경영학 석사 학위논문

Buying or Browsing? Identifying Consumers' Shopping Objectives and Their Buying Option Using Mobile App Clickstream Data

모바일 앱 클릭스트림 데이터를 활용한 소비자 구매행동 유형 및 구매상품 파악

2019년 8월

서울대학교 대학원
경영학과 마케팅전공
민 경 진

Buying or Browsing?
Identifying Consumers' Shopping Objectives and Their Buying Option Using Mobile App Clickstream Data

모바일 앱 클릭스트림 데이터를 활용한
소비자 구매행동 유형 및 구매상품 파악

지도 교수 송인성

이 논문을 경영학 석사 학위논문으로 제출함
2019년 8월

서울대학교 대학원
경영학과 마케팅전공
민경진

민경진의 석사학위논문을 인준함
2019년 8월

위원장 ___ 주우진 ___ (인)
부위원장 ___ 박기완 ___ (인)
위원 ___ 송인성 ___ (인)

Abstract

By using mobile App clickstream data, the research categorizes users by their shopping objectives based on the page content they visit. Visits are classified into four types: directed buying, hedonic buying, on going search, or knowledge building. When conducting cluster analysis, the research gets help from process mining tool to remove any irrelevant variables. The study compares the coefficients of each cluster's purchase count likelihood and sees how each visit type differs in choosing a product. In other words, the study aims to uncover which user is likely to buy one time purchases or auto-refilling plan.

Each type is driven by different motivations and varies in terms of marketing sensitivity. Identifying customer typology and its purchase count likelihood will allow marketers to segment its consumers and target them accordingly.

Keywords: user interest; clickstream data; cluster analysis; purchase count likelihood

Table of Contents

1 Introduction

In order to catch a buying public, companies tend to diversify their products and expand their category range. The challenge for marketers is how to uncover diverse consumers' interests and increase sales by efficiently handling an ever-increasing number of items and categories.

A growing literature suggests that in-store navigational clickstream aptly reflect users' preferences and forecast their next move. The posts often reflect the users' interests because they tend to click the items of their interest (Wang, She, Liu, & Fu 2009). Experienced users who already know what to buy tend to have more focused visits (e.g., comparing among different products within the same category), while novice users who want to know about the site are inclined to visit more informational related pages (Moe 2003). Since users own distinctive characteristics and take different attitudes toward the site, their browsing paths and visit density will all vary.

Bucklin and Sismeiro define clickstream data as the electronic record of a visitor's activity online—the data that tracks the path a user takes while surfing on the Web (Bucklin and Sismeiro 2009). The clickstream data record the navigation path of each user and are used to empirically analyze the user's shopping pattern (Montgomery 2001, Bucklin et al. 2002). When researchers inspect the data set of clickstream, they can easily reveal the user's activities in detail (e.g., product ratings, previous purchases, frequency of page visits, page duration time, etc).

As the mobile industry is the fastest-growing market segment, an increasing number of firms aim to provide effective marketing strategies for mobile usage as opposed to those of web browsing. Depend on PC or mobile, users behave differently in several ways. Unlike PC users, mobile users have a higher bounce rate, tend to multitask, and use during free time (Hart 2017, Enge 2019). Due to a smaller screen, mobile users have limited workspace; thus, mobile users abandon a page if the loading time takes longer than 3 seconds that its bounce rate comes in at 50% (Ibid). Also, they tend to multitask because users can freely move among apps (Ibid). Moreover, users can easily access and turn off the device due to its portability; this results in a higher bounce rate as well. PC users sometimes randomly surf on the web and not get distracted by the environment because they are usually sitting at a desk or often working; however, mobile users often look for a specific piece of information, get easily distracted by their environment, and mostly use consuming content (Hart 2017). These behavioral differences affect the types of sites users visit when they are on PC or mobile. As the mobile market is gradually outrunning the PC market, marketers need to know how to differentiate marketing approaches between mobile and web.

Marketers need to distinguish user characteristics between mobile and web more carefully, especially when marketers are analyzing mobile application (App) sites. People install mobile applications to increase accessibility and convenience. By installing mobile apps, users can easily connect to Internet services they commonly visit. Many mobile applications can be simultaneously used both in mobile web browsers and in mobile applications. Some companies only provide mobile app services; a user needs to have a mobile or tablet device in order to access the content. Increased in mobility and convenience, mobile application users tend to have more visits and longer page duration time compared to general mobile web users. People only spend 14% is mobile web surfing and spend 86% of the time in mobile applications (Hart 2017). Thus, marketers need to pay deeper attention to differentiating marketing strategies specifically for mobile application users.

Whereas the process mining has been widely applied in operations management and computer science, there is a limited study on combining process mining and data mining tool to establish marketing objectives. This article receives insights from process mining when removing irrelevant variables to lower dimension levels to get better performance in cluster analysis.

To find users' interests using mobile App clickstream data, this article defines a set of session descriptor variables and analyzes mobile app visit sessions based on proposed typology and see each cluster's purchase count likelihood to examine each user's buying pattern.

2 Literature Review

Clickstream data has become one of the most powerful resources for researchers trying to understand customers' hidden preferences and interests. The data record navigation path of each user and are used to empirically analyze the user's shopping pattern (Montgomery 2001, Bucklin et al.2002). The clickstream of users demonstrate a series of choices made within a web site (e.g., which pages to visit, how long to stay, and whether or not to make an online purchase) and across the website (Bucklin et al. 2002). Since it tracks a lot of customer's information in great detail, it may provide more detailed information regarding the choice process followed by consumers than the records contained in UPC scanner panel data sets (Ibid).

As such, many have analyzed the clickstream data collected from web browsers—PC websites—such as automotive websites (Bucklin and Sismeiro 2003), e-commerce site that sells nutrition products (Moe 2003), online bookstores (Moe and Fader 2004), and music consumption websites (Aguiar and Martens 2013). As opposed to a large literature on the analysis of PC websites' clickstream data, fewer papers have examined users' characteristics using mobile clickstream data.

Since clickstream data plays an important role in understanding behavior and choices made by individuals, many researchers challenge to scrutinize different parts of clickstream data even if analyzing the data has difficulties in filtering, cleaning, and processing (Bucklin et al. 2002). Clickstream data provide a wide range of information in a complete and timely manner that multiple features have been examined in numerous ways: order of pages viewed (Montgomery et al. 2004), product ratings (Zhao et al. 2013), purchase transactions (Moe and Fader 2004a, 2004b, Park and Park 2016), and frequency of page visits and page duration time (Bucklin and Sismeiro 2003, Johnson, Bellman, & Lohse 2003, Danaher, Mullarkey, & Essegaier 2006).

More specifically, Montgomery et al. (2004) examine users' behavior in an online bookstore. The authors assign each page to seven different categories (e.g., home, account, category, product, information, shopping cart, order, and enter/exit pages) and see how users transition from one page to another. The authors classify into two types of user behavior: deliberation and browsing. They discuss that some users can interchange their goals in the middle of the session. The article concludes that clickstream paths properly reflect user's goals helpful for forecasting user behavior and estimate purchase conversion.

Likely, Moe (2003) incorporates the content of the pages viewed to categorize shoppers' shopping objectives. For instance, Moe categorizes each page by content types (e.g., home, information related, search results, category level, brand level, product level pages, etc.), writes a list of user-specific measures (e.g., total number of shopping pages viewed, average time spent per page, maximum number of times any one product page was viewed, etc.) and conducts a cluster analysis on e-commerce site, which mainly sell nutrition products. According to her cluster analysis, Moe finds that users can be classified into four types of shopping strategies: directed buying, search/deliberation, hedonic browsers, and knowledge building. To facilitate customer relationships and refine marketing strategies to match customer expectations, user segmentation is crucial.

Similar to Moe (2003), Wu and Chou (2011) have developed a soft-clustering method to segment online customers based on their purchase records across categories along with the use of customers' demographic characteristics. Also, Su and Chen (2015) have established an improved leading clustering algorithm and weighed with a rough set theory to create users' interest patterns. Although clustering has been criticized for its ad hoc nature, clustering is one of the most frequently used data mining techniques for user segmentation as shown (Bunn 1993, Moe 2003, Wu and Chou 2011, Su and Chen 2015).

3 Typology of Shopping Objectives

Dimensions of Typology

Visual information search is a combination of two distinctive types of consumer behavior: goal-directed versus exploratory search behavior (Janiszewski 1998). Goal-directed customers utilize stored search routine to collect information more efficiently; their search is 'planned' (Ibid). The purpose of these customers is trying to gain specific information related to their consideration set. In other words, goal-directed customers' search is very focused, and they assess the performance of the item on selected attributes only (Moe 2003, Brucks 1985). On the other hand, exploratory visual search consumers' search is less focused and less likely to consider purchase (Moe 2003). Unlike goal-directed search, exploratory consumers often have less knowledge about the product or the category. They intend to gain more knowledge about the item or the site for future usage. Their search is motivated by 'moment-by-moment' activity that exploratory customers can always stop searching, or churn when the item does not appeal to them (Janiszewski 1998). These two different categories of customers have different approaches (e.g., the amount of knowledge about the item or the site itself) when they intend to make a purchase.

Table 1 gives a snapshot of four different shopping objectives of customers. As previously mentioned, search behavior can be classified under two large groups: directed versus exploratory. Whereas directed consumers—directed buying and on going search customers—have decent consideration set in mind, exploratory consumers—hedonic buying and knowledge building—do not have much knowledge about the item they want to purchase. Exploratory consumers tend to explore various categories to find what they want. This study further investigates customers' purchasing tendency. While directed buying and exploratory customers have high purchasing tendency, on going search and knowledge building customers have low purchasing tendency.

The next following section will discuss these four different types of shopping objectives in detail with examples and further examine how these four objectives can be applied in mobile App settings.

TABLE 1
Typology of Shopping Objectives

	Search Behavior	
Purchasing Tendency	*Goal-Directed*	*Exploratory Search*
High	Directed Buying	Hedonic Buying
Low	On Going Search	Knowledge Building

3.1 Directed Buying

Having a specific consideration set in mind, directed buying users already have thorough knowledge about the item they would like to buy. Thus, their search is very focused and deliberate (Moe 2003). Their decision-making process for purchase is focused and takes less time to make a final decision (Ibid). For example, when directed buying users visit a serialized fiction mobile App, users will directly search for their favorite writer and genre or search for the specific story in mind. They will go straight to their library and resume the session without exploring the App.

Directed buying users are usually loyal customers who return to App on a regular basis and are often familiar with the App interface. In other words, users are already accustomed to the App environment that they are unlikely to visit ‘explore pages,’ which are defined later in the paper. Thus, the study anticipates that directed buying customers would have low rates in genre and story variety.

Since directed buyers are loyal customers who regularly visit App, the study expects that they can be more sensitive to promotional content pages (e.g., ‘open survey,’ ‘open with push notification,’ ‘pressed notification permission allow,’ ‘press rate us yes,’ ‘show popup’ etc.). They are likely to turn on notifications or actively participate in App.

3.2 Hedonic Buying

Hedonic buying customers are driven by impulsive behavior (Moe 2003). Although these customers are not initially motivated to buy, they make an immediate purchase when the item appeals to them. For instance, these customers will explore some free content and buy the next episode impulsively when they like the content.

As previously mentioned, in mobile settings, users will abandon a page if it takes longer than 3 seconds to load (Hart 2017). Indeed, “context is everything" in mobile interaction (Braiterman and Savio 2007). Since users are value-driven, they look for hedonic consuming content during their leisure time. If the mobile App appeals enough to catch hedonic consumers, they are likely to make purchases.

Since hedonic users are likely to 'explore' the site, the study expects that the users will rate high in genre and story variety.

3.3 On Going Search

Similar to directed buyers, on going search customers have a list of items in mind but they do not have substantial knowledge about the product as much as those of directed buyers (Moe 2003). On going search customers explore the site to gather more information. They will continuously seek information on the site and compare various items to purchase in the future.

In mobile App settings, consumers may download similar Apps and compare them to find the aptest product. For example, a customer will download two different mobile serialized fiction applications and compare them. After the comparison, they will choose an application that gives the most value.

3.4 Knowledge Building

Knowledge-building consumers do not have any consideration set in mind and lack knowledge. They are usually novice users. They may purchase in the future but the rate is low.

The study believes that the knowledge-building segment will react quite differently in mobile App setting due to their innate user behavior differences. Unlike PC browsers, mobile users get easily distracted and churn right away if the App fails to meet their needs (Hart 2017, Wang et al. 2013). If the App fails to attract users in few seconds (e.g., the first content showing in the mobile App does not appeal or does not deliver hedonic value to users) the user is no longer going to remain in the site. Unlike PC browsers, who usually sit on a desk and browse in a bigger screen, will have higher page time duration to gain more information. On the other hand, mobile App users are likely to leave right away when the expectation is not met.

Table 2 demonstrates the expected characteristics of store visits based on different consumer types of shopping objectives. Directed buying users will focus on library pages (e.g., view collection, resumed episode, etc.) by resuming the last session. On going search users will focus on exploration and search pages since they are trying to gather information. Hedonic buyers are going to focus on explore pages since they anticipate finding hedonic consuming goods in the App. Lastly, knowledge-building users are going to focus on information related pages since they are fairly new to the site.

TABLE 2
Expected Patterns by Shopping Objective

	Focus of Session	*Marketing Sensitivity*	*Genre Variety*	*Story Variety*
Directed Buying	Library pages	High	Low	Low
On Going Search	Explore and Search pages	Moderate	Low	High
Hedonic Buying	Library and Explore pages	Low	High	High
Knowledge Building	Information pages	Low	Low	Low

4 Data and Measures

4.1 Data Summary

The context of my study is a mobile App that provides serialized fiction (journalism mobile platform). This mobile fiction platform is free to install but offers in-app purchases if a user wants to access extra content. The range of their product offerings provides various genres of fiction (e.g., romance, fantasy, new adult, paranormal, horror, and mystery/thriller). Although the target segment significantly depends on teenagers who mainly reside in Northern America, these shoppers vary dramatically in terms of their objectives, involvement levels, and expertise, which should lead to different fiction reading strategies as reflected by their page-to-page behavior at the site. The App has launched its service in 2016 and now has around 100,000 active users per month. It gains around 30,000 new customers every month; however, almost 70-80% of new customers leave the site after the first month.

The App sends its data in six different Software Development Kit (SDK): Amplitude, Adjust, Braze, Branch, Facebook, and Firebase. Out of six SDKs, I have mainly used Amplitude SDK[i], where all event logs are saved, Adjust SDK to see purchase-related events, and bookmark data saved in internal DB to see what genres and stories each user consumed.

The data is used in this study spans 5 weeks from April 1, 2019, to April 30, 2019. In this time, 91,773 unique visitors made 41,515,103 visit sessions. Of these 41,515,103 visit sessions, the study has excluded some negligible visit sessions (e.g., visit time less than a second, duplicated sessions,

[i] A comprehensive product analytics software for web and mobile

or null data) and used 29,381,785 visit sessions. For example, some negligible visit sessions include user sessions that have the same 'session_start' and 'session_end' activities (or less than a second), duplicated sessions, or null data where the 'user_id null' accounts for about more than 10,000 unknown users. Of these all valid visit sessions, 20,823 unique buyers have made purchases[ii].

4.2 Measures

The main objective of this article is to categorize shopping sessions. By assigning each session activity (e.g., 'press_,' 'session_end,' 'buy episode completed,' 'press themes,' etc.) to related measures, the study analyzes how four types of consumers have a different focus of the session.

The study has established three different categories of variables: *session* measures, *variety* measures, and *transaction* measures. Table 3 illustrates a description of the session measures in detail.

TABLE 3
Summary of Measures

Session measures	
HOME	% of pages that were home pages
INFOREL	% of pages that were information related pages
EXPLORE	% of pages that were explore related pages
SEARCH	% of pages that were search pages
LIBRARY	% of pages that were resumed pages
Variety measures	
DIFFGENR	Genre variety measure: % of genres that were unique
DIFFSTOR	Writer variety measure: % of writers that were unique

The study focuses on 5-page types: home pages (HOME),

[ii] The study accounts purchase for any 'buy episode' activity. Purchase does not necessarily mean purchasing the item through the App store, but it accounts for unlocking any stories that are 'paid' to be unlocked.

information related pages (INFOREL), explore-related pages (EXPLORE), search pages (SEARCH), and resumed pages (LIBRARY). These five variables demonstrate the focus of the store visit.

Applying Moe's measure idea, this study defines these variables as the percentage of all non-administrative pages spent on each type of page (Moe 2003). For example, if a visitor's session starts at the home page, information related page, home page, library page, and explore related page in chronological order. The study then characterizes this session as having five pages, 40% of which are home pages, 20% of which are information related pages, 20% of which are library pages, and 20% of which are explore related pages. Administrative pages are defined as pages that are not related to depicting users' shopping objectives. The example of administrative pages are 'change font face,' 'press privacy policy button,' 'press profile photo,' 'press send forgot password email button,' etc. By excluding administrative pages, the study can distinguish these visitor sessions from the shoppers of interest.

Although the above five-session measures give a general overview of the session in terms of the number of pages viewed by activity type, it does not depict the content of the page views in terms of the different genres and stories being viewed. Table 4 also shows a list of the variety of measures: DIFFGENR and DIFFSTOR. Another objective is to establish measures that reflect browsing across genres and stories. Within the given data frame, from April 1, 2019, to April 30, 2019, the number of genre counts for 10 and the total number of stories count for 4257. The higher percentage of measure represents more variety in genres and stories.

In the research, the transaction variable is excluded because it can exert a strong influence in forming clusters (Moe 2003).

5 Process Mining

5.1 Usage and Advantage

Process mining is an analytical tool for finding meaningful information by extracting recorded event logs (Celonis). Often used in inspecting entangled processes, process mining facilitates discovering process models in the event logs and helps to visualize the process. Moreover, process mining offers objective and systematic insights because it is derived from the actual event logs. In fact, process mining attempts to extract non-trivial and useful information from event logs (Aalst, Schonenberg, & Song 2011). The discovered process from process mining has proved to be useful for the continuous improvement of business processes (Ibid). In scholarly cases, specific applications embrace from Enterprise Resource Planning systems, Customer Relationship Management Systems, to Hospital Information Systems (Rojas, Munoz-Gama, Sepulveda, & Capurro 2016).

Whereas process mining has been widely applied in supply chain management and computer science, there is a limited study on combining process mining and data mining tool to establish marketing insights. By integrating two systems, marketers can recognize and understand real user behavior.

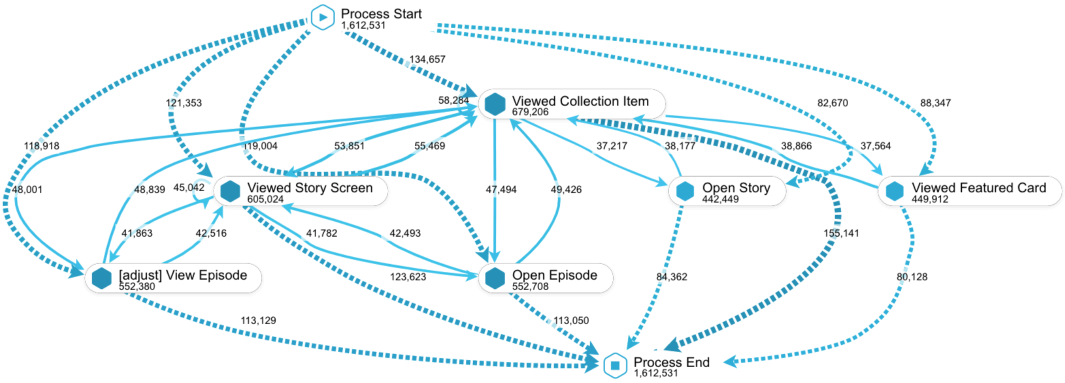
5.2 Data Application

Since clickstream data is recorded in ‘a process,’ a series of actions repeated in a progress from a defined ‘start’ to ‘end,’ the research has applied in-app clickstream data set to Celonis, an intelligent business cloud that helps to make process mining.

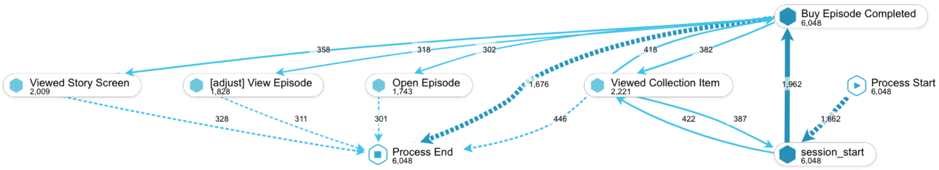
The study has first brought the App’s amplitude data into Python. Then, it has specified *N* events meaningful for business, selected only the event condition, and filtered only the rows of the condition[iii]. The study has excluded any invalid values (e.g., null data or data that does not match the type) from each row. The study has uploaded an hour data to verify the value and make sure the process is well formed[iv]; then, it has uploaded a weekly data (April 1, 2019, to April 7, 2019) to Celonis.

[iii] I first applied all 275 events; due to too many insignificant connections (lines), it was hard to see the neat, objective process map.

[iv] The data were too large to upload and it took too much time mapping in Celonis.



[Figure 1-1] Process Explorer in Celonis Process Mining
(40.7% of activities, 18.2% of connections)



[Figure 1-2] Focus on 'buy episode completed' activity

As shown in [Figure 1-1] and [Figure 1-2], process mining tool can easily adjust the process map based on the focus the researcher designates to see.

5.3 Data Application

The study has applied process mining due to three main reasons: visualizing the App's clickstream data, categorizing activity into measures, and removing any irreverent variables for clustering analysis.

The research wants to visualize and get an intuitive idea of how general users behave. Since it is hard to get a glimpse of the idea of how consumers use the App, process mining has facilitated mapping complex log data.

Second, the study aims to categorize event log data. In the previous section, it has explained how categorizing each event into a certain measure to demonstrate the focus of the store visit. There are some vague events, however. For example, the study has struggled to assign 'viewed day' as an explore page or library page. 'Viewed day' event log usually comes more from 'session_start' than 'viewed collection item.' Looking at the process mining map, the study was able to categorize the 'viewed day' event as an explore page.

Lastly, process mining helps to remove any irrelevant variables and improve performance when forming a cluster analysis. The research had too many variables when it first formed a cluster analysis. Moreover, the study needed to remove any possible outliers since it used K-means clustering. Through the process mining map, the study was able to remove variables that had low case affects and activity cases and was able to decrease dimensions of the cluster.

At first, cluster analysis has included 11 variables (PAGES, HOME, INFOREL, EXPLORE, SEARCH, LIBRARY, STORY, MKT, DIFFGENR, DIFFWRT, DIFFSTOR, and PURCH). However, after using the process mining tool, the research has removed some irrelevant variables and assigned log events to relevant variables.

The App mostly consists of romance stories. Although there are many romance writers, the audience mainly looks for one popular romance writer. In fact, the gap between number 1 and number 2 differ greatly. Due to this fact, the DIFFGENRE and DIFFWRT did not have much impact on differentiating among clusters. Also, the STORY variable did not have much differentiation between LIBRARY; thus, the variable was combined with LIBRARY.

6 Data Analysis

6.1 Clustering

The study classifies users' mobile App sessions to establish four different shopping objectives: directed buying, hedonic buying, on going search, and knowledge building.

K-means cluster is used in the analysis. As seen in [Figure 4] in the Appendix, the optimal number of cluster is 4 to 5 clusters. The study has formed two different cluster analyses and compared both; the optimal solution is the one with 4 clusters.

Cluster 1, directed buying sessions, mainly focuses on library pages—percent of pages that were resumed pages. These users return to App on a regular basis and they are often familiar with the App environment. As expected, returning users go straight to their collection to resume the fiction they have not finished.

Cluster 2, on going search sessions, mostly consists of 'explore pages' and 'search pages.' The users in these sessions spend much time searching and exploring the App. Consumers could have downloaded two different mobile serialized fiction applications and compare them to find the aptest product.

Cluster 3, Hedonic buyers' sessions are fairly dispersed compared to those of other users and have the most number of users in the cluster (*N* = 44,937). Having the most number of clusters can suggest that the App has high churn rates. Hedonic buyers like to explore and tend to visit library pages after they make purchases.

Cluster 4, Knowledge-building consumers react quite differently compared to PC browsers. While the rest of the values are all non-existent, 0%, it rates very high in information related pages, 73.7%. When the study further investigates the reason by looking at process mining analysis, it has found that only 23% read stories or explore other parts of the App after 'view first onboarding guidance'. If the App fails to attract users in a few seconds—the first content showing in the mobile App—a user is no longer going to remain in the site. Unlike PC site users who have higher tolerance in page time duration, mobile App users are likely to leave right away when their expectation is not met.

As mentioned previously, variety measures, DIFFGENR and DIFFSTOR were not significant and thus excluded. 4 Types of users have similar DIFFGENR and DIFFSTOR values. Cluster 1 account for 13.94%, 0.09%, Cluster 2 account for 14.45%, 10.9%, Cluster 3 account for 15.44%, 0.11%, and Cluster 4 account for 10.9%, 0.03%, DIFFGENR and DIFFSTOR in order.

Below table 4 shows four cluster solution results with session focus measures: HOME, INFOREL, EXPLORE, SEARCH, and LIBRARY.

TABLE 4
Four Cluster Solution

Cluster	1 *Directed Buying*	2 *On Going Search*	3 *Hedonic Buying*	4 *Knowledge Building*
N	29,618	14,414	44,937	2,804
Maxdist	165.4	102.57	258.49	258.49
Session focus measures				
HOME	6.3%	30.8%	18.1%	0.2%
INFOREL	3.2%	0.0%	1.9%	73.7%
EXPLORE	6.7%	61.5%	14.2%	0.1%
SEARCH	0.0%	46.2%	2.2%	0.0%
LIBRARY	62.8%	7.7%	17.5%	0.0%

6.2 Distribution of Purchase Options

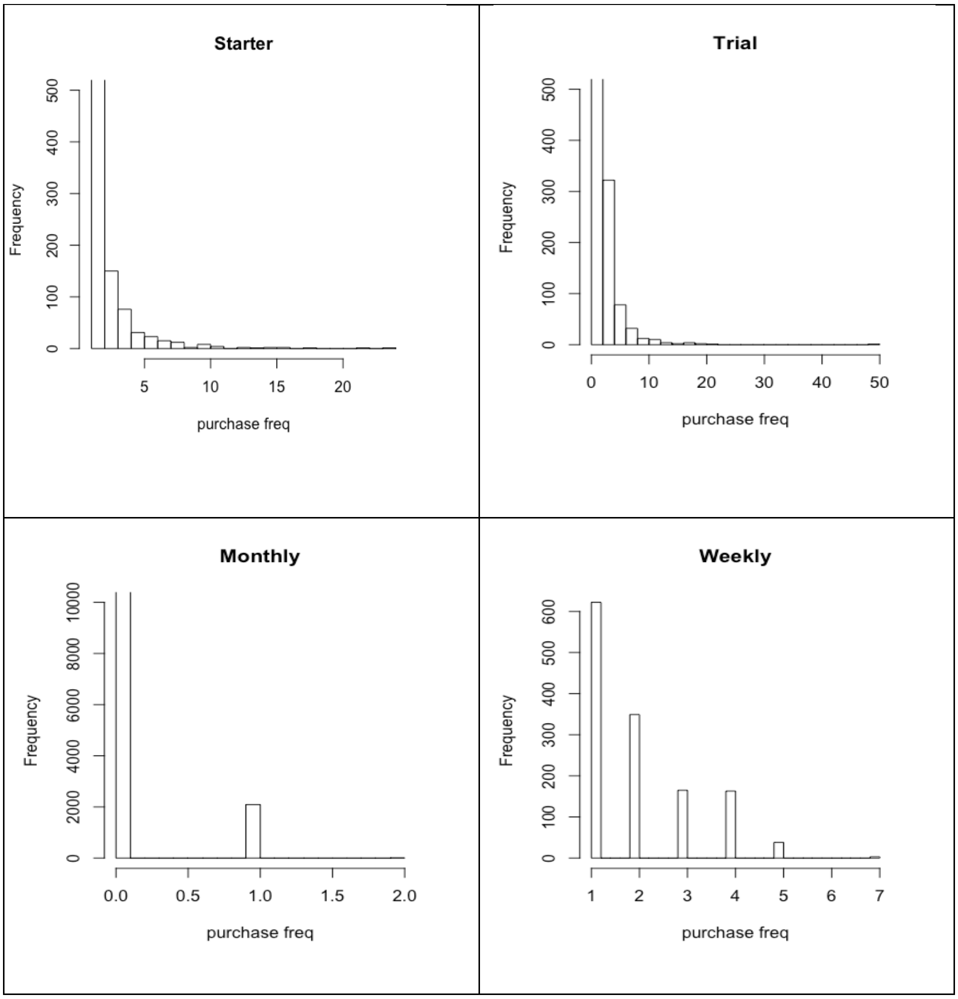


TABLE 5
Purchase Frequency

Trial	0	1	2	3	4	5	6+
Starter	77834	2094	766	237	85	48	98
Weekly	77264	3057	510	150	76	31	74
Monthly	79061	2101					

For ‘trial,’ ‘starter,’ and ‘weekly’ purchasing frequency data, the study uses Poisson distribution since these data are skewed to the left as seen in the histogram above. For ‘monthly’ subscription data, the study has adopted binary logit distribution since it is either 0 or 1. As seen in the table and the graph above, many people purchase ‘starter’ and ‘trial’ products; however, the amount of number does not accord to ‘monthly’ and ‘weekly’ products. In short, there are many hedonic users who enjoy the product once and do not make consistent purchases.

6.3 Descriptive Statistics

The first part refers to the whole number of the episode that has been viewed by all users. The second part refers to the average episode that was viewed per genre. Lastly, the third part refers to variance; the variance is large as it is shown above. That is, the number of episode deviations of the genre seen by each person is quite large and vary significantly.

The analysis is followed by two steps. When figuring out zero-inflation, its dependent variable (y_i) is either 0 or 1 (binary). For example, if the user has ever made a ‘trial’ purchase before, y_i equals to 1. If the user has ever made ‘starter’ purchase, but not ‘trial,’ y_i equals to 0. That is, all of the equations are calculated separately from each other. Its independent variables are cluster 0, cluster 1, cluster 2, cluster 3, romance, late-night, and fantasy. The romance variable refers to the number of episodes viewed by each person. Similarly, late-night and fantasy also refer to the number of episodes viewed in that specific genre. Secondly, the study builds a count model. If the user has bought three trials, $y_i = 3$. The independent variable equals the first step: cluster 0, cluster 1, cluster 2, cluster 3, romance, late-night, and fantasy.

6.4 Purchase Count Likelihood

In order to access paid content, the App offers two kinds of payment options: a coin package or auto-refilling plan. The coin package is a one-time purchase. Under coin package, a user can buy either 'Trial,' which gives 6 coins for $0.99, or 'Starter,' which gives 27 coins for $3.99. The auto-refilling plan is recurring billing that can be canceled at any time. The auto-refilling plan refills the account with the designated number of coins every week and month. So-called subscription method, auto-refilled coins have a validity period based on the plan the user chooses. Coins are valid for a week for 'weekly' plan, valid for a month for 'monthly' plan. Under the auto-refilling plan, a user has two options: weekly (35 coins per week for $3.99) and monthly (200 coins per month for $10.99). The study aims to discover which cluster is more likely to buy a coin package or likely to subscribe to the plan.

The App consists of many free users and the data is skewed to very left. Thus, the study estimated each cluster's purchase count likelihood by using zero-inflated negative binomial regression. The regression coefficients are estimated by using the method of maximum likelihood. Unlike Poisson distribution, the negative binomial is more flexible to capture more fitted distribution. Negative binomial's variance and mean do not have to match, and it is widely used for counting frequencies. Moreover, the study uses negative binomial distribution because the observed purchase data is over dispersed. In fact, the App consists of many free users who have never made a purchase.

A zero-inflated model follows two different processes. The first process is figuring out if a user has bought coins versus not bought coins during the given data period. This first part of a zero-inflated model is a binary logit model. If not bought coins, the only outcome possible is zero. If bought coins, it is then a count process. In this study, the negative binomial model is used to model the count process. The expected count is expressed as a combination of the two processes.

The zero-inflated negative binomial model's likelihood function depends on the observed value whether it is a zero or greater than zero. From the logistic model of $y_i > 1$ versus $y = 0$.

$$p = \frac{1}{1 + e^{-x_i'\beta}}$$

and

$$1 - p = \frac{1}{1 + e^{x_i'\beta}}$$

The likelihood function is written as followed.

$$\mathcal{L} =$$

$$\begin{cases} \sum_{i=1}^{n}\left[\ln(p_i) + (1-p_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\right] & \text{if } y_i = 0 \\ \sum_{i=1}^{n}\left[\ln(p_i) + \ln\Gamma(\frac{1}{\alpha}+y_i) - \ln\Gamma\,(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) + \left(\frac{1}{\alpha}\right)\ln\left(\frac{1}{1+\alpha\mu_i}\right) + y_i\ln\left(1-\frac{1}{1+\alpha\mu_i}\right)\right] & \text{if } y_i > 0 \end{cases}$$

$$where\ \mu_i = \exp\,(X_i\gamma)$$

In other words, y_i is greater than 0 if a user has ever made a purchase and is 0 otherwise. By using R, the study finds the purchase count likelihood of the users[v].

6.5 Analysis of Each Cluster's Purchase Count Likelihood

A zero-inflated model follows two different processes. Zero-inflation model coefficients demonstrate the purchase count likelihood of whether each cluster actually buys the coin or not. The count model coefficients depict the number of times the event could have happened.

In the count model, the study implies that the users are likely to make more purchases (as signified by the positive coefficient in romance independent variable) when they read romance episodes. People who read many romance episodes tend pay for weekly or monthly subscription than one-time coin package. That is, once the user starts to read romance episodes,

[v] The study has removed an intercept in R due to multicollinearity. It was better to remove intercept than to remove a cluster dummy; otherwise, the study had to compare a removed cluster with three other clusters.

the user is likely to unlock another episode (making a purchase) to read more episodes.

Unlike romance episodes, when a user sees more fantasy episode, the user is less likely to make a purchase. No matter how many fantasies the user sees, there is no effect on the probability of purchasing another episode. In other words, fantasy is not as effective as romance when the App tries to attract more loyal consumers. It can also suggest that fantasy episodes in the App are not as interesting as the romance episodes. Since it does not help the churn rate, it is recommended to increase more romance episodes. Moreover, customers whom only make one-time purchases are likely to read action and late-night novels.

Hedonic buyers are likely to purchase coin packages more than those of search users. On the other hand, directed buyers are likely to purchase an auto-refilling plan than hedonic buyers. Indeed, hedonic buyers tend to make an impulsive purchase that they are likely to make a one-time purchase. Whereas, directed buyers often have clear goals that they are likely to make an auto-refilling purchase. Hedonic users account for large proportions in the App. Thus, the study recommends to offer free romance episodes specifically to hedonic users (e.g., extra free episodes or expose more romance stories in the home page area when hedonic users are visiting) when they are first exploring the site. This can lower the churn rate.

7 Conclusion

Although process mining has been widely adopted, there is a limited study on combining process mining and data mining tool to establish marketing objectives. When conducting cluster analysis, the research uses process mining to remove any irrelevant variables to improve performance.

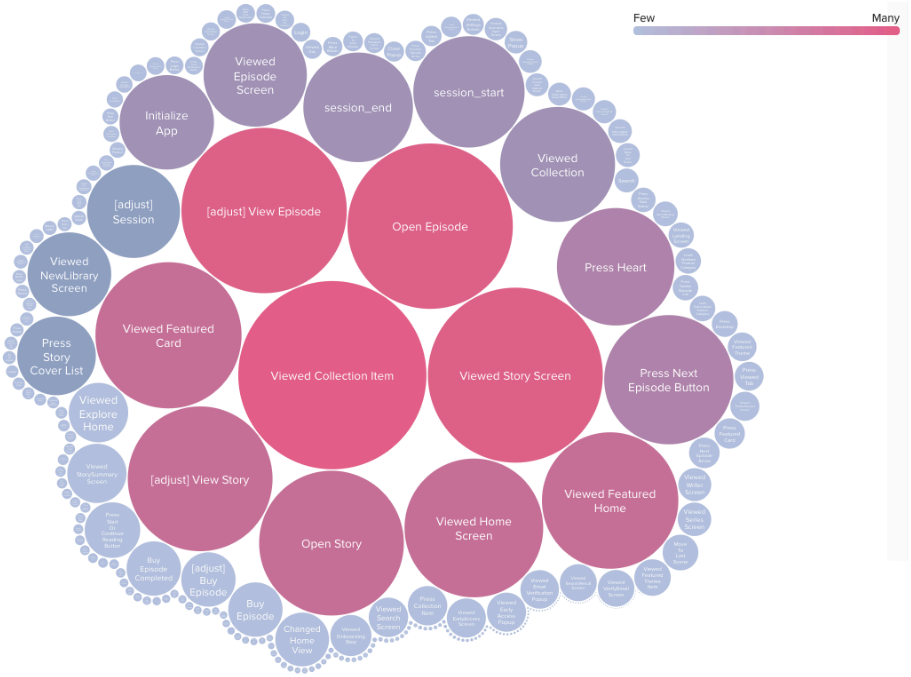
By using mobile App clickstream data, the research categorizes users by their shopping sessions based on the page content they visit. Visits are classified into four types: directed buying, hedonic buying, on going search, or knowledge building.

By using zero-inflated negative binomial regression, the study compares the purchase count likelihood of each cluster and see how likely and how many times each cluster is likely to make a purchase. In fact, hedonic buyers and on going search users are likely to purchase a coin package. On the other hand, directed buyers heavily rely on auto-refilling plans. Coin package users usually depend on search and explore content pages, whereas auto-refilling plan users often visit library pages. Unlike other purchases, knowledge buyers in mobile App settings barely make any purchases.

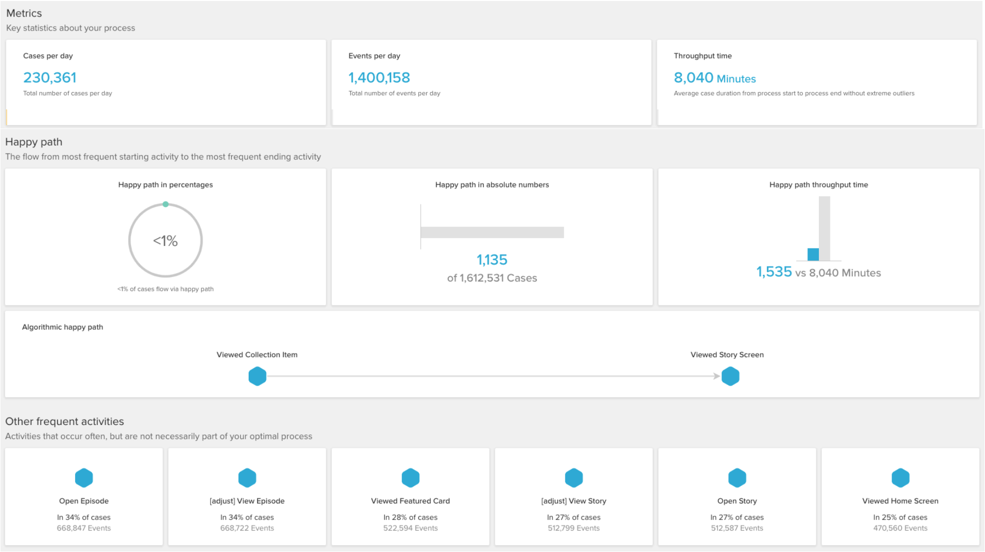
Each type is driven by different motivations and varies in terms of marketing sensitivity. Identifying customer typology and its purchase count likelihood will allow marketers to segment its consumers and target them accordingly.

Appendix

[Figure 2] Celonis : Process Overview – Activities by frequency



[Figure 3] Celonis: Process overview – Happy path (the flow from most frequent starting activity to the most frequent ending activity)



Metrics

Key statistics about your process

Cases per day

230,361

Total number of cases per day

Events per day

1,400,158

Total number of events per day

Throughput time

8,040 Minutes

Average case duration from process start to process end without extreme outliers

Happy path

The flow from most frequent starting activity to the most frequent ending activity

Happy path in percentages

<1%

<1% of cases flow via happy path

Happy path in absolute numbers

1,135

of 1,612,531 Cases

Happy path throughput time

1,535 vs 8,040 Minutes

Algorithmic happy path

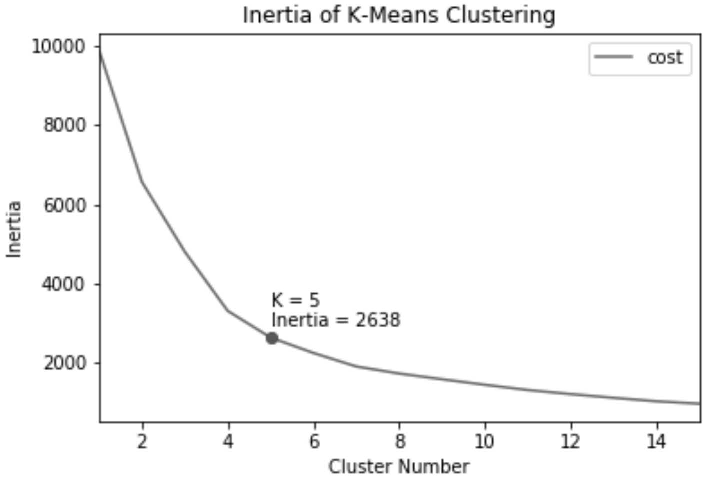
Viewed Collection Item → Viewed Story Screen

Other frequent activities

Activities that occur often, but are not necessarily part of your optimal process

Activity	Cases	Events
Open Episode	In 34% of cases	668,847 Events
[adjust] View Episode	In 34% of cases	668,722 Events
Viewed Featured Card	In 28% of cases	522,594 Events
[adjust] View Story	In 27% of cases	512,799 Events
Open Story	In 27% of cases	512,587 Events
Viewed Home Screen	In 25% of cases	470,560 Events

[Figure 4] Inertia of K-Means Clustering



References

Aalst, W.m.p. Van Der, et al. “Time Prediction Based on Process Mining.” Information Systems, vol. 36, no. 2, 2011, pp. 450–475.

Bloch, Peter H., et al. “Consumer Search: An Extended Framework.” *Journal of Consumer Research*, vol. 13, no. 1, 1986, p. 119.

Braiterman, J., & Savio, N. (2007). Design sketch: the context of mobile interaction. International Journal of Mobile Marketing, 2(1), 66-68.

Brucks, Merrie. “The Effects of Product Class Knowledge on Information Search Behavior.” *Journal of Consumer Research*, vol. 12, no. 1, 1985, p. 1.

Bucklin, Randolph E., et al. “Choice and the Internet: From Clickstream to Research Stream.” *Marketing Letters*, vol. 13, no. 3, 2002, pp. 245–258.

Bucklin, Randolph E., and Catarina Sismeiro. “A Model of Web Site Browsing Behavior Estimated on Clickstream Data.” *Journal of Marketing Research*, vol. 40, no. 3, 2003, pp. 249–267.

Bucklin, Randolph E., and Catarina Sismeiro. “Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing.” *Journal of Interactive Marketing*, vol. 23, no. 1, 2009, pp. 35–48.

Bunn, Michele D. “Taxonomy of Buying Decision Approaches.” *Journal of Marketing*, vol. 57, no. 1, 1993, pp. 38–56.

Danaher, Peter J., et al. “Factors Affecting Web Site Visit Duration: A Cross-Domain Analysis.” *Journal of Marketing Research*, vol. 43, no. 2, 2006, pp. 182–194.

Enge, Eric. “Mobile vs Desktop Usage in 2018: Mobile Widens the Gap.” *Stone Temple*, 11 Apr. 2019, www.stonetemple.com/mobile-vs-desktop-usage-study/.

Hart, Sandee. “Mobile vs. Desktop: 10 Key Differences.” *ParadoxLabs*, 2 Nov. 2017, www.paradoxlabs.com/blog/mobile-vs-desktop-10-key-differences/.

Janiszewski, Chris. “The Influence of Display Characteristics on Visual Exploratory Search Behavior.” *Journal of Consumer Research*, vol. 25, no. 3, 1998, pp. 290–301.

Johnson, Eric J., et al. “Cognitive Lock-In and the Power Law of Practice.” *Journal of Marketing*, vol. 67, no. 2, 2003, pp. 62–75.

Moe, Wendy W. “Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream.” *Journal of Consumer Psychology*, vol. 13, no. 1-2, 2003, pp. 29–39.

Moe, Wendy W., and Peter S. Fader. “Dynamic Conversion Behavior at E-Commerce Sites.” *Management Science*, vol. 50, no. 3, 2004, pp. 326–335.

Moe, Wendy W., and Peter S. Fader. “Capturing Evolving Visit Behavior in Clickstream Data.” *Journal of Interactive Marketing*, vol. 18, no. 1, 2004, pp. 5–19.

Montgomery, Alan L. “Applying Quantitative Marketing Techniques to the Internet.” *Interfaces*, vol. 31, no. 2, 2001, pp. 90–108.

Montgomery, Alan L., et al. “Modeling Online Browsing and Path Analysis Using Clickstream Data.” *Marketing Science*, vol. 23, no. 4, 2004, pp. 579–595.

Park, Chang Hee, and Young-Hoon Park. “Investigating Purchase Conversion by Uncovering Online Visit Patterns.” *SSRN Electronic Journal*, 2013.

Rojas, Eric, et al. “Process Mining in Healthcare: A Literature Review.” Journal of Biomedical Informatics, vol. 61, 2016, pp. 224–236.

Seif, George. “The 5 Clustering Algorithms Data Scientists Need to Know.” *Towards Data Science*, Towards Data Science, 5 Feb. 2018, towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

Sismeiro, Catarina, and Randolph E. Bucklin. “Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach.” *Journal of Marketing Research*, vol. 41, no. 3, 2004, pp. 306–323.

Sismeiro, Catarina, and Randolph E. Bucklin. “Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach.” *Journal of Marketing Research*, vol. 41, no. 3, 2004, pp. 306–323.

Su, Qiang, and Lu Chen. “A Method for Discovering Clusters of e-Commerce Interest Patterns Using Click-Stream Data.” *Electronic Commerce Research and Applications*, vol. 14, no. 1, 2015, pp. 1–13.

Wang, Hsiu-Yu, et al. “What Affects Mobile Application Use? The Roles of Consumption Values.” *International Journal of Marketing Studies*, vol. 5, no. 2, 2013.

Wang, Shuqing, et al. “Algorithm Research on User Interests Extracting via Web Log Data.” *2009 International Conference on Web Information Systems and Mining*, 2009.

“What Is Process Mining?” Celonis, www.celonis.com/process-mining/what-is-process-mining/?gclid=EAIaIQobChMIxMGzyIXd4gIVWKWWCh21ng8CEAAYASAAEgI1DfD_BwE.

Wicht, Luis Aguiar, and Bertin Martens. “Digital Music Consumption on the Internet: Evidence from Clickstream Data.” *SSRN Electronic Journal*, 2013.

Wu, Roung-Shiunn, and Po-Hsuan Chou. “Customer Segmentation of Multiple Category Data in e-Commerce Using a Soft-Clustering Approach.” *Electronic Commerce Research and Applications*, vol. 10, no. 3, 2011, pp. 331–341.

Zhao, Xiangyu, et al. “Interest before Liking: Two-Step Recommendation Approaches.” *Knowledge-Based Systems*, vol. 48, 2013, pp. 46–56.

Zheng, Ling, et al. “User Interest Modeling Based on Browsing Behavior.” *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, 2010.

국문초록

Buying or Browsing? Identifying Consumers' Shopping Objectives and Their Buying Option Using Mobile App Clickstream Data

민경진

경영학과 마케팅 전공

서울대학교

본 연구에서는 프로세스 마이닝 기법을 활용하여 모바일 앱 사용자들의 구매행동 패턴을 분석하는 연구를 수행하고, 군집분석과 프로세스 마이닝을 결합하여 좀 더 견고하고 성능이 향상된 군집분석을 보여준다. 모바일 앱 클릭스트림 데이터를 활용하여 사용자가 방문하는 페이지의 콘텐츠를 기반으로 각 사용자를 분류하였고, 소비자 구매행동 유형을 4가지 (목표지향적, 쾌락적, 탐색적, 혹은 지식 탐색)로 나누었다.

영과잉 음이항 회귀분석(zero-inflated negative binomial regression)을 이용하여 각 유형별 구매 상품 종류 및 횟수를 파악하였다. 영과잉 음이항 회귀분석 안에서의 두가지 과정(zero inflation model and count model)을 통하여, 어떤 유형의 고객들 혹은 어떤 페이지를 더 많이 방문하는 고객일수록 일회성 구매(one time purchase) 혹은 정기구독(auto-refilling plan)을 더 많이 신청하는지 알아보았다.

위 회귀분석을 이용하여 각 유형별 구매 확률 및 횟수를 확인하였고, 다음과 같은 소비자 구매행동을 파악할 수 있었다. 쾌락적 및 탐색적 소비자의 경우 주로 일회성 구매를 많이 한 반면, 목표지향적 구매자의 경우 정기구독을 하는 경우가 압도적으로 많았다. 또한 탐색(explore) 및 검색(search) 페이지를 많이 방문한 고객의 경우 일회성 구매가 잦았던 반면, 세션을 갱신하는 페이지를 많이 방문하고 탐색적 행동

을 자제하는 고객일 수록 정기구독에 의존했다.

비록 일회성 구매를 하는 고객들은 액션, 판타지, 레이트나잇 (late-night) 소설류를 봤지만, 정기구독자일수록 로맨스 소설류를 더 많이 구독하는 현상을 보였다.

소비자 구매행동을 구분하여 고객 유형을 분류하고 그들의 구매 확률 및 횟수를 파악함으로써 각 고객 및 고객 페이지 방문 패턴에 맞는 구매옵션을 알맞게 추천해주고 마케팅 효과를 높일 수 있다.

주요어: 소비자유형; 구매예측; 클릭스트림 데이터; 클러스터 분석
학번: 2016-20570