SI 671 Final Project:
E-commerce User Behavior Analysis and Modeling

Xinyu Zhang
{xyuzhang}@umich.edu

Abstract

Regarding the business of e-commerce platforms, every decision takes much financial and human resource costs. It is of great importance to understand how to drive their sales effectively. However, with the vast amount of user data, extracting useful information and discovering what factors influence customers' shopping habits pose significant challenges. This project introduces a data-driven pipeline that employs data mining techniques to process, analyze, and predict customer behavior. Experimental data was collected from Alibaba and contains customer behavior on the website Taobao over two weeks in 2017. The pipeline starts with data preprocessing and exploratory data analysis with visualizations. Then, customer segmentation based on RFM analysis and K-means clustering is implemented to divide customers into groups through their historical behavior. Finally, each user's behavior sequence is modeled by recurrent neural network (RNN) and its variant LSTM. The trained models can perform user behavior forecasting. This comprehensive approach can help the platform to gain valuable insights into their customer, leading to making more informed and cost-effective business decisions. The code is available at https://github.com/xinyuzhang99/E-commerce-User-Behavior-Analysis-and-Modeling. The original dataset and the processed datasets are available at https://drive.google.com/drive/folders/1OXTEbrnN_HypkcQcYYkCjugKQmtfo15E?usp=sharing

1 Introduction

The exponential evolution of the internet in the 21st century has boosted the fast growth of the online shopping industry and e-commerce platforms [Carmona et al., 2012]. Information about user behavior is stored in the web server logs in a structured way [Hernández et al., 2017]. In contrast with traditional offline shopping, online shopping provides customers with a much more convenient and flexible shopping experience. E-commerce platforms offer a vast array of products across various categories, providing users with the freedom to explore, compare, and make purchases. This shopping mode brings online shopping to a much higher profitable place.

However, gaining useful information in customer behavior data is a complex and challenging task. Each user behavior on the website, like clicks and purchases, can be collected and recorded in an extremely fast manner, thus generating a large amount of data each minute. Another challenge is data cleaning, as auto-collected data may have high noise, which will greatly interfere with the analysis process [Kohavi, 2001]. Nowadays, as the market is getting increasingly competitive and each business decision takes much financial and human resource costs, how to find what each user needs and make efficient marketing strategies are essential.

Facing a vast amount of raw data recording rich information about customers, there are several research questions worth delving into:

1. What are the peak activity times in purchasing?
2. What are the top-selling products or categories, and are the products with high sales also the ones that receive the most views?
3. Can we predict when a user is likely to stop using the platform or reduce their engagement? What are the warning signs of churn?
4. What is the next interaction of a user towards one item?

To solve the problems, data mining techniques have been widely utilized to discover patterns in user data. There has been a lot of research done in the field of e-commerce. Brijendra Singh et.al defined the user behavior analysis of the web as "web usages mining", where user data is analyzed to personalize the information to individual user [Singh and Singh, 2010]. Previous research has studied and applied approaches like association rules, classification, clustering and sequence modeling [Singh and Singh, 2010, Zhang and Segall, 2008].

In this project, a user data mining pipeline is proposed to help e-commerce platforms gain valuable insights into their customer, resulting in adjusting marketing and selling strategies to enhance user behavior as well as improve sales revenues. The whole pipeline is shown in 1. The original data was randomly collected from records of 1 million users from Taobao who have behaviors like click, purchase, adding an item to the cart and favoring an item [Zhu et al., 2018]. As the actual data collected may have redundant and invalid information, data preprocessing is essential. With the processed data, exploratory data analysis can be conducted, generating useful statistics and visualizations for intuitive findings. Then, customer segmentation is implemented to divide customers into distinct groups with different needs and behavior characteristics [Cooil et al., 2008]. Finally, user sequential modeling is conducted to predict the next user action, which the market believes is important to their revenue [Chen et al., 2018a].

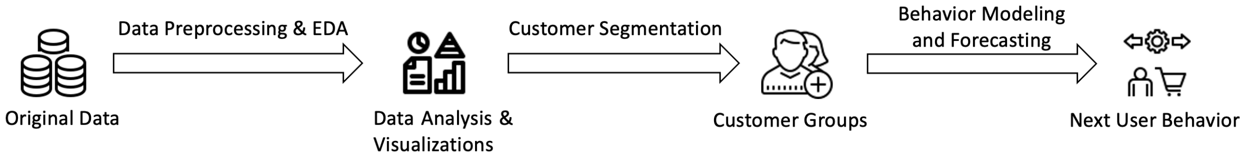


Figure 1: The block diagram of the project pipeline

2 Dataset

The data used in this project comes from an open dataset recorded by Alibaba, User Behavior Data from Taobao for Recommendation[1]. The dataset contains behaviors of approximately one million users, recorded between November 25, 2017 and December 3, 2017. There are in total four types of user behavior: product detail page views (equivalent to clicks), purchases, adding items to the shopping cart and favorating products. The dataset possesses a similar structure as MovieLens-20M, where each row represents a user behavior, consisting of user ID, product ID, product category ID, behavior type and timestamp. The description of each row is presented in Table 1 below.

[1] https://tianchi.aliyun.com/dataset/649

Table 1: The detailed description of the dataset columns

Column Name	Description
User_ID	Serialized user ID (integer)
Product_ID	Serialized product ID (integer)
Category_ID	Serialized product category ID (integer)
Behavior	pv (page view), buy (make purchases), cart (add items to the shopping cart), fav (favorite products) (string)
Timestamp	Timestamp when the behavior occurred

As there are in total more than 100 million records in the original dataset, a high requirement for computational resources is shown. Therefore for my local computer, I only sampled the records of 5% users to be experimented. Table 2 compares the major dimensions of the original and sampled dataset.

Table 2: Comparison of the original dataset and sampled dataset.

Dataset	Original Dataset	Sampled Dataset
# of Records	100150807	4952632 (-95%)
# of Users	987994	49400 (-95%)
# of Products	4162024	1099089 (-74%)
# of Categories	9439	7475 (-21%)

From Table 1, it can be found that despite sampling only 5% of the original dataset, the reduction in the number of categories isn't as substantial as the reduction in other aspects. This indicates that customers interacted with a wide range of categories in the original dataset.

3 Data Preprocessing & EDA

3.1 Data Preprocessing

After sampling the original data, data processing techniques are employed to ready it for subsequent analysis. The steps are outlined below:

1. Clean the data to remove possible null values and duplicated records.

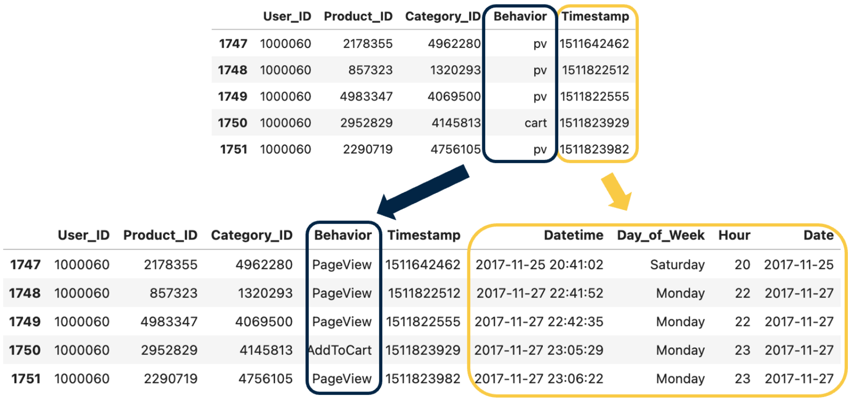
2. Filter data within an appropriate time range; modify the column "Behavior" names to be more descriptive.

3. Convert the original timestamps into datetime format and extract time features like the corresponding date, hour and day-of-week for time-series analysis.

The preprocessed dataset is shown in Figure 2.

3.2 Exploratory Data Analysis

To gain comprehensive insights from the dataset, I conducted exploratory data analysis, employing visualizations to uncover patterns. Figure 3 presents user interactions with the website, segmented by the hour of the day and day of the week. Notably, user activity peaks from 12 am to 2 pm, which is usually a midday break commonly observed during work hours. The trend remains relatively



	User_ID	Product_ID	Category_ID	Behavior	Timestamp
1747	1000060	2178355	4962280	pv	1511642462
1748	1000060	857323	1320293	pv	1511822512
1749	1000060	4983347	4069500	pv	1511822555
1750	1000060	2952829	4145813	cart	1511823929
1751	1000060	2290719	4756105	pv	1511823982

	User_ID	Product_ID	Category_ID	Behavior	Timestamp	Datetime	Day_of_Week	Hour	Date
1747	1000060	2178355	4962280	PageView	1511642462	2017-11-25 20:41:02	Saturday	20	2017-11-25
1748	1000060	857323	1320293	PageView	1511822512	2017-11-27 22:41:52	Monday	22	2017-11-27
1749	1000060	4983347	4069500	PageView	1511822555	2017-11-27 22:42:35	Monday	22	2017-11-27
1750	1000060	2952829	4145813	AddToCart	1511823929	2017-11-27 23:05:29	Monday	23	2017-11-27
1751	1000060	2290719	4756105	PageView	1511823982	2017-11-27 23:06:22	Monday	23	2017-11-27

Figure 2: The cleaned dataset after data processing techniques

stable from midnight to 10 am, with a gradual increase in user interactions starting at 11 am. After 2 pm, interactions gradually decrease, reaching their lowest point between 6 pm and 9 pm with less than 1% of overall user activity. Furthermore, a consistent trend is observed from Monday to Friday, while user engagement increases during the weekends.

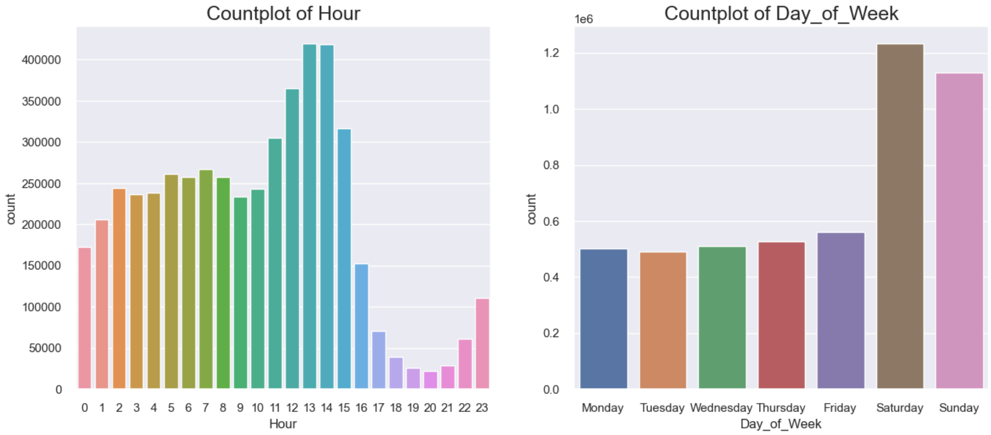


Figure 3: User behavior of each day and day of week

In Figure 4, I calculated the cumulative count of "Buy" behavior and plotted the data based on each day and day of the week. The left plot distinctly shows an increase in customer purchases from Friday to Saturday, maintaining a high count on Sunday, which indicates a preference for weekend buying. Meanwhile, the right plot visualizes the daily purchasing behavior over a two-week period. It is notable that a significant spike in user purchases occurred on December 2nd, 2017. This surge might be caused by the announcement of the Alibaba shopping carnival in December, with substantial product discounts. In addition, as November 25th and December 2nd were both Saturdays, the significant differences in the count of purchase behavior suggest that the carnival has a great influence in stimulating customer purchases. In summary, sales events and weekends appear to boost customer purchases, thus it's efficient to implement marketing campaigns and product advertisements during these periods.

Besides time series analysis, I also delved into the conversion situation for the products. The

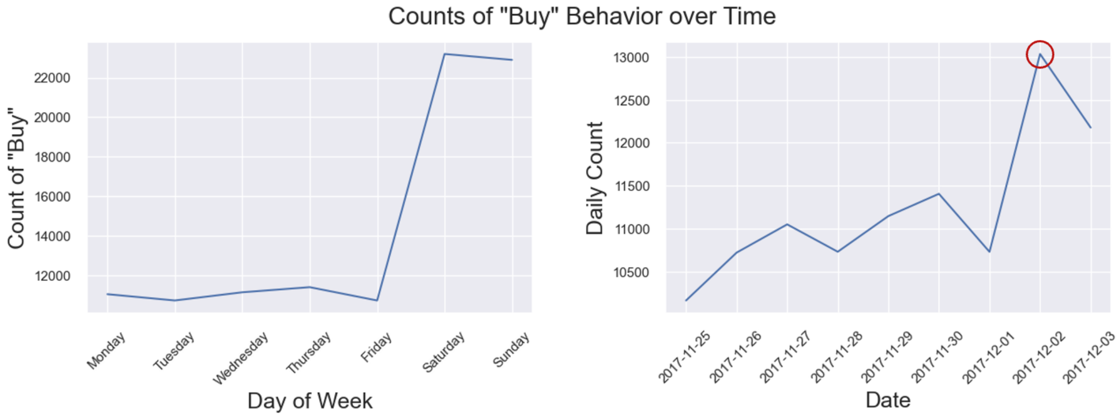
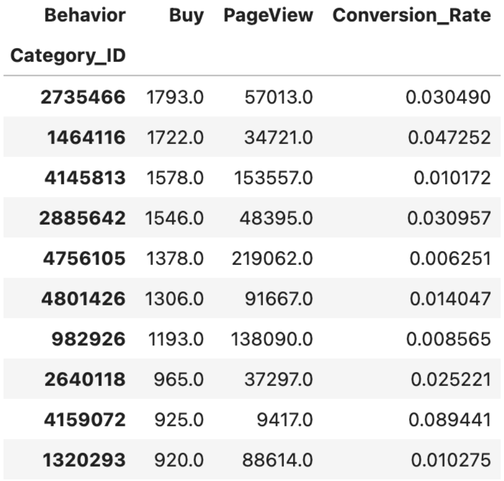


Figure 4: The cumulative count of "Buy" behavior over date and day of week

conversion rate measures the proportion of users who make a purchase compared to the total unique users visiting the platform [McDowell et al., 2016]. Although the number of interaction records on e-commerce platforms is unaccountable, only 2.3% of visits end with purchasing products as of September 2023, according to Statista [Yltävä, 2023]. That is to say, the increase in conversion rate is undoubtedly essential for a company's revenues. To check the analysis, I first calculated the conversion rate of "PageView" to "Buy" for each category, and then I displayed the conversion rates of the 10 best-sell categories. The result is presented in Figure 5.



Behavior	Buy	PageView	Conversion_Rate
Category_ID			
2735466	1793.0	57013.0	0.030490
1464116	1722.0	34721.0	0.047252
4145813	1578.0	153557.0	0.010172
2885642	1546.0	48395.0	0.030957
4756105	1378.0	219062.0	0.006251
4801426	1306.0	91667.0	0.014047
982926	1193.0	138090.0	0.008565
2640118	965.0	37297.0	0.025221
4159072	925.0	9417.0	0.089441
1320293	920.0	88614.0	0.010275

Figure 5: The conversion rates of the 10 best-selling categories

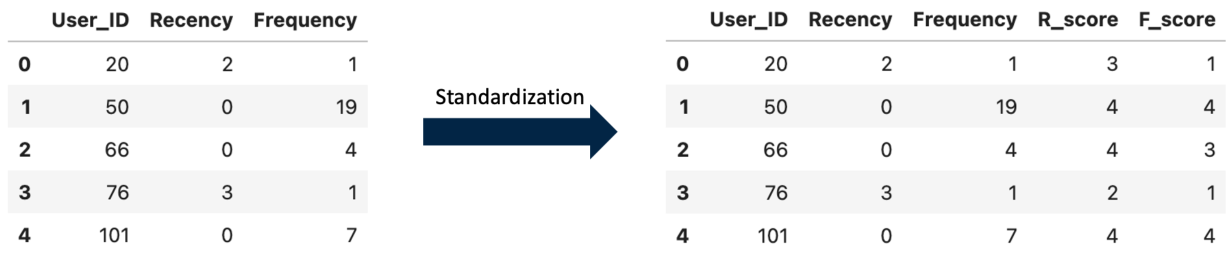
From the result above, it can be seen that even among the top-selling categories, the conversion rates remain exceedingly low. Most categories have rates below 4%, with only the second highest-selling category (1464116) and the ninth highest-selling category (4159072) exceeding this threshold. This suggests that the specific recommended items by the platform do not meet users' purchasing preferences, leading to low conversion rates after users click on the items.

4 Customer Segmentation

Each customer has their own habits during online shopping. From this perspective, it's important for an e-commerce platform to understand its customers' shopping habits and not attempt only to develop one marketing strategy for all. Instead, they need to know the shopping behaviors of different customer types and develop personalized strategies for each segmented market. Applying clustering techniques to segment customers into subgroups or market segments can be used to analyze the relationship between customers and e-commerce platforms [Wu and Chou, 2011]. The results of such analyses can offer the company insights into customer expectations when purchasing products on the website. By taking personalized actions based on these outcomes, companies can enhance customer engagement by improving their satisfaction.

The RFM model is one of the most common models used by researchers and companies to understand customers' purchasing patterns. This 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 [Miglautsch, 2000].

In this project, as monetary value is not provided in the dataset, 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 6.



	User_ID	Recency	Frequency
0	20	2	1
1	50	0	19
2	66	0	4
3	76	3	1
4	101	0	7

	User_ID	Recency	Frequency	R_score	F_score
0	20	2	1	3	1
1	50	0	19	4	4
2	66	0	4	4	3
3	76	3	1	2	1
4	101	0	7	4	4

Figure 6: The samples of calculated and standardized RFM scores

To group customers based on their recency and frequency scores, I applied the K-means algorithm, 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 [Na et al., 2010]. The formulas for the cluster assignment and center update are as below, where n denotes each data sample while k denotes each cluster.

$$r := \text{argmin}_r \ J(r, \mu); \ \mu := \text{argmin}_\mu \ J(r, \mu)$$

$$\text{where } J = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} ||x^{(n)} - \mu_k||^2$$

$$r_{nk} = \begin{Bmatrix} 1 & \text{if } k = \text{argmin}_j ||x^{(n)} - \mu_j||^2 \\ 0 & \text{otherwise} \end{Bmatrix}$$

$$\mu_k = \frac{\sum_n r_{nk} x^{(n)}}{\sum_n r_{nk}}$$

As shown above, $r_{nk} = 1$ if the nth sample is in the cluster k and $r_{nk} = 0$ for all other clusters. μ_k is each cluster center and J calculates the sum of the squared distance of data points from the center of its own cluster.

Regarding 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 [Bholowalia and Kumar, 2014]. 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.

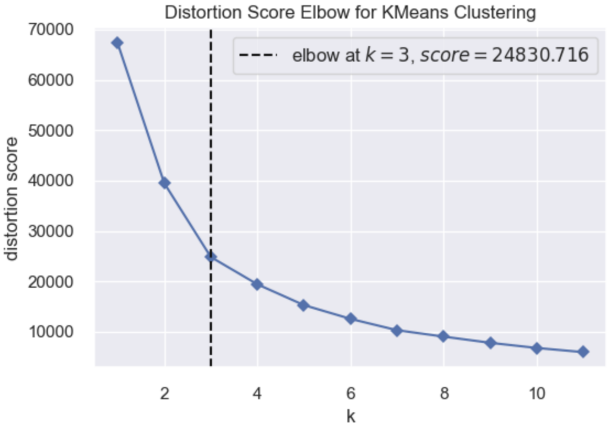
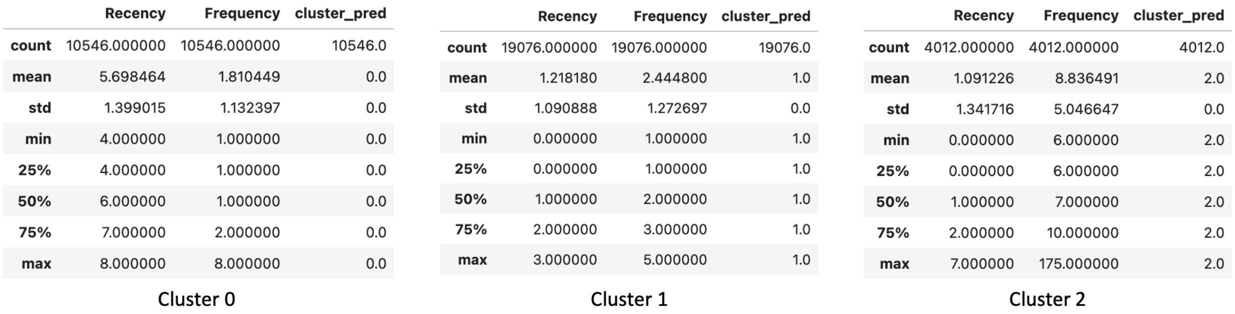


Figure 7: The samples of calculated and standardized RFM scores

From Figure 7, the elbow method divides the customers into 3 groups based on their recency and frequency scores.



	Recency	Frequency	cluster_pred
count	10546.000000	10546.000000	10546.0
mean	5.698464	1.810449	0.0
std	1.399015	1.132397	0.0
min	4.000000	1.000000	0.0
25%	4.000000	1.000000	0.0
50%	6.000000	1.000000	0.0
75%	7.000000	2.000000	0.0
max	8.000000	8.000000	0.0

Cluster 0

	Recency	Frequency	cluster_pred
count	19076.000000	19076.000000	19076.0
mean	1.218180	2.444800	1.0
std	1.090888	1.272697	0.0
min	0.000000	1.000000	1.0
25%	0.000000	1.000000	1.0
50%	1.000000	2.000000	1.0
75%	2.000000	3.000000	1.0
max	3.000000	5.000000	1.0

Cluster 1

	Recency	Frequency	cluster_pred
count	4012.000000	4012.000000	4012.0
mean	1.091226	8.836491	2.0
std	1.341716	5.046647	0.0
min	0.000000	6.000000	2.0
25%	0.000000	6.000000	2.0
50%	1.000000	7.000000	2.0
75%	2.000000	10.000000	2.0
max	7.000000	175.000000	2.0

Cluster 2

Figure 8: K-Means results of the 3 clusters

From the summary of clustering results shown in 8, 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.

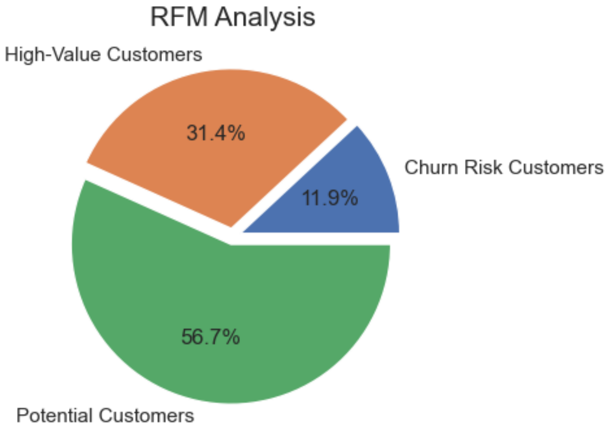


Figure 9: The percentage of each customer group

From Figure 9, the majority of customers are potential customers, with a percentage of 56.7%. It is also a good sign that high-value customers also take a large portion, nearly three times the portion of churn-risk customers. For the 11.9% of churn-risk customers, platforms should provide special offers or send tailored communications to reduce the potential loss. Keeping high-value customers may involve VIP benefits and exclusive offers to foster loyalty. As the largest segment, potential customers may benefit from personalized recommendations aligned with their browsing history to enhance conversion rates.

5 Predictive Behavior Modeling

As user behavior is tracked alongside timestamps, it is intuitive to organize user behavior into a sequence of actions. Previous research has found that developing models for predicting the next user actions is of significant benefit for marketers in enhancing user experience [Lee, 2002, Chen et al., 2018b]. Different users exhibit varying shopping habits - users view a product detail page just once before purchasing directly, others frequently favorite items without completing transactions, and some add items to their shopping carts in a regular manner. Leveraging the meaningful sequential connections between these actions enables the mining of individual user interests in online shopping behavior and predictions of their subsequent actions.

Recurrent neural network (RNN) with its temporal structure, has been widely used to capture sequential dependencies in data. The process involves computing a new state based on the input and the old state [Vu et al., 2016].

$$\begin{aligned} h_t &= f_W(h_{t-1}, x_t) && \text{Update hidden state} \\ y_t &= W_{hy} h_t + b_y && \text{Calculate the output} \end{aligned}$$

Although vanilla RNNs are simple, the backward flow of gradients in RNN can explode or vanish, resulting in poor performance. Therefore it's common to use its variant, long short term memory (LSTM), with additive interactions to improve gradient flow [Hochreiter and Schmidhuber, 1997].

The method overview in this part is illustrated in Figure 10 and the model is implemented using PyTorch. Each user behavior towards one item is input in order and calculates the current state. By maintaining block states and utilizing gating mechanisms, LSTM layers learn to model sequential relations and output the predicted event.

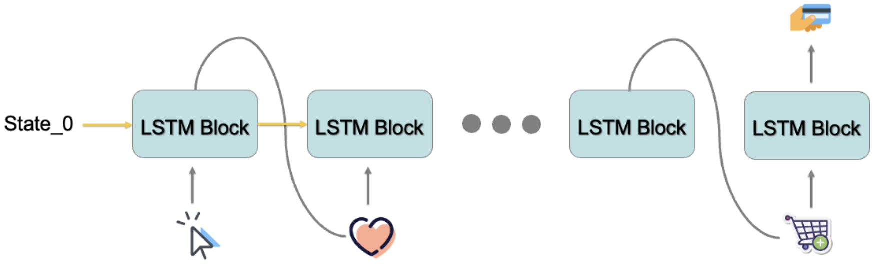


Figure 10: The overview of the method used in the project

The training loss curve is shown in Figure 11, which provides a visual representation of how the loss of the LSTM model changes over the training process. The curve has a clear decreasing trend throughout the whole steps, indicating the model is learning and improving its ability to predict user actions.

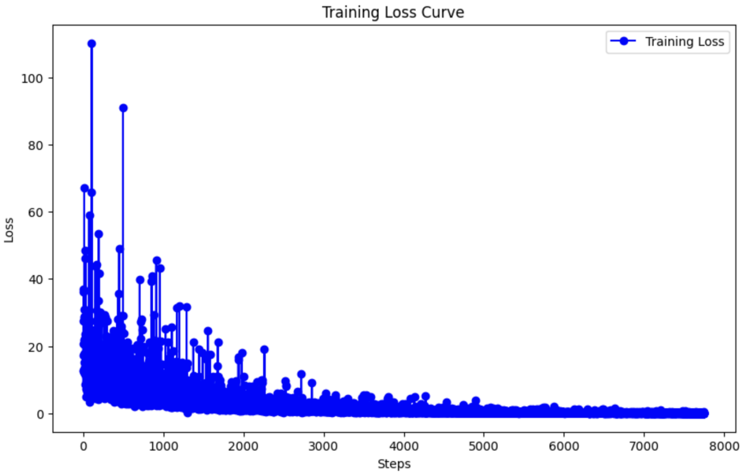


Figure 11: The training loss curve of lstm

To compare with the performance of our LSTM model, a vanilla RNN is also trained. The accuracy results are presented below in Table 3. Compared with randomly selecting one of the four behaviors as the lower bound threshold, the accuracy is much higher using recurrent neural networks. Meanwhile, LSTM has outperformed RNN in predicting users' next action by 22%.

Table 3: The experimental results of the three methods

Methods	**Test Accuracy**
Random Choice	0.25
RNN	0.73
LSTM	0.95

6 Conclusion & Discussion

In this project, how to mine user interests and extract useful information through a large amount of user data on e-commerce platforms has been researched. Given the raw data, companies are eager to discover user behavior patterns and customer groups to adjust their marketing strategies. To help solve the problem, a pipeline has been designed, which is a full investigation of data processing and analysis, customer segmentation and predictive behavior modeling.

The research questions mentioned in Section 1 Introduction can be answered as follows:

- The interest of user purchasing greatly increases during weekends and also boosts when there are shopping festivals with large discounts. In terms of each day, the peak activity hours are from 12 am to 2 pm, usually a midday break during work hours.
- The top-10 best-selling categories are shown in Figure 5. After investigating the conversion from page views to purchases, it's surprising that even among the top-selling categories, the conversion rates remain exceedingly low, mostly below 4%. This suggests that the recommended items by the platforms do not meet users' purchasing preferences, and the corresponding algorithms need to be improved.
- The RFM analysis combined with K-means clustering is applied to group customers. The metrics measured are the recency and frequency of a user's historical purchase behaviors. The warning signs of churn are high recency and low frequency. Through this method, the result shows around 11.9% of users are churn-risk users, who were inactive and would likely stop using the platform. Tailored strategies are required to avoid the loss. Fortunately, the high-value customers and potential customers still take up a very large portion, which can contribute positively to Taobao's operations.
- To predict the next interaction of a user towards one item, their past activities can be converted into a sequence. Then, the sequential patterns inside can be learned and discovered automatically through recurrent neural networks. The LSTM model predicts the next user behavior with an accuracy of 95%, which presents the feasibility of the model.

7 Future Study

Although the project has conducted some research on the pipeline of user behavior analysis, there are still some limitations and practical problems to be further studied. They are listed as follows.

1. Due to the limitations of local computer resources, only 5% data of the original dataset is sampled, thus the data analysis may encounter some bias and information loss. Distributed

computing frameworks like Spark and computing resources like GPU can be used to process the whole dataset for a more comprehensive analysis.

2. Regarding predictive behavior modeling, this project only considers predicting the next user action towards one item. However, the RNN and LSTM models can be further applied with recommendation systems to predict the user preferences of different products and categories.

References

[Bholowalia and Kumar, 2014] Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

[Carmona et al., 2012] Carmona, C., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M., and García, S. (2012). Web usage mining to improve the design of an e-commerce website: Orolivesur.com. *Expert Systems with Applications*, 39(12):11243–11249.

[Chen et al., 2018a] Chen, C., Kim, S., Bui, H., Rossi, R., Koh, E., Kveton, B., and Bunescu, R. (2018a). Predictive analysis by leveraging temporal user behavior and user embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2175–2182, New York, NY, USA. Association for Computing Machinery.

[Chen et al., 2018b] Chen, C., Kim, S., Bui, H., Rossi, R., Koh, E., Kveton, B., and Bunescu, R. (2018b). Predictive analysis by leveraging temporal user behavior and user embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2175–2182, New York, NY, USA. Association for Computing Machinery.

[Cooil et al., 2008] Cooil, B., Aksoy, L., and Keiningham, T. L. (2008). Approaches to customer segmentation. *Journal of Relationship Marketing*, 6(3-4):9–39.

[Hernández et al., 2017] Hernández, S., Álvarez, P., Fabra, J., and Ezpeleta, J. (2017). Analysis of users' behavior in structured e-commerce websites. *IEEE Access*, 5:11941–11958.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

[Kohavi, 2001] Kohavi, R. (2001). Mining e-commerce data: The good, the bad, and the ugly. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, page 8–13, New York, NY, USA. Association for Computing Machinery.

[Lee, 2002] Lee, P.-M. (2002). Behavioral model of online purchasers in e-commerce environment. *Electronic Commerce Research*, 2:75–85.

[Maryani et al., 2018] Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno, and Pratama, E. A. (2018). Customer segmentation based on rfm model and clustering techniques with k-means algorithm. In *2018 Third International Conference on Informatics and Computing (ICIC)*, pages 1–6.

[McDowell et al., 2016] McDowell, W. C., Wilson, R. C., and Kile, C. O. (2016). An examination of retail website design and conversion rate. *Journal of Business Research*, 69(11):4837–4842.

[Miglautsch, 2000] Miglautsch, J. R. (2000). Thoughts on rfm scoring. *Journal of Database Marketing & Customer Strategy Management*, 8:67–72.

[Na et al., 2010] Na, S., Xumin, L., and Yong, G. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 63–67.

[Singh and Singh, 2010] Singh, B. and Singh, H. K. (2010). Web data mining research: A survey. In *2010 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–10.

[Vu et al., 2016] Vu, N. T., Adel, H., Gupta, P., and Schütze, H. (2016). Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv:1605.07333*.

[Wu and Chou, 2011] Wu, R.-S. and Chou, P.-H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3):331–341.

[Yltävä, 2023] Yltävä, L. (2023). Global e-commerce conversion rate by device 2023.

[Zhang and Segall, 2008] Zhang, Q. and Segall, R. S. (2008). Web mining: a survey of current research, techniques, and software. *International Journal of Information Technology & Decision Making*, 7(04):683–720.

[Zhu et al., 2018] Zhu, H., Li, X., Zhang, P., Li, G., He, J., Li, H., and Gai, K. (2018). Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1079–1088.