-driven, cost-effective decision-making in e-commerce environments.

The experimental results demonstrate clear insights into user behavior. User activity peaks during midday hours and weekends, especially around promotional events, highlighting the importance of time-aware marketing strategies. Conversion analysis shows that even among the top-selling product categories, the conversion rate remains mostly below **5%**, reflecting a gap between browsing and actual purchases. Customer segmentation using the RFM model with K-means further revealed that approximately **60.1%** of users are at risk of churn. Finally, sequential modeling demonstrated that the LSTM achieved a test accuracy of **86.3%**, outperforming the RNN at **85.2%**, confirming the effectiveness of deep learning in predicting user actions.

II. Related Work

In [2], Zhang *et al.* proposed a prediction model called RNN-NB, which combines a Recurrent Neural Network (RNN) with a Naïve Bayes (NB) classifier to predict online purchase behavior. Their goal was to improve recommendation systems by capturing the sequence of user actions over time. They used a dataset from the Ali Tianchi platform with over 3 million records [1], and focused on user, product, and interaction features. The model showed better accuracy than using NB alone. However, the study had some limitations, such as using only one dataset and ignoring other important factors like user reviews or anonymous sessions. In our study, we aim to solve the same problem-predicting customer behavior more accurately but with a more complete and advanced approach. Our method includes several steps: data preparation, exploration, model building, evaluation, and improvement. We also use the same dataset [1]. For modeling, we use both RNN and LSTM. LSTM is an improved version of RNN that works better with long user behavior sequences and avoids common training issues. We also compare results with XGBoost, a powerful machine learning model. Our study will explore whether LSTM can improve prediction results compared to the RNN-NB model in [2].

In [3] Chen *et al.* proposed the Behavior Sequence Transformer (BST), a recommendation model that leverages the Transformer architecture to capture the sequential patterns in user behavior on the Taobao platform. Their objective was to improve click-through rate (CTR) prediction by modeling users' historical interactions using self-attention mechanisms. The model was trained on a large-scale dataset consisting of over 47 billion user behavior logs [1]. Experimental results showed that BST achieved superior performance compared to traditional models such as Wide & Deep Learning (WDL) and Deep Interest Network (DIN), with noticeable improvements in offline AUC and online CTR. However, the study was limited in scope, focusing solely on click prediction and relying on a narrow set of features. Important behavioral signals such as user reviews, merchant ratings, and anonymous session data were not considered. Furthermore, the study did not explore alternative sequential models such as RNN or LSTM, which are effective in modeling long-term dependencies. To address these limitations, our research aims to predict actual purchase decisions rather than clicks, by applying both RNN and LSTM to better capture extended behavioral patterns. In addition, we incorporate enriched contextual features to improve model generalization and practical impact in e-commerce recommendation systems.

In [4] Chen *et al.* proposed a big data analysis framework for e-commerce user behavior, combining Apache Spark for real-time distributed processing with XGBoost for purchase behavior prediction, using the "Alibaba Tianchi" [1] dataset. The study included challenge of efficiently processing massive volumes of user interaction data to improve recommendation accuracy and business decision-making. The methodology included large-scale data collection, preprocessing to handle missing values, feature engineering to represent user activity, real-time processing with Spark Streaming, and model training and evaluation using precision, recall, and F1-score. Results viewed improved prediction accuracy and stability compared to traditional methods, demonstrating the effectiveness of integrating distributed computing with machine learning in e-commerce scenarios. However, the study was limited to click-independent purchase prediction, used a single data source, and lacked richer contextual features such as user reviews, merchant ratings, and anonymous session data, while also omitting sequential deep learning models like RNN or LSTM that could capture long-term behavioral dependencies. In this research, will applying RNN and LSTM to model the temporal sequence of user behavior, by view product and purchase, add cart, and favorite behavioral features, and focusing on predicting realistic purchase decisions, to enhancing predictive performance and the of recommendation purchase product.

In [5] Zhou, Zhu, Song, *et al.* proposed the Tree-based Deep Model for Recommender Systems to address the computational and accuracy challenges of large-scale recommender systems. by reorder all candidate items into a hierarchical multi-level tree, where each non-leaf node represents an aggregated category and each leaf node corresponds to an individual item. Recommendation is performed as a top-down search in the tree, where at each level, a deep neural network is trained to predict the most relevant child nodes for a given user representation. This hierarchical retrieval strategy significantly reduces the search space from millions of items to a logarithmic number of nodes to evaluate, thereby improving retrieval efficiency while maintaining competitive accuracy. In experiments conducted on Taobao's dataset, TDM achieved substantial improvements over strong baselines such as YouTube DNN and Item-CF, with reported gains in precision and recall of up to several percentage points, while also demonstrating scalability to billions of items. this study limitations for not explicitly capture the sequential or temporal dependencies in user behavior, which are critical for accurately modeling purchase decision processes. Furthermore, the model relies on static tree structures that may not adapt quickly to evolving user behavior in real time. In my research, will based-on efficiency of TDM by merge sequential deep learning architectures, namely Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, to capture long-term behavioral patterns and temporal dependencies in user behaver sequences. This approach addresses the limitations of TDM by combining hierarchical candidate retrieval with temporal modeling and user behavior, making the recommendation process both scalability and compatibility with business e-commerce.

In [6] Lu, Deng, and Lu propose an efficient multi-behavior recommendation framework that fuses graph neural networks with multi-head attention (MGAT) to jointly model different user behaviors (e.g., click, add-to-cart, purchase) for improved purchase prediction. Their pipeline includes robust preprocessing and feature construction, multi-task optimization to balance heterogeneous behavior signals, and evaluation on real-world e-commerce datasets (Taobao and Beibei) using HR@K and NDCG@K. Empirically, their MGAT-based approach outperforms strong baselines such as BPR, LightGCN, and NMTR, highlighting the benefit of explicitly modeling cross-behavior relations. However, their method does not explicitly model temporal/sequential dependencies in user activity; instead, it focuses on multi-relation structure over user–item graphs. Moreover, while they consider multiple behaviors, richer contextual signals (e.g., textual reviews) are not integrated. In contrast, our work retains rigorous preprocessing and feature engineering but moves to sequential deep learning—using RNN/LSTM—to capture temporal dynamics of user behavior sequences (e.g., favorites, add-to-cart, product views) and predict purchase decisions over time, thereby complementing multi-behavior modeling with explicit sequence modeling.

III. Methodology

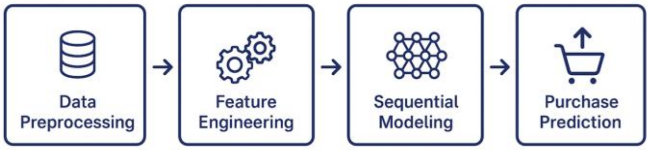


Fig. 1. The block diagram of the project pipeline

A. *Data Preprocessing*

Dataset Descriptions

This report Consumers the User Behavior Data from Taobao (Alibaba) [1], a benchmark big data dataset widely used in e-commerce research. It represents large-scale, real-world user interactions and is particularly suited for behavior modeling, recommendation systems, and purchase prediction. It contains random select about 1 million users who have behaviors including click, purchase, adding item to shopping cart and item favoring during November 25 to December 03, 2017. each line represents a specific user-item interaction, which consists of user ID, item ID, item's category ID, behavior type and timestamp, separated by commas. The dataset was specifically designed for research in user behavior modeling, recommendation systems, and purchase prediction in large-scale e-commerce platforms.

TABLE I
DESCRIPTION OF THE DATASET COLUMNS

#	Field	Description
1	user_id	Serialized user ID (integer)
2	item_id	Serialized product ID (integer)
3	category_id	Serialized product category ID (integer)
4	behavior_type	pv (page view), buy (make purchases), cart (add items to the shopping cart), fav (favorite products) (string)
5	timestamp	Timestamp when the behavior occurred

Dimensions of the dataset contain ~100 million interaction records, covering a diverse range of product categories and behavioral patterns, as explained in **Table 2**.

TABLE 2
DESCRIPTION DIMENSIONS OF THE DATASET

#	Dimension	Number
1	users	987,994

2	items	4,162,024
3	categories	9,439
4	interactions	100,150,807

Dataset Advantages as Big Data:

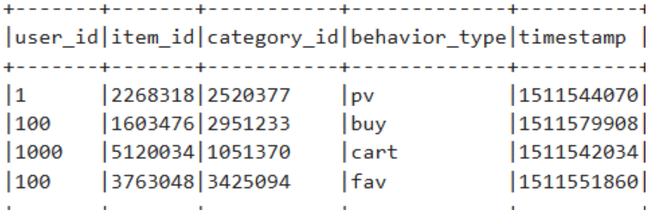
1) Variety: The dataset captures diverse user actions (view, cart, favorite, buy), covering the full customer journey.
2) Velocity: Events are collected continuously with timestamps, enabling real-time and sequential pattern detection.
3) Volume: With more than 100M records from millions of users, it reflects real-world large-scale data.
4) Value & Veracity: The inclusion of direct “buy” events makes it highly valuable for purchase prediction, while its origin from Alibaba ensures reliability as a benchmark dataset.

Data Clean

After sampling the original dataset, the data were carefully cleaned to ensure accuracy and consistency. preparing the dataset for reliable analysis.

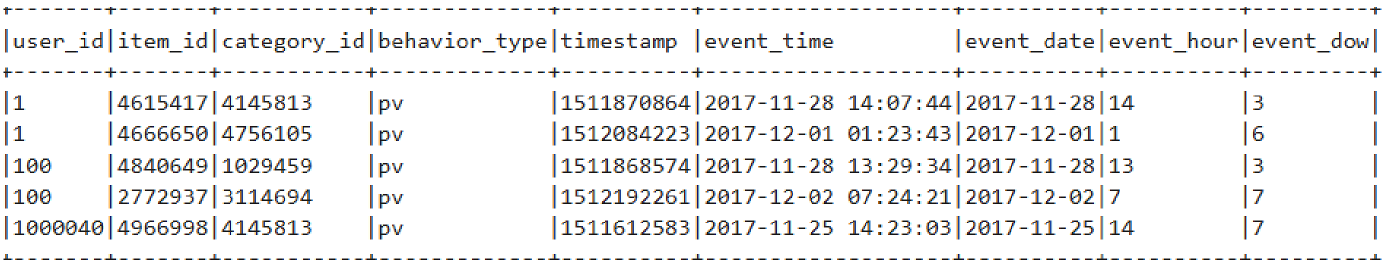
The steps are outlined below:

1) Clean the dataset by removing null values, duplicated rows, irrelevant rows, and inconsistencies.
2) Filter data within an appropriate time range.
3) modify the column "Behavior" names to be more descriptive.
4) Convert the original timestamps into human-readable datetime format.
5) From these timestamps, extract time features such as date, hour, and day-of-week were extracted. These features are crucial for identifying daily and weekly usage patterns in time-series analysis, and for understanding the purchasing behavior of consumers.



```
+-------+-------+-----------+-------------+----------+
|user_id|item_id|category_id|behavior_type|timestamp |
+-------+-------+-----------+-------------+----------+
|1      |2268318|2520377    |pv           |1511544070|
|100    |1603476|2951233    |buy          |1511579908|
|1000   |5120034|1051370    |cart         |1511542034|
|100    |3763048|3425094    |fav          |1511551860|
```

Fig. 2. Sample dataset before data processing techniques



```
+-------+-------+-----------+-------------+----------+-------------------+----------+----------+---------+
|user_id|item_id|category_id|behavior_type|timestamp |event_time         |event_date|event_hour|event_dow|
+-------+-------+-----------+-------------+----------+-------------------+----------+----------+---------+
|1      |4615417|4145813    |pv           |1511870864|2017-11-28 14:07:44|2017-11-28|14        |3        |
|1      |4666650|4756105    |pv           |1512084223|2017-12-01 01:23:43|2017-12-01|1         |6        |
|100    |4840649|1029459    |pv           |1511868574|2017-11-28 13:29:34|2017-11-28|13        |3        |
|100    |2772937|3114694    |pv           |1512192261|2017-12-02 07:24:21|2017-12-02|7         |7        |
|1000040|4966998|4145813    |pv           |1511612583|2017-11-25 14:23:03|2017-11-25|14        |7        |
+-------+-------+-----------+-------------+----------+-------------------+----------+----------+---------+
```

Fig. 3. Sample dataset after data processing techniques

B. *Feature Engineering*

Feature engineering was conducted through Exploratory Data Analysis (EDA). The exploratory analysis of the Taobao user behavior dataset produced the following key findings:

User Behavior for Hours and Day of Week:

As part of feature engineering, the timestamp column was transformed into derived features such as hour and day-of-week. These time-based features were then explored through EDA to identify user activity patterns. shows **Figure 2** presents patterns in user behavior by user interactions with the platform segmented by the hour of the day and the day of the week. The analysis revealed that user activity peaks between 8:00 pm and 10:00 pm, aligning with evening leisure hours when users are more likely to browse and purchase. During the morning and early afternoon hours (8:00 am – 2:00 pm), activity remains moderate, while the lowest engagement occurs in the early morning hours (2:00 am – 6:00 am). Additionally, weekday patterns remain consistent, showing relatively stable interactions from Monday to Friday. However, a noticeable increase in engagement occurs during weekends, suggesting that users are more active in browsing and purchasing when they have more free time. This temporal analysis highlights the importance of time-aware recommendation strategies, where promotions and targeted advertisements can be optimized during evening peaks and weekends to maximize user engagement and potential sales.

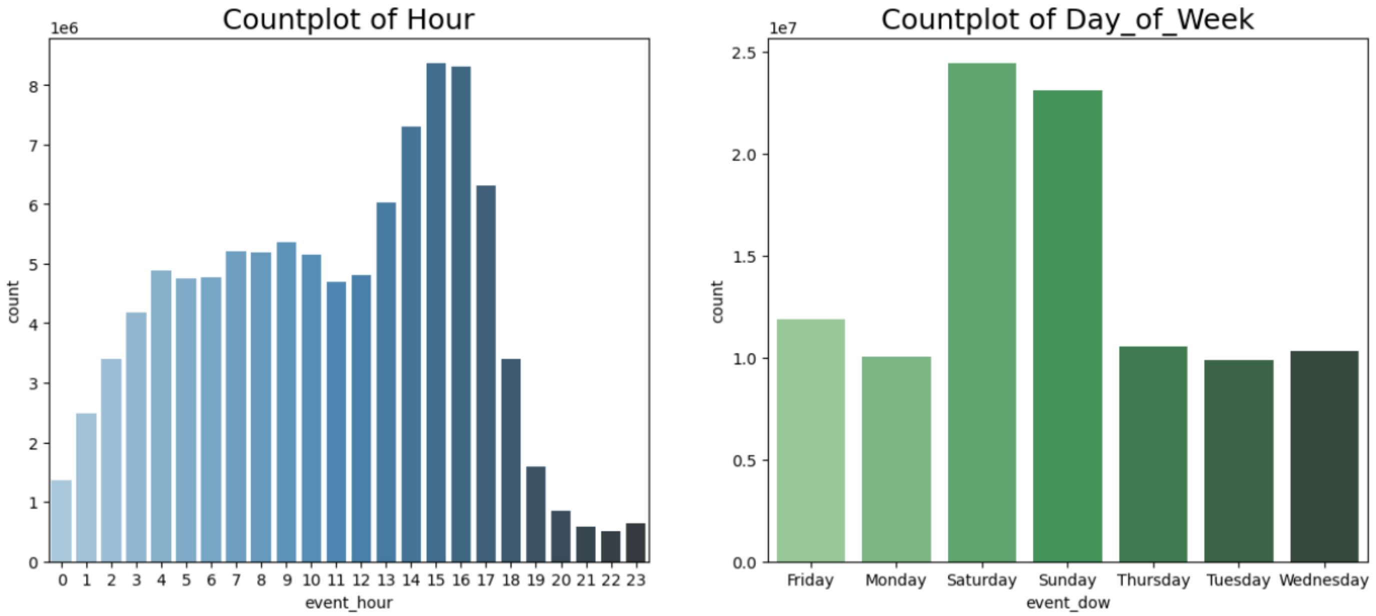


Fig. 2. User behavior of each day and day of week.

Customer Purchase Trends Over Time

In **Figure 3**, the cumulative count of "Buy" behaviors was analyzed and plotted by both day of the week and daily activity over two weeks. The left plot shows that customer purchases increase significantly from Friday to Saturday, with high counts sustained on Sunday, reflecting a strong weekend shopping preference. The right plot displays the daily purchasing trend, highlighting a major spike on December 2nd, 2017. This surge is strongly linked to the Alibaba shopping carnival, during which substantial discounts were offered. Interestingly, both November 25th and December 2nd fell on Saturdays, yet the drastic difference in purchase counts confirms the impact of promotional events on stimulating buying behavior.

Overall, these results indicate that weekends and special sales campaigns are key drivers of purchase activity, suggesting that targeted marketing strategies during these periods can maximize sales outcomes. Although the observed spike corresponds to events in **2017**, the dataset remains a widely used benchmark in e-commerce research due to its scale and richness, and the insights such as the influence of weekends and promotions are generalizable to modern e-commerce contexts.

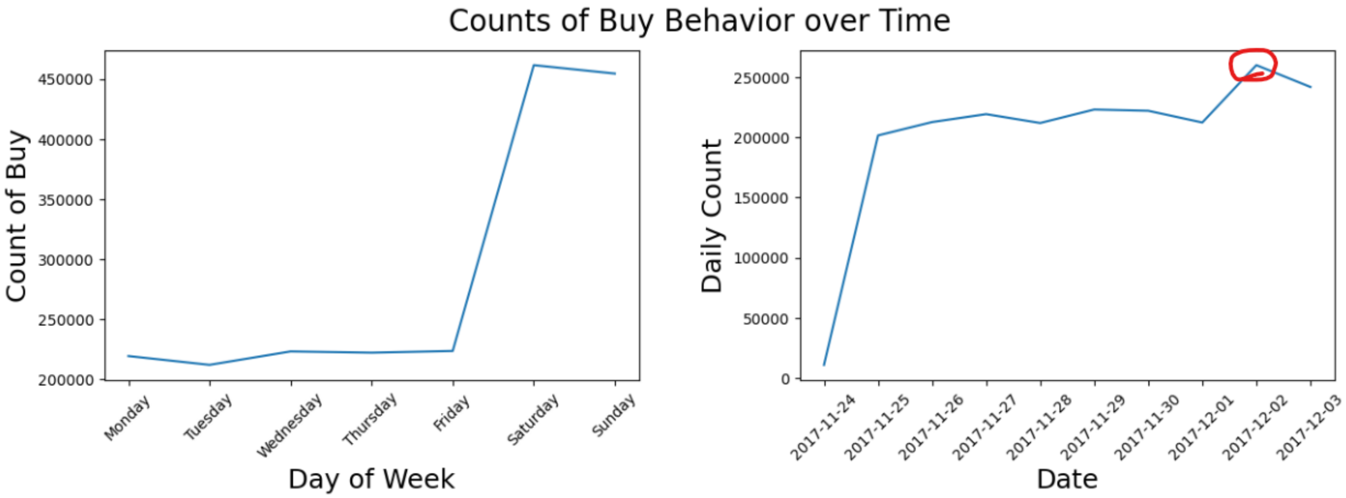
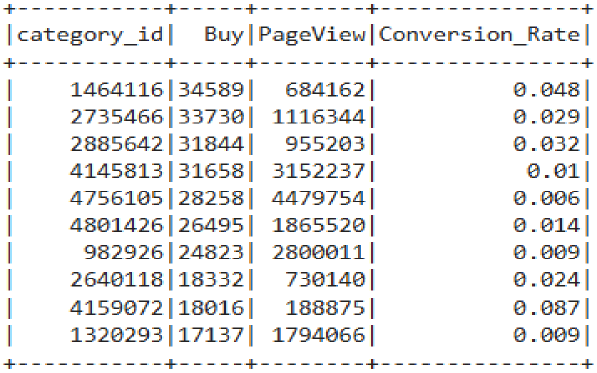


Fig. 3. The cumulative count of "Buy" behavior over date and day of week.

Product Conversion Rates

In Figure 4, the conversion rates were calculated to evaluate how efficiently page views translated into purchases across product categories. Conversion rate, defined as the ratio of purchases to total unique views [16], is a critical performance metric for e-commerce platforms.

The analysis revealed that, despite massive user interactions, only a small fraction of visits resulted in purchases in our dataset. This aligns with global benchmarks, where just 2.3% of visits end in purchases as of September 2023 [17]. Figure 4 visualizes the top 10 best-selling categories, showing that although their conversion rates outperform others, purchase likelihood remains uneven across product groups. The results demonstrate that conversion rates remain relatively low, even among the top-selling categories. Most categories achieved less than 4% conversion, with only two categories significantly standing out: category 1464116 (4.8%) and category 4159072 (8.7%). Notably, category 4159072 achieved the highest conversion efficiency, despite recording far fewer page views compared to others (188,875 views), suggesting that certain niche products attract more decisive buyers. Conversely, highly popular categories such as 4145813 (3.1M views, 1.0% conversion) and 4756105 (4.4M views, 0.6% conversion) exhibit substantial user interest but very low purchase follow-through. This indicates that high visibility does not guarantee high sales, and mismatches may exist between what is promoted and what customers actually intend to buy. Overall, the findings highlight a critical gap between product exposure and actual purchase decisions. This suggests that current recommendation strategies may not fully align with customer preferences and that optimizing personalization and targeting could significantly improve conversion efficiency.



```
+-----------+-----+--------+---------------+
|category_id|  Buy|PageView|Conversion_Rate|
+-----------+-----+--------+---------------+
|    1464116|34589|  684162|          0.048|
|    2735466|33730| 1116344|          0.029|
|    2885642|31844|  955203|          0.032|
|    4145813|31658| 3152237|           0.01|
|    4756105|28258| 4479754|          0.006|
|    4801426|26495| 1865520|          0.014|
|     982926|24823| 2800011|          0.009|
|    2640118|18332|  730140|          0.024|
|    4159072|18016|  188875|          0.087|
|    1320293|17137| 1794066|          0.009|
+-----------+-----+--------+---------------+
```

Fig. 4. The conversion rates of the 10 best-selling categories.

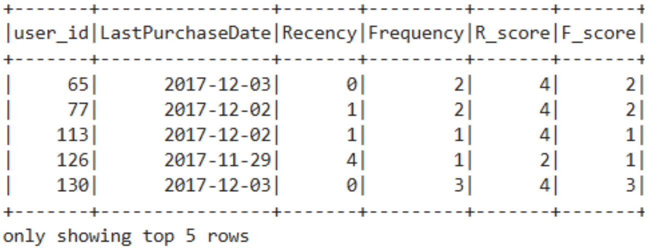
C. *Sequential Modeling*

Each customer has unique habits during online shopping. From this perspective, it is essential for e-commerce platforms to understand these diverse behaviors rather than applying a single marketing strategy for all users. Instead, businesses need to identify different customer types and tailor personalized strategies to meet the needs of each segment. Applying clustering techniques to segment customers into subgroups or market segments enables a deeper analysis of the relationship between consumers and the platform, providing valuable insights into customer expectations when purchasing products [18]. By taking personalized actions derived from such analyses, companies can enhance engagement and improve overall customer satisfaction. To achieve this, the RFM (Recency, Frequency, and Monetary/Behavioral value) model is one of the most widely adopted approaches by both researchers and practitioners. In its standard form, RFM evaluates customer behavior across three dimensions: Recency (R), which represents the time elapsed since the last purchase; Frequency (F), which captures how often a customer makes purchases; and Monetary (M), which reflects the financial value of their spending [19]. However, since the Taobao dataset does not provide explicit monetary information, this study adapts the third dimension to represent a behavioral value the intensity of user interactions (such as repeated clicks, cart additions, favorites, or purchases). This adaptation ensures that the RFM model remains applicable to the dataset while still capturing variations in customer engagement levels.

In this study, the adapted RFM model is used as the foundation for customer segmentation, followed by the application of the **K-means clustering algorithm** on Apache Spark to categorize users into distinct groups. This approach allows e-commerce platforms to uncover hidden behavioral patterns, distinguish between inactive users, occasional buyers, and loyal customers, and ultimately design more effective recommendation and marketing strategies.

RFM-based clustering process through four main steps:

1) Samples of calculated and standardized RFM scores:
The RFM model measures three dimensions of a customer: recency, frequency and monetary. Recency is defined as days since the last purchase, frequency is the total number of purchases and monetary is the total money this customer spent [19]. I only calculate the recency and frequency of each user’s purchase behavior. The scores are then standardized for a better comparison of the relative performance of customers, as shown in **Figure 5**.



```
+-------+----------------+-------+---------+-------+-------+
|user_id|LastPurchaseDate|Recency|Frequency|R_score|F_score|
+-------+----------------+-------+---------+-------+-------+
|     65|      2017-12-03|      0|        2|      4|      2|
|     77|      2017-12-02|      1|        2|      4|      2|
|    113|      2017-12-02|      1|        1|      4|      1|
|    126|      2017-11-29|      4|        1|      2|      1|
|    130|      2017-12-03|      0|        3|      4|      3|
+-------+----------------+-------+---------+-------+-------+
only showing top 5 rows
```

Figure 5: The samples of calculated and standardized RFM scores.

2) Group customers using K-means algorithm:
K-means algorithm is an unsupervised clustering approach, to identify customers with similar behavior patterns. With the chosen value of k, k cluster centers (centroids) are initialized arbitrarily. Next, each data object is assigned to the nearest center using the Euclidean distance. Then, the average of the clusters is calculated. The iteration repeats until convergence [20]. The formulas for the cluster assignment and center update are as below, where n denotes each data sample while k denotes each cluster. The number of clusters k, researchers have proposed to use the elbow method. The elbow method measures the proportion of variance explained as a function of the number of clusters. The optimal value of k is selected where an additional cluster doesn’t significantly enhance the data modeling [21]. To implement the elbow method, I used KElbowVisualizer, a tool to fit the model with a range of values of K and select the one with the best performance. As shown in **Figure 6**, the elbow method

divides the customers into 3 groups based on their recency and frequency scores.

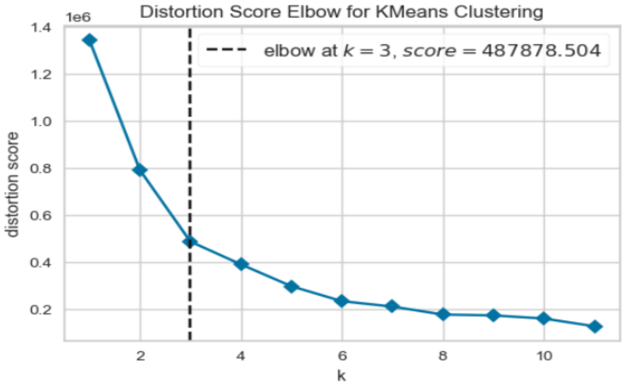
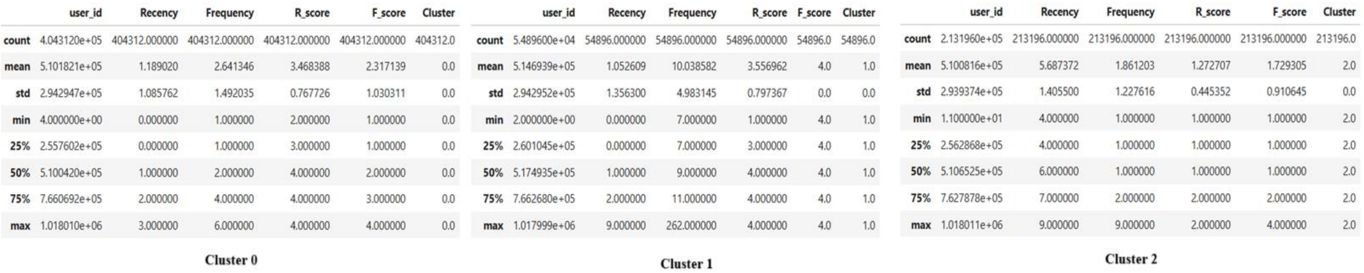


Fig. 6. Distortion Score ELBOW for KMeans Clustering.

3) K-Means results of the 3 clusters:
In **Figure 7** shown summary of clustering results the customers can be categorized into the following groups:

- Churn Risk Customers (high recency and low frequency): They have relatively low interactions or purchases. They could be at risk of churning or reducing their engagement.
- Potential Customers (moderate recency and frequency): They have interacted recently but don't have a high frequency of purchases. They could be potential customers exploring or making occasional purchases.
- High-Value Customers (low recency with high frequency): They have both recent purchases and a significantly higher frequency of transactions. They are highly engaged and loyal customers, contributing frequently to the e-commerce platform.



	user_id	Recency	Frequency	R_score	F_score	Cluster
count	4.043120e+05	404312.000000	404312.000000	404312.000000	404312.000000	404312.0
mean	5.101821e+05	1.189020	2.641346	3.468388	2.317139	0.0
std	2.942947e+05	1.085762	1.492035	0.767726	1.030311	0.0
min	4.000000e+00	0.000000	1.000000	2.000000	1.000000	0.0
25%	2.557602e+05	0.000000	1.000000	3.000000	1.000000	0.0
50%	5.100420e+05	1.000000	2.000000	4.000000	2.000000	0.0
75%	7.660692e+05	2.000000	4.000000	4.000000	3.000000	0.0
max	1.018010e+06	3.000000	6.000000	4.000000	4.000000	0.0

Cluster 0

	user_id	Recency	Frequency	R_score	F_score	Cluster
count	5.489600e+04	54896.000000	54896.000000	54896.000000	54896.0	54896.0
mean	5.146939e+05	1.052609	10.038582	3.556962	4.0	1.0
std	2.942952e+05	1.356300	4.983145	0.797367	0.0	0.0
min	2.000000e+00	0.000000	7.000000	1.000000	4.0	1.0
25%	2.601045e+05	0.000000	7.000000	3.000000	4.0	1.0
50%	5.174935e+05	1.000000	9.000000	4.000000	4.0	1.0
75%	7.662680e+05	2.000000	11.000000	4.000000	4.0	1.0
max	1.017999e+06	9.000000	262.000000	4.000000	4.0	1.0

Cluster 1

	user_id	Recency	Frequency	R_score	F_score	Cluster
count	2.131960e+05	213196.000000	213196.000000	213196.000000	213196.000000	213196.0
mean	5.100816e+05	5.687372	1.861203	1.272707	1.729305	2.0
std	2.939374e+05	1.405500	1.227616	0.445352	0.910645	0.0
min	1.100000e+01	4.000000	1.000000	1.000000	1.000000	2.0
25%	2.562868e+05	4.000000	1.000000	1.000000	1.000000	2.0
50%	5.106525e+05	6.000000	1.000000	1.000000	1.000000	2.0
75%	7.627878e+05	7.000000	2.000000	2.000000	2.000000	2.0
max	1.018011e+06	9.000000	9.000000	2.000000	4.000000	2.0

Cluster 2

Fig. 7. K-Means results of the 3 clusters.

4) Customer Segmentation Insights:

In **Figure 8**, shown the largest segment of customers falls under the churn-risk category, representing 60.1% of the total. This poses a significant challenge for platforms, as losing such a large portion could lead to considerable revenue decline. The second-largest group is high-value customers, accounting for 31.7%. Retaining this segment is vital, as they contribute substantially to profitability, and can be engaged through VIP programs, loyalty benefits, and exclusive offers. Finally, potential customers make up 8.2% of the total. Although the smallest segment, they represent opportunities for future growth, where personalized recommendations and targeted marketing strategies can help increase their conversion rates.

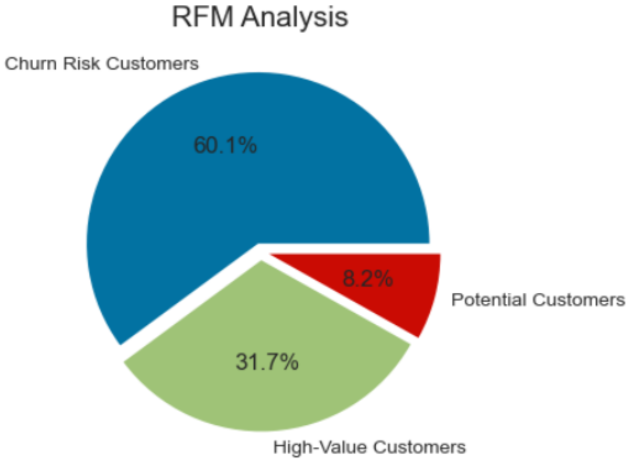


Fig. 8. percentage of each customer group.

D. *Purchase Prediction*

As user behavior is tracked alongside timestamps, it is intuitive to represent user interactions as sequences of actions. Previous studies have highlighted that modeling sequential patterns of user activity enables accurate prediction of future interactions, which in turn benefits marketers by improving personalization and enhancing customer experience [13]. Different users exhibit unique shopping habits; some purchase directly after a single product view, others frequently favorite items without completing transactions, while some habitually add items to their shopping cart. By leveraging temporal dependencies between these actions, it becomes possible to mine customer interests and predict their next likely interaction.

To process this, sequential deep learning models were employed. Recurrent Neural Networks (RNN) are designed to capture temporal structures by computing new states based on current input and past states [22]. However, vanilla RNNs are prone to vanishing or exploding gradients, which can hinder long-term dependency learning. To mitigate this, the Long Short-Term Memory (LSTM) architecture introduces gating mechanisms and additive state interactions, enabling more robust modeling of sequential behavior [23].

Data Preparation for Sequential Modeling

Before training the models, the dataset was preprocessed to fit sequential modeling requirements. Each user's interaction history was ordered chronologically, and behavior types (PageView, Cart, Favorite, Buy) were encoded into numerical indices. Sessions were segmented to form fixed-length sequences, where each element represented an interaction event with its corresponding timestamp. The target variable was defined as the next user action, allowing the models to learn a mapping from historical interactions to future behavior. This sequential representation ensured that both short-term and long-term dependencies could be captured effectively.

LSTM Model

This model was implemented using PyTorch, with an embedding layer to represent user actions, followed by LSTM layers that maintained hidden and cell states across sequences. The gating mechanisms (input, output, and forget gates) enabled the network to retain relevant information across long behavior chains and discard irrelevant patterns. A fully connected layer with a Softmax activation was employed to output probabilities across the four possible user actions, as shown in **Figure 9** architecture of the LSTM model.

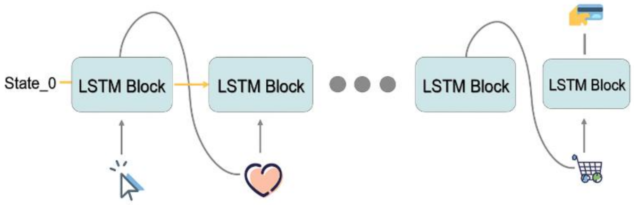


Fig. 9. Overview of the method used in LSTM.

Experimental results showed that the LSTM model achieved a Test Accuracy **of 0.8634**, confirming its effectiveness in modeling sequential dependencies and predicting next-step user behavior with high precision.

RNN Model

This model was also trained with the same data preprocessing pipeline. The RNN architecture used recurrent layers that updated hidden states iteratively based on each new input and previous state. While capable of capturing short-term patterns, its performance was limited by gradient vanishing issues in longer sequences, as shown in **Figure 10** structure of the RNN model is presented.

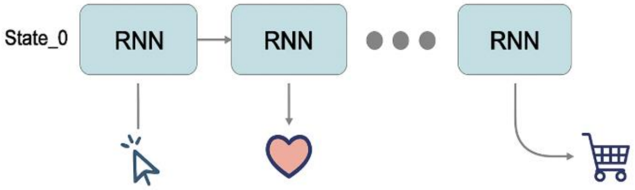


Fig. 10. Overview of the method used in RNN.

Despite these challenges, the RNN model still performed competitively, achieving a Test Accuracy of **0.8520**, though slightly lower than the LSTM. This confirmed that while RNNs provide a baseline for sequential modeling, LSTM outperforms them in handling complex, long-term behavior patterns.

IV. CONCLUSION & DISCUSSION

In this report, a comprehensive big data driven pipeline was developed to analyze and predict user behavior in e-commerce. Starting with user interaction logs from Taobao, the pipeline included data preprocessing, exploratory data analysis, customer segmentation using the RFM model combined with K-means clustering, and predictive modeling through sequential deep learning.

The report summary is as follows:

1) User purchasing behavior over time:
Results demonstrated that purchasing activity peaks during weekends and is further boosted during shopping events with large discounts. Daily activity patterns also indicated increased interactions during midday hours (12 PM to 2 PM).
2) Top-selling categories and conversion rates:
The analysis revealed that even among the top-selling categories, conversion rates remained below 9% in most cases, with only category 4159072 exceeding this threshold (8.7%). This indicates that recommended products often do not align with user purchase preferences, suggesting the need for improved recommendation algorithms.
3) Customer segmentation, churn-risk identification:
Applying RFM with K-means clustering, customers were segmented into distinct groups based on their recency and frequency. The results showed that a considerable proportion of users presented churn risk due to high recency and low frequency. However, high-value and potential customers still represented a dominant share, highlighting opportunities for targeted marketing strategies.
4) Sequential modeling for purchase prediction:
To predict the next user interaction, sequences of past behaviors were modeled using RNN and LSTM.

Experimental results demonstrated that LSTM achieved higher predictive accuracy (**86.34%**) compared to RNN (**85.20%**), confirming its effectiveness in capturing long-term dependencies in user behavior as shown in **Figure 11**.

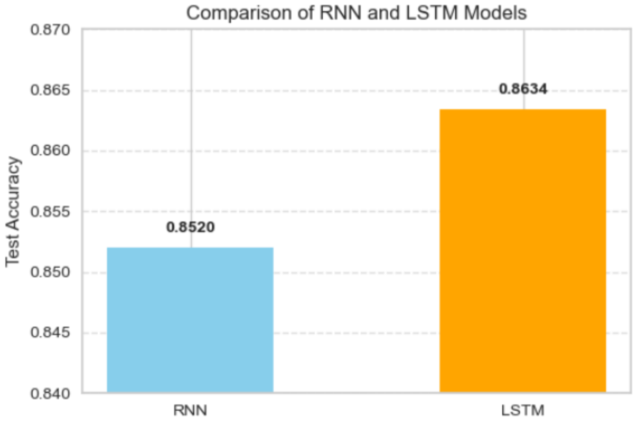


Fig. 11. Comparison of RNN and LSTM Models.

Overall, this report has shown the importance of combining big data analytics with deep learning for e-commerce. While clustering techniques can help identify user groups and churn risks, sequential models provide valuable insights into individual-level predictions.

Performance Metrics and Parallel Efficiency

The performance evaluation compared Pandas (single-node sequential execution) with Apache Spark (distributed parallel execution). While Pandas processes operations linearly, Spark leverages distributed memory and parallel task scheduling, which significantly reduces execution times. In this study, Spark was configured with increased driver and executor memory (8 GB each), four executor cores, and optimized result size limits to balance memory usage and task distribution efficiently.

The experiments yielded the following results:

1) Events per hour: reduced from 8.16s (Pandas) to 0.65s (Spark), achieving a 12.5× speedup.
2) Events per weekday: reduced from 1.22s to 0.13s, achieving a 9.6× speedup.
3) User-item behavior aggregation: reduced from 27.6s to 0.60s, achieving a 46.1× speedup.

These results emphasize several important aspects of parallel performance:

1) Sources of overhead: Spark introduces scheduling and communication overhead, but the benefits of parallel execution outweigh these costs.
2) Granularity: fine-grained tasks (hourly counts) benefit from efficient task distribution, while coarse-grained tasks (user-item aggregation) achieve the highest relative speedups.
3) Scalability: Spark scales efficiently with larger datasets, whereas Pandas performance degrades rapidly under the same conditions.
4) Execution efficiency: Spark achieves near cost-optimal runtimes by effectively utilizing memory and CPU cores.

Overall, Spark achieved up to 46× faster execution compared to Pandas, demonstrating its effectiveness for large-scale e-commerce analytics and confirming the importance of parallel frameworks in real big data environments. The comparison is illustrated in **Figure 12**, which highlights the significant runtime differences across different operations.

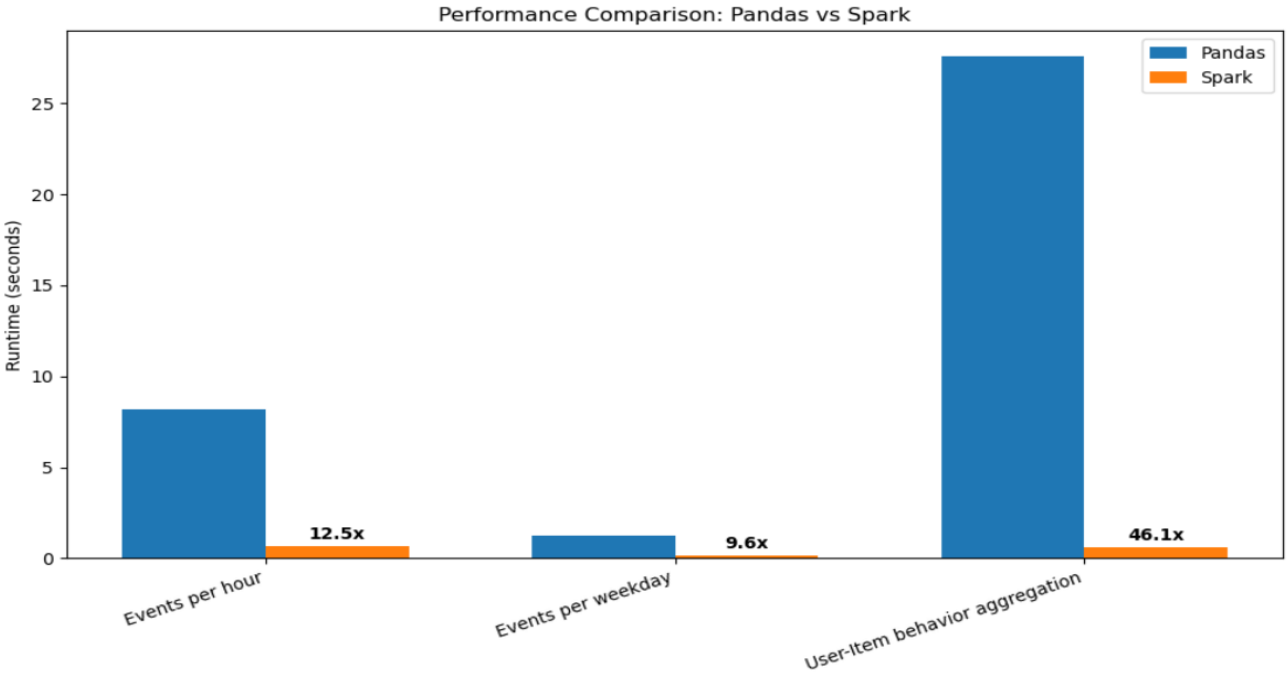


Fig. 12. Performance Comparison: Pandas vs Spark.

V. FUTURE STUDY

A. Integration with recommendation systems

The models can be combined with recommendation systems to not only predict the next action but also suggest products more accurately and improve conversion rates.

B. Incorporating additional features:

Can add more features like user location, device type, or product details to improve clustering and prediction results.

C. Exploring advanced models:

Can test newer models like Transformers, which may perform better than RNN and LSTM in predicting user behavior.

D. Churn prediction and retention strategies:

Can build models to predict churn more accurately and help design better retention plans.

E. Real-time prediction and deployment:

Can test real-time use of the pipeline to detect user intent instantly and improve recommendations.

APPENDIX

The implementation code for this report is available at: https://github.com/ityasser/Analysis-Prediction-E-Commerce-Big-Data.git

REFERENCES

[1] Alibaba Group, "User Behavior Data from Taobao for Recommendation," Alibaba Cloud Tianchi, May 10, 2018. Available: https://tianchi.aliyun.com/dataset/649.

[2] C. Zhang, J. Liu, and S. Zhang, "Online Purchase Behavior Prediction Model Based on Recurrent Neural Network and Naive Bayes," J. Theor. Appl. Electron. Commer. Res., vol. 19, no. 4, pp. 3461–3476, Dec. 2024, doi: 10.3390/jtaer19040168.

[3] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, "Behavior Sequence Transformer for E-commerce Recommendation in Alibaba," arXiv preprint arXiv: 1905.06874.

[4] H. Yuan, "Research on Big Data Analysis of E-commerce User Behavior," Proc. 2025 Int. Conf. on E-Commerce and Internet Technology (ECIT), Jan. 2025, pp. 113–118, doi: 10.1109/ECIT60527.2025.00025.

[5] G. Zhou, X. Zhu, C. Song, et al., "Learning Tree-based Deep Model for Recommender Systems," in Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD), London, UK, Aug. 2018, pp. 1079–1088, doi:10.1145/3219819.3219826.

[6] H. Lu, X. Deng, and J. Lu, "Research on Efficient Multi-Behavior Recommendation Method Fused with Graph Neural Network," Electronics, vol. 12, no. 9, Art. no. 2106, May. 2023, doi:10.3390/electronics12212106.

[7] C. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. del Jesus, and S. García, "Web usage mining to improve the design of an e-commerce website: Orolivesur.com," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11243–11249, Sep. 2012. doi:10.1016/j.eswa.2012.03.046.

[8] S. Hernández, P. Álvarez, J. Fabra, and J. Ezpeleta, "Analysis of users' behavior in structured e-commerce websites," *IEEE Access*, vol. 5, pp. 11941–11958, 2017. doi:10.1109/ACCESS.2017.2707600.

[9] R. Kohavi, "Mining e-commerce data: The good, the bad, and the ugly," in *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2001, pp. 8–13. doi:10.1007/3-540-45357-1_2

[10] Q. Zhang and R. S. Segall, "Web mining: A survey of current research, techniques, and software," *Int. J. Inf. Technol. Decis. Making*, vol. 7, no. 4, pp. 683–720, Dec. 2008. doi:10.1142/S0219622008003150.

[11] B. Singh and H. K. Singh, "Web data mining research: A survey," in *Proc. 2010 IEEE Int. Conf. Computational Intelligence and Computing Research*, Coimbatore, India, Dec. 2010, pp. 1–10. doi:10.1109/ICCIC.2010.5705856.

[12] B. Cooil, L. Aksoy, and T. L. Keiningham, "Approaches to customer segmentation," *J. Relationship Marketing*, vol. 6, no. 3–4, pp. 9–39, 2008. doi: 10.1300/J366v06n03_02

[13] C. Chen, S. Kim, H. Bui, R. Rossi, E. Koh, B. Kveton, and R. Bunescu, "Predictive analysis by leveraging temporal user behavior and user embeddings," in *Proc. 27th ACM Int. Conf. Information and Knowledge Management*, Torino, Italy, Oct. 2018, pp. 2175–2182. doi: 10.1145/3269206.3272032.

[14] T. Tekdoğan and A. Cakmak, "Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification," Sept. 2022. doi: 10.1145/3481646.3481649.

[15] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in Proc. 4th Int. Conf. Learning Representations (ICLR), San Juan, Puerto Rico, May 2016, pp. 1–10. Doi: https://doi.org/10.48550/arXiv.1511.06939.

[16] W. C. McDowell, R. C. Wilson, and C. O. Kile, "An examination of retail website design and conversion rate," Journal of Business Research, vol. 69, no. 11, pp. 4837–4842, Nov. 2016, doi: 10.1016/j.jbusres.2016.04.040

[17] SellersCommerce, "Mobile devices have a better conversion rate than desktops; tablets lead at 3.1%," SellersCommerce, Feb. 2025. Available: https://www.sellerscommerce.com/blog/ecommerce-statistics.

[18] R.-S. Wu and P.-H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," Electronic Commerce Research and Applications, vol. 10, no. 3, pp. 331–341, May–Jun. 2011, doi: 10.1016/j.elerap.2010.11.002.

[19] J. R. Miglautsch, "Thoughts on RFM scoring," Journal of Database Marketing & Customer Strategy Management, vol. 8, pp. 67–72, 2000, doi: 10.1057/palgrave.jdm.3240019.

[20] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in Proc. 2010 Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), Jinggangshan, China, Apr. 2010, pp. 63–67, doi: 10.1109/IITSI.2010.74.

[21] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," International Journal of Computer Applications, vol. 105, no. 9, pp. 17–24, Nov. 2014, doi: 10.5120/18405-9674.

[22] N. T. Vu, H. Adel, P. Gupta, and H. Schütze, "Combining recurrent and convolutional neural networks for relation classification," May 2016, doi: 10.48550/arXiv.1605.07333.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.