Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

Identify Potential Loyalists in Shopping Festival: Repeat Buyer Prediction for E-Commerce Based on Feature Engineering and Ensemble Learning

Nishuang Yue (556184ny@student.eur.nl)

Supervisor:	Pieter Schoonees
Second assessor:	Sean Brüggemann
Date final version:	25th August 2024

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

This study aims to identify key features from user log data to predict potential repeat buyers following shopping festival promotions and to develop effective prediction models. Using real user log data from the 2017 T-mall “Double-Eleven” Shopping Festival and the preceding six months, comprehensive feature engineering and selection processes were undertaken. The study constructed six feature profiles: User Profile, Seller Profile, User-Seller Profile, User-Item/Brand/Category Profile, Seller-Item/Brand/Category Profile, and Bipartite Graph Profile, of which 116 features were selected for model training. Machine learning methods, including ensemble learning techniques, were employed to build predictive models. The imbalanced dataset was handled using the SMOTE method, and models such as Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost were developed. Bayesian Optimization and 5-fold cross-validation were used for parameter tuning. The XGBoost model achieved the highest AUC value of 0.6521, and further enhancement through model stacking of XGBoost and CatBoost resulted in an improved AUC of 0.6554. To determine the most critical features in predicting repeat buyers, SHAP values were calculated for each feature using the best-performing XGBoost model. The findings revealed that customer consumption diversity at a seller, indicated by the number of product categories, was the most significant predictor. Additionally, metrics of user engagement levels, particularly those during the “Double-Eleven” Shopping Festival, were among the most dominant features.

Keywords: repeat buyer; shopping festival; e-commerce; feature engineering; ensemble learning

Contents

Chapter 1

Introduction

1.1 Research Background

The development of the internet has witnessed significant changes in people's lifestyle over the past few decades. E-commerce, as a derivative of the internet, is changing consumers' shopping habits, redefining business models, and having a substantial impact on the global economy. According to Statista, global online shopping sales went up to about 5.8 trillion U.S. dollars in 2023, marking a 9.4% growth from the previous year's 5.3 trillion U.S. dollars (Statista, 2024). In China, online retail sales for 2023 reached 15.4 trillion RMB, posting an 11% increase from 2022 and maintaining its status as the world's largest e-commerce market for the 11th consecutive year (Ministry of Commerce, People's Republic of China, 2024).

To further encourage consumption and boost sales, e-commerce platforms have been organizing shopping festivals on specific dates (e.g., Black Friday, Cyber Monday), where consumers are incentivized to shop within a limited period with significant discounts, offered at "the lowest prices of the year". On November 11th, 2009, Tmall, a Chinese e-commerce platform under the Alibaba Group, launched the "Double-Eleven Shopping Festival". Initially, only 57 merchants participated in this festival, and the promotions offered were limited, but the sales far exceeded expectations. According to recent disclosures from Tmall, the 2021 "Double-Eleven" sales reached 540.3 billion RMB, with over a thousand brands achieving sales in just one hour that surpassed the entire day's sales on "Double-Eleven" in 2020. Witnessing the huge success of Tmall, other Chinese e-commerce platforms all followed the same practice, turning the "Double-Eleven Shopping Festival" into an annual celebration for the entire Chinese e-commerce industry, which gradually influences the global e-commerce sector.

Despite the promising turnover posted after e-commerce shopping festivals like "Double-Eleven", negative impact of these shopping festivals has been exposed. On the one hand, the Chinese e-commerce industry has exhibited an unhealthy growth pattern with a high dependency on festival promotions. Merchants blindly follow low-price strategies, ensnared in a price war of homogenized products. Such low-price strategies are unsustainable to the long-term development of e-commerce merchants. Firstly, price wars hinder businesses from investing in innovation and quality improvements, eventually leading to low product quality and consumer dissatisfaction (Heil & Helsen, 2001). Secondly, these strategies also undermine customer trust and loyalty, as customers attracted solely by low prices may easily switch to competitors (Dick & Basu, 1994).

On the other hand, constrained by limited growth resulting from the low-price strategies, merchants have resorted to inventing various complicated promotional rules that confuse customers, attempting to gain profits through false advertising and consumer deception (Xie et al., 2023). The once simple and straightforward sales promotion has evolved into a complex mathematical problem, greatly diminishing the shopping experience for consumers during e-commerce shopping festivals. According to a report published by the Consumers Association of China in 2017, among the 539 items claimed to be on sale during the "Double-Eleven" promotion, 78.1% of them could be purchased outside the shopping festival at "Double-Eleven" prices or even lower prices. Issues such as increasing the original price before applying any discount, displaying a fake original price alongside the discount, and arbitrary price labeling are particularly prominent (Ministry of Commerce, People's Republic of China, 2018).

Plagued by shopping festival scams, consumers are becoming fatigued with various "festival promotions". Market competition in the e-commerce industry is intensifying, and customer acquisition costs are increasing. How to attract new customers and, more importantly, convert these new customers into loyal repeat customers has become a challenge for e-commerce businesses.

1.2 Research Relevance

Studies on sales promotion across industries have consistently concluded that it can increase the purchase conversion rate (Becerril-Arreola et al., 2013; Wicks & Schuett, 1991; Zhang et al., 2013), helping merchants acquire new customers. However, many of these attracted new customers are one-time buyers who do not make additional purchases after the "good deals" provided in the shopping festival. Promotions targeting these one-time buyers will not contribute to future sales for the store. To address this issue, merchants must identify individuals who have the potential to become repeat buyers. Through targeted marketing aimed at these potential loyal customers, businesses can enhance the long-term return on investment (ROI) for the store.

This study is relevant in the following three aspects. Firstly, it is a widely accepted business wisdom that acquiring a new customer is five to ten times more costly than retaining an existing one, while the profit contributed by a loyal customer is 16 times that of a new customer, revealing the importance and promising ROI of identifying and retaining potential loyal customers. Furthermore, current marketing strategies are shifting from product sales, which focus on one-time revenue generation, to long-term customer relationship management, which emphasizes the continuous revenue growth of the business (Ataman et al., 2010). Therefore, developing a good relationship with customers in the long run has become an important strategy not only in a financial sense but also for the reputation of businesses. Finally, a company's marketing resources are limited. Customer relationship management requires businesses to carefully assess the costs and benefits of investments and determine the optimal allocation of resources to marketing and sales activities over time (Venkatesan & Kumar, 2004), with identifying the most valuable customer group being a good starting point for tailoring positioning strategies.

Based on comprehensive feature engineering work and state-of-the-art machine learning algorithms, this study aims to identify potential loyal customers after shopping festival promotions for Tmall, one of the largest e-commerce platforms in China, to help it better segment, target, and position its customers. The study uses user log data, also known as clickstream data, to explore factors that impact customers making repeat purchases at a certain merchant. Specifically, this study will try to answer the following research questions: *What are the important features that can be extracted from user log data to identify potential repeat buyers after shopping festival promotions, and how to predict these repeat buyers effectively?*

The result of this study is relevant not only for the local Tmall e-commerce platform but also for other e-commerce platforms that collect customer clickstream data on a similar scale. On the one hand, Alibaba Group recently announced its plans to introduce the Tmall platform into the European market. On the other hand, the concept of “shopping festivals” has become a global phenomenon for many e-commerce businesses. For example, Amazon’s “Prime Day” is a notable case.

Chapter 2

Literature Review

This study will leverage large-scale user clickstream data and machine learning algorithms to understand a customer's buying behavior and to predict whether a customer will become a repeat buyer after certain promotions. Therefore, the literature review of this study will focus on two research streams: user behavior research based on clickstream data and repeat buyer prediction based on machine learning algorithms.

2.1 User behavior research based on clickstream data

In today's digital world, users generate a vast amount of data through their interactions with companies across various channels. These interactions include a wide range of activities, such as browsing websites, making purchases, leaving reviews, and engaging with content. Each of these interactions leaves a digital footprint, collectively forming a rich source of data that companies can leverage to gain insights into user behavior. Krafft et al. (2021) classified customer data disclosure into two types: customer-initiated data disclosure and passive data disclosure, also known as declarative data and non-declarative data (Blasco-Arcas et al., 2022). According to Blasco-Arcas et al. (2022), declarative data are data that voluntarily and consciously shared by customers with the company, for example, through consumer survey, or user-generated content such as product review or comments posted on social media. Non-declarative data consist of observable customer behaviors, either digital or physical, that companies can monitor, process, and analyze using specific technologies to gain insights into consumer behavior. Typically, non-declarative data are generated without the customer's awareness, are large in volume, and can be either structured, such as transaction histories, or highly unstructured, such as website navigation paths (Balducci & Marinova, 2018).

Recent decades have seen a bloom of marketing research on user behavior using unstructured non-declarative data due to their ease to obtain and availability in volume (Balducci & Marinova, 2018), notably clickstream data. Clickstream data refers to the electronic record of a user's internet activity, tracking the sequence of actions a visitor performs while browsing the Web (Bucklin & Sismeiro, 2009). Early research mainly uses clickstream data to gain insights behind users' website browsing patterns. The study of Bucklin and Sismeiro (2003) reveals that a visitor's level of interest in online purchasing can be reflected by the number of pages viewed and the time spent on the site. Moe (2003) further considered the content of the pages viewed and suggested that visits can be categorized as a buying, browsing, searching, or knowledge-building visits, whose purchasing likelihood varies. Moe and Fader (2004) modelized the visit-to-purchase conversion based on the pattern of previous visits and purchases for each site visitor. Sismeiro and Bucklin (2004) developed a model to predict individual online buying behavior by studying visitors' time and page views, repeat visits, use of interactive decision aids, data input effort, and information gathering and processing. Their study concluded that the frequency of repeat visits does not necessarily indicate the likelihood to make a purchase and the presence of advanced decision-making tools does not assure conversion.

Later studies have made richer use of clickstream data, specifically related to personalized advertising and recommendation system design to improve users' online shopping experience. Manchanda et al. (2006) studied the individual-level data on the amount of banner advertising exposures, the variety of websites and pages the consumer encounter advertisements, and the diversity of advertisements encountered, suggesting that the former three factors all have a positive effect on repeat purchase probabilities. Unlike previous studies that only used customer purchase data, Kim and Yum (2011) developed a recommendation system based on customers' navigational and behavioral patterns on e-commerce sites. These patterns include product clicks, basket placements, length of reading time on clicked products, number of visits, types of clicks, and actions like printing and bookmarking product information. The system estimates product preference levels for clicked but unpurchased items and uses collaborative filtering based on these preferences to make recommendations. Su and Chen (2015) mined and analyzed customers' category-level browsing behavior data, such as visiting sequence, frequency, and time spent on each category, through the URLs recorded on an e-commerce website. Their analysis revealed differences in customers' interest patterns and provided insights into web-page optimization and personalized recommendations.

Another popular area of clickstream research is to predict users' online purchasing behavior for e-commerce using data mining and machine learning algorithms, either before the session, for the current session, or for the future sessions. Esmeli et al. (2022) focused on the early stage of customers landing on a site when their navigational data are still unavailable. They suggested a framework to forecast customers' early purchase intentions using contextual and loyalty features. They found that the number of users' past purchases and visits are the most important factors influencing early purchase decisions in new sessions, providing insights for personalized content design and early pricing strategies. Purchase prediction for the current session is also known as purchase intent prediction, which is to use a customer's navigation patterns of a certain session to predict whether this session will end with a purchase, as a customer can engage in online shopping either for utilitarian benefits or hedonic benefits (Bridges & Florsheim, 2008). Mokryn et al. (2019) presented a method to predict the shopping intent of anonymous visitors to a site during a certain session, utilizing the temporal information of the session, its duration, and the recent trendiness of products clicked on in that session. Their findings can be employed to develop an innovative real-time recommendation system, potentially converting browsing users with low purchase intent into buyers. To describe the customer's behavior in a particular month and predicting whether they will make a purchase in the following month, Martínez et al. (2020) generated a large amount of customer features from historical transaction data that are related to purchase time, purchase value, and demographics. Their findings shed light on businesses' inventory planning at the warehouse and customer churn identification.

In summary, plenty of research has studied customers' purchasing behavior based on their clickstream data. Previous studies have shown that customers' transaction history and navigational patterns are predictive of their buying behavior, revealing the close relationship between click and purchase. However, due to the different nature of the tasks, the features used or concluded significant in these studies can vary depending on the research objective, the intended use of study results, and the data availability, with feature engineering establishing the essential foundation for the analysis and conclusion. Regarding this study, the goal of the task is to predict the repurchase behavior of newly acquired customers for a certain merchant after the shopping festival, where the customer-merchant interaction pattern and the temporal nature of the observed navigational behavior (before and during the shopping festival) can play an important role, differentiating the task from other purchasing behavior prediction tasks. So far, customer behavior studies under the scenario of shopping festival mainly focus on consumption motivations and are mostly based on questionnaires (Akram et al., 2018; Chen & Li, 2020; Yang et al., 2018). Only limited attention has been paid to the user behavior modeling and purchase prediction using clickstream data (Zeng et al., 2019; Zhao et al., 2019). Therefore, this study aims at conducting and reporting a comprehensive feature engineering using the clickstream data to reveal the factors that drive the repurchase behavior of new customers acquired during the online shopping festival.

2.2 Purchase and repurchase prediction based on machine learning algorithms

This section offers a systematic overview of relevant research with focus on studies that utilized clickstream data and machine learning algorithms for repeat purchase prediction.

In recent years, many scholars have conducted research on predicting user purchase or repeat purchase behavior using clickstream data. Based on the different nature of the data at hand, such prediction can be classified as either sequence labeling or non-sequence labeling (Graves & Graves, 2012; Koehn et al., 2020). In the context of customer purchase prediction, a sequence labeling task can be a session segmentation task, where the goal is to predict whether each user session ends with a purchase or not. The input sequence can consist various events such as page views, product clicks, and time spent on pages. A sequence labeling task often show the sequential nature of the input events or the dependency of one event on previous events, as the occurrence of a purchase at any point during a user session is often influenced by the sequence of events that occurred earlier in the session. Such dependency is also known as user behavior patterns. For example, the path of a buyer may start by browsing product categories pages, then view individual product pages, add items to the shopping cart, and finally proceed to checkout. Alternatively, a customer may exit the website after viewing a few pages for a short time without making a purchase. These behavior patterns are reflected in the sequence of events captured in the clickstream data. A non-sequence labeling task, on the other hand, can be the prediction for a user's future purchase. It is considered non-sequence labeling because rather than making predictions for the current session, the goal is to predict whether the user is likely to make a purchase within a specific time frame (e.g., within the next month). The features used for a non-sequence labeling task may include user demographics, browsing history, past purchase behavior, etc., but the prediction is made at the user level without considering the sequence of events.

To capture the complex sequential structure of clickstream data in sequence labeling tasks, advanced learning algorithms are needed. Early research tried to customize probabilistic models for the task (Montgomery et al., 2004; Sismeiro & Bucklin, 2004), or leveraged Markov models (Chan, 2014; Lakshminarayan et al., 2016), while recent studies derived better model performance through deep learning such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gate Recurrent Unit (GRU), or the combination of different kinds of Neural Networks (NN). (Toth et al., 2017; Wu et al., 2015)

On the other hand, the clickstream data used in non-sequence labeling tasks does not present a complex sequential nature but requires proper transformation of the unstructured clickstream data to make the task a standard classification or regression problem. In most cases, this transformation is accomplished through comprehensive feature engineering, which requires domain knowledge and can be time-consuming. The purpose of feature engineering is to extract meaningful features that include all relevant information describing customer's latent characteristics from its past purchase behavior. Feature engineering can be primarily done through computing summary statistics of the clickstream data for each customer. For example, aggregation feature (e.g., number of purchases made), average feature (e.g., average number of products viewed within a category), or ratio feature (e.g., share of a certain category viewed among all categories). In addition to basic summary statistics, more advanced feature generation methods have also been employed. For example, Suh et al. (2004) used association rule mining to extract purchase patterns, while Shapoval and Setzer (2018) introduced two novel unsupervised mechanisms to aggregate similar sequences to generalized buying types.

After transforming the unstructured clickstream data into structured predictors, the non-sequence labeling task for buying decision can be now approached with any binary classification algorithm. Early research employed multiple individual prediction models, for example, Logistic Regression (LR), Support Vector Machine (SVM), and Decision Tree (DT) to predict customers' purchases. However, using a single machine learning algorithm can lead to sub-optimal performance. It may have limited capacity to capture complex patterns in the data, especially when relationships are nonlinear and interactions between features are important, or it may suffer from high bias or high variance issues. Therefore, recent studies have proposed various ensemble models to pursue better prediction performance, among which the most popular ones are Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGBoost). Study results show that the ensemble method, especially the boosting algorithm family, outperforms the individual prediction model in terms of accuracy and robustness (Cen et al., 2024; Hwang et al., 2020; Kumar et al., 2019; Martínez et al., 2020).

To further improve the model prediction performance, some research also adopted the Voting or Stacking technique. Voting is an ensemble method where multiple base models are trained independently on the same dataset, and their predictions are combined to make a final prediction. Two main types of voting technique are hard voting, where the final prediction is determined by majority vote or the average of the individual predictions; and soft voting, where the final prediction is determined by the class with the highest average probability or the weighted average of predicted values. For example, Zhao et al. (2019) use Light Gradient Boosting Machine (LightGBM) and XGBoost models to obtain the primary probability respectively and then get an intermediate value of each probability using Sigmoid inverse function. The final repeat buyer probability was then obtained by Sigmoid function after calculating the mean value of the intermediate values. Stacking is another ensemble learning technique that involves training multiple base models and combining their predictions by training a meta-model on the predictions of base models to make the final prediction. Stacking technique is widely used due to its better performance than simple voting. Liu et al. (2020) selected LR and DT based XGBoost model. They then used the model fusion algorithm to avoid the shortcomings of the linear model and the over-fitting of the DT model, further improving the prediction result. Dong et al. (2022) proposed a BERT-MLP prediction model that uses the idea of "large-scale data unsupervised pre-training with small amount of labeled data fine-tuning", whose accuracy is proved to be better than the baseline model. The study of Cen et al. (2024) showed that the stacked RF-LightGBM-LR model and the stacked RF-XGBoost-LR model shows better performance in predicting user repurchase behavior than single prediction models.

In summary, while studies on customer purchase prediction are increasing, limited attention has been paid on repurchase behavior prediction. A literature overview about repurchase behavior prediction can be found in Table 2.1. In general, repurchase prediction can be seen as a complementary of purchase prediction research, where they share the similar strategy to recognize user behavior pattern from their history and to predict their future action. Therefore, the machine learning algorithms used in repurchase prediction are similar to those used in purchase prediction tasks. However, repurchase prediction is different in a sense that, the tasks are typically non-sequence labeling tasks, given the fact that the main objective of the task is to predict whether a customer will make a repeat purchase within a specific time frame instead of in the current session. Thus, repurchase prediction can propose different models and feature selection from purchase prediction.

Table 2.1: *A summary of literature using machine learning algorithms for repeat purchase prediction, with few studies conducted in the context of shopping festivals.*

Reference	**ML Algorithms Used**	**Algorithms Optimizing**	**Shopping Festival**
Kumar et al. (2019)	DT, AdaBoost, RF, SVM, NN	×	×
Zhao et al. (2019)	LR, SVM, FM^a, XGBoost, LightGBM	Soft Voting	√
Liu et al. (2020)	LR, SVM, DT, XGBoost	Stacking	×
Zhang and Wang (2021)	LR, RF, SVM, KNN^b, CNN^c, DF^d	Deep Forest	×
Dong et al. (2022)	LR, RF, GBDT, KNN, XGBoost, MLP^e, BERT^f	Stacking	√
Cen et al. (2024)	LR, RF, LightGBM, XGBoost	Stacking	×

Note: ^{a} Factorization Machine, ^{b} K-Nearest Neighbors, ^{c} Convolutional Neural Network, ^{d} Deep Forest, ^{e} Multilayer Perceptron, ^{f} Bidirectional Encoder Representations from Transformers

Chapter 3

Data

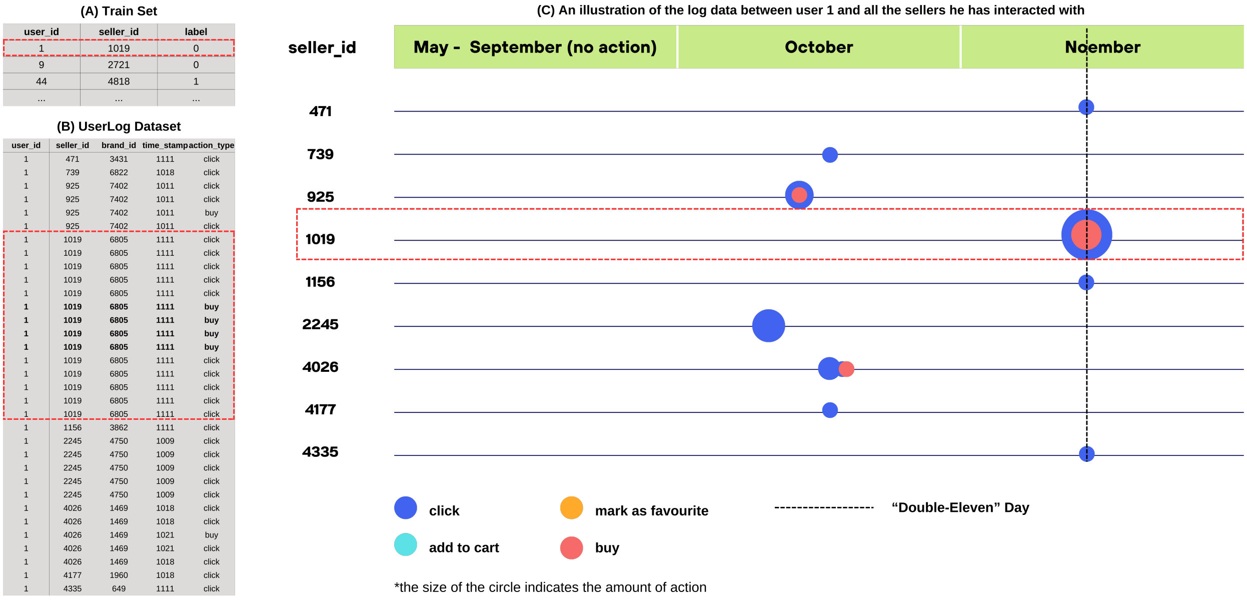
3.1 Data Source and Problem Definition

The data used in this study is publicly available on Tianchi Platform, an online platform developed by Alibaba Cloud for data competitions and collaboration. The dataset comprises anonymized shopping logs of Tmall users from the six months leading up to and including the "Double-Eleven" days of 2017, as well as labels indicating repeat buyers. Due to privacy concerns, the sampling method is biased, so the statistical result derived from this dataset would deviate from the actual result of Tmall. However, this deviation does not affect the applicability of solutions.

The original datasets provided on Tianchi Platform contain four parts: user log information, user demographics, training set, and test set. An overview of each dataset can be found below in Table 3.1. Due to privacy concerns, all log data has been encoded to make the actual user, seller, product, brand, and category information anonymous.

Table 3.1: *An overview of datasets used in this study. The label in the TestSet is muted, as it was originally used for generating results in Tianchi Platform competition. Consequently, the TestSet will only be used for taking subsets of the UserLog and UserInfo datasets in this study, but not for model training or testing.*

Data Set Name	Variable Name	Variable Description
UserLog	user_id	1 – 6 digits unique identification of the buyer
	seller_id	1 – 4 digits unique identification of the seller
	item_id	1 – 7 digits unique identification of the product
	cat_id	1 – 4 digits unique identification of the category that the product belongs to
	brand_id	1 – 4 digits unique identification of the brand that the product belongs to
	action_type	0 = click, 1 = add to shopping cart, 2 = buy, 3 = mark as favorite
	time_stamp	The date that the action took place in the format of mm/dd, from 0511 to 1112
UserInfo	user_id	1 – 6 digits unique identification of the buyer
	age_range	Age group of the buyer. 1 = [0-17], 2 = [18-24], 3 = [25-29], 4 = [30-34], 5 = [35-39], 6 = [40-49], 7 and 8 = [50+], 0 and NULL = Unknown
	gender	Gender of the buyer. 1 = female, 2 = male, 0 and NULL = Unknown
TrainSet	user_id	1 – 6 digits unique identification of the buyer
	merchant_id	Same as seller_id. 1 – 4 digits unique identification of the seller
	label	Repeat buyer label. 0 = non-repeat buyer, 1 = repeat buyer
TestSet	user_id	1 – 6 digits unique identification of the buyer
	merchant_id	Same as seller_id. 1 – 4 digits unique identification of the seller
	label	NULL



(A) Train Set

user_id	seller_id	label
1	1019	0
9	2721	0
44	4818	1
...	...	...

(B) UserLog Dataset

user_id	seller_id	brand_id	time_stamp	action_type
1	471	3431	1111	click
1	739	6822	1018	click
1	925	7402	1011	click
1	925	7402	1011	click
1	925	7402	1011	buy
1	925	7402	1011	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	**1019**	**6805**	**1111**	**buy**
1	**1019**	**6805**	**1111**	**buy**
1	**1019**	**6805**	**1111**	**buy**
1	**1019**	**6805**	**1111**	**buy**
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1019	6805	1111	click
1	1156	3862	1111	click
1	2245	4750	1009	click
1	2245	4750	1009	click
1	2245	4750	1009	click
1	2245	4750	1009	click
1	2245	4750	1009	click
1	4026	1469	1018	click
1	4026	1469	1018	click
1	4026	1469	1021	buy
1	4026	1469	1021	click
1	4026	1469	1018	click
1	4177	1960	1018	click
1	4335	649	1111	click

(C) An illustration of the log data between user 1 and all the sellers he has interacted with

*the size of the circle indicates the amount of action

Figure 3.1: *An illustration of the connection between TrainSet and UserLog Dataset using user 1 as an example. (A) TrainSet, which contains the user-seller pairs that need to be predicted. (B) UserLog Dataset, which contains the log data between the target users as shown in TrainSet and all the sellers with whom they have interacted. (C) An illustration of the log data between user 1 and all the sellers with whom he has interacted. User 1 pairs with Seller 1019, but not other sellers, to be part of the prediction task, is because User 1 made purchases from seller 1019 for the first time on the "Double-Eleven" Day, which makes User 1 to be considered as a new buyer of Seller 1019. The label of User 1 in the TrainSet is 0, indicating that User 1 is not a repeat buyer of Seller 1019 after 6 months.*

The objective of this study is to predict whether a new buyer of a certain seller will become a repeat buyer for that seller within six months after the "Double Eleven" shopping festival. Specifically:

- A new buyer of a certain seller is defined as a buyer who made the "buy" action on any product from this seller on and only on the 11th of November, 2017;
- A repeat buyer of a certain seller is defined as a buyer who, again, made the "buy" action on any product from this seller within the 6 months after the 11th of November, 2017.

In this study, both new and repeat buyers are pre-identified and reflected in the TrainSet/TestSet. In other words, users appearing in the TrainSet/TestSet are all new users for a certain seller, and their labels indicate whether they became repeat buyers for that seller.

An illustration of the task can be found in Figure 3.1 above. The illustration uses the log data of a buyer with user_id = 1, who is a 25-29-year-old male. As shown in the chart, user 1 interacted with 9 stores during the study period, with all interactions occurring in October and November. On November 11th, user 1 made three purchases from seller 1019. Since user 1 had not made any purchases from seller 1019 before, he is considered a new buyer for seller 1019. Consequently, the "user 1 – seller 1019" pair will appear in the TrainSet/TestSet for the prediction task, and the label "0" indicates that he is not a repeat buyer. Since user 1 did not make any other purchases from any other sellers on November 11th, they will not be part of the prediction task. However, their interactions will be used to study user 1's shopping behavior.

3.2 Descriptive Analysis

3.2.1 Full Dataset

The original user log dataset contains 54,925,330 log entries of 424,170 customers from 12th of May to 12th of November. An overview of the UserLog dataset and a statistic summary of the variable *action_type* can be found below in Table 3.2 and Table 3.3, respectively.

Table 3.2: *A statistic summary of variables in the full UserLog Dataset*

#Log	#User	#Seller	#Product	#Brand	#Category	#Day
54,925,330	424,170	4,995	1,090,390	8,444	1,658	186

Table 3.3: *A statistic summary of the variable "action_type" in the full UserLog Dataset*

#Click	#Add to cart	#Buy	#Mark as favourite
48,550,713	76,750	3,292,144	3,005,723
88.39%	0.14%	5.99%	5.47%

The dataset was originally published for competition purpose and had been split into TrainSet and TestSet, with the label muted in the Test Set. A statistic summary of the full TrainSet and TestSet can be found below in Table 3.4.

Table 3.4: *A statistic summary of the full TrainSet and TestSet*

Dataset	#User	#Merchant	#User-Merchant Pair	#Repeat Buyer	%Repeat Buyer
TrainSet	212,062	1,993	260,864	15,952	6.12%
TestSet	212,108	1,993	261,477	NULL	NULL

As the labels in the TestSet are unavailable, this study took subsets of the UserLog and UserInfo datasets to only include users present in the TrainSet. Furthermore, there are 45,364 (0.16%) observations with missing *brand_id* in the subset of UserLog dataset. Since the amount of missing data is insignificant and *brand_id* is not considered crucial for the analysis, this study removed these observations. Lastly, to manage computation time due to the large dataset, this study randomly sampled 10% of the users in the TrainSet and further took subsets of the UserLog and UserInfo datasets accordingly. The descriptive analyses of the final datasets are presented in the following sections.

3.2.2 Subset: User log data

The subset of UserLog dataset contains 2,727,610 log entries from 21,206 unique customers, ranging from the 12th of May to the 11th of November. An overview of the subset and a statistical summary of the variable *action_type* can be found in Table 3.5 and Table 3.6, respectively.

Table 3.5: *A statistic summary of the subset of UserLog Dataset*

#Log	#User	#Seller	#Product	#Brand	#Category	#Day
2,727,610	21,206	4,995	395,648	6,864	1,269	185

Table 3.6: *A statistic summary of the variable "action_type" in the subset of UserLog Dataset*

#Click	**#Add to cart**	**#Buy**	**#Mark as favourite**
2,416,446	12	162,675	148,477
88.59%	0.00%	5.96%	5.44%

Taking a closer look at the distribution of days on which buyers' actions occurred, Figure 3.2 indicates that a significant part of actions took place in November, particularly on the 10th and 11th (19.3% and 24.6%, respectively).

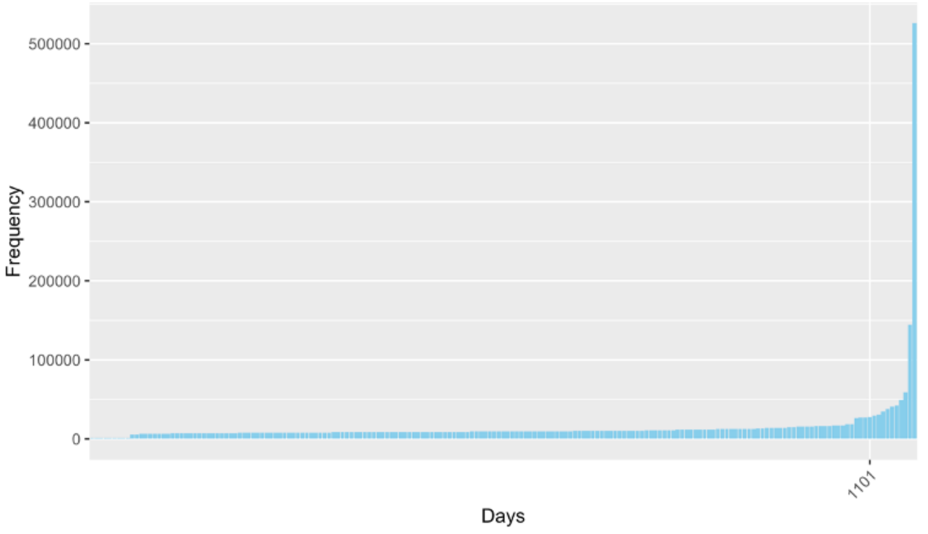


Figure 3.2: *The date distribution of user actions. The time span is from May 12th to November 11th. Most user actions occurred in November, with actions on November 10th and November 11th account for 43.9% of the total user actions.*

Regarding the number of actions taken per customer, the distribution is right-skewed, as shown in Figure 3.3. This indicates that the majority of customers have a total action count of fewer than 100. Specifically, the minimum, mean, and maximum numbers of actions per customer are 2, 129, and 3,596, respectively.

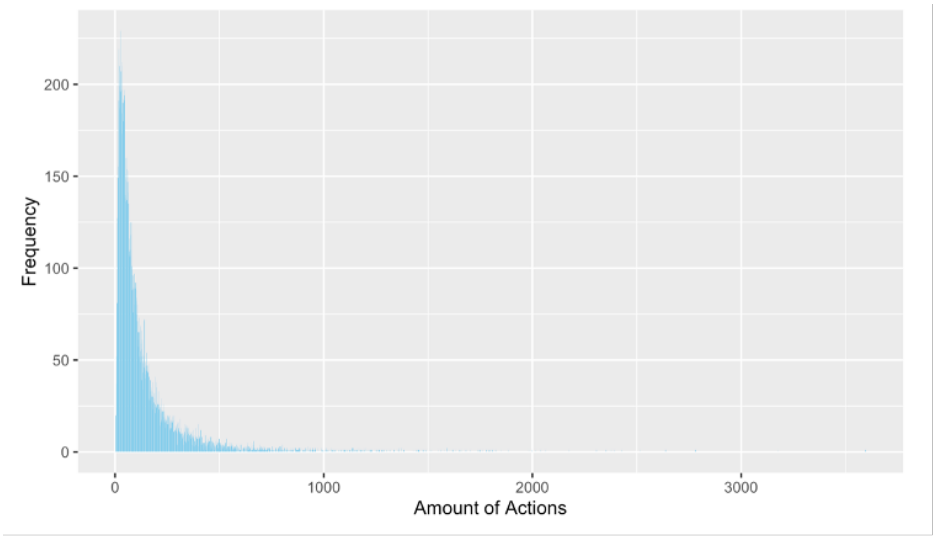


Figure 3.3: *The distribution of the amount of user actions per user. The average is 129 times.*

3.2.3 Subset: User Demographics

The subset of UserInfo contains the age range and gender information of 21,062 unique customers. A statistical summary and distribution plots can be found in Table 3.7, Table 3.8, and Figure 3.3.

Table 3.7: *A statistic summary of the variable "age_range" in the subset of UserInfo Dataset*

#[0-17]	#[18-24]	#[25-29]	#[30-34]	#[35-39]	#[40-49]	#50+	#Unknown
2	2,605	5,617	3,923	2,019	1,758	409	4,873
0.01%	12.28%	26.49%	18.50%	9.52%	8.29%	1.93%	22.98%

Table 3.8: *A statistic summary of the variable "gender" in the subset of UserInfo Dataset*

#Female	#Male	#Unknown
14,282	6,121	803
67.35%	28.86%	3.79%

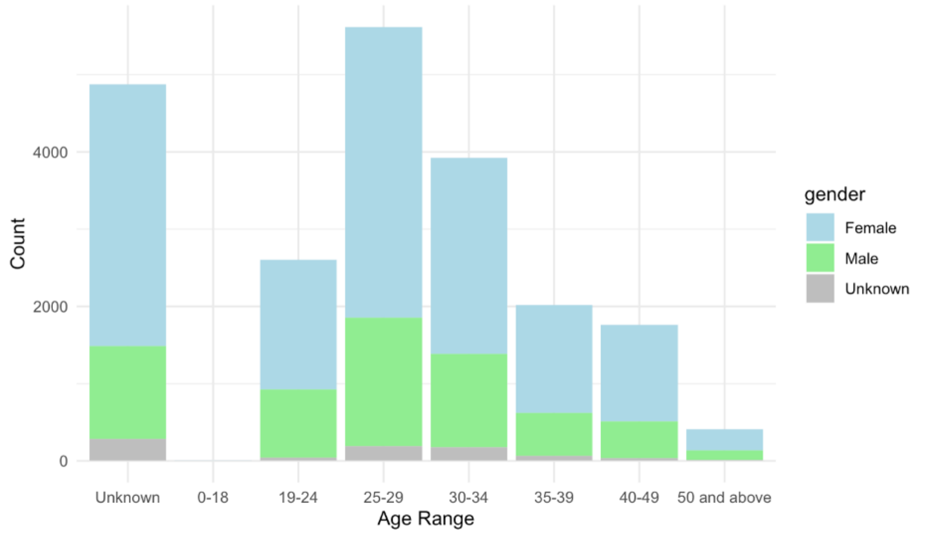


Figure 3.4: *Gender composition by age range. Females and young adults (25-34 years old) are dominant in the sample. The absence of users in the age group of 0-18 is likely due to restrictions on setting up online payment accounts for minors in China.*

Considering the significant number of missing values in *age_range* and *gender* variables, this study chose not to remove records of customers with unknown age or gender, because eliminating these observations would significantly reduce the sample size available for this study and discard the hidden information contained within these samples. Instead, this study retained the "unknown" feature, treating these customers as a group with strong privacy concerns regarding their personal information on Tmall. When registering as a customer or making a purchase on Tmall, providing age and gender is not mandatory. Therefore, data is not randomly missing in the dataset but is missing due to users' concerns about privacy or lack of obligation to provide complete data. This is especially true for optional data requested by the system (Sim et al., 2015).

Chapter 4

Methodology

This section introduces the methodologies used in this study, including feature scaling, imbalanced data processing, feature selection methods, prediction models, model evaluation metrics, and feature importance evaluation methods.

4.1 Feature Scaling

Feature scaling is important in data preprocessing to ensure that all features contribute equally to the model's learning process. Many machine learning algorithms, especially distance-based ones, are sensitive to features with large scales, which can dominate the training process and lead to biased outcomes. Feature scaling mitigates this issue by standardizing the range of independent variables, improving the performance and training stability of machine learning models, and generating more comparable results.

Two of the most common feature scaling methods are Min-Max Scaling and Z-Score Scaling. Min-Max Scaling typically re-scales variables to the range of [0, 1] by subtracting the minimum value of the variable from each data point and then dividing by the range of the variable, mapping the smallest value to 0 and the largest value to 1. However, Min-Max Scaling is sensitive to outliers because extreme values can stretch or compress the range of the remaining data points, potentially distorting the scaling of the rest of the data.

The Z-Score method standardizes the variables by subtracting the mean and dividing by the standard deviation, transforming the data to have a mean of 0 and a standard deviation of 1. Mathematically, the scaled value z_i of a feature x_i is computed as:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i is the original feature value, μ is the mean of the feature, and σ is the standard deviation.

Compared to Min-Max Scaling, Z-Score Scaling is less sensitive to outliers as it standardizes data by mean and standard deviation, which moderates extreme values. Considering the fact that outliers are not defined and addressed in this study, Z-Score Scaling is employed, as it is more robust in the presence of extreme values.

4.2 Data imbalance processing

When training a predictive model on an imbalanced dataset, where some classes are underrepresented, the model can become biased, favoring the majority class while neglecting the minority class. Several techniques can be used to address this bias, including oversampling, undersampling, and synthetic data generation. Oversampling works by duplicating observations in the minority class, while undersampling works by removing observations in the majority class.

In this study, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) will be applied. SMOTE mitigates the data imbalance issue by generating synthetic samples for the minority class. It works by selecting a sample from the minority class and finding its k-nearest neighbors, then generating synthetic samples along the line segments connecting the minority sample and its neighbors. Specifically, if x_i is a minority class sample and $x_{i1}, x_{i2}, \ldots, x_{ik}$ are its k-nearest neighbors, a synthetic sample x_{new} is generated as:

$$x_{\text{new}} = x_i + \delta \times (x_{in} - x_i)$$

where x_{in} is a randomly chosen neighbor and δ is a random number between 0 and 1.

SMOTE ensures the model to learn equally from both classes. By generating synthetic samples rather than simply duplicating or reducing, SMOTE reduces the risk of overfitting compared to traditional oversampling and retains valuable information as well as the dataset size compared to undersampling.

4.3 Feature selection method

Feature selection is important for improving model performance and interpretability. By reducing data dimensionality and removing redundant or noisy features, feature selection enhances model training efficiency and prevents overfitting. Mainstream feature selection methods include filter methods, wrapper methods, embedded methods, and other methods such as feature importance from tree-based algorithms. This study employs a combination of filter methods and tree-based feature importance to perform feature selection.

- *Variance Threshold.* One Filter method is to remove features with low variance, assuming that they contain little information. Two hyperparameters determine whether a variable has low-to-near-zero variance. The first is Frequency Cutoff, which determines the ratio of the frequency of the most common value to the second most common value. The second is Uniqueness Cutoff, which defines the percentage of unique values a variable must have relative to the total number of samples.

- *Correlation Coefficient.* Another Filter method is to select features that are highly correlated with the target variable while removing redundant features that are highly correlated with each other, based on their correlation coefficients.

- *Feature importance from tree-based methods.* Variable importance can be calculated using tree-based methods, such as Random Forest, where Gini Importance is employed by assessing how much each feature reduces the Gini impurity across all trees. When a feature splits a node, it decreases the Gini impurity, indicating higher homogeneity. The Mean Decreased Gini is the average of these reductions for each feature across all trees in the forest. Features with higher Mean Decreased Gini values are more important, as they contribute more to improving the purity of the splits and, consequently, to the model's predictive power.

4.4 Prediction models

This section introduces the machine learning models employed in this study. Specifically, we focus on the application of ensemble methods, which are known for their superior prediction performance compared to single models due to their ability to reduce model variance and improve generalization. Main types of ensemble methods includes:

- *Bagging*, which is to aggregate the predictions of multiple models, typically by averaging or voting, that are trained on different subsets of the whole data set with replacement.
- *Boosting*, which is to sequentially train models where each new model corrects the mistakes made by the previous ones, focusing on observations that are hard to predict.
- *Stacking*, which is to train a meta-model that learns how to optimally combine the outputs of multiple base models to make the final prediction.

We start with a single model, Logistic Regression, to establish a straightforward and interpretable performance benchmark for this study. Then, considering the complexity and high-dimensionality of user log data, we explore different ensemble learning methods, including Bagging (Random Forest), Boosting (XGBoost, LightGBM, and CatBoost), and Stacking to comprehensively capture the complex behavior of repeat buyers, leveraging the best practices in ensemble learning for optimal performance.

4.4.1 Logistic Regression (LR)

Logistic Regression is a parametric algorithm that can predicts the probability of a binary outcome by applying the logistic function to transform a linear combination of predictor variables. The (binary) LR model predicts the probability that the dependent variable Y equals 1 (the outcome of interest) given the predictor variables X. The model can be expressed using the following equation for the log-odds of the outcome:

$$\text{logit}(P(Y=1|X)) = \log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where $P(Y=1|X)$ is the probability of the outcome occurring. β_0 is the intercept. $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the predictor variables $X_1, X_2, \ldots, X_k$.

The method is grounded in Maximum Likelihood Estimation, where hyperparameters are optimized to maximize the likelihood of observed outcomes. LR offers advantages such as simplicity, interpretability, and efficiency for linear relationships, but it is sensitive to outliers and relies on assumptions of linearity. Thus, validating model assumptions is important for accurate predictions. Nevertheless, due to its simplicity, LR is employed as the baseline model in this study.

4.4.2 Random Forest (RF)

To address the limitations of linear functional form assumptions and sensitivity to outliers in binary logistic regression, non-parametric classification methods like Random Forest are employed. Random Forest (Breiman, 2001) is an ensemble learning technique that trains multiple decision trees and combines their predictions to enhance accuracy and robustness. It utilizes bagging by creating diverse trees from random subsets of data and features. Each tree then contributes to the final prediction by voting for a class, with the forest predicting the class that receives the most votes.

RF offers advantages such as high predictive performance, resilience to overfitting, and effective handling of complex relationships. However, it may require significant computational resources and can be less interpretable due to its ensemble nature. Therefore, thorough model validation is important to mitigate potential drawbacks.

4.4.3 Extreme Gradient Boosting (XGBoost)

In addition, Extreme Gradient Boosting (XGBoost) introduced by Chen and Guestrin (2016) is employed. As a boosting method, XGBoost improves model accuracy by combining weak learners to form a strong predictor. It addresses model limitations by assigning higher weights to misclassified instances. Unlike bagging methods such as RF, XGBoost prioritizes challenging instances, thereby creating a robust learner with enhanced performance. The objective function of XGBoost combines a loss function and a regularization term:

$$\text{Objective} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

where $l(y_i, \hat{y}_i)$ is the loss function that measures the difference between the predicted value $\hat{y}_i$ and the actual value y_i. $\Omega(f_k)$ is the regularization term for the complexity of the k-th tree f_k. n is the number of training examples. K is the number of trees.

XGBoost was selected in this study due to its demonstrated success in terms of predictive accuracy across various competitions and its capability of effectively handling large datasets and complex relationships. Its regularization techniques mitigate overfitting issue, while parallel processing functions enhance speed, making it a powerful and scalable solution compared to traditional boosting methods such as gradient boosting or Adaboost. Nevertheless, careful tuning is necessary to strike a balance between complexity and performance.

4.4.4 Light Gradient-Boosting Machine (LightGBM)

LightGBM (Ke et al., 2017) is another GBDT model developed by Microsoft. LightGBM is designed to be more efficient than XGBoost by improving GBDT models in several ways. Firstly, it uses a histogram-based algorithm to convert continuous variables into discrete bins, reducing both computation time and memory usage, and helps prevent overfitting. For sampling, LightGBM employs the GOSS (Gradient-based One-Side Sampling) algorithm, which speeds up calculations by excluding less important samples, improving overall model performance. Furthermore, LightGBM uses the EFB (Exclusive Feature Bundling) algorithm, which combines mutually exclusive features into a single new feature, reducing dimensionality and thus memory usage and computational time.

Finally, for tree splitting, LightGBM grows trees leaf-wise, splitting only the leaf with the highest information gain, whereas the traditional level-wise splitting, as used in XGBoost algorithm, splits all nodes in a level. This strategy reduces unnecessary splits and achieves better optimization of the loss function with the same number of splits. Advantages of LightGBM include faster training and lower memory consumption. However, it may be less robust than XGBoost on small datasets or datasets with fewer features.

4.4.5 Categorical Boosting (CatBoost)

CatBoost (Prokhorenkova et al., 2018) is an open-source gradient boosting algorithm developed by Yandex. As the name suggested, Catboost is specifically designed to handle categorical features effectively. Nevertheless, CatBoost is known for its robust performance across various types of datasets, not just those with categorical variables. CatBoost uses "Ordered Boosting" by leveraging the natural order of feature values to enhance tree building and utilizes "Minimal Variance Sampling" to selects data points that contribute the most to prediction variance during training, reducing overfitting and improving model robustness. These techniques make CatBoost efficient in handling categorical features and improving model performance on diverse datasets. Moreover, CatBoost is user-friendly as it requires minimal intervention for data preprocessing and parameter tuning.

4.5 Hyperparameters Tuning

Model hyperparameters tuning is necessary for all the above-mentioned algorithms to optimize hyperparameters to improve model performance. Common hyperparameters tuning methods include Random Search and Grid Search, where Random Search samples hyperparameters randomly and Grid Search exhaustively searches predefined hyperparameter spaces. Both methods explore the entire range of possible hyperparameter values independently, without considering previous outcomes, thus can be inefficient. This study employed Bayesian Optimization for hyperparameters tuning. The key idea behind Bayesian Optimization is to evaluate a surrogate function, typically a Gaussian Process, instead of the true objective function, to reduce the training cost, while optimizing an acquisition function to iteratively measure which point in the predefined space to evaluate next, reducing searching cost (Snoek et al., 2012).

Common acquisition functions for Bayesian Optimization are:

- *Probability of Improvement (PI)*, which selects the next evaluation point based on the likelihood of it improving upon the current best-known result.

- *Expected Improvement (EI)*, which selects the next evaluation point by considering both the expected improvement over the current best result and the likelihood of achieving it.

- *Upper Confidence Bound (UCB)*, which dynamically selects the next evaluation point by balancing the predicted performance of the model with the uncertainty of that prediction.

Among the three acquisition functions, UCB is is employed in this study due to the fact that it is the most flexible and can be the most explorative. UCB finds a balance between exploring new areas of the hyperparameter space (exploration) and exploiting areas that are known to perform well (exploitation) based on both the predicted mean and the uncertainty of the surrogate model. Mathematically, UCB method can be expressed as

$$\mathrm{UCB}(x) = \mu(x) + \kappa\sigma(x),$$

where $\mu(x)$ is the predicted mean of the objective function at point x, $\sigma(x)$ is the predicted standard deviation (uncertainty) at point x, and κ is the hyperparameter that controls the trade-off between exploration and exploitation. High values of $\mu(x)$ indicate high expected performance, which are preferred. High values of $\sigma(x)$ indicate regions that have not been explored much, which are also preferred. κ, with a typical value of 2.576, balances between these two factors, with higher values encouraging more exploration by placing more weight on the uncertainty term $\sigma(x)$.

To increase the robustness of the tuning result, 5-fold cross-validation method is incorporated in the evaluation of the objective function. This works by splitting the data into 5 folds, using 4 folds as train set and the remaining 1 fold as validation set. Then, rotate the 5 folds to make sure each fold to be used once as the validation set. Finally, record the performance for each fold and the mean performance across all 5 folds. This mean performance represents the performance of the model with the given hyperparameters. In summary, an illustration of Bayesian Optimization can be found below in Figure 4.1:

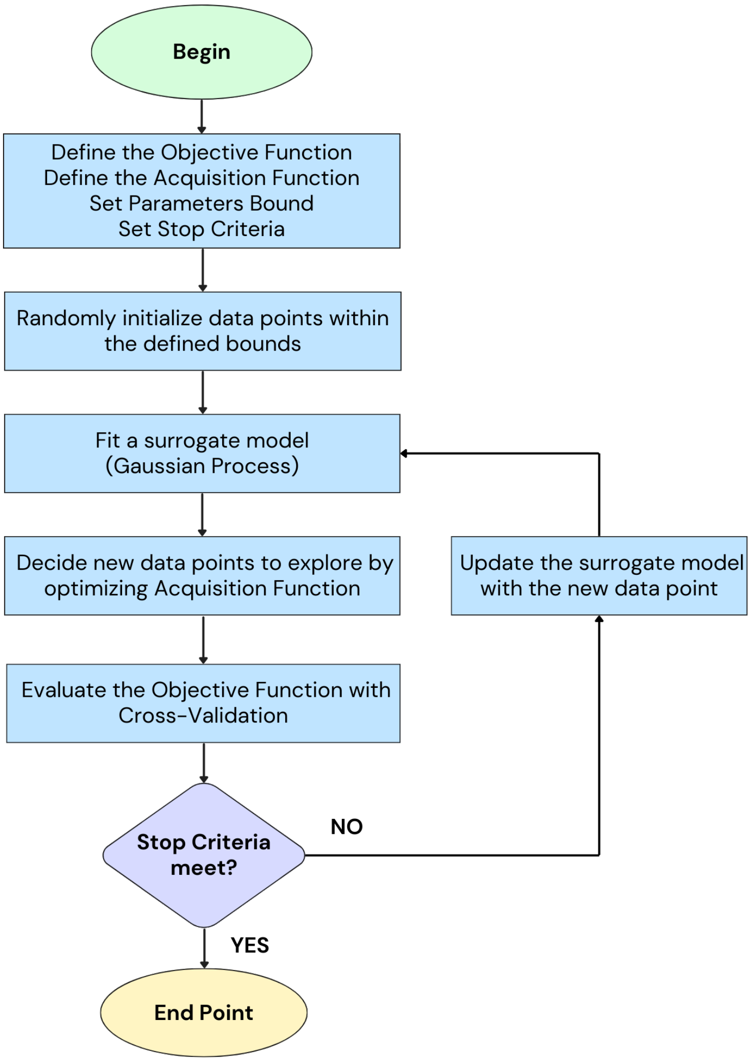

Begin
Define the Objective Function
Define the Acquisition Function
Set Parameters Bound
Set Stop Criteria
Randomly initialize data points within the defined bounds
Fit a surrogate model (Gaussian Process)
Decide new data points to explore by optimizing Acquisition Function
Update the surrogate model with the new data point
Evaluate the Objective Function with Cross-Validation
Stop Criteria meet?
NO
YES
End Point

Figure 4.1: *The work flow of Bayesian Optimization for model hyperparameters tuning*

4.6 Model Evaluation Metrics

4.6.1 Confusion Matrix

To evaluate a model's performance, the confusion matrix and its derived metrics are employed. A confusion matrix is a table that summarizes the performance of a classification model, with elements defined as follows: True Positive (TP) are instances that are actually positive and predicted as positive. True Negative (TN) are instances that are actually negative and predicted as negative. False Positive (FP) are instances that are actually negative but predicted as positive. False Negative (FN) are instances that are actually positive but predicted as negative. Important metrics derived from the confusion matrix include:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

4.6.2 Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

ROC curve and AUC are commonly used metrics to evaluate the performance of binary classifiers. The ROC curve is a graphic plot that illustrates the trade-off between the True Positive Rate (TPR, also expressed as Sensitivity) and False Positive Rate (FPR, also expressed as 1 - Specificity) of a binary classification model, while the AUC represents the area under the ROC curve, quantifying the overall performance of the model. AUC ranges from 0 to 1, where a value closes to 1 indicates that the model has excellent discrimination ability while a value close to 0.5 suggests that the model's predictions are not much better than random guessing. When dealing with imbalanced datasets, metrics such as accuracy can be misleading as they favor the majority class. AUC shows advantages over other evaluation metrics because it considers the trade-off between TPR and FPR comprehensively, making it less sensitive to imbalanced class distributions and providing a more unbiased assessment of a model's predictive performance than other metrics. In summary, AUC's straightforward interpretability and robustness to class imbalance make it the preferred metric for assessing model performance in this study.

4.7 Feature Importance Evaluation Metrics

Finally, to evaluate variable importance, SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) are employed. SHAP values provide a model-agnostic method to quantify the contribution of each feature to a model's prediction. The core idea is to calculate the average marginal contribution of each feature across all possible subsets of features, offering a consistent and comprehensive measurement of feature importance. Features with higher SHAP values are considered more important, indicating a greater impact on the model's predictions. Moreover, SHAP values come with highly interpretable visualization tools (e.g., SHAP summary plots), which help in communicating the importance and influence of features in a visually intuitive manner.

Chapter 5

Feature Engineering and Selection

This section details the feature engineering process used in this study. A transaction record involves five entities: user, seller, brand, category, and product. For each entity, there are five components to build their features: click, add-to-cart, buy, mark-as-favorite, and the timestamp (on a daily level) of these actions. Additionally, for users, information of their gender and age range are available. This study created feature profiles for each entity and combinations of entities. Since the goal of this study is to predict whether a customer will become a repeat buyer for a certain seller in the future, the User Feature Profile, Seller Feature Profile, and User-Seller Feature Profile are expected to significantly influence the results. Figure 5.1 below provides an overview of the feature profiles generated in this study.

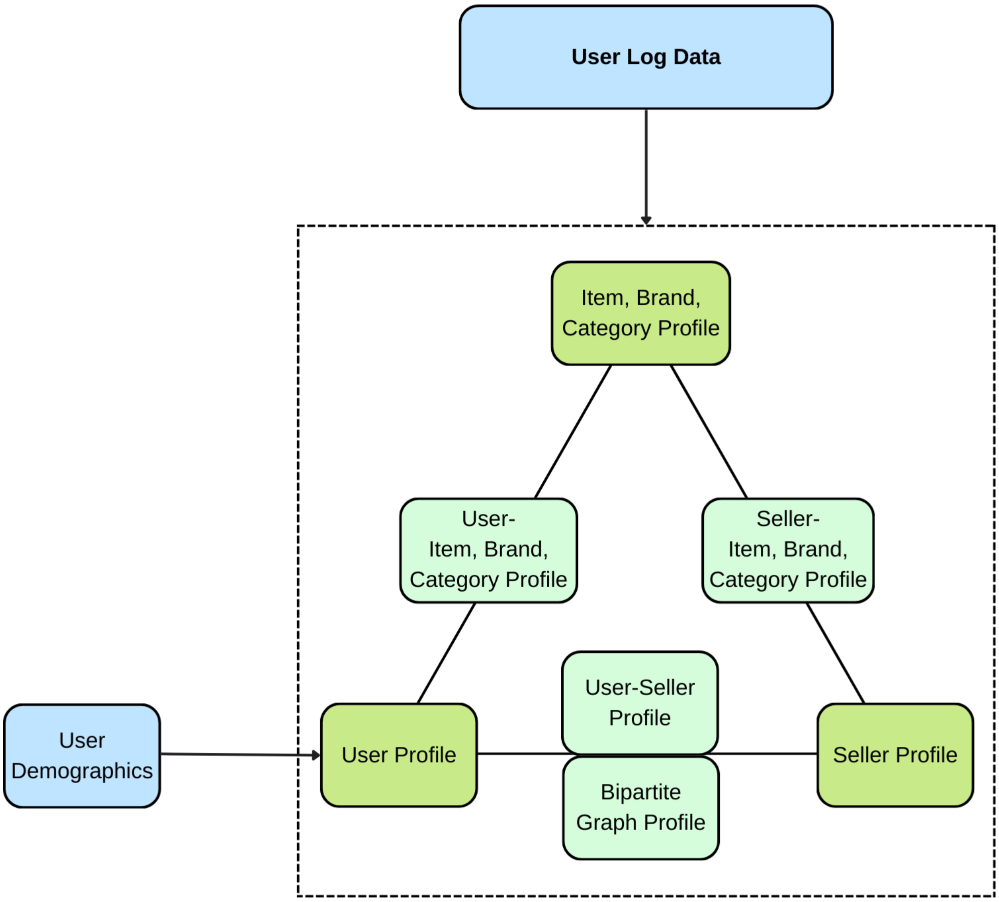

User Log Data
Item, Brand, Category Profile
User- Item, Brand, Category Profile
Seller- Item, Brand, Category Profile
User Demographics
User Profile
User-Seller Profile
Bipartite Graph Profile
Seller Profile

Figure 5.1: *An overview of feature profiles generated through feature engineering*

5.1 User Feature Profile

As the primary entity in purchasing, the buyer's behavior characteristics directly impact purchasing outcomes. This section examines user demographics and constructs features from the buyer's perspective by grouping behavior in the UserLog dataset by *user_id*.

5.1.1 User Demographics

Intuitively, user gender indicates different product preferences. For instance, female customers are generally the main buyers of cosmetics and clothing, whereas electronic stores attract more male customers. Additionally, young people spend more time shopping online compared to older adults. Studies show that age and gender influence customer satisfaction and repurchase intentions by moderating perceptions of product quality and popularity. Specifically, quality influences satisfaction more for males and those over 40, while popularity does for females. Innovativeness impacts satisfaction and repurchase intention more for consumers in their 20s (Chiu & Cho, 2021).

5.1.2 User Engagement Level: Count and Ratio of User Actions

The frequency of user actions records how often a user clicks, marks as favorite, adds to cart, or buys products over the entire data period and on a monthly level. This frequency reflects a customer's engagement in online shopping, with higher participation positively affecting repurchase intention (Kim & Hyun, 2022).

The ratio of user actions calculates the proportion of each action type in a user's total actions, over the whole data period and on a monthly level. These ratios reflect buying behavior. For example, a high click ratio combined with a low buy ratio may indicate that a user is an "information seeker" or "window shopper", with a lower willingness to spend compared to users with a high buy ratio, thus indicating a lower probability of repurchase.

5.1.3 Regular or Occasional Buyer: User Active Days

User active days measures how many days a user has action records on the shopping platform, both overall and monthly. Active days reflect shopping patterns. For example, a user who is active on the platform every month is considered a regular user, while a user who is active only a few days in a single month is considered an occasional user.

5.1.4 Consumption Diversity: User Interaction with Unique Sellers, Items, Brands, and Categories

If a user clicks, marks as favorite, adds to cart, or buys a product, it is considered an interaction with that product and its corresponding brand, category, and seller. This study calculated the number of unique sellers, items, brands, and categories a user interacted with, overall and monthly. This feature reflects consumption diversity. A wide range of products suggests diverse consumption needs and a willingness to experiment, therefore encouraging repeat purchases. Conversely, limited shop visits indicate high consumption stickiness, where users are "risk-averse" and prefer familiar shops and brands, reducing the likelihood of repeated purchases in new shops.

5.1.5 Aggregated Monthly User Features

So far, this study has measured the above user features both overall and monthly from May to November. Due to the large number of monthly features, this study used the mean, max, median, and standard deviation of these monthly features instead. Since November is the month of the "Double-Eleven" shopping festival, when users are particularly active, this study excluded data from November and considered only the six months from May to October. Specifically, the standard deviation of monthly features reflects the stability of a user's behavior outside the shopping festival period.

5.1.6 User Preference for Items, Brands, and Categories

User preference for a product, brand, or category is measured by the number of clicks, add-to-cart actions, mark-as-favorites actions, or purchases divided by the total number of these actions for all items, brands, or categories. A strong preference for purchased products increases the likelihood of future repurchases from the seller. User preferences were added to the feature profile based on the product purchased on "Double-Eleven" day from the seller, including corresponding brands and categories. For users who bought multiple items on "Double-Eleven" day, the average preference values are used.

5.1.7 User Consumption Trend

User consumption trend features were created based on the previously generated monthly features. This feature records the trend in the number of clicks and purchases per user per month. For example, the user click trend is defined as the slope of the linear regression function of the number of clicks against time periods. An example of user consumption trend using the click data of user 9 is shown in Figure 5.2. Since click and buy counts are exponentially high in November due to the shopping festival, this study excluded November data and calculate the trend feature over the six months from May to October. User consumption trends capture overall changes in users' willingness to shop and spend. An increasing trend in browsing or purchasing suggests users' growing interest in online shopping, making repeat purchases from a seller more likely.

5.1.8 User Engagement on "Double-Eleven" Days

As November data was excluded from aggregation and trend features due to bias issues, this study constructed "Double-Eleven" days features separately. User Engagement on "Double-Eleven" days records the proportion of clicks, add-to-cart actions, mark-as-favorite actions, and purchases made by each user on "Double-Eleven" day, during the week before "Double-Eleven" day, and during the month before "Double-Eleven" day, relative to the total of these actions made by each user over the entire data period. These features were created to reflect how likely a user is a deal-hunter by comparing their engagement during the shopping festival to that on normal days. If a user shows a high level of engagement on "Double-Eleven" days but a low level of general engagement, they are likely to be a one-time buyer seeking good deals.

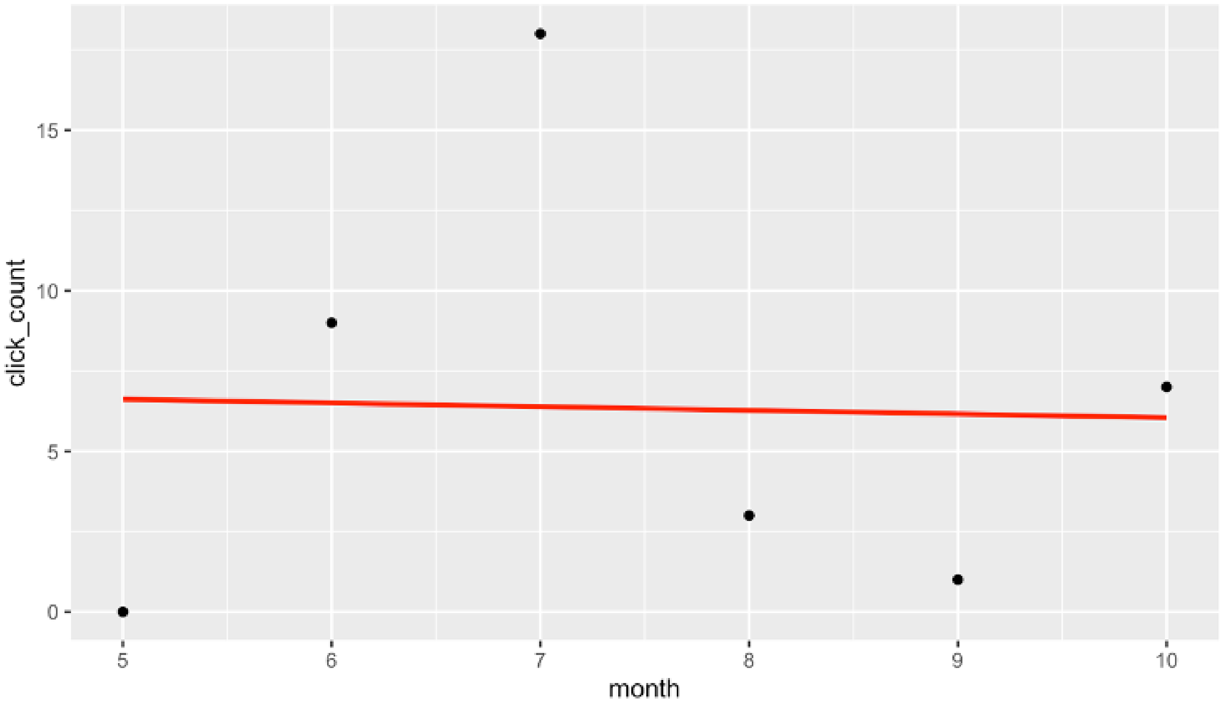


Figure 5.2: *An illustration of User Consumption Trend using user 9's click activity as example. User 9's consumption is declining, as evidenced by the variation in the amount of click between May and October.*

5.2 Seller Feature Profile

As providers of products, sellers can influence users' willingness to repurchase through aspects such as service quality and store reputation (Lowry et al., 2008; Sullivan & Kim, 2018). This section introduces how to build the seller feature profile in terms of store popularity, product diversity, and store market share by grouping the behavior in the UserLog database by *seller_id*.

5.2.1 Store Traffic: Number of Unique Users and Count and Ratio of User Actions

The traffic of a store is measured by the size of its customer base and how many times its products were clicked, added to cart, marked as favorite, or bought over the entire data period and on a monthly level. This study measures the customer base of a seller by the number of unique users that have ever interacted with this seller. The number of times a store is browsed can indicate its popularity in the market. A higher number of clicks or purchases may result from trending items sold in the store or the store's good reputation, which, in any case, can affect the probability of consumers making repeat purchases.

In addition to the count of actions, this study also considered the proportion of each of the four types of actions in the total number of actions for a certain seller over the entire data period and on a monthly level.

5.2.2 Stability of the Store Traffic: Aggregated Monthly Seller Features

Similarly, instead of adding all the monthly features to the profile, this study used the mean, max, median, and standard deviation of each monthly features across 6 months from May to October.

Particularly, the standard deviation of the monthly features can reflect the stability of a store's traffic outside the shopping festival period.

5.2.3 Product Diversity: Unique Items, Brands, and Categories in the Store

The product diversity of a seller is measured by the number of unique items it sells, as well as the number of unique brands and categories its products belong to. The product diversity of a seller can to some extent reflect its business model. A small number of products, brands, and categories may indicate a small store with a limited selection for its customers. However, product diversity can stimulate customers making repeat purchases by improving their shopping experience and value perception of the store (Maisarah & Yani, 2022).

5.2.4 Seller's Market Share

Due to the absence of price or sales value information, this study measures a seller's market share based on user count and four types of actions from three perspectives:

- *Overall market share of a seller* is measured by the traffic a seller receives compared to the total traffic. It is calculated by dividing the number of users of a seller by the total number of users, and by dividing the number of certain actions on a seller by the total number of those actions.
- *Seller's market share on brands* measures a seller's importance to a brand, based on the traffic a brand receives from that seller. It is calculated by dividing the number of users of a brand from a seller by the total number of users of that brand, and by dividing the number of certain actions on a brand from a seller by the total number of those actions on that brand.
- *Seller's market share on categories* measures a seller's importance to a category, based on the traffic a category receives from that seller. Similar to the above two market share features, it is calculated using the ratio of users and actions.

5.2.5 Aggregated Traffic and Market Share of Seller's Items, Brands, and Categories

Following the analysis of a seller's market share, this study also considered the aggregated traffic and market share of the items, brands, and categories that a seller offers. The aggregated features were calculated by first determining the traffic and market share in terms of user count and four types of actions for each item, brand, and category. Then, for each seller, this study calculated the mean, maximum, median, and standard deviation of the traffic and market share of the items, brands, and categories they offer.

Since sellers typically offer a variety of products across multiple brands and categories, each product's market performance varies, with differences shown between bestsellers and less popular items. These combined effects impact the overall store performance, making it meaningful to aggregate the market performance of different products, brands, and categories to predict customer repurchases.

5.2.6 Store Traffic Trend

The store traffic trend records the monthly number of clicks and purchases each seller receives over 6 months, from May to October. Increasing traffic over time indicates growing popularity and attractiveness of the store to customers, which contributes to the likelihood of new customers becoming repeat buyers.

5.2.7 Store Traffic on "Double-Eleven" Days

Store Traffic on "Double-Eleven" days records the proportion of clicks, add-to-cart actions, mark-as-favorite actions, and purchases received by each seller on "Double-Eleven" day, during the week before the "Double-Eleven" day, and during the month before the "Double-Eleven" day, relative to the total clicks, add-to-cart actions, mark-as-favorite actions, and purchases received by each seller over the whole data period. Store traffic on "Double-Eleven" days reflects the effectiveness of a store's promotional efforts by comparing traffic during the shopping festival to normal days. A store with significantly higher traffic on "Double-Eleven" days is likely offering substantial discounts to attract customers. However, these customers may only be interested in the limited-time deals and may not repurchase when the discounts are unavailable.

5.3 User-Seller Feature Profile

As the goal of this study is to predict whether a user will become a repeat buyer for a certain seller, features at the user-seller pair level are crucial. Similarly, this study derived features from UserLog dataset by grouping behavior by user_id and seller_id from the following perspectives.

- *User Engagement Level: Count and Ratio of User Actions for A Seller*
- *Regular or Occasional Buyer: User Active Days for A Seller*
- *Consumption Diversity: User Interaction with Unique Items, Brands, and Categories from A Seller*
- *Aggregated Monthly User Features for A Seller*
- *User Preference for A Seller*

5.3.1 Aggregated User-Seller Features for The Seller

Beyond individual user-seller pairs, aggregating features for user-seller pairs can provide valuable insights for both users and sellers. Therefore, this study added the following aggregated features to the profile:

- *Aggregated Engagement Level: Count and Ratio of Actions of All Users for A Seller*

  After calculating the user engagement level for each user-seller pair, this study focused on a single seller and calculated the mean, maximum, median, and standard deviation of user engagement levels over all users for that seller, over the entire data period.

- *Aggregated Active Days of All Users for A Seller*

  This study measured active days in five different terms: overall active days, active days with clicks, active days with add-to-cart actions, active days with mark-as-favorite actions, and active days with purchases. Then, this study calculated the mean, maximum, median, and standard deviation of user active days in these terms over all users for that seller, over the entire data period.

- *Aggregated Consumption Diversity: Interaction with Unique Items, Brands, and Categories from All Users for A Seller*

  The mean, maximum, median, and standard deviation of user consumption diversity over all users for that seller, over the entire data period.

- *Aggregated Preference of All Users for A Seller*

  The mean, maximum, median, and standard deviation of user preference over all users for that seller, over the entire data period.

- *Aggregated Engagement Level on "Double-Eleven" Days of All Users for A Seller*

  The mean, max, median, and standard deviation of user engagement level on "Double-Eleven" Days over all users for that seller, over the entire data period.

The rationale behinds aggregating user features for a seller is to summarize store performance at an average level. For example, if a seller's customers frequently visit or buy products, or have high conversion rates or preference scores, then new customers of this seller are more likely to return in the future.

5.3.2 Aggregated User-Seller Features for The User

Similarly, this study focused on a single user and aggregated their features over all sellers with whom they have interacted. The purpose is to summarize the shopping behavior at an average level for each user. For example, if a user frequently visits or buys products from online stores, or has a high conversion rate, they are more likely to return to a store they have visited before. Therefore, this study added the following aggregated features to the profile:

- *Aggregated Engagement Level: Count and Ratio of Actions of A User Over All Sellers*
- *Aggregated Active Days of A User Over All Sellers*
- *Aggregated Consumption Diversity: Interaction with Unique Items, Brands, and Categories of A User Over All Sellers*
- *Aggregated Preference of A User Over All Sellers*
- *Aggregated Store Traffic on "Double-Eleven" Days for A User Over All Sellers*

5.3.3 User-Seller Similarity Score

The similarity score between a user-seller pair is defined as the inner product of the user preference vector and the seller market share vector. Given the brand and category collection of this study, for each user, this study generated a preference vector according to their preference towards each brand and category, based on the "click" and "buy" action. Similarly, for each seller, this study generated a market share vector according to their market share on each brand and category, based on the "click" and "buy" action. For example, if there are 5 brands included in this study, and user 1's preference towards them measured by "click" action are (0.1, 0.2, 0, 0.6, 0.1), and seller 1's market share for these brands measured by "click" action are (0.3, 0.2, 0.5, 0, 0), the similarity score between user 1 and seller 1 will be $0.1 \times 0.3 + 0.2 \times 0.2 + 0 \times 0.5 + 0.6 \times 0 + 0.1 \times 0 = 0.07$

The User-Seller similarity score measures the match between a user's needs and a seller's product portfolio. A high similarity score indicates that the seller is offering what the user wants, increasing the likelihood of repeat purchases.

5.3.4 Repeat Buyers

Repeat buyers are defined as customers who have made purchases from a seller on at least two different days. Initially, repeat buyers are identified on the user-seller pair level. Subsequently, these buyers are aggregated for each user and seller. For a user, this study calculated the total number and ratio of sellers from whom the user is a repeat buyer. Similarly, for a seller, this study calculated the total number and ratio of repeat buyers. The rationale behinds these repeat buyer features is that if a buyer typically repurchases from a familiar seller, the likelihood of them becoming a repeat buyer of a seller they bought from during the "Double-Eleven" shopping festival increases. Additionally, if a seller already has a substantial group of repeat buyers, the new customers acquired on "Double-Eleven" day are more likely to become repeat buyers.

5.3.5 Item, Brand ,and Category Feature Profile

A customer's intention to repurchase from a seller depends not only on the customer's or seller's attributes but also on the features of the item bought, as well as the brand and category to which the item belongs. According to Chiu and Cho (2021), factors such as brand quality, value, innovativeness, and popularity shape customers' perceived brand leadership, positively influencing satisfaction and significantly affecting repurchase intention. Furthermore, value and popularity directly enhance repurchase intention. Customers may develop loyalty to certain brands or prefer specific product categories, leading them to favor sellers that offer those brands or categories, thus driving their decision to repurchase.

In this section, feature profiles for items, brands, and categories, as well as the pair features of brands and categories with users and sellers are developed. The following features are generated for items, brands, and categories:

- *Item, Brand, and Category Traffic: Number of Unique Users and Count and Ratio of User Actions*
- *Item, Brand, and Category Traffic Trend*
- *Item, Brand, and Category Traffic on "Double-Eleven" Days*

- *Item, Brand, and Category Market Share*
- *Repeat Buyers of an Item, Brand, and Category: Count and Ratio*
- *Aggregated Active Days with Purchases of All Users for A Brand and Category*

Due to various factors such as different store size, different marketing and promotion strategies, or different user experience and interface design, products from the same brand or the same categories can have different sales performance in different online stores. Therefore, developing features on the brand-seller and category-seller levels is meaningful to capture the heterogeneity among sellers. The following features were added to the profile:

- *Brand's and Category's Market Share within A Seller*
- *Seller-Brand Traffic: Number of Unique Users and Count and Ratio of User Actions*
- *Seller-Brand Traffic Trend*
- *Seller-Category Traffic: Number of Unique Users and Count and Ratio of User Actions*
- *Seller-Category Traffic Trend*
- *Repeat Buyers of A Seller-Brand and A Seller-Category: Count and Ratio*
- *Aggregated Active Days with Purchases of All Users for a Seller-Brand and a Seller-Category*

5.4 Bipartite Graph Features

A bipartite graph is a type of graph that can be split into two distinct sets of vertices, or nodes, with all edges connecting vertices from one set to the other set, and no edges within the same set. In this case, the structure of the log data allows each user-seller pair to be represented as a bipartite graph. One set represents users, and the other set represents sellers, with the edges between them indicating interactions such as click, add-to-cart, mark-as-favorite, or buy. An example of the underlying bipartite graph of UserLog dataset can be found below in Figure 5.3, where the upper set are user nodes, the lower set are seller nodes, and the lines in between are edges. There are a few important concepts in bipartite graph theory for analyzing the structure of the graph, as well as the roles and significance of nodes and edges within the graph.

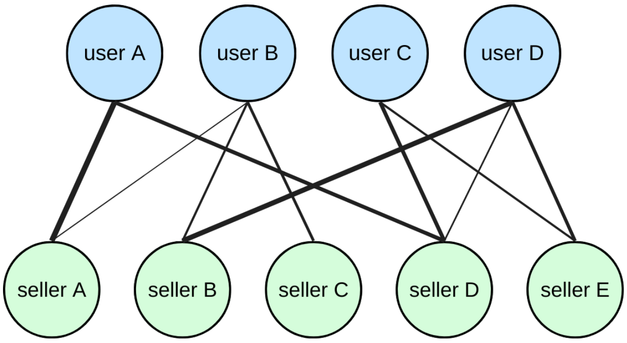


Figure 5.3: *An illustration of the underlying bipartite graph in the user log data. The blue nodes represent users, the green nodes represent sellers, and the lines connecting them, with varying thickness, represent different frequencies of interactions between buyers and sellers.*

- *The Degree of Nodes* is the number of edges connected to it. In this case, since the two sets of nodes are users and sellers, the degree of nodes measures how many sellers that a user had interaction with and vise versa. The Consumption Diversity feature in section 5.1.3 can be regarded a measurement of the degree of nodes.

- *The Weight of an Edge* represents the strength or frequency of interactions between a buyer and a seller, illustrated by the thickness of the line in between user and seller nodes, with higher weights suggesting stronger relationships. In this case, the weight of an edge can be calculated based on the frequency sum-up of the four different actions between a user-seller pair. Since different actions make up different proportions of a customer's overall behavior and reflect different likelihoods of making a final purchase, they should be weighted differently when determining the strength of the edge in the graph. The "click" action has respectively little impact on the customer's final purchase, so it will be assigned a low weight. "mark-as-favorite" and "add-to-cart" both show the strong interest of a customer buying a product, while "add-to-cart" will be weighted slightly higher than "mark-as-favorite". Finally, "buy" action should have an exponentially higher weight. Furthermore, the mean frequency of the four actions over all buyers in the dataset also exhibits exponential difference (on average 114 times of click, 13 times of mark-as-favorite, 1 times of add-to-cart, and 0.23 times of buy). Based on this logic, this study assigned the weight to the four actions based on the exponential function e^x, where x equals to 0, 1, 2, 3 when it is a "click", "mark-as-favorite", "add-to-cart", and "buy", respectively. For example, user 9 has 14 clicks and 1 buy with seller 2721, while 0 mark-as-favorite and add-to-cart. Therefore, the weight of the edge between user 9 and seller 2721 will be $14 \times e^0 + 0 \times e^1 + 0 \times e^2 + 1 \times e^3$, which is approximately 34.

- *Community Detection* is to identify groups of nodes (communities) that are more densely connected internally than with the rest of the network. A popular method for detecting communities is to optimize modularity, which measures how well a network is divided into communities. Modularity optimization is also known as Louvain method (Blondel et al., 2008). It maximizes the modularity score Q, defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

  where A_{ij} is the adjacency matrix of the graph, k_i and k_j are the degrees of nodes i and j, respectively, m is the total number of edges in the graph, and $\delta(c_i, c_j)$ is the Kronecker delta function, which is 1 if nodes i and j are in the same community, and 0 otherwise.

- *PageRank* (Page et al., 1999) is an algorithm developed by Google to rank web pages in search results. It is originally used to measure the importance of each page based on the number and quality of links to it. In the context of this study, PageRank can measure the importance of users or sellers based on their interaction history. Take seller PageRank for example, The PageRank PR of a seller i is defined as:

$$\mathrm{PR}(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{\mathrm{PR}(j)}{L(j)}$$

where d is the damping factor with a typical value of 0.85, N is the total number of sellers, $M(i)$ is the set of users who have interacted with seller i, and $L(j)$ is the number of sellers a user j has interacted with. This iterative process ranks sellers based on their potential to attract repeat buyers. Similarly, the PageRank algorithm can be used on users to rank their potential of becoming a repeat buyer.

- *Edge Betweenness* measures the importance of the connection between a buyer and a seller. It quantifies how frequently the buyer-seller relationship acts as a bridge for other interactions. The formula for edge betweenness EB(e) of an edge e is:

$$\mathrm{EB}(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths between all pairs of nodes s and t, and $\sigma_{st}(e)$ is the number of shortest paths passing through edge e. High edge betweenness suggests that the buyer-seller relationship is crucial for connecting other nodes, therefore influencing the likelihood of repeat purchases.

Based on these bipartite graph statistics, this study constructed the following features for each user, seller, and user-seller pair, taking the weight of their edge into consideration.

- *User/Seller Community* were generated as high-level features that summarizes complex behavior patterns among users and sellers. These communities often indicate groups of loyal customers or preferred sellers, which can be strong indicators of repeat buying behavior. For example, users within the same community often share similar behavior patterns or preferences. If some users in a community are repeat buyers, it's likely that others in the same community might have similar behavior and might also become repeat buyers.
- *User/Seller PageRank* can be used to identify influential users who are more likely to interact with multiple sellers and influential sellers who attract more users. A high PageRank for a seller indicates that the seller is more likely to receive repeated interactions from many users, and a high PageRank for a user indicates that the user is more likely to be a repeat buyer across various sellers.
- *User-Seller Edge Betweenness* can reveal edges which represent crucial interactions between users and sellers that facilitate connectivity and influence. High edge betweenness can indicate interactions that are important in maintaining the structure of the network. This can represent influential users, who might drive repeat purchases, or sellers, who attract loyal customers.

5.5 Feature Selection

In total 956 features were generated, with a summary of the feature profiles and feature groups to which they belong can be found in the table below in Table 5.1.

Table 5.1: *A summary of the features generated through feature engineering. A total of 956 features were created, which can be categorized into 7 distinct feature profiles and 54 feature groups.*

Profile Name	Feature Group Name	Count
User Features	User Demographics	2
	User Engagement Level: Count and Ratio of User Actions	8
	Regular or Occasional Buyer: User Active Days	1
	Consumption Diversity: User Interaction with Unique Sellers, Items, Brands, and Categories	4
	Aggregated Monthly User Features	52
	User Preference for Items, Brands, and Categories	15
	User Consumption Trend	2
	User Engagement on “Double-Eleven” Days	24
Seller Features	Store Traffic: Number of Unique Users and Counts and Ratio of User Actions	12
	Stability of the Store Traffic: Aggregated Monthly Seller Features	40
	Product Diversity: Unique Items, Brands, and Categories in the Store	4
	Seller’s Market Share	15
	Aggregated Traffic and Market Share of Seller’s Items, Brands, and Categories	120
	Store Traffic Trend	2
	Store Traffic on “Double-Eleven” Days	24
User - Seller Features	User Engagement Level: Count and Ratio of User Actions for A Seller	8
	Regular or Occasional Buyer: User Active Days for A Seller	1
	Consumption Diversity: User Interaction with Unique Items, Brands, and Categories from A Seller	3
	Aggregated Monthly User Features for A Seller	48
	User Preference for a Seller	4
	User-Seller Similarity Score	4
	User-Seller Consumption Trend	2
	Repeat Buyers	6
	Aggregated Engagement Level: Count and Ratio of Actions of a User Over All Sellers	32
	Aggregated Active Days of a User Over All Sellers	24
	Aggregated Consumption Diversity: Interaction with Unique Items, Brands, and Categories of a User Over All Sellers	48
	Aggregated Preference of a User Over All Sellers	16
	Aggregated Store Traffic on “Double-Eleven” Days for a User Over All Sellers	60
	Aggregated Engagement Level: Count and Ratio of Actions of All Users for A Seller	32
	Aggregated Active Days of All Users for A Seller	24
	Aggregated Consumption Diversity: Interaction with Unique Items, Brands, and Categories from All Users for A Seller	48
	Aggregated Preference of All Users for A Seller	16
	Aggregated Engagement Level on “Double-Eleven” Days of All Users for A Seller	60
Item, Brand, and Category Features	Item, Brand, and Category Traffic: Number of Unique Users and Count and Ratio of User Actions	30
	Item, Brand, and Category Traffic Trend	6
	Item, Brand, and Category Traffic on “Double-Eleven” Days	72
	Item, Brand, and Category Market Share	15
User - Item, Brand, and Category Features	Repeat Buyers of an Item, Brand, and Category: Count and Ratio	9
	Aggregated Active Days with Purchases of All Users for A Brand and Category	8
	Repeat Buyers of a Seller-Brand and a Seller-Category: Count and Ratio	6
	Aggregated Active Days with Purchases of All Users for a Seller-Brand and a Seller-Category	8
Seller - Item, Brand, and Category Features	Brand’s and Category’s Market Share within A Seller	10
	Seller-Brand Traffic: Number of Unique Users and Count and Ratio of User Actions	11
	Seller-Brand Traffic Trend	2
	Seller-Category Traffic: Number of Unique Users and Count and Ratio of User Actions	11
	Seller-Category Traffic Trend	2
Bipartite Graph Features	User Community	1
	Seller Community	1
	User PageRank	1
	Seller PageRank	1
	User-Seller Edge Betweenness	1

To balance the training cost and the model's performance, this study selected features that contribute the most to the model's predictive power. Firstly, features with low variance were removed. A typical Frequency Cutoff value of 95/5 was set, which indicates that if a variable has one value that appears 95% of the time and the second most frequent value appears 5% of the time, it will be considered low variance. Additionally, a Uniqueness Cutoff of 10 was set, which means a variable must have at least 10% unique values relative to the total number of samples to be considered for inclusion. Next, features that were highly correlated with each other were removed, using Pearson correlation with a threshold of 90%. Finally, features with low variable importance were removed. In this case, the Mean Decreased Gini based on a Random Forest model was calculated for all features, and features with a Mean Decreased Gini below the average (7.772) were removed.

In summary, a total of 668 features were removed due to near-zero variance (372 features) and high correlation (296 features), while 116 features with an above-average mean decreased Gini have remained in the dataset. An overview of the distribution of the final feature profiles can be found in the Figure 5.4 below.

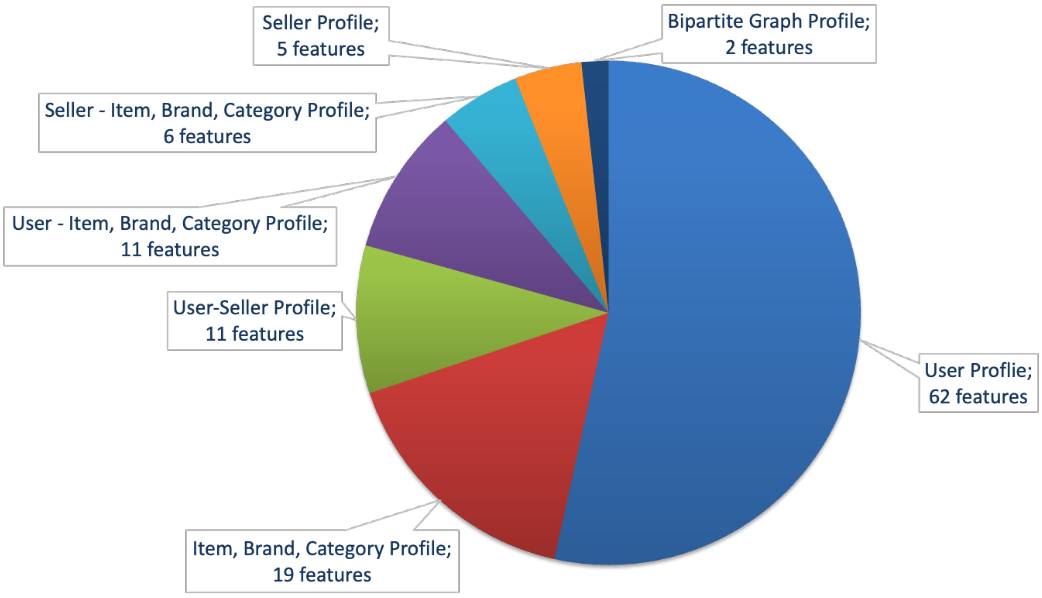


Figure 5.4: *The profile distribution of the 116 features from feature selection. Features from User Profile constitute the majority (62 features, 53.4% of the total) the selected important features.*

Chapter 6

Analyses and Results

This section firstly reports the training process and the predictive performance of the models. Then, how to leverage the Stacking method to improve prediction accuracy is discussed. Finally, variable importance is reported based on the best-performing model to discuss the impact of factors on whether new customers make repeat purchases, aiming to identify the most influential factors among the many possible features.

6.1 Create Training Set and Test Set

Firstly, all the independent variables were scaled using Z-Score method. Then, the original 26,036 samples were divided into training and test sets in a ratio of 80/20. For the training set, the SMOTE method was applied to address the class imbalance issue. Using the K-Nearest Neighbors method with a k = 5, synthetic minority samples were generated to achieve a majority/minority class ratio of 1.05/1 in the training set. In the following sections, the model will be trained on the balanced training set while the model performance will be evaluated on the original imbalanced test set. The workflow of creating training set and test set for this study can be found below in Figure 6.1.

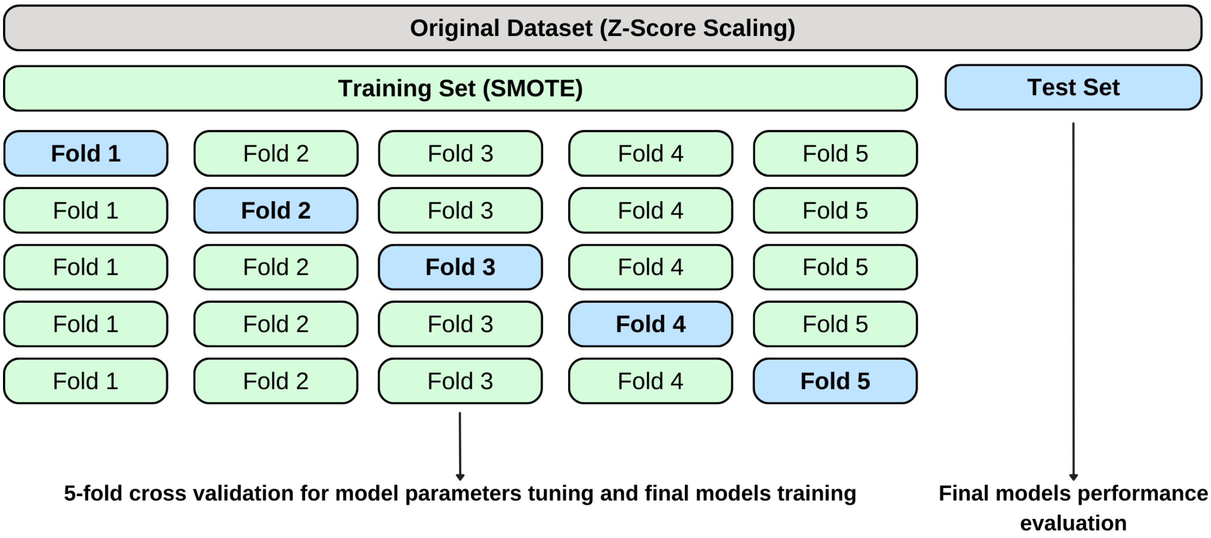

Original Dataset (Z-Score Scaling)
Training Set (SMOTE)
Test Set
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
5-fold cross validation for model parameters tuning and final models training
Final models performance evaluation

Figure 6.1: *The workflow of creating training set and test set for this study*

6.2 Model Construction and Hyperparameters Tuning

The models used in this study include Logistic Regression, Random Forest XGBoost, LightGBM, and CatBoost. For model hyperparameters tuning, a combination of Bayesian Optimization and 5-fold cross-validation was employed. Bayesian Optimization was used to generate the hyperparameter combinations to be tested, with an initial set of 10 random data points and a stopping criterion of 30 iterations, and 5-fold cross-validation was used to evaluate the performance of the models corresponding to these hyperparameter combinations. The range of hyperparameters to be tuned was determined based on the common practice from existing studies. The best numbers of boosting iterations were found using an early stopping rounds of 10. The optimal hyperparameters will be those that maximize the AUC of the model. Table 6.1 lists the hyperparameters to be tuned for all models, while Table 6.2 shows the tuning results and model AUCs on the test set.

Table 6.1: *An overview of hyperparameters to be tuned for each machine learning algorithm*

Model	Hyperparameters	Hyperparameters Definition
LR	alpha	penalty term. 1 (L1 regularization), 0 (L2 regularization), or no penalty
RF	ntree	number of trees
	mtry	number of variables randomly sampled at each split point in each tree
	nodesize	the minimum number of samples required to be at a terminal node (leaf) of a tree
XGBoost	max_depth	the maximum depth of a tree
	subsample	the fraction of samples to be used for training each tree
	colsample_bytree	the fraction of features to be used for training each tree
	min_child_weight	the minimum sum of instance weight needed in a child
	eta	the contribution of each tree to the ensemble
	nrounds	the total number of trees to be built
LightGBM	max_depth	the maximum depth of a tree
	bagging_fraction	the fraction of samples to be used for training each tree
	feature_fraction	the fraction of features to be used for training each tree
	num_leaves	the maximum number of leaves in one tree
	lambda_l2	L2 regularization term on leaf weights
	learning_rate	the contribution of each tree to the ensemble
	nrounds	the total number of trees to be built
CatBoost	depth	the maximum depth of a tree
	border_count	number of splits for numerical features
	bagging_temperature	the amount of randomness introduced into the selection of samples for each tree
	l2_leaf_reg	L2 regularization term on leaf weights
	learning_rate	the contribution of each tree to the ensemble
	iterations	the total number of trees to be built

Table 6.2: *An overview of hyperparameters tuning results for each machine learning algorithm*

Model	Hyperparameters	Tuning Space	Optimal Results	Test AUC
LR	alpha	1, 0, no penalty	no penalty	0.6250
RF	ntree	[10, 500]	500	0.6328
	mtry	[1, 50]	11	
	nodesize	[1,20]	1	
XGBoost	max_depth	[3, 50]	13	0.6521
	subsample	[0.1, 1]	0.2	
	colsample_bytree	[0.1, 1]	0.6	
	min_child_weight	[1, 10]	7	
	eta	[0.01, 0.3]	0.01	
	nrounds	[2, 3000]	531	
LightGBM	max_depth	[3, 15]	15	0.6458
	bagging_fraction	[0.5, 1]	0.5	
	feature_fraction	[0.5, 1]	0.5	
	num_leaves	[20, 150]	50	
	lambda_l2	[0, 10]	5	
	learning_rate	[0.01, 0.3]	0.2	
	nrounds	[2, 3000]	1167	
CatBoost	depth	[3, 10]	5	0.6467
	iterations	[100, 1000]	1000	
	border_count	[32, 255]	254	
	bagging_temperature	[0, 1]	0.1	
	l2_leaf_reg	[1, 10]	1	
	learning_rate	[0.01, 0.3]	0.03	
	iterations	[2, 3000]	606	

To explore the possibility of further improving the model's predictive ability, the Stacking method was employed. Stacking is an ensemble learning technique where predictions of multiple single models, known as Base Models (Level-0 Models), are used as new features to train a Meta Model (Level-1 Model), which generates the final predictions. Best practices for constructing a stacking model include choosing diverse algorithms as base models to capture a wide range of patterns and using a simpler model like Linear Regression or Logistic Regression as the meta-model to avoid overfitting. This study constructed a stacking model using Random Forest, XGBoost, LightGBM, and CatBoost as base models and Logistic Regression as the meta model. The stacking flow is illustrated in Figure 6.2.

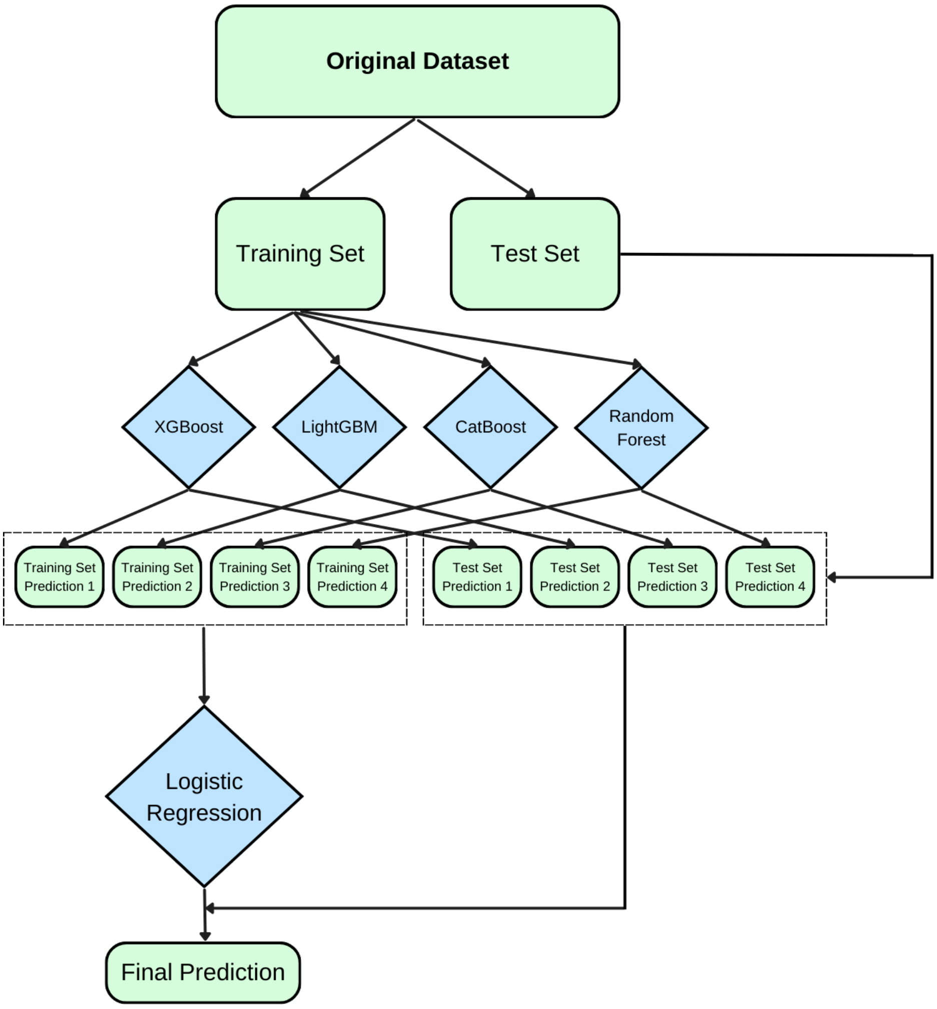


Figure 6.2: *The workflow of model stacking in this study*

Since the XGBoost model exhibited the best performance, it was considered the baseline model and was included as a base model. Considering that the RF model performed the worst among all four tree-based models, this study first stacked the RF model with XGBoost to investigate its influence. The AUC after stacking XGBoost with RF is 0.6398, which is lower than the AUC of XGBoost as an independent model. Therefore, this study proceeded without considering RF as a base model but explored all other possible model combinations in order to achieve the best AUC performance. Table 6.3 shows that having XGBoost and CatBoost as base models achieves the highest AUC of 0.6554.

Table 6.3: *The comparison of AUC values between XGBoost (the baseline model) and all stacking models. The stacking model incorporating both XGBoost and CatBoost achieves the most significant improvement, with an AUC of 0.6554, representing an increase of 0.0033 over the baseline model's AUC of 0.6521.*

Stacking Models	Test AUC	Test AUC Improvement
XGBoost (Baseline Model)	0.6521	/
XGBoost + RF	0.6398	-0.0123
XGBoost + LightGBM + CatBoost	0.6553	+0.0032
XGBoost + LightGBM	0.6526	+0.0005
XGBoost + CatBoost	**0.6554**	**+0.0033**
LightGBM + CatBoost	0.6552	+0.0031

6.3 Important Features

This section reports factors that significantly impact repeat purchases by new customers. Based on the best-performing model, XGBoost, this study firstly obtained the top 10 features with the highest SHAP values globally. Then, for each of the six feature profiles, the top 3 important features within each profile were reported, except for those already reported at a global level. This study employed 5-fold cross-validation to calculate feature importance, with the value being the mean importance of features across 5 folds. Figure 6.2 presents the top 10 features, and Tables 6.4 and 6.5 provide variable definitions of these important features at both global and profile levels.

As indicated by the SHAP values, *us_unique_categories* is the top feature, representing the diversity of product categories purchased by the user from the seller. A higher variety of categories consumed strongly indicates that the user will become a repeat buyer for this seller. Features specifically related to "Double-Eleven" days also significantly impact predicting repeat buyers. *us_de_click_ratio* represents the proportion of clicks a user made on a specific seller on "Double-Eleven" day relative to their total clicks on that seller, reflecting the significance of the user's traffic on "Double-Eleven" day to that seller. A higher standard deviation of *us_de_click_ratio* across all possible sellers for that user indicates a lower likelihood of the user becoming a repeat buyer. Conversely, a high average value of *us_de_click_ratio* across all possible sellers for that user suggests a greater likelihood of repeat purchases.

Besides aggregated store traffic on "Double-Eleven" days, item traffic (*item_deweek_like_count*) and user engagement level during the shopping festival (*user_de_buy_count*) are also strong indicators of repeat buying. Regarding the user-brand and user-seller relationship, *ub_buy_count* (the number of purchases the user made from the brand) and *us_day_count* (days the user is active at the seller) are also indicative features, with higher values indicating a higher likelihood of the user becoming a repeat buyer. *category_repeat_buyer_ratio*, the ratio of repeat buyers to all buyers for the category, is a straightforward indicator suggesting that if a category already attracts repeat buyers, a new buyer is also likely to become a repeat buyer for that category. Finally, two other aggregated user engagement level measurements, *seller_clickratio_sd* and *seller_buyratio_max*, are also identified as top features.

Table 6.5 shows the top 3 most important features within each profile, excluding those already reported in Table 6.4. Again, *us_unique_items*, ranking 11th globally, shows the user-seller consumption diversity as a strong indicator of repeat buying. Item and category traffic on "Double-Eleven" day, measured by sales volume (*item_de_buy_count*) and number of clicks (*cat_de_click_count*), also show importance at both global and profile levels. Regarding user relationships with items, brands, or categories, study results reveal that user preference towards a category (*uc_click_count*) indicates whether this user will become a repeat buyer from the seller offering this category. For seller relationships with items, brands, or categories, the traffic trend of the brand being sold by the store (*sb_buy_trend*) is a top indicator of repeat buying, followed by two market share indicators measured by the mutual importance of sellers and categories (*sc_like_share* and *cs_user_share*). Finally, in the Seller Profile, aggregated store traffic (*user_buyratio_mean* and *su_day_count_sd*) and the ratio of existing buyers for the seller (*seller_repeat_buyer_ratio*) are important predictors of repeat buyers.

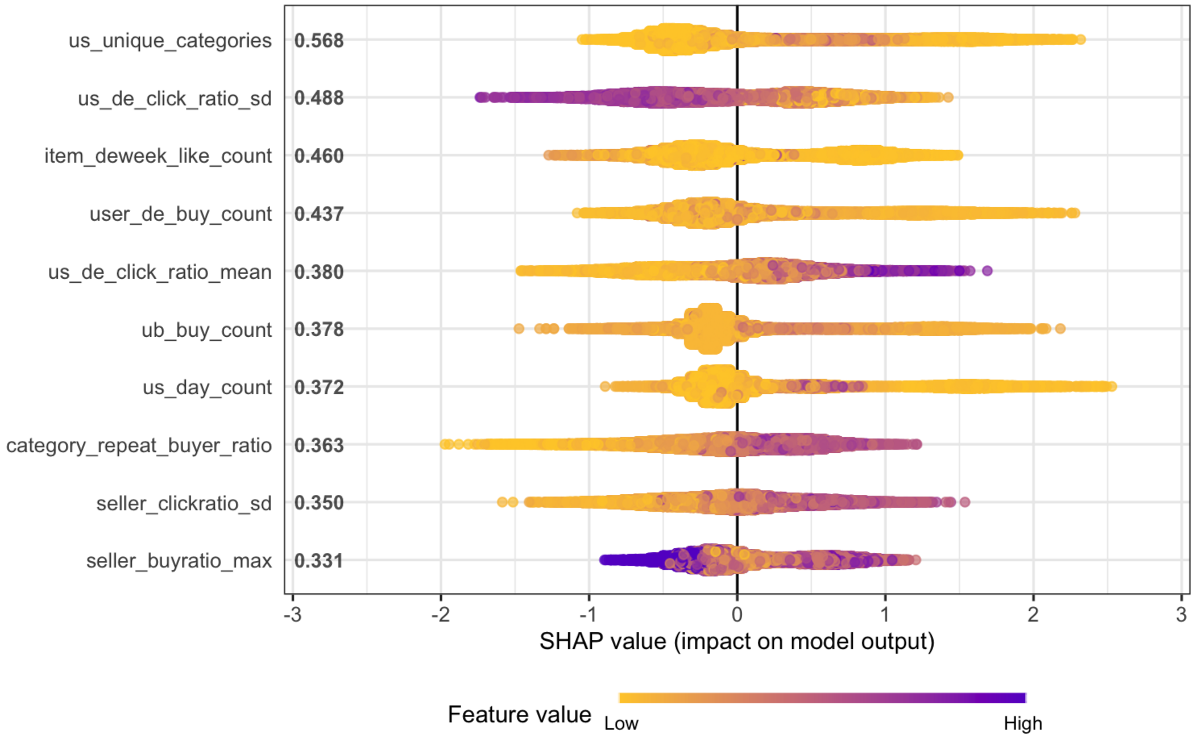


Figure 6.3: *The top 10 most important features with the highest SHAP values from the XGBoost Model. The color gradient indicates feature values, with purple representing low values and yellow representing high values. The width of each violin plot reflects the distribution of SHAP values for varying feature values. The feature "us_unique_categories" has the highest SHAP value of 0.568, indicating its strong influence on the prediction of repeat buyers.*

In summary, the results conclude that customer consumption diversity at a seller, measured by the number of product categories purchased, is the most significant predictor of repeat buying. This finding is in line with the theory of Customer Relationship Management (CRM), which is designed to foster long-term relationships between customers and companies and where cross-buying behavior, as reflected by diverse product categories consumption in this study, is a key indicator of successful CRM (Blattberg, 2001; Bolton et al., 2004). Metrics related to "Double-Eleven" days also play a crucial role, showing that customers' shopping behavior during the shopping festival differs from that in normal days and is worth researching separately. Moreover, this study shows that aggregated features, along with those from item, brand, and category profiles, are more dominant predictors than specific user-seller pair features. Therefore, when predicting whether a customer will become a repeat buyer for a particular seller, it is important to consider not only the relationship between the user and the seller but also the user's inherent shopping habits and his or her exposure to certain brands or categories, as these factors collectively influence the customer's repurchasing decision.

Table 6.4: The variable definition, feature profile, and feature group of features with top 10 SHAP values globally across all feature profiles.

Global Rank	Variable Name	Feature Profile	Feature Group	Variable Definition
1	us_unique_categories	User-Seller	User-Seller Consumption Diversity	The number of distinct product categories that the customer has purchased from the seller
2	us_de_click_ratio_sd	User	Aggregated User Engagement Level on "Double-Eleven" Days	The standard deviation of the ratio of a user's clicks on a specific seller on 1111 day to their total clicks on that seller, calculated across all sellers the user interacted with.
3	item_deweek_like_count	Item, Brand, Category	Item Traffic on "Double-Eleven" Days	The amount of likes that the product received during the week before 1111 day
4	user_de_buy_count	User	User Engagement Level on "Double-Eleven" Days	The amount of purchases that the user made on 1111 day
5	us_de_click_ratio_mean	User	Aggregated User Engagement Level on "Double-Eleven" Days	The mean of the ratio of a user's clicks on a specific seller on 1111 day to their total clicks on that seller, calculated across all sellers the user interacted with.
6	ub_buy_count	User - Item, Brand, Category	User Preference for Brand	The amount of purchases that the user made from the brand
7	us_day_count	User-Seller	User-Seller Engagement Level	Days that the user is active at the seller
8	category_repeat_buyer_ratio	User - Item, Brand, Category	Repeat Buyer	The ratio of repeat buyers to all buyers for the category
9	seller_clickratio_sd	User	Aggregated User Engagement Level	The standard deviation of the ratio of a user's clicks on a specific seller to their total actions on that seller, calculated across all sellers the user interacted with.
10	seller_buyratio_max	User	Aggregated User Engagement Level	The maximum of the ratio of a user's purchases on a specific seller to their total actions on that seller, calculated across all sellers the user interacted with.

Table 6.5: The variable definition, feature profile, and feature group of features with top 3 SHAP values locally in each feature profile.

Feature Profile	Global Rank	Variable Name	Feature Group	Variable Definition
User-Seller	11	us_unique_items	User-Seller Consumption Diversity	The number of distinct product that the user purchased from the seller
Item, Brand, Category	12	item_de_buy_count	Item Traffic on "Double-Eleven" Days	The sales volume of the product on 1111 day
	14	cat_de_click_count	Category Traffic on "Double-Eleven" Days	The amount of clicks that the category received on 1111 day
User - Item, Brand, Category	16	uc_click_count	User Preference for Categories	The amount of clicks that the user made on the category
Seller - Item, Brand, Category	23	sb_buy_trend	Seller-Brand Traffic Trend	The sales trend of the brand from the seller, from May to October
	36	sc_like_share	Seller's Market Share Within The Category	The ratio of the likes a category received from a specific seller to the total likes that category received
	41	cs_user_share	Category's Market Share Within The Seller	The ratio of customers of a specific category within a seller to the total number of customers of that seller
Seller	24	user_buyratio_mean	Aggregated Store Traffic	The average ratio of a seller's purchases from a specific user to the total actions by that seller, calculated across all users the seller interacted with
	27	su_day_count_sd	Aggregated Store Traffic	The standard deviation of user active days at a specific seller, calcualted across all users the seller interacted with
	29	seller_repeat_buyer_ratio	Repeat Buyer	The ratio of repeat buyers of a seller to the total users of that seller

Chapter 7

Conclusion and Discussion

7.1 Conclusion

With the development of e-commerce, “shopping festivals” have become increasingly common. Many new customers are typically acquired through promotional activities during these events. However, many of these new buyers, attracted by discounts, tend to be one-time deal seekers who do not return once the promotions end, which limits the store’s long-term profit growth. Consequently, identifying potential loyal customers and improving long-term ROI through effective customer relationship management with these potential loyalists has become a key concern for e-commerce businesses.

This study aims to address this concern by answering two main research questions: What important features can be extracted from user log data to identify potential repeat buyers after shopping festival promotions, and how can these repeat buyers be effectively predicted? Using real user log data collected during the 2017 Tmall “Double-Eleven” Shopping Festival and the preceding six months, this research conducted systematic feature engineering and selection. Machine learning methods, particularly ensemble learning, were applied to build models that predict which new customers acquired during the shopping festival will become repeat buyers for a specific seller after six months.

Specifically, based on the entities and their relationships presented in the user log data, this study constructed six feature profiles: User Profile, Seller Profile, User-Seller Profile, User-Item/Brand/Category Profile, Seller-Item/Brand/Category Profile, and Bipartite Graph Profile. A total of 956 features were generated, with 116 selected for the final model training. After handling the imbalanced dataset using the SMOTE method, Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost models were developed. Bayesian Optimization with 5-fold cross-validation were used to determine the optimal hyperparameters. The results showed that the XGBoost model achieved the best performance, with an AUC of 0.6521. Furthermore, efforts to enhance predictive accuracy revealed that model stacking, combining XGBoost and CatBoost, achieved the highest predictive power, with an AUC of 0.6554.

Finally, to identify the most important features in predicting repeat buyers, SHAP values were calculated for each feature using the best-performing XGBoost model. The results indicated that customer consumption diversity at a seller, measured by the number of unique product categories purchased, is the most significant predictor. Additionally, metrics related to user engagement levels, measured through various methods, particularly during the “Double-Eleven” Shopping Festival, were found to be among the most important features for predicting repeat buyers.

7.2 Managerial Implication

This study provides the following actionable management insights:

- *Leveraging Data-driven Customer Analysis.* This study demonstrates how e-commerce businesses can analyze user behavior and predict potential repeat buyers through systematic feature engineering, feature selection, and machine learning methods. Since user log data is non-declarative, large in amount, and easy to collect, e-commerce businesses should fully leverage this available information for customer relationship management while adhering to consumer data protection regulations. Analyzing user log data is more efficient and faster than traditional surveys or interviews. The feature engineering process and key features identified in this study can be directly adopted or used as a reference by e-commerce businesses that collect data at a similar scale, offering valuable insights for customer analysis.

- *Improving Customer Targeting.* The machine learning methods employed in this study allow businesses to predict the likelihood of each customer becoming a repeat buyer. These quantified results can be used as a valuable reference for customer targeting, enabling e-commerce businesses to focus primarily on customers with mid-to-high probabilities of repeat buying. Depending on the business model, e-commerce businesses can tailor product recommendations, marketing campaigns, or loyalty programs to these customer segments.

- *Optimizing Product Selection and Customer Positioning.* The important features identified by this study also provide guidance for optimizing product selection and customer positioning strategies. For instance, the top feature, *us_unique_categories*, suggests that customers who purchase from a variety of product categories are more likely to become repeat buyers. Stores can consider expanding their product offerings to encourage customers to explore and purchase across different categories. Additionally, user engagement during shopping festivals is a strong predictor of repeat buying. Businesses can segment customers based on their festival engagement and target them with personalized post-festival promotions and communications to sustain their interest. Furthermore, user-item and user-brand relationships have shown strong predictive power. Customers who engage with popular items or have a history of purchasing from specific brands are more likely to buy again. Therefore, stores should promote popular items, offer special deals for these products, and provide brand-related promotions and personalized recommendations based on previous purchases.

7.3 Limitations and Future Research

The study has the following limitations, primarily due to time and computational resource constraints:

- Due to the large size of the original dataset and limited computational resources, this study used only 5% of the original user log data for feature engineering and model training. Future research could use the full dataset to capture more comprehensive information, which may enhance model performance.

- This study employed Bayesian Optimization for hyperparameter tuning, which can introduce some randomness in the results. Future research could expand the search space and consider using Grid Search to determine optimal hyperparameters, potentially improving the model's performance.
- This study focused on ensemble learning methods. Future research could explore more diverse experimental settings, such as incorporating Neural Network models, to compare the performance of different models on this dataset in greater depth.

References

Akram, U., Hui, P., Khan, M. K., Hashim, M., Qiu, Y., & Zhang, Y. (2018). Online impulse buying on "double eleven" shopping festival: An empirical investigation of utilitarian and hedonic motivations. *Proceedings of the Eleventh International Conference on Management Science and Engineering Management 11*, 680–692.

Ataman, M. B., Van Heerde, H. J., & Mela, C. F. (2010). The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, *47*(5), 866–882.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*, 557–590.

Becerril-Arreola, R., Leng, M., & Parlar, M. (2013). Online retailers' promotional pricing, free-shipping threshold, and inventory decisions: A simulation-based analysis. *European Journal of Operational Research*, *230*(2), 272–283.

Blasco-Arcas, L., Lee, H.-H. M., Kastanakis, M. N., Alcañiz, M., & Reyes-Menendez, A. (2022). The role of consumer data in marketing: A research agenda. *Journal of business research*, *146*, 436–452.

Blattberg, R. C. (2001). Customer equity: Building and managing relationships as valuable assets. *Harvard Business School Publishing Corporation.*

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The theoretical underpinnings of customer asset management: A framework and propositions for future research. *Journal of the academy of marketing science*, *32*(3), 271–292.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. `https://doi.org/10.1023/A:1010933404324`

Bridges, E., & Florsheim, R. (2008). Hedonic and utilitarian shopping goals: The online experience. *Journal of Business research*, *61*(4), 309–314.

Bucklin, R. E., & Sismeiro, C. (2003). A model of web site browsing behavior estimated on clickstream data. *Journal of marketing research*, *40*(3), 249–267.

Bucklin, R. E., & Sismeiro, C. (2009). Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive marketing*, *23*(1), 35–48.

Cen, X., Chen, Z., Chen, H., Ding, C., Ding, B., Li, F., Lou, F., Zhu, Z., Zhang, H., & Hong, B. (2024). User repurchase behavior prediction for integrated energy supply stations based on the user profiling method. *Energy*, *286*, 129625.

Chan, T. (2014). Predictive models for determining if and when to display online lead forms. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(2), 2882–2889.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, C., & Li, X. (2020). The effect of online shopping festival promotion strategies on consumer participation intention. *Industrial Management & Data Systems*, *120*(12), 2375–2395.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chiu, W., & Cho, H. (2021). E-commerce brand: The effect of perceived brand leadership on consumers' satisfaction and repurchase intention on e-commerce websites. *Asia Pacific Journal of Marketing and Logistics*, *33*(6), 1339–1362.

Dick, A. S., & Basu, K. (1994). Customer loyalty: Toward an integrated conceptual framework. *Journal of the academy of marketing science*, *22*, 99–113.

Dong, J., Huang, T., Min, L., & Wang, W. (2022). Prediction of online consumers' repeat purchase behavior via bert-mlp model. *Journal of Electronic Research and Application*, *6*(3), 12–19.

Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2022). An analyses of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain. *Journal of Business Research*, *147*, 420–434.

Graves, A., & Graves, A. (2012). *Supervised sequence labelling*. Springer.

Heil, O. P., & Helsen, K. (2001). Toward an understanding of price wars: Their nature and how they erupt. *International Journal of Research in Marketing*, *18*(1-2), 83–98.

Hwang, S., Kim, J., Park, E., & Kwon, S. J. (2020). Who will be your next customer: A machine learning approach to customer return visits in airline services. *Journal of Business Research*, *121*, 121–126.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kim, S.-J., & Hyun, B.-H. (2022). Effects of psychological variables on the relationship between customer participation behavior and repurchase intention: Customer tolerance and relationship commitment. *Economies*, *10*(12), 305.

Kim, Y. S., & Yum, B.-J. (2011). Recommender system based on click stream data using association rule mining. *Expert Systems with Applications*, *38*(10), 13320–13327.

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, *150*, 113342.

Krafft, M., Kumar, V., Harmeling, C., Singh, S., Zhu, T., Chen, J., Duncan, T., Fortin, W., & Rosa, E. (2021). Insight is power: Understanding the terms of the consumer-firm data exchange. *Journal of Retailing*, *97*(1), 133–149.

Kumar, A., Kabra, G., Mussada, E. K., Dash, M. K., & Rana, P. S. (2019). Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention. *Neural computing and Applications*, *31*, 877–890.

Lakshminarayan, C., Kosuru, R., & Hsu, M. (2016). Modeling complex clickstream data by stochastic models: Theory and methods. *Proceedings of the 25th International Conference Companion on World Wide Web*, 879–884.

Liu, C.-J., Huang, T.-S., Ho, P.-T., Huang, J.-C., & Hsieh, C.-T. (2020). Machine learning-based e-commerce platform repurchase customer prediction model. *Plos one*, *15*(12), e0243105.

Lowry, P. B., Vance, A., Moody, G., Beckman, B., & Read, A. (2008). Explaining and predicting the impact of branding alliances and web site quality on initial consumer trust of e-commerce web sites. *Journal of Management Information Systems*, *24*(4), 199–224.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

Maisarah, I., & Yani, A. S. (2022). The effect of customer trust and product diversity on shopee users repurchase intention with customer satisfaction as a moderating variable. *IJHCM (International Journal of Human Capital Management)*, *6*(2), 32–40.

Manchanda, P., Dubé, J.-P., Goh, K. Y., & Chintagunta, P. K. (2006). The effect of banner advertising on internet purchasing. *Journal of Marketing Research*, *43*(1), 98–108.

Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, *281*(3), 588–596.

Ministry of Commerce, People's Republic of China. (2018). Consumers association of china released the "double 11" online shopping experience report in 2017. `https://www.gov.cn/xinwen/2018-02/08/content_5264957.htm`

Ministry of Commerce, People's Republic of China. (2024). The head of the ministry of commerce's e-commerce department introduces the development of china's e-commerce in 2023. `https://www.gov.cn/lianbo/fabu/202401/content_6927101.htm`

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, *13*(1-2), 29–39.

Moe, W. W., & Fader, P. S. (2004). Dynamic conversion behavior at e-commerce sites. *Management Science*, *50*(3), 326–335.

Mokryn, O., Bogina, V., & Kuflik, T. (2019). Will this session end with a purchase? inferring current purchase intent of anonymous visitors. *Electronic Commerce Research and Applications*, *34*, 100836.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing science*, *23*(4), 579–595.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* (tech. rep.). Stanford infolab.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Shapoval, K., & Setzer, T. (2018). Next-purchase prediction using projections of discounted purchasing sequences. *Business & information systems engineering*, *60*(2), 151–166.

Sim, J., Lee, J. S., & Kwon, O. (2015). Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Mathematical problems in engineering*, *2015*(1), 538613.

Sismeiro, C., & Bucklin, R. E. (2004). Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of marketing research*, *41*(3), 306–323.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

Statista. (2024, February). Global retail e-commerce sales 2014-2027 [Infographic] [Accessed on February 6, 2024]. `https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/`

Su, Q., & Chen, L. (2015). A method for discovering clusters of e-commerce interest patterns using click-stream data. *electronic commerce research and applications*, *14*(1), 1–13.

Suh, E., Lim, S., Hwang, H., & Kim, S. (2004). A prediction model for the purchase probability of anonymous customers to support real time web marketing: A case study. *Expert Systems with Applications*, *27*(2), 245–255.

Sullivan, Y. W., & Kim, D. J. (2018). Assessing the effects of consumers' product evaluations and trust on repurchase intention in e-commerce environments. *International Journal of Information Management*, *39*, 199–219.

Toth, A., Tan, L., Di Fabbrizio, G., & Datta, A. (2017). Predicting shopping behavior with mixture of rnns. *ecom@ sigir*.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of marketing*, *68*(4), 106–125.

Wicks, B. E., & Schuett, M. A. (1991). Examining the role of tourism promotion through the use of brochures. *Tourism Management*, *12*(4), 301–312.

Wu, Z., Tan, B. H., Duan, R., Liu, Y., & Mong Goh, R. S. (2015). Neural modeling of buying behaviour for e-commerce from clicking patterns. In *Proceedings of the 2015 international acm recommender systems challenge* (pp. 1–4).

Xie, J., Yoon, N., & Choo, H. J. (2023). How online shopping festival atmosphere promotes consumer participation in china. *Fashion and Textiles*, *10*(1), 5.

Yang, S., Li, L., & Zhang, J. (2018). Understanding consumers' sustainable consumption intention at china's double-11 online shopping festival: An extended theory of planned behavior model. *Sustainability*, *10*(6), 1801.

Zeng, M., Cao, H., Chen, M., & Li, Y. (2019). User behaviour modeling, recommendations, and purchase prediction during shopping festivals. *Electronic Markets*, *29*, 263–274.

Zhang, W., & Wang, M. (2021). An improved deep forest model for prediction of e-commerce consumers' repurchase behavior. *Plos one*, *16*(9), e0255906.

Zhang, Z., Zhang, Z., Wang, F., Law, R., & Li, D. (2013). Factors influencing the effectiveness of online group buying in the restaurant industry. *International Journal of Hospitality Management*, *35*, 237–245.

Zhao, B., Takasu, A., Yahyapour, R., & Fu, X. (2019). Loyal consumers or one-time deal hunters: Repeat buyer prediction for e-commerce. *2019 International Conference on Data Mining Workshops (ICDMW)*, 1080–1087.