Consumer Behavior Analysis and Repeat Buyer Prediction for E-commerce

Dissertation
for the award of the degree

Doctor of Philosophy(Ph.D.)
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral Program in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

submitted by
Bo Zhao

from Henan, China
Göttingen, 2019

Thesis Committee:

Prof. Dr. Ramin Yahyapour
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Members of the Examination Board/Reviewer:

Reviewer:
Prof. Dr. Ramin Yahyapour
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Institut für Informatik, Georg-August-Universität Göttingen

Second Reviewer:
Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Further members of the Examination Board:
Prof. Dr. Marcus Baum
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Winfried Kurth
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Carsten Damm
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Lutz M. Kolbe
Faculty of Economic Sciences, Georg-August-Universität Göttingen

Date of the oral examination: 02.12.2019

Acknowledgement

During my Ph.D. study, I got a lot of help from so many people. I would say that I could not finish my study without their help.

I will express my great thanks and gratitude to my supervisor Prof. Dr. Ramin Yahyapour for his countless advice, the inspiring discussions and his encouragements. His ample knowledge and experiences give me a deep impression. Many thanks are also given to Prof. Dr. Xiaoming Fu for his kind supervision and the interesting and informative discussions.

I am grateful to Dr. Philipp Wieder, Martina Brücher, Dr. Edwin Yaqub, Dr. Song Yang and Dr. Fei Zhang for the help and advice during my study. I also thank my colleagues from the eScience group of the GWDG for providing the interesting research environment.

My study is impossible without the financial support from the “China Scholarship Council (CSC)". Best wishes to my country and the people. Also, many thanks are given to the kind Germans and the beautiful German sceneries. They left me many precious memories and I had a lot of fun during leisure time.

Last but not least, I owe my great thanks to my family and my parents for their endless love and encouragements which are always the motivations make me go forward.

Abstract

The proliferation of mobile devices especially smart phones brings remarkable opportunities for both industry and academia. In particular, the massive data generated from users' usage logs provide the possibilities for stakeholders to know better about consumer behaviors with the aid of data mining. In addition, with the popularization of the mobile Internet and the prevalence of delivery service, Online Takeout Ordering & Delivery (OTOD) using Apps from smart phones or websites from PC has become an emerging service and prosperous industry(e.g., KFC delivery). Merchants sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, "Black Friday" or "Double 11 (Nov 11th)", in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI).

Firstly, we studied the consumers' short-term and long term behavior across different platforms comprehensively. Then we tried to find a series of features to deal with the problem of repeat buyer prediction in E-commerce.

For the consumer behavior analysis, we examine the consumer behaviors across multiple platforms based on a large-scale mobile Internet dataset from a major telecom operator, which covers 9.8 million users from two regions among which 1.4 million users have visited e-commerce platforms within one week of our study. We make several interesting observations and examine users' cultural differences from different regions. Our analysis shows among the multiple e-commerce platforms available, most mobile users are loyal to their favorable sites and people (60%) tend to make quick decisions to buy something online, which usually takes less than half an hour. Furthermore, we find that people in residential areas are much easier to perform purchases than in business districts and more purchases take place during non-work time. Meanwhile, people with medium socioeconomic status like browsing and purchasing on e-commerce platforms, while people with high and low socioeconomic status are much easier to conduct purchases online directly. We also show the predictability of cross-platform shopping behaviors with extensive experiments on the basis of our observed data.

In order to improve the quality of service and recommendation personalization, we tried to find the key factors leading to a successful purchasing of takeout food in this paper. We collected Internet access records related to OTOD service of 34,845 users with a time duration of nearly four months. At first, we did a preliminary study on users' daily and periodic purchasing behaviors of takeout food. Then we combine the demographic information and location information with the purchasing activities to find the most potential purchasing groups of takeout food. Based on the features extracted from historical purchasing records, demographic information and location information, we use several popular machine learning methods to predict the future purchasing activities within a specific time. The experiments show that our extracted features can be well used for the takeout food purchasing prediction problem.

It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. With the long-term user behavior log accumulated by Tmall.com, we get a set of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day. Our goal is to predict which new buyers for given merchants will become loyal customers in the future. To achieve this goal, we did a comprehensive feature engineering to find the key factors influencing consumers' repeat purchasing in the future. Based on the features, we build a merged machine learning model to predict the repeat buyer and achieve a roc-auc score with 0.697.

Contents

Chapter 1

Introduction

Online shopping and e-Commerce have become a significant part of the global economy and their applications are becoming a primary vehicle for people to find, compare, and ultimately purchase products. By understanding trends in online shopping and how each part of an e-Commerce site from customer reviews to social media links can affect the traffic and conversion rate can help online business better serve their customers and increase their revenue. One of the fundamental questions that arises in e-commerce is to characterize, understand, and model user behavior and purchasing intent, which is important as it allows for personalized and context relevant e-commerce services.

E-commerce has drastically changed traditional buyer-seller relationships, as well as the shopping process for many consumers[5]. Nowadays, consumers are able to browse and compare various product catalogs, save favorite items, and enjoy powerful features such as search, personalized recommendation, and the benefits of social networks[9, 21, 32, 52]. As the complexity of online shopping behaviors has increased, it has become increasingly important to understand and characterize consumer online purchasing behavior. In particular, it is essential to understand how user activity might build up over time into purchase intent, and ultimately, a purchase. Here, purchase intent is defined as a predictive measure, at a given time, of subsequent purchasing behavior[40]. Besides of that, there are various e-commerce platforms and plenty of merchants on a platform. Consumers may move across different online platforms to search for their ideal products by considering complex factors, such as nice price, good service or sales. However, due to the limitations of lack of data, previous work mainly focused on user behavior analysis of single e-commerce platforms[69]. Little work has been done to indicate whether people will move across different shopping platforms and even why and how the users jump from one platform to the next. China is the well deserved global e-commerce leader according to its volume and growth rate[1]. Consequently, we choose datasets of Chinese e-commerce for analysis. Consumers' behaviors may be diverse due to their different background, such as culture and religion, etc. However,

[1] https://www.emarketer.com/content/global-ecommerce-2019

our research methods and proposed models are general for such kinds of problems and can be used to other datasets.

While e-commerce is rapidly spreading around the world, the food delivery industry also ushers in the spring. Companies have changed their traditional business strategies into online marketing to suit customer needs and taste at any time[2]. Although that the products provided in E-commerce platforms and online food service platforms have great difference, the most important roles in the transactions are the same: consumers. Online food ordering and delivering is another interesting research field about consumer behavior analysis and repeat purchase prediction.

Understanding the consumer behaviors in online shopping, we can do a lot of work on the basis to improve the service quality and profit of the merchants. This can benefit both the sellers and buyers, as well as the third party participants, such as e-commerce platforms and delivery providers. One of the most interesting and positive research direction is repeat buyer prediction. Based on the primary analysis of consumer behavior, we can extract as many as possible features and use machine learning methods to identify the potential repeat buyers in the future. Further more, we can give more precised advertisements and more personalized recommendations to these potential repeat buyers, which can reduce the recommendation cost greatly and effectively.

1.1 Motivation

1.1.1 Consumer Behavior Analysis in E-commerce

Modeling and recognizing purchase patterns is vital for providing better services, more usable e-commerce platforms, and improved personalization in content and search result rankings, as well as advertising. There are three main components for online shopping, namely seller, buyer and action. The profiles of sellers(e.g., historical sell records and comments, etc.) and buyers(e.g., age and gender, etc.) should have influence on the purchase actions intuitively. Considering the action takes place between seller and buyer, spatiotemporal factors(e.g., weekdays or weekends, residence or work places, etc.) may also influence the consumers' actions. However, there are several challenges in studying the purchase patterns of online users. Generally, most prior work has examined short-term user activity and considered predicting whether a given user session will result in a purchase[29, 37, 55, 58, 45]. But the purchase preference of a consumer may change with many factors, such as recommendation from friends and special promotions by the merchants. Furthermore, traditional studies often examine user behavior on a single e-commerce platform, while users may use several different services and move across e-commerce platforms when deciding which product to purchase and where. Thus, what is missing from past research is a cross-platform analysis of how user purchase actions change across different platforms. To this end, it is important to analyze consumer behavior across various e-commerce platforms, and then also identify how purchasers' on-line behavior changes over time from the norm as a result of impending purchases. To solve this problem, we made a preliminary analysis of consumer behavior and a comprehensive cross-platform comparing analysis to identify the most popular consuming patterns that lead to successful purchases.

1.1.2 Consumer Behavior Analysis and Prediction of Takeout Food Purchasing

While the online food delivery market has seen rapid expansion, there is still room for businesses to grow as food delivery accounts for a relatively small portion of the total catering industry. Identifying the most potential customers make great sense for the platforms and merchants to enlarge their market share and profit. Intuitively, we believe that demographic factors(e.g., gender, age and occupation, etc.) and spatiotemporal factors(e.g., weekdays or weekends, home or office, etc.) have great influence on the takeout food purchasing since different groups have different concern and attitude to the takeout food. Understanding the consumer behaviors about takeout food purchasing can help the merchants better carry out their market strategies and improve their market share and profit. Naturally, the consumers themselves will also benefit from the more personalized service.

1.1.3 Repeat Buyer Prediction

Merchants sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, "Black Friday" or "Double 11 (Nov 11th)", in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI). It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. An important part of this research lies in the feature engineering and model training. Even a small improvement on the prediction performance can bring about great market profit in macro view.

1.2 Contributions

1.2.1 Consumer Behavior Analysis in E-commerce

For the consumer behavior analysis in e-commerce, our main contribution is the long-term consumer behavior analysis a across different e-commerce platforms. Based on our observations and analysis, we mainly tried to answer the following questions:

(1) How spatiotemporal factors influence users' shopping behaviors.

(2) How users' shopping behaviors vary in different functional zones.

(3) Whether users' profile (e.g., app usage behaviors) and socioeconomic status would influence their shopping decisions.

(4) How do people make their shopping decisions.

(5) Whether users exhibit signs of loyalty to certain shopping platforms.

(6) Whether users' cross-platform shopping behaviors are predictable.

1.2.2 Consumer Behavior Analysis and Prediction of Takeout Food Purchasing

Our main contribution in this part is try to find the key factors influencing the takeout food purchasing. To the best of our knowledge, this is the first work that thoroughly studies the consumer behavior analysis and prediction problems in the takeout food industry. The features we generated can be used in purchasing behavior prediction and product recommendation and our work could be valuable for data science practitioners, who need to develop solutions for prediction and recommendation tasks in takeout food markets.

In general, our contributions of this paper are as following:

(1) We present a statistic results of consumers' long-term purchasing behaviors related to takeout food using data mining. We collected nearly 4 months takeout food access and purchasing records of more than 10, 000 users and extract the purchasing actions from them.

(2) We try to find the relationship between the demographic factors(e.g., gender and age,etc.) and purchasing actions of takeout food.

(3) We extract the location information embedded in the records of takeout food purchasing activities to infer the possible occupations and then study their different attitude and purchasing actions of takeout food. On the basis, we find the most potential groups tending to purchase takeout food.

(4) We use machine learning to predict the future repeat purchasing of takeout food. We combine the demographic features, historical records and spatiotemporal features together to predict consumers' future purchasing actions within one week, two weeks, three weeks and one month.

1.2.3 Repeat Buyer Prediction

Based on the consumer behavior analysis above, we use engineering technique to extract as many features as possible for the repeat buyer prediction work. Our main contributions are as following.

(1) We use feature engineering technique to find a series of features that can be used in our training and testing models.

(2) We propose a weighted merged machine learning model of different classification models for the repeat buyer prediction task, which can outperforms each single model separately.

(3) We propose a weighted merged machine learning model of different lightGBM models with different parameter sets for the repeat buyer prediction task, which can bring about great performance improvement.

1.3 Dissertation Structure

The contents of this dissertation are organized as follows:

- Chapter 1 introduces the motivations behind our study and the contributions of this dissertation regarding the targeted problems.
- Chapter 2 provides a preliminary study on users' daily and periodic online shopping behaviors, as well as the influence of special online shopping events and gender factors.
- Chapter 3 examines the consumer behaviors across multiple platforms based on a large-scale mobile Internet dataset and analyzes the various consumption patterns.
- Chapter 4 analyzes consumer behavior from another aspect of online shopping service, takeout food purchasing. Different from the products in e-commerce platforms, food is more regularly consuming products in our daily life. We analyzes the consumer behavior in detail and demonstrate the predictability of takeout food purchasing.

- Chapter 5 extracts features that influence a consumer's repeat purchasing in the future to provide more personalized recommendation to its potential consumers for a merchant at first. On the basis, we proposed a merged model for the repeat buyer prediction and make comprehensive experiments to validate our ideas.

- Chapter 6 summarizes the work in this dissertation and gives an outlook of future research topics based on the contents of this dissertation.

Chapter 2

A Preliminary Study of E-commerce User Behavior Based on Mobile Big Data

The rapid popularity of mobile devices especially smart phones has changed human life style greatly. In this chapter, we examine the consumer behaviors on several e-commerce platforms based on a large-scale dataset of mobile internet access records for about 3.5 months from a major telecom operator in China, which covers 126,388 users from Shanghai. We provide a preliminary study on users' daily and periodic online shopping behaviors, as well as the influence of special online shopping events and gender factors. These findings may be exploited by e-commerce providers e.g., for developing personalized recommendation systems to improve their service quality and profit.

2.1 Introduction

In the past few years mobile phones have experienced a remarkable evolution and explosive popularization [46]. Meanwhile, e-commerce has a prosperous development and drastically changes traditional commercial relationships, as well as the shopping process for the fast-growing online shoppers [5]. With a smart phone at hand, a consumer can check the details of products, compare the prices across various e-commerce platforms, save items into charts and enjoy a number of benefits such as personalization from merchants and recommendation from social networks [21, 32, 69]. As more and more people purchase online, understanding consumers' online behaviors becomes more and more important. Based on the behavior analysis, e-commerce companies may enforce corresponding marketing strategies to improve their service quality to keep and gain more consumers.

For online shoppers, searching for ideal products also takes plenty of time and energy, since many of them would purchase products based on their own budget.

Facing diverse e-commerce platforms and many merchants, different consumers may exhibit different behaviors because of their diversities in economic status, personal preference and social influence etc. For e-commerce providers, the shortened lifecycles of products and intensified market competition lead to an imperative need to study the consumer purchasing behavior in order to make appropriate marketing strategies, such as improving their personalized service and catching consumers' attention and trust.

Unlike most work of identifying purchasing intent, our work focuses on analyzing consumer online shopping behavior with implicit purchasing intent. Which time period do most consumers make their purchasing decisions? Are there differences between the behavior of male and female consumers? Do the promotion periods such as November 11 ("11-11") and December 12 ("12-12") have special influence on consumer behavior? By analyzing an anonymized dataset from a major telecom operator in China, we try to answer such questions in this chapter, shedding light for e-commerce providers and merchants to improve their service quality and profit.

We observed the consumer behavior difference within a day and a week as well as studied the influence of special shopping festivals. In summary, the main contributions of this chapter include:

- An overview about the visiting and purchasing fluctuation with two different time frames (hour and day).
- Empirical evidence about the different consumer behavior considering the factors of gender.
- An observation about the influence of special shopping events on consumer behavior, such as “11-11" and “12-12" as well as the new year, Chinese spring festival and Valentine's Day.

2.2 Dataset

2.2.1 Data Collection

The dataset contains complete anonymized Internet access records of mobile users in cellular environments, which is provided by one of the three major mobile telecom operators in China. We collected the anonymized mobile Internet access data for 126,388 users from Shanghai which is the commercial and financial center of China from November 1, 2016 to February 11, 2017. Because of the popularity of WiFi in Shanghai, mobile users can access Internet using WiFi rather conveniently and our Internet access records cannot cover all the Internet access activities but can help to analyze user behavior under cellular environments. Each record contains the following information of an Internet access: anonymized ID of the mobile user, start time of the Internet access, destination URL and reference URL of the access.

2.2.2 Data Pre-processing

The collected data is heterogeneous and noisy, including all the active and passive Internet access records. In order to study consumer behavior using these various mobile Internet access records, we need to do data cleaning first. There are a lot of e-commerce platforms in China and for simplified analysis, we chose the 5 most popular ones, which are Taobao (taobao.com), JD (jd.com), Suning (suning.com), Dangdang (dangdang.com) and Vip(vip.com). Taobao and JD are the largest two comprehensive online shopping platforms while Suning, Dangdang and Vip are mainly corresponding to electronics, books and fashionable products respectively. We focused on all the users who have actions on these platforms and extracted all online shopping records at first. Due to the multiple interaction rounds and references of web service requests and response queries on various platforms, there are plenty of redundant records. To identify the unique actions from many redundant interaction records, we identified the item IDs and order IDs and only counted each page visit once for the same item or order. After eliminating redundant records, we obtained 0.4 million unique browsing and purchasing records, related to 28,752 users.

2.2.3 E-commerce Platforms

Taobao is a Chinese online shopping website similar to eBay and Amazon and is operated in China by Alibaba Group. Founded by Alibaba Group on May 10, 2003, Taobao Marketplace facilitates consumer-to-consumer (C2C) retail by providing a platform for small businesses and individual entrepreneurs to open online stores that mainly face to consumers in Chinese-speaking regions (Mainland China, Hong Kong, Macau and Taiwan) and now also expands its business abroad. Consumers can almost buy whatever they want, while they have to face the diversity of products and merchants so as to "tao"(Chinese word which means "buy") an ideal commodity.

Jingdong, formerly 360buy, is a Chinese electronic com-merce company headquartered in Beijing. As a major competitor to Alibaba-run Tmall, it is the largest business-to-consumer(B2C) online retailers in China by transaction volume and revenue. Jingdong launched its English website on October 18, 2012, aiming at expanding worldwide shipping.The company was founded in July 1998 and its B2C platform went online in 2004. Founded as an online magneto-optical store, it soon diversified, with products from electronics and mobile phones to general merchandise, covering almost all kinds of products desired by consumers. Jingdong Mall changed the domain name to 360buyimg.com and JD.com in 2007 and 2013, respectively.

Dangdang is a Chinese electronic commerce company, founded by Peggy Yu and Li Guoqing in 1999. It is headquartered in Beijing and its main competitors are Amazon.cn (or Amazon China, formerly Joyo.com) and JD.com (or Jingdong, formerly 360buyimg.com). The competition escalated into a price war in December 2010, with each retailer marking down a wide range of items, especially books. DangDang made an IPO on the NYSE in November 2010, estimated at approximately $1 billion. Dangdang's main product categories include household merchandise, home appliances, cosmetics, digital, books, audio, clothing and child categories etc. while consumers mainly buy books from it. There are over 10 million new registered customers per year in Dangdang. There are about 30 million people browse different kinds of products each month and its monthly sale of goods is over 20 million.

Suning is one of the largest privately owned retailers in China, headquartered in Nanjing, Jiangsu. Suning has more than 1600 stores covering over 700 cities of Mainland China, Hongkong and Japan. Its e-commerce platform, Suning.com ranks among top three Chinese B2C companies. The operation categories include physical merchandise, such as home appliances, 3C products, books, general merchandise, household commodities, cosmetics and baby care products, content products and service merchandise with the total number of SKU exceeding 3 million.

Vip is a leading online discount retailer for brands in China. The Company offers high quality and popular branded products to consumers throughout China at a significant discount from retail prices. As compared to conventional on-line marketplaces or large-scale multi-category online retailers, Vip has successfully created a third e-commerce model and proven that it can provide tremendous scale and profitability. By providing special offers and deep discounts on branded products, the Company has pioneered the online discount retail model in China and become the expert and leader trusted by its customers and brand partners alike. Since its founding in August 2008, the Company has rapidly built a large-scale and growing base of customers and brand partners.

2.3 Consumer Behavior Analysis

2.3.1 Consumer Behavior within a Day

Different consumers have different online shopping time within a day according to their preference and available time. In total, the access peak, purchase peak and successful purchasing ratio peak all occur at 10:00 in the morning, which is at the beginning of work time for most people, as shown in Fig. 2.1. It seems that many people prefer to do some personal business such as online shopping before work. This maybe also have some relationship with the delivery strategy of logistics companies because people tend to have their goods delivered as soon as possible and orders paid in the morning usually have priority to be delivered. Another possible explanation is that

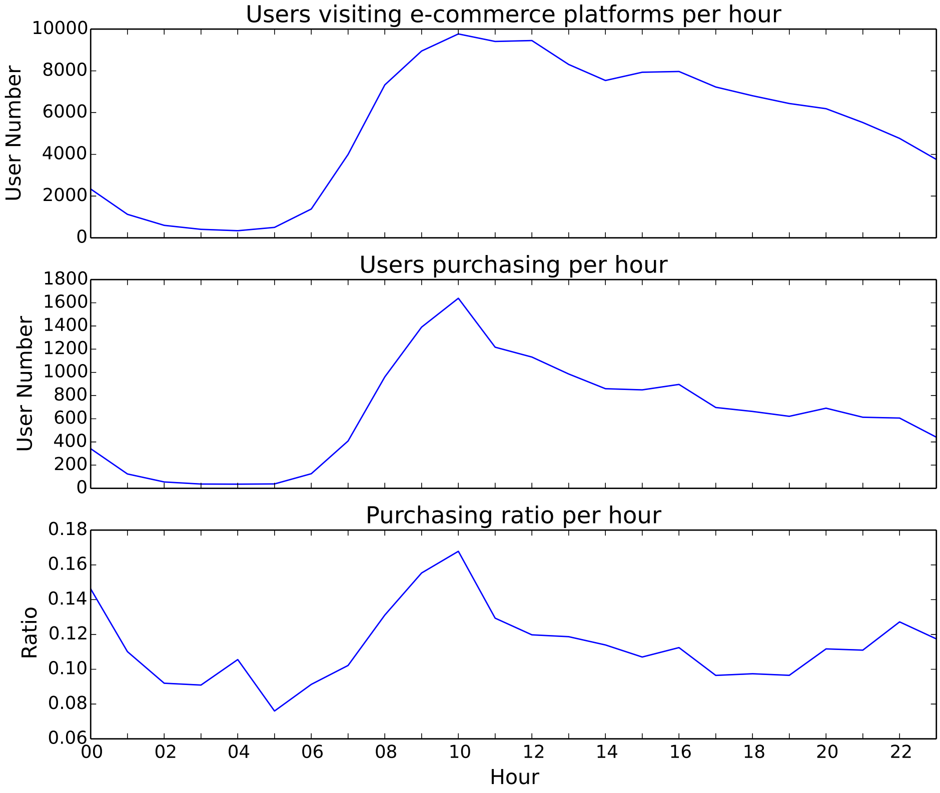


Figure 2.1: Distribution of consumer actions within a day

some users are still on their way to work and therefore have time to visit e-commerce platforms.

However, men and women have quite different consumer behaviors, as shown in the second part of Fig. 2.1. An obvious purchasing peak occurs in the very early morning around 6:00 for female consumers, which maybe because women have more passion for shopping early. Male consumers tend to finish their online shopping in the morning while female consumers keep browsing and buying nearly throughout the whole day. Moreover, women have more passion for online shopping in the afternoon and an empirical observation is that women tend to be more easily attracted by online shopping, children and small talks etc. in the afternoon in China. Online retailers can carry out more promotion online shopping activities oriented to women consumers to attract their attention and actions.

2.3.2 Consumer Behavior within a Week

We tried to find the user behavior difference between workdays and weekends, as shown in Fig. 2.3. As a whole, users tend to visit and finish purchase on weekdays. The Chinese delivery market is fiercely competitive and thus, the delivery time is quite short

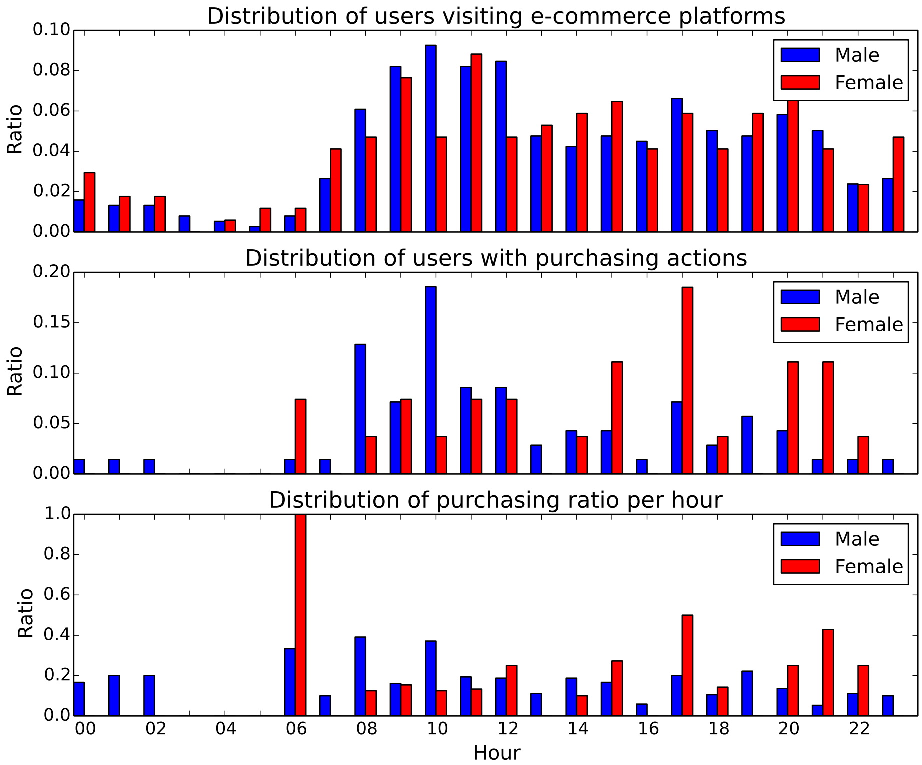


Figure 2.2: Distribution of consumer actions considering genders within a day

for satisfying and winning more customers. On average, an order can be delivered in two workdays. Influenced by the delivery situation, it is quite reasonable to tell why few people choose to buy products on Thursdays. As shown in Fig. 2.4, male and female users can be divided into two main groups: early workday shoppers and Friday shoppers. Users prefer to do online shopping from Monday to Wednesday probably because they want to receive their products on workdays in their companies, while users finishing their online shopping on Friday tend to receive their orders on weekends at home. Weekends are usually used for entertainment and outdoor activities and users tend to spend less time on Internet usage. In addition, users seem to visit e-commerce platforms mainly by wifi which cannot be traced and therefore our result is probably a biased statement.

Since consumers browse and purchase more on weekdays, online retailers can adapt their sale strategies to this phenomenon and organize more promotion activities to attract more consumers and improve their profit.

2.3.3 Influence of Special Events

In this section, we investigate the influence of special shopping festivals on consumer behavior in order to aid retailers in their development of marketing programs that can

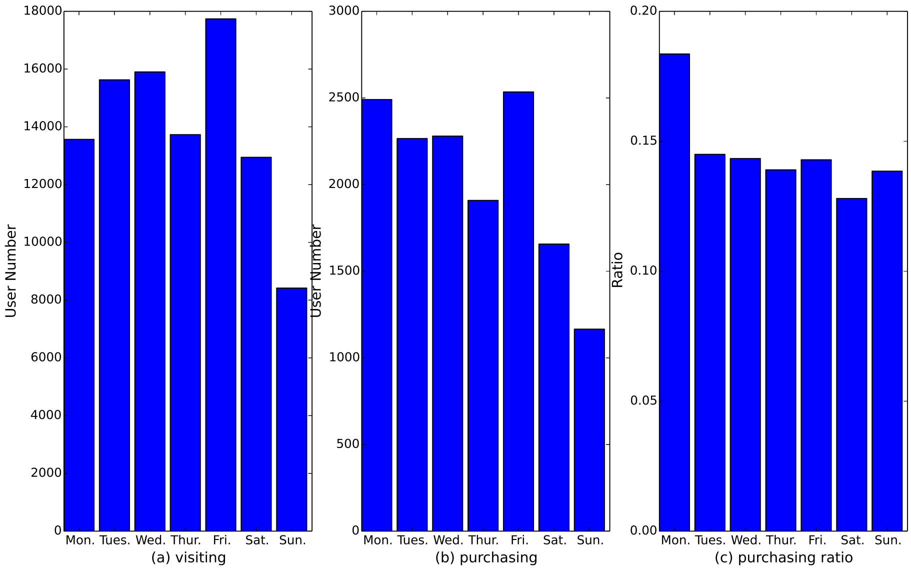


Figure 2.3: Distribution of consumer actions within a week

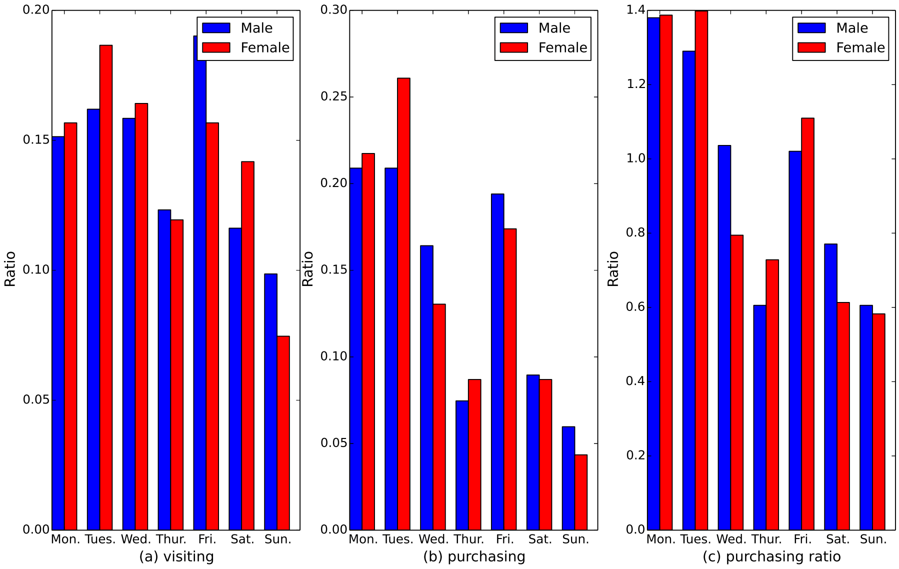


Figure 2.4: Distribution of consumer actions considering genders within a week

help increase shopping festival sales as well as the total profit throughout the whole year. U.S. Retailers consider two major holiday shopping days as their most profitable: the Friday after Thanksgiving, Black Friday, and the Monday after Thanksgiving, Cyber Monday. Inspired by this, Alibaba held the first "11-11" shopping promotion day on Taobao.com in 2009, storming the online shopping for the very first time. Big promotions in the name of celebrating Nov. 11 Bachelor's Day usually start at the very beginning of November with huge discounts and give always lined up. There are some other smaller shopping promotion days compared with "11-11", such as "12-12" and "6-18" as well as some traditional festivals, such as new year, Chinese spring festival, Valentine's Day etc. Our dataset covers a period including "11-11" and "12-12" of 2016 as well as New Year and Chinese Spring Festival of 2017. Consequently, we can have a rather comprehensive observation about consumer behavior around these festivals and investigate the influence.

From our evaluation, obvious access and purchase peaks occur around "11-11" and "12-12", as shown in Fig. 2.5. As mentioned before, "11-11" is an online shopping festival starting several days before the very date. Accessing peaks occur from the beginning of November and in order to decrease the browsing and purchasing pressure of the very day of Nov. 11, many retails choose to bring forward their promotion activities. Considering the purchasing ratio, a higher successful level occurs from Jan. 14, 2017, which is about two weeks before the Chinese Spring Festival. A possible reason is that consumers tend to make some special purchases for the Spring Festival and the need is stronger than usually.

When considering the gender, the result is quite complex and irregular. In average, female consumers have higher accessing ratio while male consumers have higher successful purchasing ratio. More interestingly, both male and female consumers are interested in visiting e-commerce platforms while only male consumers purchase before Valentine's Day. This phenomenon is inline with Chinese traditional concept of value that a man should buy gifts for his girl friend or wife.

As analyzed above, online retailers can adjust their market strategies to attract more attention from their potential consumers and make their total profit maximum. Special online shopping events are very good opportunities for merchants to finish their annual sale goals while the competition is also very fierce. Proper adjustment for the date maybe make the online retailers benefit from the special online shopping promotion days as well as maximally avoid competition with other shops.

2.3.4 Consumer Clustering

Empirically, consumers tend to have different online shopping preference and habits influenced by various factors, such as occupation, socioeconomic status and education background etc. Some users just have interest to have a look about the details(e.g., price, size, function etc.) online while prefer to buy products from physical stores.

Understanding consumer behavior difference can help online retailers to design specific strategies for different consumer groups to maximum their profit. In this chapter, we observed that about 85% users have access records to e-commerce websites while have no purchasing actions. Some consumers tend to buy stuff according to their real need and their shopping records are random. In addition to the two kinds of consumers

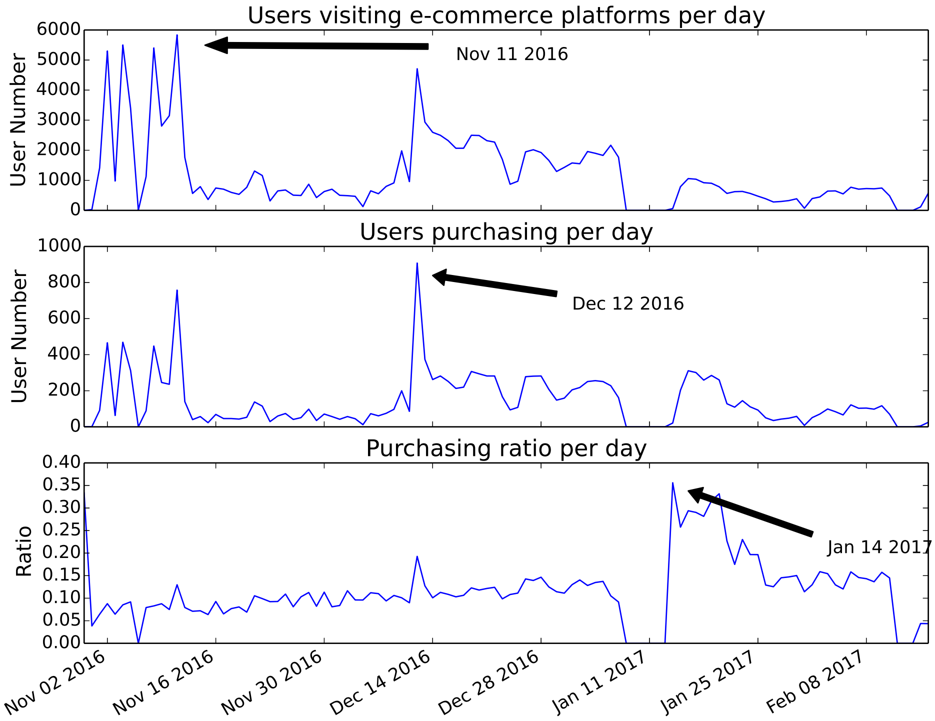


Figure 2.5: Distribution of consumer actions across the whole period

above, another group of users tend to buy plenty of goods during the special shopping events such as "11-11" and "12-12" because of the huge discount. We take the actions of the telecom users in our dataset as features to cluster the online shopping users into several groups.

For consumer clustering, we use $U = \{u_1, u_2, ..., u_M\}$ to represent the consumers access to e-commerce platforms, in which M means the total number of users who have access records to e-commerce platforms. We use $F = \{B, S, P, N\}$ to represent the behavior patterns for each consumer, in which B, S, P, N represent Both(scan and purchase), Scan(no purchase), Purchase(directly purchase without scan) and None(no scan or purchase) respectively. It is easy to understand the consumer behavior of B, S and N, while P is also very common for some consumers who prefer to add products into shopping chart first and then need some time for final purchasing decision. In this section, we use K-means clustering algorithm and the input is an array with $D = 110$ dimensions. $\forall u_i \in U$, the corresponding input array is $a_i = [f_1, f_2, ..., f_D]$, in which $f_j \in F = \{B, S, P, N\} (j \in [1, D])$ means the consumer behavior throughout the whole observation period. The number of consumers in our dataset with online shopping actions is 28,752 and the clustering result is shown as Fig. 2.7(a) when the clusters number is set as 4, in which the X-axis means the number of items scanned and Y-axis

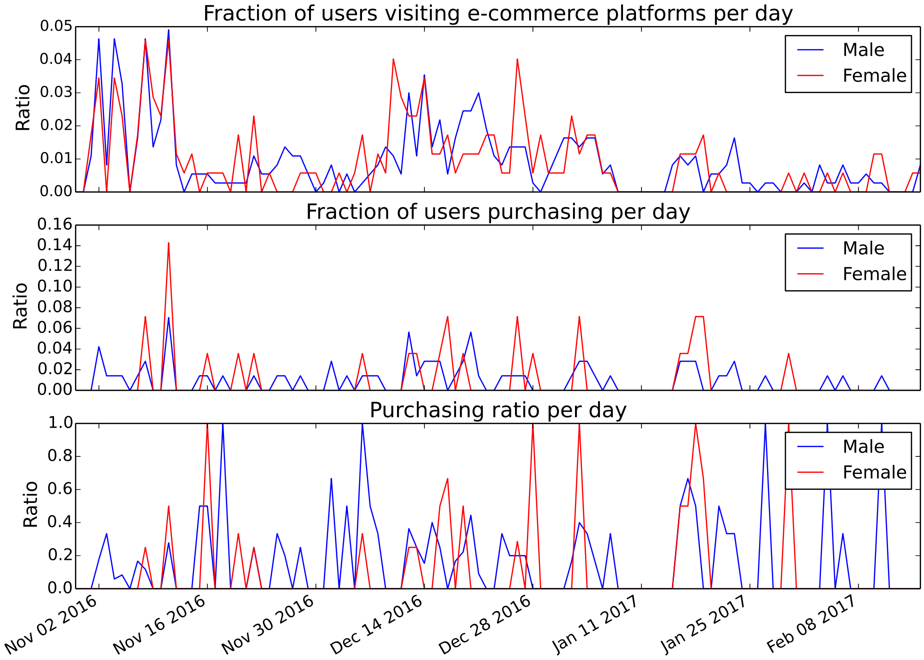
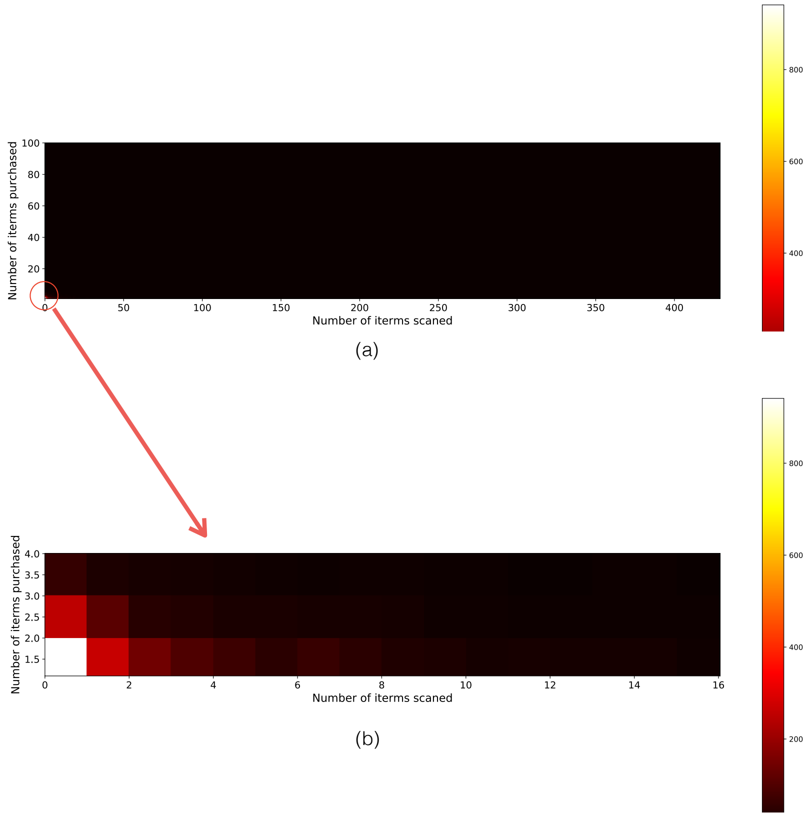


Figure 2.6: Distribution of consumer actions considering genders across the whole period

means the number of items bought for each user. In Fig. 2.7(a), the darker the block is, the fewer the number is. Based on the clustering result, most users(78%) only have scanning but no purchasing records. As for the remaining users(22%) who have purchasing actions, 10% only have purchasing records without scanning product details, which is quite normal during the special shopping festivals such as ""11-11", shown as the white block in Fig. 2.7(b). 8% users prefer to finish their purchasing operations after a plenty of scanning actions to have a comprehensive understanding about the products themselves and different prices across different e-commerce platforms, shown as the red blocks close to the X-axis in Fig. 2.7(b). The last 4% users tend to make purchasing decisions very quickly after a few scanning operations, shown as the red blocks close to the Y-axis in Fig. 2.7(b).

2.4 Related Work

For the past few years mobile phones have a remarkable evolution and explosive popularization [46]. Meanwhile, e-commerce also has a prosperous development and drastically changed traditional commercial relationships, as well as the shopping process for the fast-growing online shoppers[5]. With a smart phone at hand, the consumer can check the details of products, compare the prices across various e-commerce platforms, save items into carts and enjoy a great many benefits such as



(a)

(b)

[!tb]

Figure 2.7: User clustering based on access and purchase records

personalization from merchants and recommendation from social networks [9, 21, 32, 69]. The complexity of users' online behaviors is increasing and understanding consumer online behavior becomes more and more important to know the buildup of successful purchases. Understanding the consumer buying process can make a great difference between success and failure in consumer marketing strategies [56].

Research surrounding online shopping analysis has a large body of work [21, 32, 69, 36], dating back to the early research of purchasing behavior on the Web [5, 43, 31]. The beginning research work focused on the intention identification of users using web service, such as search and browsing [31, 14], which is helpful to improve the quality of a search engine's results or the attraction of a special website. With the prosperous development of e-commerce and rapid popularization of smart phones, more and more attention are attracted to the user behavior analysis within online shopping. The past research investigated a series of factors leading to successful purchasing results, including motivations, recommendations [49, 26, 50, 53], personalization, as well as demographic factors, such as gender, age and residence [23]. This is very useful for e-commerce providers to improve their service quality and competition ability as a result. Our work mainly analyzes user online shopping behaviors using the most popular five e-commerce platforms based on the dataset of telecom, considering factors of gender, workdays and special shopping festivals. In addition, we also studied the purchasing results of the whole day using a time unit of hour and find some interesting results about the most possible "successful" purchasing time periods.

Special shopping festivals have great influence on consumers' shopping enthusiasm and bring huge profit to online or offline retailers. Esther Swilley et al. [59] examined attitudes and behaviors of shoppers for these two shopping occasions, the Friday after Thanksgiving, Black Friday and the Monday after Thanksgiving, Cyber Monday to help retail managers have a better opportunity to market on these two days with an understanding of consumer intentions for these major shopping occasions based on their findings. Jasmin H. Kwon et al. [70] studied the value of collaborative research on seasonal shopping events and behavior and took Black Friday as a case for study. Chinese online shopping festivals came into being quite later while the influence grows very fast. Juan Liu [27] took a case study of T-Mall "Double Eleven" online shopping event to introduce the change of "11-11" from festive ceremony culture to marketing. Xi et al. [63] tried to analyze the consumer behavior and bandwagon effect with the binary choice model using 1,811 college students as the research objects based on the micro survey data of the "double eleven" online shopping. Our dataset chooses 126,388 telecom users in Shanghai randomly and the results are more general.

2.5 Conclusion

The popularity of smart phones and prosperity of e-commerce platforms have changed human life style greatly. Meanwhile, the massive mobile data generated brings remarkable opportunities for consumer behavior analysis with the aid of data mining. In this chapter, we examine the consumer behaviors using various platforms based on a large-scale mobile Internet dataset from a major telecom operator, which covers about 126 thousand users from Shanghai among which nearly half of the users have visited e-commerce platforms within nearly 3.5 months of our study. From our preliminary analysis, we see that male and female online shoppers have quite different behavior and shopping preference. Interestingly, most online shoppers choose to make their purchase at around 10 am., which is the really beginning work time for most people. In addition, we observed that special online shopping festivals such as "11-11" and "12-12" have great influence on consumer behavior in both searching and purchasing products from e-commerce platforms. These findings can be used by e-commerce providers for personalized recommendation system to improve their service quality and profit.

For further work, we currently plan to carry out the research in three aspects. Firstly, we will consider the influence of occupation on consumer behavior. Empirically, people with different occupations have different life styles and social economic status, therefore their attitude and preference to online shopping are also various. Secondly, we will try to find the consumer behavior differences across different regions since different development level and strategies will also have influence on e-commerce market. Finally, we will consider the influence of social relationship on consumer behavior of online shopping since we friends should have similar interests and life styles and we will have more confidence on a product recommended by our friends. We will try to have a comprehensive understanding about consumer behavior and preference when shopping online and then build a recommendation system for different e-commerce retailers to better carry out their market strategies to attract more consumers and gain more profit.

Chapter 3

A Cross-Platform Consumer Behavior Analysis of Large-Scale Mobile Shopping Data

The proliferation of mobile devices especially smart phones brings remarkable opportunities for both industry and academia. In particular, the massive data generated from users' usage logs provide the possibilities for stakeholders to know better about consumer behaviors with the aid of data mining. In this chapter, we examine the consumer behaviors across multiple platforms based on a large-scale mobile Internet dataset from a major telecom operator, which covers 9.8 million users from two regions among which 1.4 million users have visited e-commerce platforms within one week of our study. We make several interesting observations and examine users' cultural differences from different regions. Our analysis shows among the multiple e-commerce platforms available, most mobile users are loyal to their favorite sites; people (60%) tend to make quick decisions to buy something online, which usually takes less than half an hour. Furthermore, we find that people in residential areas are much easier to perform purchases than in business districts and more purchases take place during non-work time. Meanwhile, people with medium socioeconomic status like browsing and purchasing on e-commerce platforms, while people with high and low socioeconomic status are much easier to have successful purchases online. We also show the predictability of cross-platform shopping behaviors with extensive experiments on the basis of our observed data. Our discoveries in this chapter is a sufficient supplementation for the last chapter and could be a better guide for e-commerce future strategy making.

3.1 Introduction

With the development of smart phones and mobile applications, people are spending more and more time on mobile devices. According to a recent survey, nearly 75 percent of US adults will use a smartphone in 2017. On average people spend 3 hours and 15 minutes per day on a mobile device[1]. In November 2016, the mobile Internet usage even surpassed desktop usage for the first time[2]. The proliferation of mobile usage has already shaped our lives (e.g., conquered our wallets) and dramatically changed the business models for numerous enterprises. A study shows that the majority of online shopping sales in the UK are now conducted through smartphones and tablets, instead of traditional computers or laptops[3].

The popularity of mobile devices and the massive data generated from mobile usage offers the research community unprecedented opportunities to study mobile user behavior patterns, which were previously difficult to explore due to a lack of sufficient data. A better understanding of user behavior and underlying usage patterns can allow a mobile service provider to define effective marketing strategies for attracting more users and maintaining current users, eventually increasing its profit. An example is the story

[1] http://www.geomarketing.com/us-mobile-usage-in-2017-stats-you-need-to-know

[2] http://bgr.com/2016/11/02/internet-usage-desktop-vs-mobile/

[3] http://www.telegraph.co.uk/news/shopping-and-consumer-news/12172230/ Are-mobiles-changing-how-we-shop.html

of beer and diapers[4] which suggests an innovative marketing strategy when analyzing supermarket consumer behavior data. For individual users, a better understanding of their own temporal behavior patterns can help them better plan their own household budgets and make better use of the provider's marketing strategies.

With the emergence and ever increasing number of online shopping platforms, users have more possibilities to do their shopping online. They may move across different online platforms to search for their ideal products with considering complex factors, such as nice price, good service or sales. However, due to the limitations of lack of data, previous work has mainly focused on user behavior analysis of single e-commerce platforms[69]. It is still unclear whether people will move across different shopping platforms and even why and how the users jump from one platform to the next.

In addition, users' profiles such as their culture, social and ethical and as well as the functional regions they belong to would also influence their behaviors[56]. Researchers pay more attention to users profiling[4, 16, 22, 24, 69] and apply them in many areas, such as personalization and recommender systems[1, 28, 39, 54]. Whether and how users' profile (e.g., app usage behaviors), their functional zones and socioeconomic status would influence their shopping decisions will also provide useful insights.

Thanks to the e-commerce big data associated with smart phones, it is now possible to correlate a single user's shopping behavior across multiple platforms and with large-scale mobile usage logs, we are able to access all the platforms that users have visited although it also brings us challenges during accessing and processing the data. For instance, the size of compressed mobile Internet data usage records including active online shopping activities for 10 million mobile phone users during one week could easily exceed 40 TB, which were used in the scenario of this chapter.

In this chapter, employing a large mobile communication data from a major telecom provider in two populous regions in China over a period of one week as the basis in our study, we systematically investigate the problem of cross-platform and cross-region consumer shopping behaviors. We first try to answer the following 6 questions:

- How spatiotemporal factors influence users' shopping behaviors;
- How users' shopping behaviors vary in different functional zones;
- Whether users' profile (e.g., app usage behaviors) and socioeconomic status would influence their shopping decisions;
- How do people make their shopping decisions;
- Whether users exhibit signs of loyalty to certain shopping platforms;

[4] https://www.theregister.co.uk/2006/08/15/beer_diapers/

- Whether users' cross-platform shopping behaviors are predictable.

We made several interesting observations. For example, among the multiple e-commerce platforms available, most mobile users are loyal to their favorable sites; people (60%) tend to make quick decisions to buy something online, which usually takes less than half an hour. People in residential areas are much easier to make purchases and they prefer to purchasing during non-work time. Furthermore, people with medium socioeconomic status like browsing and purchasing on e-commerce platforms, while people with high and low socioeconomic status are much easier to conduct purchases online.

Based on the observations, we further examine the predictability of cross-platform shopping behaviors. We build a framework with four types of features: temporal feature, loyalty feature, profiling feature and demographic feature. The prediction results show that consumers' cross-platform shopping behaviors are predictable and our prediction performance is as high as 94% in terms of both F1 and accuracy.

3.2 Dataset

3.2.1 Data Collection

The dataset is drawn from a log of anonymized browsing records of mobile usage in cellular environments provided by China Telecom, which is one of the three major mobile telecom operators in China. By the end of 2017, China telecom shared 17% of Chinese mobile market[5]. There are five main things to consider when choosing a telecom operator, namely network security and reliability, service offerings and support, costing and profitability, technology and scalability, customization, respectively[6]. Because of the mature technique and transparent competition in telecom market, the other two telecom operators are not much different with China Telecom except for the number of base stations and market shares. The user distribution of these three Chinese telecom operators is consistent with the population structure. Currently we only have the dataset from China Telecom ant it is typical for consumer behavior analysis for online shopping. In the future, we will try to have collaborations with the other two telecom operators to make more general analysis and get more analysis about consumers' preference to different telecom providers. The dataset for short-term analysis contains the mobile usage data for over 9,700,000 users from two populous regions over a period of roughly one week each: one is Shanghai, the most populous metropolitan in the world (and also the commercial and financial center of mainland China), between April 20 and April 26, 2016 and the other is Shandong province, the second most populous province

[5] https://www.chyxx.com/industry/201711/581711.html

[6] https://www.sifytechnologies.com/blog/5-things-to-consider-when-choosing-a-telecom-provider/

Table 3.1: Dataset Statistics

Item	Shandong	Shanghai
Time period	Aug 6th - Aug 14th, 2016	Apr 20th - Apr 26th, 2016
#Mobile User	5,461,244	4,309,914
#Average user per day	2,827,771	2,914,294
#Online shopper	301,426	233,537
#Average shopper per day	45,481	47,579
#Purchaser	33,189	35,041
#Average purchaser per day	3,970	5,454
#Browsing records	156,019	135,154
#Average browsing records per day	17,335	19,308
#Purchase	40,753	54,453
#Average purchase per day	4,528	7,779

of China, with only 45% of per capita disposable income of Shanghai[7], between August 6 and August 14, 2016. Each of these records contains the anonymized ID of the mobile device and the start time for each action, as well as browsing records. Part of these records contain geo-location information in the forms of longitude and latitude where the action was performed.

3.2.2 Data Pre-processing

The collected data is heterogeneous and noisy. In order to study consumer behavior using these vast mobile browsing records, we need to begin by cleaning the data.

We analyzed the 5 most popular Chinese B2C e-commerce platforms, which are Taobao (taobao.com), JD (jd.com), Suning (suning.com), Dangdang (dangdang.com) and Vip (vip.com). We focused on all users who browsed or purchased on these platforms, and extracted all browsing and purchasing records. Due to the multiple interaction rounds of web service requests and response queries on various platforms, a single browsing or purchasing action needed to be identified from many redundant interaction records. To make it simple, we only counted each page visit once. After eliminating redundant records, we obtained 386,379 unique browsing and purchasing records. The detailed data statistics is shown in Table 3.1.

3.3 How Spatiotemporal Factors Influence Users' Shopping Behaviors?

In this part, we will examine how spatiotemporal factors (e.g., time, regions and platforms) influence users' shopping behaviors, i.e., product browsing or purchasing.

[7] http://www.yicai.com/news/5229761.html

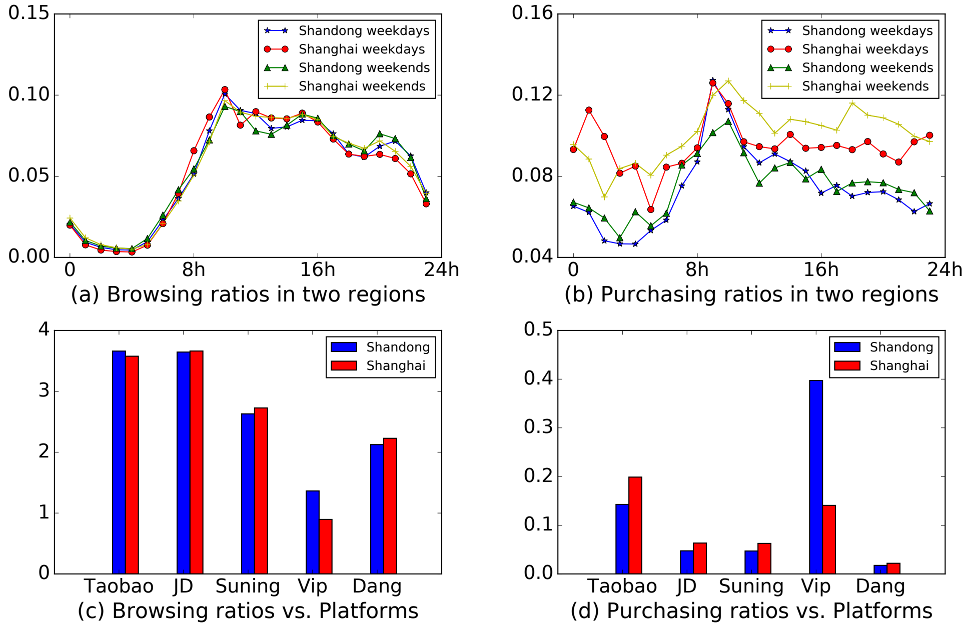


Figure 3.1: How spatiotemporal factors influence users' shopping behaviors? (a) and (b) show the browsing and purchasing ratios in two regions respectively, with the X-axis being the hour in a day, and Y-axis being browsing or purchasing ratios where the ratio means the percentage of purchases within one hour to the total number browses within a day. (c) and (d) show the browsing and purchasing ratios on diverse platforms, with X-axis unit being diverse platforms and Y-axis being browsing or purchasing ratios. Notify that in order to make the figure more readable, in (c) we make the Y-axis as $log10$ (actual browsing ratios*10000).

Influence of Time. People usually have different time schedules on weekdays and weekends in different regions. Fig. 3.1(a) and (b) show users' browsing and purchasing behaviors separately in Shanghai and Shandong during different time periods. From Fig. 3.1(a), we can see that people tend to have the similar browsing behaviors on both weekdays and weekends. For example, people in both Shandong and Shanghai are willing to browse shopping websites during the morning coffee break (i.e., around 10:00).

In addition, people prefer to browse shopping pages during working hours (8:00 – 17:00). In terms of purchasing behavior, people are more willing to pay for their orders around 11:00 in the morning, which is right after the time most people spend browsing. Moreover, people prefer to place their orders on weekends versus weekdays according to Fig. 3.1(b).

Influence of Platforms. We focus primarily on the 5 most popular Chinese e-commerce platforms. Here we will examine users' shopping behaviors over each platform separately. Fig. 3.1 (c) and (d) show users' browsing and purchasing behaviors on each platform.

Table 3.2: Identification of Functional Zones Based on POI.

Zones	POI labels
Business	government; education; hospital; company; etc.
Residence	town; village; villa; realty; etc.
Leisure	hotel; sport; scenery; restaurant; shopping; etc.
Others	others.

From Fig. 3.1(c), we can see that Taobao and JD are the most popular platforms, which are the two largest and most comprehensive online shopping platforms in China, making up 74.0% of the browsing records and 93.9% of the purchases from our dataset. Interestingly, we find that people are more willing to purchase on Taobao and Vip, as shown in Fig. 3.1(d).

Influence of Regions. From Fig. 3.1(a), we can see that people in different regions tend to have similar browsing behaviors. However, they react quite differently when making purchases. From Fig. 3.1(b), we can see that users from Shanghai are more likely to carry out online shopping purchases than people in Shandong. This might be due to that people from less developed regions are more concerned about spending their hard earned money. Furthermore, Shanghai consumers tend to carry out purchases late at night or in the early morning hours, versus consumers in Shandong, which might be a reflection of Shanghai's socioeconomic situation, as the business and financial center of China.

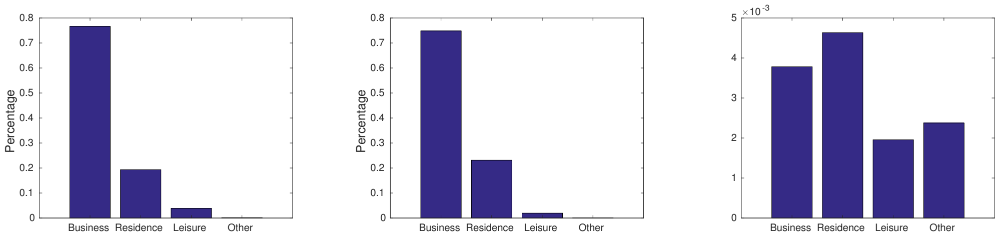
3.4 How Users' Shopping Behaviors Vary in Functional Zones?

The modern civilization and urbanization fosters functional zones in a city[68] and people behave differently in various zones. In this section, we examine whether users' shopping behaviors vary in functional zones.

In this chapter, we divide a city into four types of functional zones: business districts, residential areas, leisure areas and others. Since we only have users' geo-locations, we determine these functional zones according to the Point of Interest (POI) associated with these locations[8]. POI labels associated with each functional zones are shown in Table 3.2.

We now check users' shopping behaviors in each type of functional zones. From Figure 3.2, we can see that people in business districts perform the highest number of browsing and purchasing activities. However, people in residential areas are mostly like to make purchase decisions. This is partly because in business districts, people are more likely to visit e-commerce platforms in cellular environments, thus having

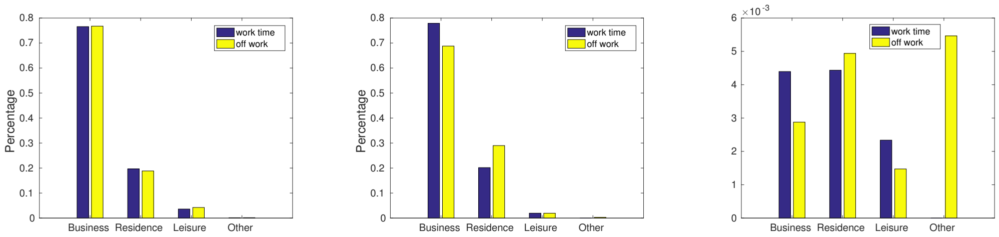
[8] The POI dataset is public under the link: http://pan.baidu.com/s/1pKCL6YZ.



(a) Browsing (b) Purchasing (c) Purchasing/Browsing

Figure 3.2: Users' shopping behaviors in 4 functional zones. (a) browsing behavior; (b) purchasing behavior; (c) the ratio of purchase.

more browsing and purchasing records. In addition, during work time in the day, consumers can talk with his colleagues about the products, which reflects the power of recommendation from social network. Yet, it is easier for people to make purchase decisions in residential areas.



(a) Browsing (b) Purchasing (c) Purchasing/Browsing

Figure 3.3: Users' shopping behaviors in work time vs. off work in 4 functional zones. (a) browsing behavior; (b) purchasing behavior; (c) the ratio of purchase.

3.4.1 Users' Shopping Behavior in Work Time vs. Off Work Time

In addition, we check users' shopping behaviors in each functional zone in different time slots. We divide a day into work and off work time according to their working state (i.e., 9 am - 6 pm is for work time and the others for off work time.). From Fig. 3.3, we can see that people in business districts tend to have more browsing and purchasing activities in work time while people in residential areas tend to have more purchases in off work time, which is quite coincident with our intuition.

3.4.2 Users' Shopping Behaviors vs. Socioeconomic Status

In this part, we examine whether users' socioeconomic status will influence their shopping behaviors. Socioeconomic status is the social standing or class of an indi-

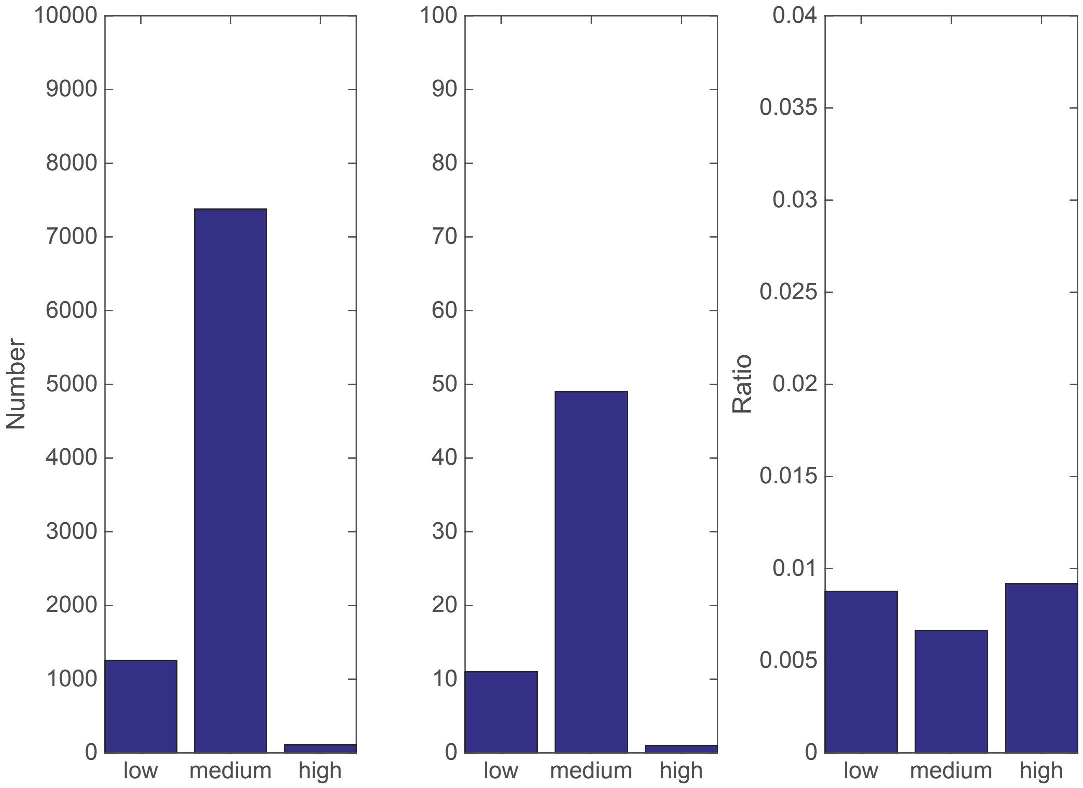


Figure 3.4: Users' shopping behaviors vs. their socioeconomic status. We divide people's socioeconomic status into three categories: high, medium and low, which is shown in x-axis. Left: browsing behavior; Middle: purchasing behavior; Right: the ratio of purchase.

vidual or group. It is often measured as a combination of education, income and occupation[62]. However, due to privacy reasons, it is difficult to map the education, income and occupation information of an individual to the online shopping data. Hence, we use the housing price[9] associated with a user's residential address (geo-location) to approximate the user's socioeconomic status, and divide people's socioeconomic status into three rough categories: high, medium and low[10]. Figure 3.4 shows the results. From the figure, we can see that people with medium socioeconomic status like browsing and purchasing on e-commerce platforms, while people with high and low socioeconomic status are much easier to conduct purchases online.

3.5 Do Users' App Usage Behaviors matter?

Users have their own shopping behaviors and preference towards the usage of apps on smart phones. Usually, a user's app usage behaviors could well profile and characterize the user. In this section, we will examine whether users' apps usage behavior will influence their shopping decisions. We will first classify all users' app

[9] The house price is crawled from Lianjia (lianjia.com), one of the most famous real estate agency platforms in China.

[10] Price lower than 40,000 CNY per m2 as low, higher than 70,000 as high and others as medium.

usage behaviors and then check the correlation between app usage behaviors and shopping behaviors.

What Apps Do People Always Use? According to a report by Nielsen[11], users spent most of their time (84%) on smart phones on just 5 non-native apps. The five apps vary from person to person and show personality of this user. For some user, their top five could include social media or gaming, while others may spend more time in instant messaging. To this end, we analyze one user's app usage behavior using his five most frequent used ones.

However, due to the limitation in data availability, we cannot directly know which apps people are using with users' mobile visiting records, based on the available information concerning the urls of web visiting with domain names. Thus, we need first parse these records with domains names in order to understand the apps people are using.

In total, we obtain 8,898 unique domain names for various apps from 12,385 mobile users. We then cluster these domain names into several clusters using DBSCAN method[17] under the Levenshtein Distances[33].

The 8,898 domain names are clustered into 393 clusters. We then manually labeled these clusters with the reference to app names in Xiaomi app store, one of the largest mobile app store in China. In this way, we get the apps people use from their mobile internet usage records.

Users' App Usage Behaviors vs. Shopping Behaviors. Based on the apps' functionality, we classify all users' apps usage into three categories: Human-oriented apps, Utility-oriented apps, and Entertainment-oriented apps. Human-oriented apps represents apps that serves people's basic needs in daily lives such as "Shopping", "Health" and "Lifestyle", etc. Utility-oriented apps are for utility perspective such as "Travel" and "Photography", while entertainment-oriented apps included apps for leisure such as "Games". We then check whether people's app usage behaviors will influence their shopping behaviors.

Since JD and Taobao are the two major online shopping platforms in China, we only consider the correlation between users' apps usage behaviors and their shopping decisions on these two platforms. Fig. 3.5 shows the results. From this figure, we can see that users who prefer human-oriented apps are more likely to buy goods online as they pay more attention to "Shopping".

[11] https://techcrunch.com/2015/06/22/consumers-spend-85-of-time-on-smartphones- in-apps-but-only-5-apps-see-heavy-use/

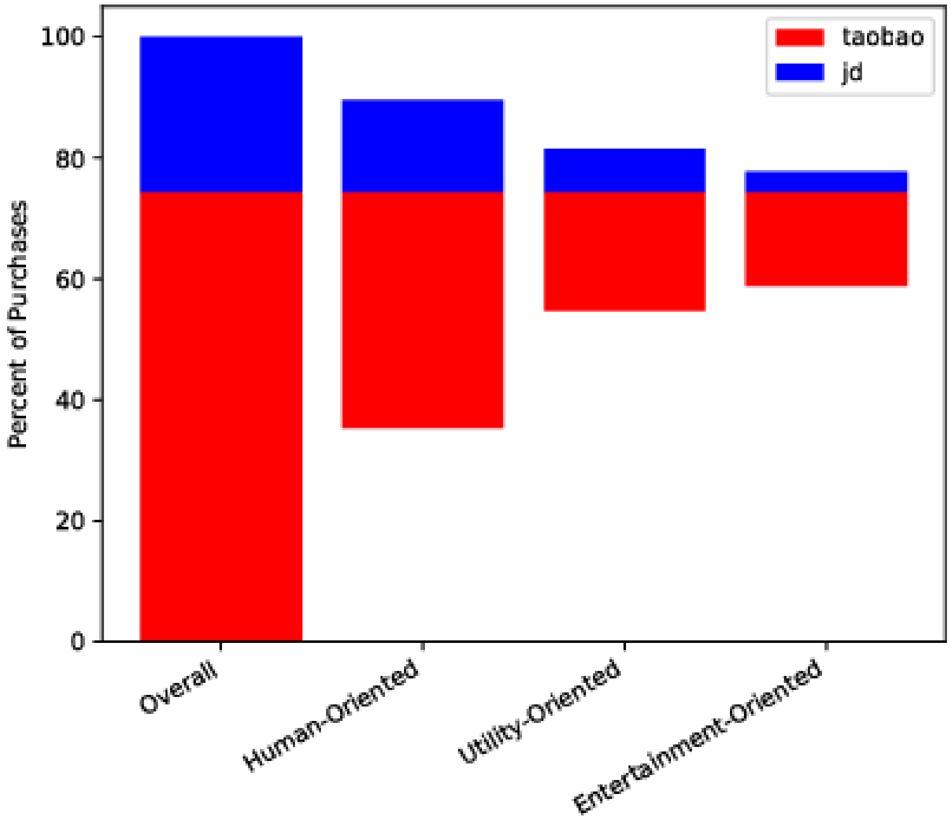


Figure 3.5: Correlations between users' apps usage behaviors and their shopping decisions. X-axis: overall cases and three categories. Y-axis: Percent of purchases.

3.6 How Long It Takes a User to Make His Decision to Purchase?

In this part, we would like to focus on the question regarding "how long it takes one user to buy a product?". For simplicity's sake, we assume that each purchase is independent. That is, a user will start a new purchase only after he ends the last one. Based on our observations, more than 96% of purchases take less than 4 days to carry out, so that we have focused on a 4-day time period before each purchase. We divide the time period into several time frames, namely 0-10m, 10m-30m, 30m-1h, 1h-2h, 2h-4h, 4h-8h, 8h-16h, 16h-32h, 32h-64h, 64h-96h. We then observe users' browsing records in each time frame. For case study, we only consider the largest two shopping platforms, Taobao and JD.

3.6.1 Time for Decision Making

Fig. 3.6 shows users' page visiting counts in each time frame. From the figure we can see that when people do shopping on Taobao, they visit the shopping pages

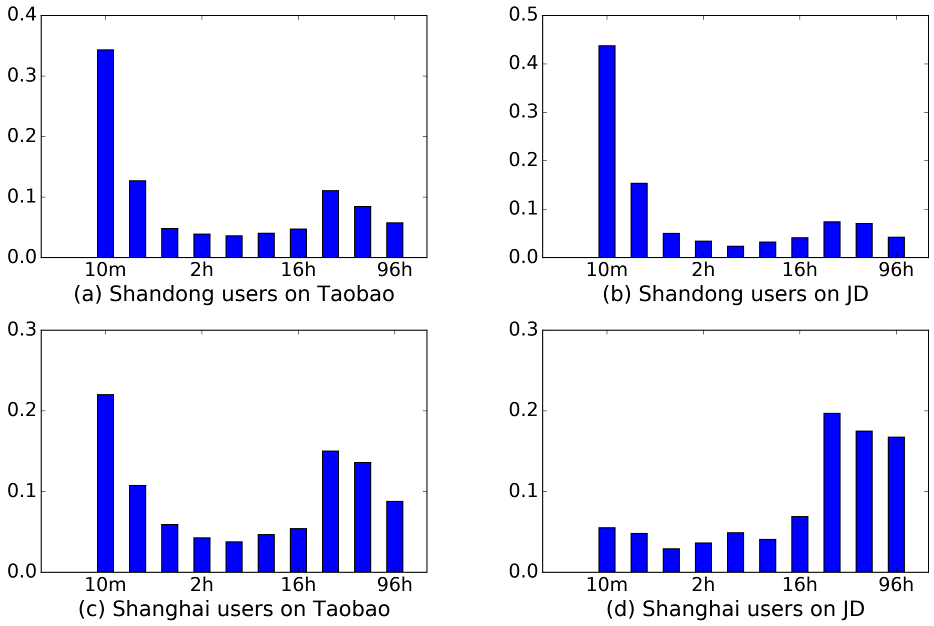


Figure 3.6: How much time it takes you for one successful purchase? X-axis: time frame 0-10m, 10m-30m, 30m-1h, 1h- 2h, 2h-4h, 4h-8h, 8h-16h, 16h-32h, 32h-64h, 64h-96h; Y-axis: browsing ratios.

frequently, which is shortly before each purchase. If a user only browses product pages shortly before his final purchase without any previous visits, we can say that he is quick purchaser as he usually spends little time on thinking about the purchase.

However, when shopping on JD, Shandong users tend to make quick decisions, whereas Shanghai users tend to spend more time thinking about their shopping purchases.

3.6.2 Group of People that Have Similar Shopping Decision Making Behaviors

We already know that some users are spend little time on making purchases while others need more time to think about their purchases. For sellers, this information can assist them in further developing more personalized marketing strategies. According to our observations, if a consumer tends to visit a web page more frequently, he is likely to make a final purchase more successfully. In order to explore this feature more closely, we divide the 4-day time period into the following time frames in an exponentially increasing manner 0-10m, 10m-20m, 20m-30m, 30m-40m, 40m-50m, 50m-1h, 1h-2h, 2h-4h, 4h-6h, 6h,12h, 12h-18h, 18h-24h, 24h- 36h, 36h-48h, 48h-60h, 60h-72h, 72h-84h, 84h-96h. Drawing on these browsing behaviors in each time frame,

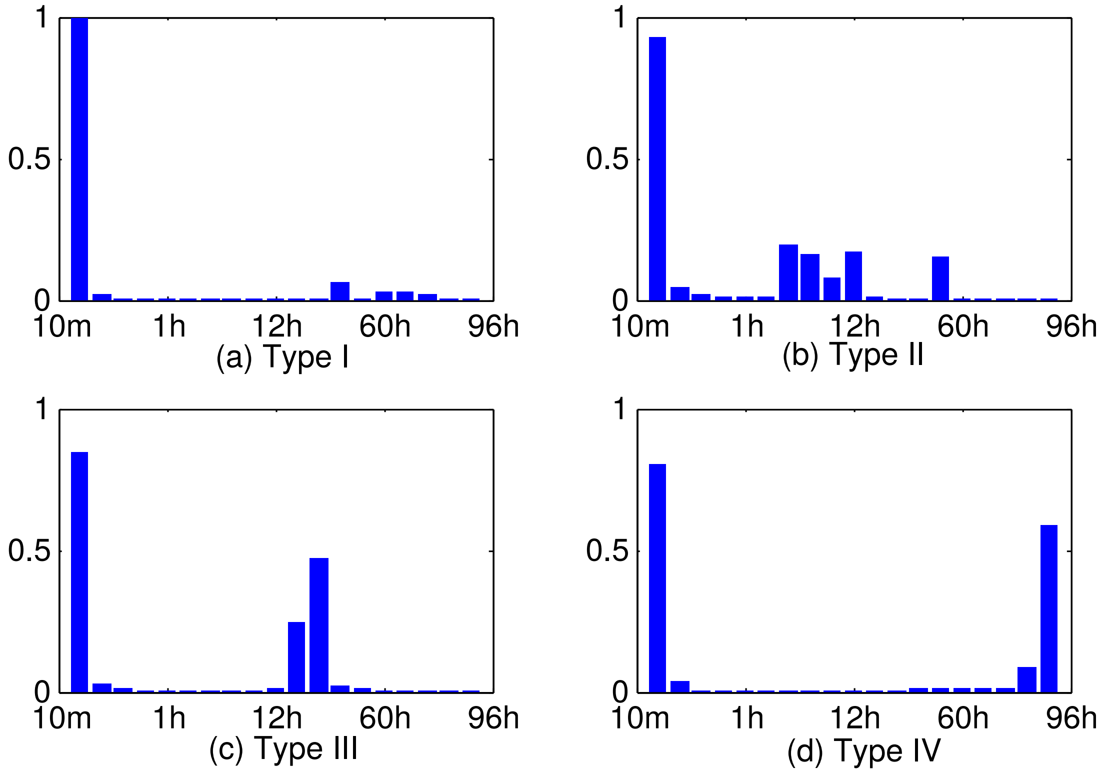


Figure 3.7: Illustration of four types of users based on their browsing behaviors. X-axis: time frame 0-10m, 10m-20m, 20m-30m, 30m-40m, 40m-50m, 50m-1h, 1h-2h, 2h-4h, 4h-6h, 6h,12h, 12h-18h, 18h-24h, 24h-36h, 36h-48h, 48h-60h, 60h-72h, 72h-84h, 84h-96h; Y-axis: browsing ratios. (a) Quick purchaser; (b) Hesitant user; (c) Short-term decision-maker; (d) Long-term decision-maker.

we can cluster users into several groups. We use K-Spectral Centroid (K-SC) clustering algorithm [29] and set the number of cluster as 4. Fig. 3.7 illustrates these clusters.

As illustrated in the figure, the vast majority of users tend to visit selected pages for 30 minutes before a final purchase. This makes sense since we often check the status of a product before a final order is placed.

The four types of users are as follows:

Type I – users will browse the pages for less than 30 minutes before a purchase. We refer to these users as quick purchasers. Over 60% of users are such "quick purchasers".

Type II – users will sporadically keep returning to shopping pages, as they have a hard time during making decisions. We refer to these users as "hesitant users".

Type III – users generally browse the pages for 12 hours before their purchase, so that we might assume that they spend about half a day to make a decision. We refer to these users as "short-term decision-makers".

Table 3.3: Types of Users

Type ID	Type Name	Percentage
I	Quick purchaser	61.20%
II	Hesitant user	21.93%
III	Short-term decision-maker	9.53%
IV	Long-term decision-maker	7.24%

Type IV – users who need nearly 4 days to make a decision, thus we refer to them as "long-term decision-makers".

Among all 4 types, quick decision-makers comprise the largest group, which is around 60% of users, followed by hesitant users, which comprises about 20% of consumers. The remaining 20% of users belong to either of the other two user types. These four clusters of users and their distribution is summarized in Table 3.3.

3.7 Are Consumers Loyal?

Many users choose to visit their preferred shopping platforms and do not want to try others, while other users will move across different platforms to search for the best deals. If users continually visit their preferred shopping platforms, we refer to these users as loyal users. In this section, we will answer the question "Are users loyal to certain shopping platforms and to which extent?". In other words, to which extent do users use the same shopping platform and to which extent do they move across different platforms in search of the best deal?

What do people usually do during one purchase? According to users' shopping behaviors, it is possible to identify multiple behavior patterns that take place during purchases. Unlike traditional frequent pattern mining scenarios where each item in one transaction may appear only once, online shopping behaviors tend to repeat themselves. For example, before a user purchases on Taobao, he browses 20 pages on Taobao and 10 pages on JD. Mining frequent patterns with repeated items makes this problem more complicated to analyze.

Leveraging the EFIM (EFficient high-utility Itemset Mining) algorithm[72], an efficient solution to one of the extension problems of frequent pattern mining at linear time with low memory, we are able to discover high-utility item-sets (i.e., group of items) in our mobile shopping transaction data containing utility information. The utility information usually refers to quantities and unit price for each item.

In this case, we consider browsing and purchasing behaviors as items with quantities, and the unit price for each item is the same. That means, it is possible to determine multiple behavior patterns which occur most often.

Table 3.4: The top 12 behavior patterns

Pattern ID	Behaviors patterns
1	JD_browsing
2	JD_browsing, JD_purchasing
3	Taobao_browsing, Taobao_purchasing
4	Taobao_browsing
5	Suning_browsing, Suning_purchasing
6	Vip_browsing
7	Suning_browsing
8	Taobao_browsing, JD_browsing
9	Taobao_browsing, Taobao_purchasing, JD_browsing
10	Vip_browsing, JD_browsing
11	Vip_browsing, Vip_purchasing
12	Suning_browsing, JD_browsing

Table 3.4 shows the results. We list the top 12 behavior patterns that users tend to exhibit. From the table, we can see that most users remain on the same platform, which demonstrates a certain amount of loyalty to certain shopping platforms. In addition, some users will browse shopping pages without purchasing anything, especially on JD and Taobao. There are also plenty of users who browse pages across multiple platforms to select the best products.

3.7.1 Are Users Loyal to Shopping Platforms?

Due to the existence of numerous online shopping platforms, people now have multiple choices and would either choose different shopping platforms due to complex reasons, such as nice price, good service and sales, or just stay in one platform. Here we will look from the distribution of how many shopping platforms a user will use to simply answer the question whether users are loyal to shopping platforms first.

From Fig. 3.8, we can see that around 67% users only visited one shopping platform in about one week, which also shows users' loyalty to shopping platforms to some extent.

3.7.2 To Which Extent Are Users Loyal to Shopping Platforms?

From the previous subsection, we know that to some extent users are loyal to shopping platforms. But the degree of loyalty to these platforms is still unknown. In order to answer this query, we will attempt to build a model to calculate users' loyalty to shopping platforms.

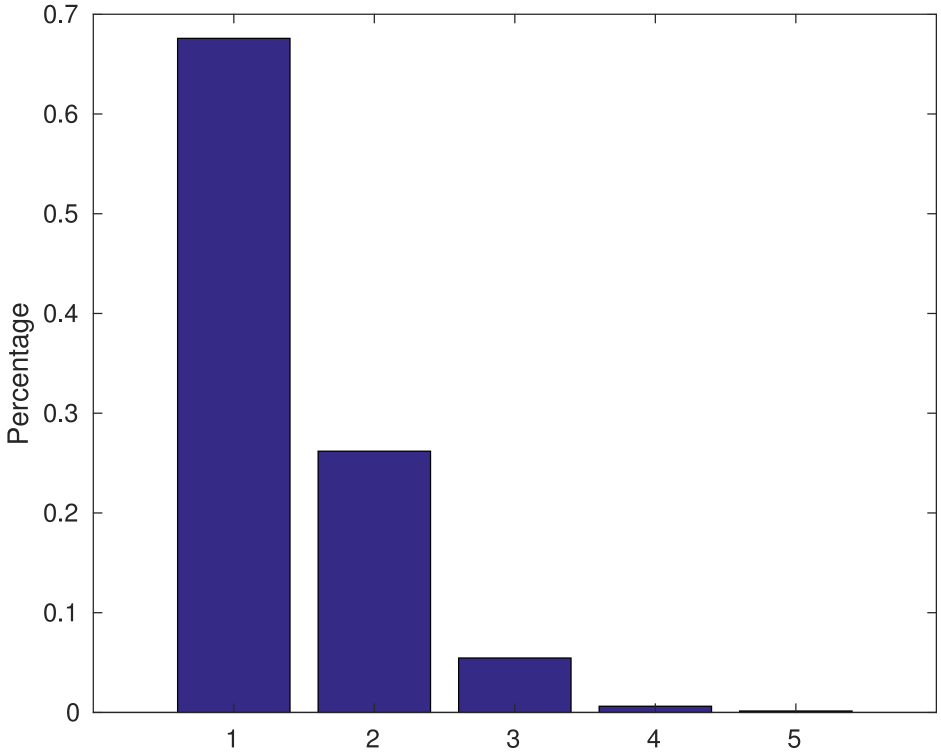


Figure 3.8: Distribution of users who have cross-platform behaviors. X-axis: The number of platforms a user has visited. Y-axis: Percentage.

Oliver[42] defines brand loyalty as a deeply held commitment to re-buy or re-patronize a preferred product/service consistently in the future, thereby causing repetitive same brand or same brand-set purchasing, despite situational influences and marketing efforts having the potential to cause switching behavior. Thus we consider two types of loyalty during the whole process according to this definition. Firstly, a user's browsing loyalty for each purchase. Before each purchase, a user is free to browse pages on any platform. The browsing loyalty refers to whether a user will only browse a single platform for a purchase or whether a user may move across different platforms to search for the best deals. Secondly, a user's purchasing loyalty in all his purchases, that is to say, whether a user would buy products on the same platform, or whether a user will buy goods from different platforms, such as on Taobao the first time and later place his order on JD.

In order to model a user's browsing loyalty, we consider the pages a user visits during a purchase. We denote pin as the probability that a user is browsing on the platform i for the purchase n, whereas p_i can be calculated directly from a user's browsing history. With respect to a user's purchasing loyalty, we denote q_i as the probability that a user purchases on the platform i. As a result, we can define user loyalty L to shopping platforms as follows:

$$L = \frac{1}{N}\sum_{n=1}^{N} q_i p_{in} \tag{3.1}$$

where n means the nth purchase, N is the total number of purchases a user has and i is the platform the user makes his purchase on. For example, if a user has two successful purchases, he made the first purchase on Taobao and the second one on JD. For the first purchase, he browsed 10 shopping pages in total, among which he browsed 2 pages on Taobao. For the second purchase, he also browsed 10 pages and 4 pages of them are on JD. So his loyalty is $1/2 \times (0.5 \times 0.2 + 0.5 \times 0.4) = 0.15$.

We consider two extreme cases to validate our model. In the first case, we presume that a user is quite loyal to one shopping platform and that he carries out all his browsing and purchases on the same platform. Thus, his loyalty is 1. In the second case, we presume that a user carries out 5 successful purchases, once each on the platforms mentioned above. For each purchase he browsed 10 pages, but only 1 page on the platform where he made his purchase. Thus his loyalty is $0.2 \times (0.1 \times 0.2) \times 5 = 0.02$, which is quite low. This confirms our assumption.

According to the loyalty definition, we calculate all users' loyalty in our data. We find that more than 99% of users are loyal with loyalty greater than 0.99.

3.8 The Predictability of Consumers' Purchasing Behaviors

Experimental Setup. In this section, we explore whether the consumer's shopping behaviors are predictable. From previous sections, we have known that most of the users are loyal to the platforms they visited. We can see in Fig. 3.8 that around 67% users have only visited one platform. In other words, if we assume that one user would choose to use the most frequent visited platform, we can get the prediction performance with accuracy up to at least 67%. Could the consumer behaviors be better predictable?

With characteristics that lead to consumers' purchase learned from previous sections, we build a prediction model to predict which platform a user would like to purchase on at certain time in this section.

Our problem can be formalized in the following way: Given a number of users who have scanned or purchased in shopping platforms such as "Taobao", "Jingdong", "Dangdang", "Suning", "Vip", we have all their past browsing and purchasing records, our goal is to predict which platform the user will use next time to make his purchase?

Table 3.5: Types of Users

Algorithm	Parameter Settings
J48	confidence factor C =0.5, instance leaf M = 20%
RandomForest	trees = 100, features per tree = 6
NaiveBayes	Default
SVM	cost=256, gamma=0.00048
LSTM	learning rate = 0.02, # neurons = 256, batch size = 50, loss function = Softmax cross entrop, optimizer = Adam optimizer

To address this issue within our dataset, we extract 265,619 records for more than 65,000 users from Apr. 20, 2016 to Aug. 14, 2016. After eliminating the records with no purchases or only with one purchase, we get a sample dataset that contains 102,517 records of 12,384 users.

We randomly select half of all users as training and validation set, which uncovers the most suitable experimental parameters shown in Table 5. We then use the rest half of the users as a test dataset. That is to say, there are 6,192 users in the training set, and the left 6,192 users in the testing set for all experiments. We have utilized Weka and Tensorflow to train and predict using following algorithms: J48 (C4.5), RandomForest, NaiveBayes, SVM and Long Short-Term Memory (LSTM) Network. All experiments are performed on a PC running ubuntu 16.04 with an Intel Core i5 CPU (2.8GHz) and 8GB memory.

Features. The features used in our model are extracted based on the observations from previous sections. They can be summarized as follows, which can be divided into four types of features.

- **Temporal feature:** We build a temporal feature to show the time correlation. In practise, we split a day into three periods: sleeping hours (1am - 9am), active hours (9am - 5pm) and spare hours (5pm - 1am). Then we consider users' scan and purchase behaviors in each periods as features.
- **Loyalty feature:** We construct a loyalty feature follows Equation 3.1 as described in 3.7.
- Profiling feature: We extract the most frequently used apps to profile users' usage behaviors, and use them as profiling features. In practice, users' usage behaviors have been classified into three categories: Human-oriented, Utility-oriented and Entertainment-oriented.
- Demographic feature: We consider users' demographic as one feature, such as his location.

Table 3.6: Prediction performance

Algorithm	Precision	Recall	F1	Accuracy
J48	0.982	0.978	0.979	0.978
RandomForest	0.980	0.980	0.980	0.979
NaiveBayes	0.944	0.943	0.943	0.942
SVM	0.975	0.975	0.975	0.976
LSTM	0.974	0.974	0.974	0.977

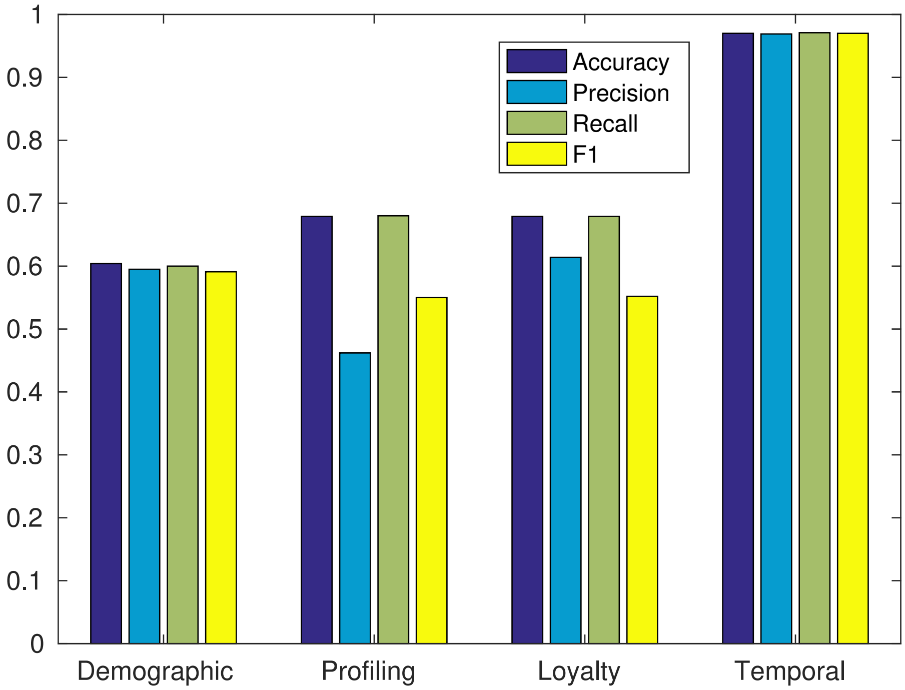


Figure 3.9: Factor contribution analysis. X-axis represents the four types of features we considered in our prediction frame- work. Y-axis means the prediction performance.

Prediction Performance. The prediction results are shown in Table 3.6. As we can see, all tested algorithms perform similar performance with F1 score higher than 0.9, which shows that consumers' cross-platform shopping behaviors are predictable and our prediction performance is sound (far better than random guess). We examine the contribution of four types of features and check each features' prediction power separately. Fig. 3.9 shows the results. We can see from the figure that temporal feature has the strongest prediction power.

3.9 Related Work

In the past few years mobile phones have witnessed a remarkable evolution and explosive popularization[46]. Meanwhile, e-commerce also has a prosperous development and drastically changed traditional commercial relationships, as well as the shopping process for the fast-growing online shoppers[5]. With a smart phone at hand, the consumer can check the details of products, compare the prices across various e-commerce platforms, save items into charts and enjoy a great many benefits such as personalization from merchants and recommendation from social networks[9, 21, 32, 69]. At the age of information explosion, the complexity of users' on- line behaviors is increasing while understanding the targets and preferences behind a user's online behavior can provide valuable information for content providers, such as improving user satisfaction by personalizing their contents in search engines and e-commerce web sites[11, 14]. Understanding the consumer buying process can make a difference between success and failure in consumer marketing strategies[57]. On the basis, e-commerce companies can improve their service quality to keep competition ability.

There is plenty of work investigating factors that contribute to successful purchasing results, such as motivation[41] and demographics (e.g. gender, age and residence)[23, 26, 49, 50]. One important research issue is to identify consumers' purchasing intentions from multiple datasets, such as web search contents, social network posts and mobile data cookies. Early researches focused predominantly on identifying web search goals in order to derive commercial information. Dai et al.[14] focused on users' commercial intention from search queries and webpages, i.e., when a user submitted a query or browsed a webpage, whether he/she was about to commit or in the middle of a commercial activity, such as purchase, auction, selling, paid service, etc. Guo et al.[20] proposed an improved method for the detection of a searcher's intention and studied an important practical application predicting ad clicks for a given search. Guo et al.[21] studied the relationship between social networks and consumer behavior in order to know how individuals' commercial transactions were embedded in their social graphs. Zhang et al.[69] provided an extensive analysis on how users' Facebook profile information correlated to their purchases on eBay, and analyzed the performance of different feature sets and learning algorithms for the task of purchase behavior prediction.

Most previous researches only focused on single e-commerce platform[69]. However, users usually choose to move across various e-commerce platforms to compare the quality, price and service etc. to make their best choices[61], which haven't yet been well explored. The dramatic increase in mobile datasets provide new potential in identifying consumers' purchasing intentions and modelling their behaviors. One related work is[36], where Caroline et al. performed a large-scale cross-platform longitudinal study of user purchase intent and how it evolved. More specifically, the authors tried to understand consumer behaviors leading to successful purchasing across different platforms. However, as the dataset they used was entirely based on Pinterest, a content discovery application which allows users to share their shopping fruits on Pinterest rather than real-world e-commerce platforms, their study may not directly reflect the users' shopping behaviors and therefore reflect biased results, since not everyone is willing to share with others all his shopping behaviors and fruits.

Moreover, users' profiles such as their culture, social and ethical would also influence their behaviors [57]. Researchers pay more attention to users profiling[4, 16, 22, 24, 69] and apply them in many areas, such as personalization and recommender systems[1, 28, 39, 54]. In this chapter, we also consider whether users' app usage behaviors would influence their shopping decisions.

To the best of our knowledge, our study is the first one to explore real-world consumer behaviors across diverse e-commerce platforms in depth to identify users' commercial intention and purchasing patterns. We take telecom mobile usage data as our source dataset which comprises comprehensive records of users' shopping platform usage. By analyzing this data, we are able to analyze users' online shopping behaviors across different e-commerce platforms.

3.10 Conclusion

In this section, we conducted a comprehensive study on cross-platform mobile shopping behaviors using a real-world, large-scale mobile dataset. We found that most people are loyal to the shopping platforms they visit and they do not move across platforms to select goods they want to buy. In addition, most people are quick purchaser who complete a purchase in less than 30 minutes after first browsing for the item of purchase. Besides, with the patterns learned from this paper, we examine the predictability of users' shopping behaviors on multiple platforms. These findings could provide useful insights for future e-business strategies.

To the best of our knowledge, this is the first work studying cross-platform consumer behavior in depth. Future studies could analyze users' social relationships and how they influence users' shopping decisions.

3.11 Acknowledgements

The work of this section is supported in part by the National Key Research and Development Program of China under grant No.2016QY02D0302, and National Natural Science Foundation of China under grant No.61433019 and No.U1435217.

Chapter 4

Consumer Behavior Analysis and Prediction of Takeout Food Purchasing

With the popularization of the mobile Internet and the prevalence of delivery service, Online Takeout Ordering & Delivery (OTOD) using Apps from smart phones or websites from PC has become an emerging service and prosperous industry(e.g., KFC delivery). In order to improve the quality of service and recommendation personalization, we tried to find the key factors leading to a successful purchasing of takeout food in this paper. We collected Internet access records related to OTOD service of 34,845 users with a time duration of nearly four months. At first, We did a preliminary study on users' daily and periodic purchasing behaviors of takeout food. Then we combine the demographic information and location information with the purchasing activities to find the most potential purchasing groups of takeout food. Based on the features extracted from historical purchasing records, demographic information and location information, we use several popular machine learning methods to predict the future purchasing activities within a specific time. The experiments show that our extracted features can be well used for the takeout food purchasing prediction problem.

4.1 Introduction

In recent years, with the prevalence of the Internet and the increasing life pace, more and more people tend to use the Online Takeout Ordering & Delivery (OTOD) service through Apps from smart phones or websites from PCs to buy food for lunches or dinners. The online food delivery market hits 204.6 billion Chinese yuan ($31.9 billion) in 2017, 23 percent more than the previous year, according to a report by Meituan Waimai[1], a major food delivery firm in China. Using the OTOD service, users could receive their takeout food delivered by the restaurant staff very quickly and

[1] http://waimai.meituan.com/

conveniently after placing the orders. Consequently, some new platforms are developed to provide the OTOD service, such as the three most popular platforms, Baidu Waimai[2], Meituan Waimai and ele.me[3] which occupy nearly 90% of the takeout food market share in China. In general, the OTOD service is convenient and time-saving especially for people who are busy or just want to stay at home or in office. Identifying the potential successful purchasers of takeout food and making personalized recommendation on the basis can help the merchants better to prepare the food and improve the delivery efficiency.

While the online food delivery market has seen rapid expansion, there is still room for businesses to grow as food delivery accounts for a relatively small portion of the total catering industry. Identifying the most potential customers make great sense for the platforms and merchants to enlarge their market share and profit. Intuitively, we believe that demographic factors(e.g., gender, age and occupation, etc.) and spatiotemporal factors(e.g., weekdays or weekends, home or office, etc.) have great influence on the takeout food purchasing since different groups have different concern and attitude to the takeout food. On the basis of these assumptions, we analyzed our data and found some quite interesting phenomenons which seem to be different from our inertial thinking.

Repeat purchasing prediction [34, 66] and recommendation systems [15, 71] have been widely researched with the prosperous development of E-commerce. The repeat buyer prediction problem can be formulated as a typical classification problem and model training of this task is not much different from that of other classification tasks. Instead, feature engineering is the main component. As an emerging industry, the prediction and recommendation problems for OTOD service are quite different from those for traditional E-commerce because of the property difference of the products. You may do only one shopping for clothes, shoes or electronics within several months while you have to solve the meal problem everyday. Our main contribution is try to find the key factors influencing the takeout food purchasing. We consider the historical records as well as consumers' profiling information for the future purchasing prediction for OTOD service.

To the best of our knowledge, this is the first work that thoroughly studies the consumer behavior analysis and prediction problems in the takeout food industry. We will describe how to generate various types of features from user activity log data and study the importance of these features through extensive experiments. The features we generated can be used in purchasing behavior prediction and product recommendation. We hope that our work can be valuable for data science practitioners, who need to develop solutions for prediction and recommendation tasks in takeout food markets.

[2] http://waimai.baidu.com/waimai?qt=find

[3] https://www.ele.me/home/

In general, our contributions of this paper are as following:

(1) We present a statistic results of consumers' long-term purchasing behaviors related to takeout food using data mining. We collected nearly 4 months takeout food access and purchasing records of more than 10, 000 users and extract the purchasing actions from them.

(2) We try to find the relationship between the demographic factors(e.g., gender and age,etc.) and purchasing actions of takeout food.

(3) We extract the location information embedded in the records of takeout food purchasing activities to infer the possible occupations and then study their different attitude and purchasing actions of takeout food. On the basis, we find the most potential groups tending to purchase takeout food.

(4) We use machine learning to predict the future repeat purchasing of takeout food. We combine the demographic features, historical records and spatiotemporal features together to predict consumers' future purchasing actions within one week, two weeks, three weeks and one month.

4.2 Related Work

4.2.1 Sales Forecasting

Traditional purchase prediction task has aimed to forecast future sales in offline stores. For example, sales of newspapers [3], restaurant food [73, 38], and theater ticket [6] are estimated to identify their daily demands. For that, they model past purchase patterns in repetitive daily routines [25, 7] and leverage surrounding features correlated with purchase such as day of the week and weather. Since micro blogging and social network services have been popular, web-based features have been also adopted as a predictive evidence for purchase prediction. For example, assuming that product sales are correlated with product popularity on the web, the future sales of movies [65] and smart phones [30] are predicted by using micro blog posts, online reviews, and ratings. However, these traditional approaches cannot be applied to the online purchase prediction problem in e-commerce sites. First, existing features extracted from the physical world become far less correlated with online purchases as e-commerce sites are easier to visit than offline stores. For example, day of the week in online purchases [44] are not anymore as predictive as those in offline purchases. Second, the local nature of e-commerce sites makes it difficult to use general web signals, such as reviews [13] and ratings obtained from external web databases. Such signal within an e-commerce site suffers from its sparseness and thus cannot be assumed to exist in all e-commerce sites, in retargeting prediction. Recently, researchers investigated a series of factors leading to successful purchasing results in E-commerce, including motivations, recommendations [51, 64], personalization [36], as well as demographic factors, such as gender, age and income [23].

4.2.2 Takeout food Purchasing

Classical research about takeout food mainly focused on the nutrition and safety of the food [8]. The past five years have witnessed the prosperous development of the food delivery service in China. As mentioned before, the online food delivery market hit $38.411 billion in 2018, nearly 23% more than the previous year, according to a report by Meituan Waimai[4], a major food delivery firm in China. Currently, studies are mainly related to food delivery network construction and optimization. The problem is formulated into the object delivery problem and a series of research has been done [60, 12, 35]. Yeo *et al.* [67] examined the structural relationship between convenience motivation, post-usage usefulness, hedonic motivation, price saving orientation, time saving orientation, prior online purchase experience, consumer attitude and behavioral intention towards online food delivery services systematically. The study proposes an integrative theoretical research model based on the Contingency Framework and Extended Model of IT Continuance. Though there have already been many studies indicating people have great interest on takeout food in spite of some health considerations, little work has been done about the system research of consumer behavior analysis and repeat prediction at the level of single user.

Due to the regular purchasing behavior(three meals per day), repeat buyer prediction can be done on the basis. Considering the attitudes to takeout food are different among different user groups, we can further analyze the purchasing behaviors of different groups of people. For example, doctors and students have different attitudes to takeout food and daily timetable. As a result, their purchasing behaviors to takeout food are assumed to be different empirically. However, the actual result may be different considering various factors. Take more information about takeout food purchasing into consideration, such as location and user profiles, much work can be done to analyze the consuming behaviors of different people and predict repeat purchasing actions in the future. Unfortunately, the studies related to such aspects are limit. Our work can be seen as an attempt to analyze consumer behaviors and make repeat purchasing predictions for specific consumers based on the analysis.

4.3 Dataset

4.3.1 Data Collection

The dataset is collected from a log of anonymous browsing records of mobile usage in cellular environments provided by one of the three major mobile telecom operators

[4] http://waimai.meituan.com/

Table 4.1: Users selected for the analysis and prediction

Category	Count	Comments
Telecom Users	125,753	Randomly selected in Shanghai
Sina Weibo Users	125,753	Related to telecom users
Selected Telecom Users	34,845	Purchasing takeout food
Selected Sina Weibo Users	57,643	Profile information completed
Combined Users	16,840	Intersection of selected users
Final Users	11,265	Used for training and testing

in China. It contains the mobile usage data for 36,325 users from Shanghai, the most populous metropolitan in the world (and also the commercial and financial center of mainland China), over a period of roughly 3.5 months from November 1, 2016 to February 11, 2017. Each record contains all the information of an Internet access and we just abstract the anonymous ID of the mobile user, start time of the Internet access, destination URL and reference URL of the access.

As for the demographic information, we collected the Sina Weibo profiles of the users whose Weibo IDs are combined with the mobile phone numbers. The profiles contain the demographic information provided by the users themselves, including gender, age, address, number of posts, number of fans and number of followees, etc.

4.3.2 Data Preprocessing

Among the randomly selected 125,753 users, there are 34,845 users who have takeout purchasing experience, which indicates that more than 25% people using mobile phones purchase takeout food within nearly four months at least once. As for demographic information, some users' age information in the Sina Weibo profiles is missing or fault obviously. We get 57,643 Sina Weibo users whose profile information is complete. Combining the Telecom users who have takeout food purchasing experience and the Sina Weibo users whose profile information is complete, we get 16,840 users for the final analysis and experiments. The statistical results of the data set is shown as Table 4.1

The collected data is heterogeneous and noisy, including all the active and passive Internet access records. In order to study consumer behavior using these various mobile Internet access records, we need to do the data cleaning work first. There are mainly three takeout food ordering platforms in China, namely Meituan Waimai, Baidu Waimai and Ele Me. Due to the multiple interaction rounds and references of web service requests and response queries on various platforms, there are plenty of redundant records. To identify the unique actions from many redundant interaction records, we abstracted the order IDs and only counted each page visit once for the same order. We also extracted the location information embedded in the URL record and decoded it into various function zones, such as hotel, office and residence, etc.

4.3.3 Feature Extraction

In order to explore the latent relationship between the user profiles and their behavior in booking takeout food, we selected some features extracted from both Sina Weibo data and mobile user logs for takeout food browsing. We categorized the location of user into residence or other places and flag the residence as 1. To make prediction for where are the users booking takeout food, empirically, we considered it is related to the time, weekday and job of users. As time and date are contained in the browsing logs, we partitioned the time into three time frame breakfast, lunch and dinner flagged as numbers. For weekday, we transformed the date and also flagged it from 0 to 6. We didn't have the data of users' job, so we chose age as the feature which can partly represent the status of people. To predict the possibility of future takeout food booking, we select the former purchasing in the last 70 days as feature. Then we also want to know the influence of users' age and gender. As age is in a huge scale, however the past booking data is binary, we applied min-max scaling method to standardize the age in 0 to 1.

4.4 Demographic Factors

Among the selected 16,840 users, the ratio of female consumers is 56.26% while the ratio of male consumers is 43.74%. It seems that more women have interest in purchasing takeout food according to the gender ratio in our randomly selected consumers and the purchasing ratio is exactly a little higher. The 58.03% female consumers made 60.88% successful purchasing while the 41.97% male consumers contributed to 39.12% successful purchasing. The statistic results indicate that women have more interests in takeout food and tend to more likely to make successful purchasing. We did an interview with randomly selected 20 people(including 10 men and 10 women, respectively) and most of them think that women care more about food than men. The gender distribution of takeout food websites access and purchase records is shown in Fig. 4.1. Female consumers have obvious more access and purchasing records for lunch while the access and purchase counts are close to each other for dinner. According to the feedback of our randomly selected interviewees, men tend to go to the company canteen for lunch while women prefer to order some takeout dishes which are more attractive and tasty. As for dinner, the gender factor has less influence on the takeout food access, while the successful purchase ratio of female consumers is still a little higher than that of male consumers. It indicates that women prefer to purchase takeout food during work time as well as tend to choose takeout food after work in the evening.

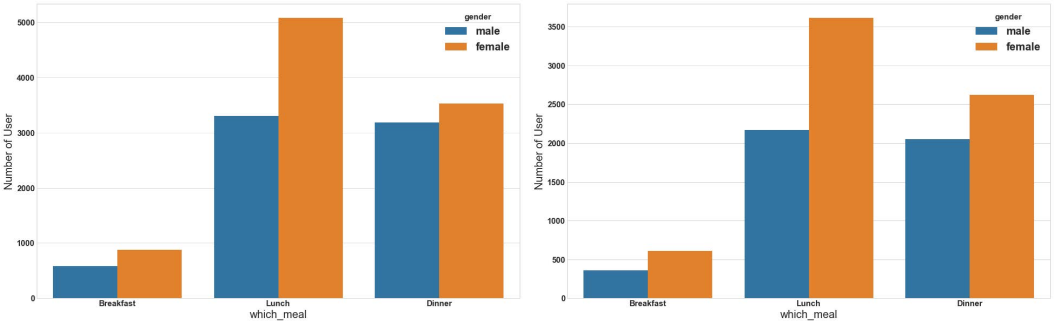


Figure 4.1: Gender distribution of selected users. The left is the distribution of users with access actions while the right is the distribution of users with purchase actions.

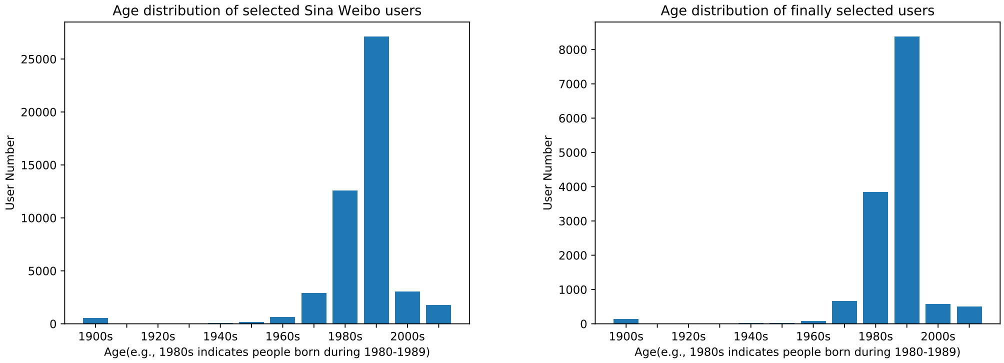


Figure 4.2: Age distribution of selected users. The left is the distribution of selected Sina Weibo users with age information while the right is the distribution of the finally selected users(combined Sina Weibo users and takeout food purchasing users) with age information.

As shown in Fig. 4.2, the age distributions of randomly selected Sina Weibo users and the finally selected users who have takeout food purchasing experience are similar, which indicates that the age distribution has no obvious influence on the takeout food purchasing. The 1980s and 1990s are the most active groups using Internet in daily life and they are the busiest working groups in the meantime. Students or career starters tend to work harder and care less about what they eat during tense study or work time. They spend more time on the virtual world of network and also enjoy its convenience. It is a great attract that they just move a figure on their mobile phones or desktops and then wait for the food delivered to their hands.

We divide the users into four main groups according to their ages, which are 15-22, 23-25, 26-28, 29-32 and 33+ years old, respectively.The access trend of these five groups are close to each other while the main difference occurs from Jan 15, 2017 to Feb 8, 2017, which is the winter holiday for undergraduate students. Most undergraduates choose to go back to home during the winter holidays, which has great influence on the takeout food purchasing behaviors. The results are shown as Fig. 4.3.

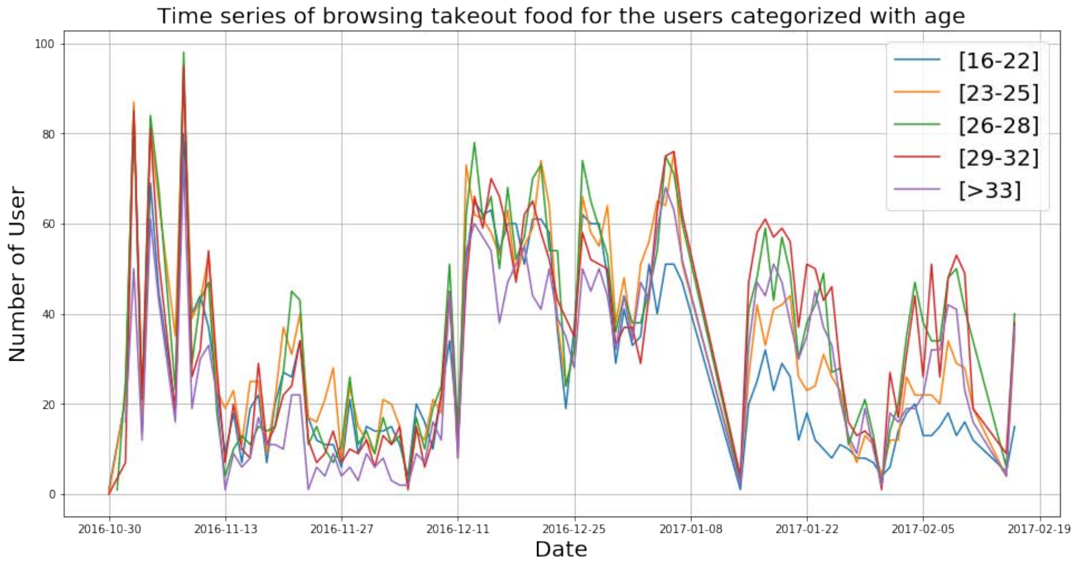


Figure 4.3: Distribution of users with different age labels during the whole period.

4.5 Spatiotemporal Factors

4.5.1 Statistic results and analysis

Our data collection lasts 102 days, including 16 Sundays, 15 Mondays and Tuesdays, as well as 14 Wednesdays, Thursdays, Fridays and Saturdays. Without loss of generality, we use the average user numbers to explain the results. As shown in Fig. 4.4, the accessing patterns for weekdays and Saturdays are quite similar while there is an obvious decrease on Sundays. With the popularization of double day weekend system in China, it seems that takeout food is more attractive to the people who work longer one week. The initially selected 34,845 users and the finally selected 16,840 users who have takeout food platforms access experience show similar purchasing trends, as shown in Fig. 4.4. There are two access peaks in a day, which are related to lunch and dinner. Especially for the lunch, many people would like to have the takeout food delivery service to enjoy the convenience and time saving advantages. As for dinner, some of the people who work late in the office or people who don't want to make dinner themselves at home also tend to have great interest in takeout food purchasing. However, more and more young people tend to enjoy the convenience instead of cooking by themselves. According to our interview, some takeout food lovers really have no time for cooking while more people just have no interest to make food by themselves.

We further analyzed the consumption behavior of three meals per day considering the weekdays and weekends and find an interesting phenomenon, as shown in Fig. 4.5. Consumers have more access and purchasing actions during weekends at home, which indicates that people have more interest in takeout food at home during weekends

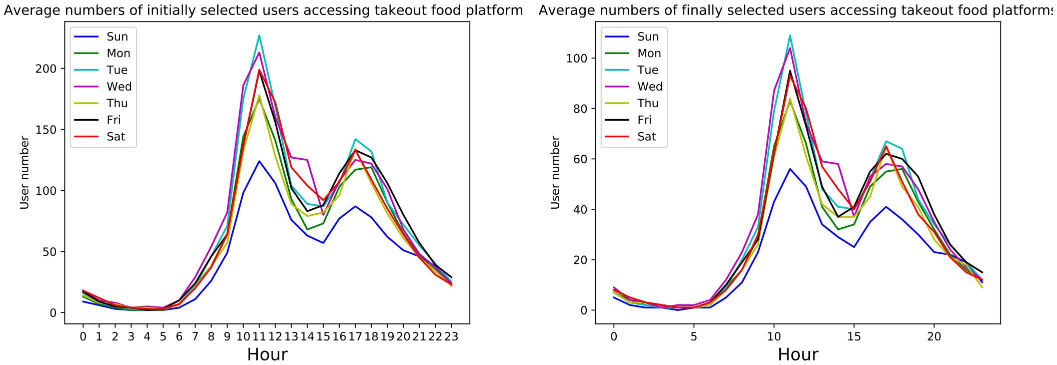


Figure 4.4: Average numbers of consumers accessing takeout food platforms

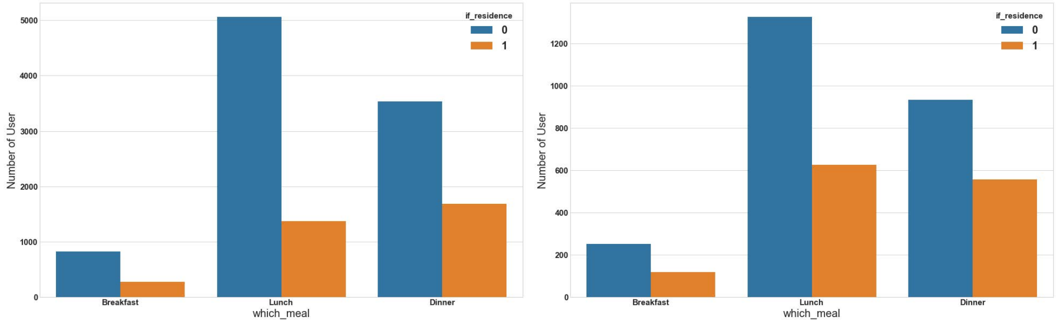


Figure 4.5: Gender distribution of selected users. The left is the distribution of users during weekdays while the right is the distribution of users during weekends, in which 1 means access and purchase at home while 0 is the opposite.

even though they have time for cooking. We also add this into the survey in our work and the results show that people tend to have lunch or dinner with their friends or partners in restaurants only when having outdoor activities during weekends. It seems that homebodies always have more interest in takeout food.

We use the POI classifications from Baidu Map[5], one of the most popular map provider in China. The 18 primary classifications and their corresponding key words are shown as Table 4.2. We made a small modification about the classifications by dividing the realty cluster into residence and office buildings. In addition, we delete the natural features cluster because there are very few such locations related to it in our dataset. The location distribution of takeout food purchasers on weekdays and weekends are shown as Fig. 4.6 and Fig. 4.7. The locations labeled "others" are those locations which are too indistinct or difficult to label them with the other location labels. For example, some locations only include the information of a road while no house number or more detailed information.

As shown in Fig. 4.6, users tend to have more takeout food purchasing actions for lunch at working places, especially for the office staff. Their work time is more regular

[5] http://lbsyun.baidu.com/index.php?title=lbscloud/poitags

Table 4.2: Purchasing location classifications based on Baidu map API

Classifications	Key words
Food	Restaurant, fast food, snack, coffee, bar, cake, dessert
Hotel	Hotel, hostel
Shopping	Shopping mall, store, supermarket, shop, market
Life Services	Communications office, post, delivery, ticket, photo studio, agent, maintenance
Beauty	Beauty, hairdressing, nail, body care
Scenic spot	Park, zoo, botanic garden, amusement, museum, church, historical relics, scenic,cultural relics, aquarium, beach
Entertainment	Resort, farmyard, cinema, KTV, theater, dance hall, bath, massage, square, games
Fitness	Stadium, gym, fitness, playground
Education & Training	University, college, school, kindergarten, education, training, library, research, study
Cultural medium	News media, publication, radio, television, exhibition, culture, art gallery
Medical treatment	Hospital, clinic, pharmacy, drugstore, medical, emergence, disease illness
Auto Service	Auto Sales, auto maintenance, auto beauty, auto parts, car rental, auto testing
Transportation	Airport, station, railway, bus, parking, port, service area, refueling
Finance	Bank, ATM, credit, investment, finance, pawn
Office	Office building
Residence	Residence, dormitory
Enterprise	Company, enterprise, factory, mine
Government	Central authority, government, administration, police, court, procuratorate, political, welfare

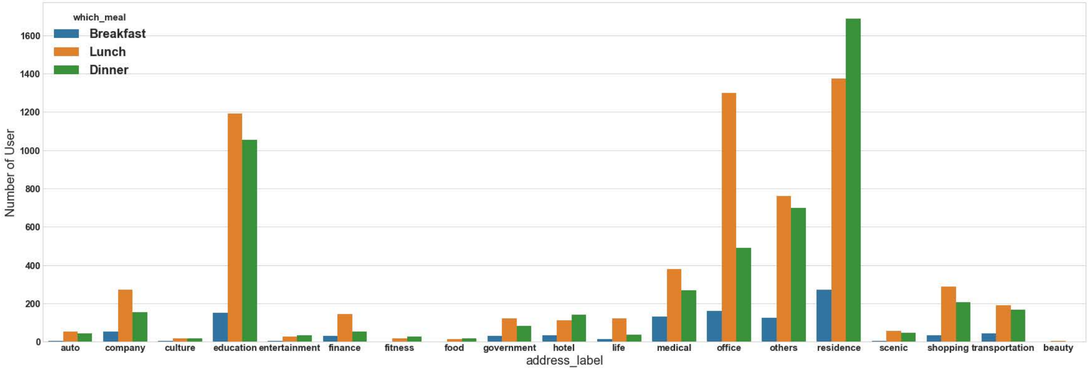


Figure 4.6: Purchasing location distribution of weekdays

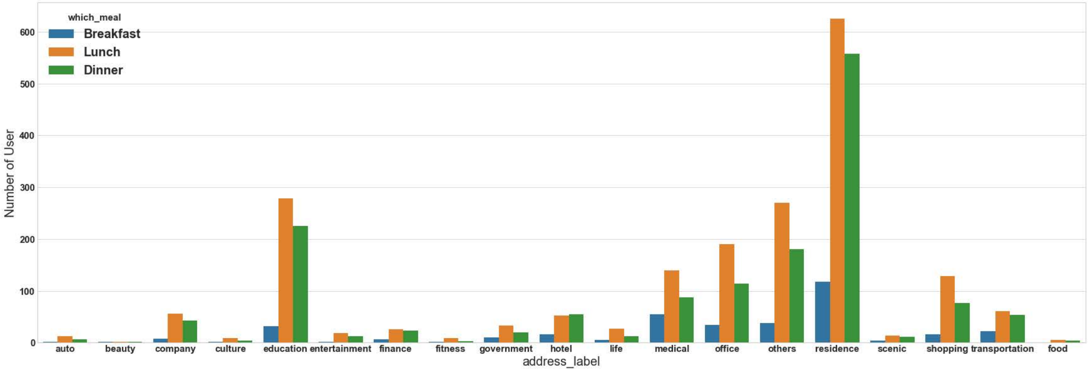


Figure 4.7: Purchasing location distribution of weekends

Table 4.3: Prediction accuracy for different time interval using various machine learning algorithms

Algorithms	k-NN	NB	DT	SVM	RF	LR
Prediction accuracy	66.00%	72.12%	71.61	72.05%	72.18%	72.13%

and their meal time is also regular consequently. For the labeled locations, residence, office and education institutions are the top 3 popular places where takeout food are purchased. As introduced in the demographic section, the main consumers of takeout food are youths under 35 years old, who have high work pressure and tend to enjoy the convenience of network and food delivery service. According to the age information, consumers in the education institutes tend to be students, which indicate that college students have special interest on takeout food even though they have enough time for meals. It seems that the university canteens need to pay more attention to attract more students by improving the food taste since they already have the advantage of price comparing with takeout food providers.

4.5.2 Purchasing location prediction

Intuitively, consumers tend to purchase lunch at working places while purchase dinner at home. Based on the purchasing day labels(weekdays and weekends) and meal labels(breakfast, lunch and dinner), we make a preliminary prediction about the purchasing location. At first, we divide the purchasing locations into two groups, residence and working places. We tried several supervised learning algorithms to validate the prediction accuracy of different time intervals, including k-nearest neighbors algorithm(k-NN), naive bayes(NB), decision tree(DT), support vector machines(SVM), random forest(RF) and logistic regression(LR). The prediction results are shown as Table 4.3. The best prediction accuracy is around 72%. It is helpful to know the people distribution to provide more personalized relevant service beyond takeout food delivery service. In the future, We will collect more data from more users and try to predict the more detailed locations of purchasers to have further understanding about consumer behaviors and better serve the potential consumers of takeout food.

Table 4.4: Prediction accuracy for different time interval using various machine learning algorithms

Algorithms	Prediction accuracy for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	82.73%	76.04%	70.20%	64.58%
NB	81.66%	76.73%	73.56%	70.39%
DT	79.24%	74.02%	70.79%	68.07%
SVM	83.00%	77.38%	74.88%	70.15%
RF	83.32%	76.82%	73.80%	69.47%
LR	83.84%	78.35%	75.09%	70.47%
Algorithms	Recall score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	54.28%	55.22%	54.59%	55.29%
NB	**64.45%**	**63.26%**	**62.24%**	**62.48%**
DT	59.57%	59.19%	58.85%	60.09%
SVM	50.78%	54.70%	57.52%	58.88%
RF	52.09%	52.93%	54.04%	56.17%
LR	58.24%	59.70%	59.69%	60.17%
Algorithms	F1 score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	54.13%	54.46%	53.87%	54.72%
NB	**65.41%**	**64.46%**	**63.17%**	**63.00%**
DT	60.29%	59.94%	59.36%	60.35%
SVM	47.02%	52.80%	56.73%	57.89%
RF	49.69%	49.43%	50.40%	52.82%
LR	59.96%	60.77%	60.10%	59.88%

4.6 Machine learning for repeat purchasing prediction

As mentioned before, we combined the Sina Weibo users with complete profile information with the users with takeout food purchasing experience to get the finally selected 16,840 users. The time interval of these final users is 100 days. According to the purchasing records, we further filter the users to adapt to the machine learning models and finally we select 11,265 users who have purchasing records in the first 70 days for the training and testing. We extracted the historical purchasing records of each user as a feature for our prediction model, shown as following.

$\mathcal{H} = \{h_1, h_2, ..., h_{100}\}, h_i \in \{0, 1\}, 1 \leq i \leq 100$,where $h_i, 1 \leq i \leq 100$ indicates the purchasing record of user i, in which $h_i = 1$ if the user has at least one purchasing record at the ith day and $h_i = 0$ otherwise. We extracted the front 70 days' records as the historical records to predict the repeat purchasing in one week, two weeks, three weeks and one month. The statistic results show that 20.17% users have repeat purchasing actions in one week, 29.49% in two weeks, 35.06% in three weeks and 44.57% in one month. We tried several supervised learning algorithms to validate the prediction accuracy of different time intervals, including k-nearest neighbors algorithm(k-NN), decision tree(DT), support vector machines(SVM), random forest(RF) and logistic regression(LR). We divided the dataset into two parts, 67% used for training and the rest used for testing. The experimental results are shown as Table 4.4. The base line is the result only using the historical records for prediction. Because of the imbalance of the dataset, we use the F1 score to measure different machine learning methods except for the prediction accuracy. Since our goal is to find as many potential purchasing consumers as possible to increase the total profit of takeout food providers, we choose Recall as an important criteria to judge the performance of machine algorithms. As shown in Table 4.4, NB outperforms the other algorithms with the best Recall and F1 Scores. Meanwhile, the prediction accuracy of NB method is also acceptable comparing with the other models.

According to the statistic results of the access and purchase actions of female and male consumers, we empirically assume that gender factor should make sense to predict repeat buyer in the future. However, the experimental results are not positive enough to support our assumption. Gender factor has quite little influence on the repeat purchasing prediction, as shown in Table 4.5. The difference between the experimental results with and without considering the gender factor is quite little, which indicates that female and male consumers seem to have similar consumption customs and it is very difficult to find their difference.

We also tried to improve the prediction performance by adding more and more features into the base line methods. The prediction performance is a little higher than that of the baseline while the improvement is not so obvious as we imagined empirically. The experimental results are shown as Table 4.6.

The experimental results shown in Table 4.7 and Table 4.4 demonstrate that demographic factors have little influence on the repeat purchasing prediction. Though different groups of users have different purchasing ratios, their trends are close to each other. It is quite difficult to improve the prediction performance by adding the demographic factors.

4.7 Conclusion

With the popularization of the mobile Internet and the prevalence of delivery service, Online Takeout Ordering & Delivery (OTOD) using Apps from smart phones or websites from PC has become an emerging service and prosperous industry(e.g., KFC delivery). In order to improve the quality of service and recommendation personalization, we tried to find the key factors leading to a successful purchasing of takeout food in this paper. We collected Internet access records related to OTOD service of 34,845 users with a time duration of nearly four months. At first, We did a preliminary study on

Table 4.5: Prediction accuracy for different time interval using various machine learning algorithms(+gender)

Algorithms	Prediction accuracy for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	81.98%	74.77%	70.68%	65.95%
NB	81.60%	76.71%	73.61%	70.60%
DT	78.21%	73.75%	71.09%	67.24%
SVM	83.03%	77.35%	74.85%	70.39%
RF	83.22%	76.87%	73.72%	69.50%
LR	83.86%	78.27%	75.20%	70.47%
Algorithms	Recall score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	53.52%	54.42%	54.81%	55.94%
NB	**64.48%**	**63.28%**	**62.19%**	**62.78%**
DT	58.64%	59.09%	59.40%	59.66%
SVM	50.80%	54.65%	57.41%	59.08%
RF	51.84%	53.04%	53.93%	56.19%
LR	58.14%	59.64%	59.74%	60.30%
Algorithms	F1 score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	53.08%	53.64%	54.03%	55.13%
NB	**65.40%**	**64.47%**	**63.13%**	**63.34%**
DT	59.15%	59.80%	59.97%	59.91%
SVM	47.03%	52.71%	56.55%	58.11%
RF	49.23%	49.63%	50.21%	52.84%
LR	59.82%	60.70%	60.17%	60.09%

Table 4.6: Prediction accuracy for different time interval using various machine learning algorithms(+age)

Algorithms	Prediction accuracy for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	82.65%	77.08%	73.32%	66.22%
NB	81.68%	76.73%	73.56%	70.39%
DT	76.79%	72.49%	69.04%	66.03%
SVM	83.03%	77.38%	75.04%	70.46%
RF	83.32%	77.00%	73.86%	69.63%
LR	84.08%	78.38%	75.15%	70.36%
Algorithms	Recall score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	54.42%	56.21%	56.58%	56.76%
NB	**64.53%**	**63.26%**	**62.24%**	**62.48%**
DT	59.27%	60.60%	59.31%	60.32%
SVM	50.80%	54.70%	57.78%	59.28%
RF	51.85%	53.09%	54.11%	56.22%
LR	58.70%	60.06%	59.82%	60.15%
Algorithms	F1 score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	54.38%	55.74%	55.82%	56.29%
NB	**65.49%**	**64.46%**	**63.17%**	**63.00%**
DT	59.25%	60.98%	59.66%	60.59%
SVM	47.03%	52.80%	57.11%	58.41%
RF	49.15%	49.62%	50.50%	52.74%
LR	60.60%	61.24%	60.30%	59.88%

Table 4.7: Prediction accuracy for different time interval using various machine learning algorithms(+age+gender)

Algorithms	Prediction accuracy for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	82.38%	76.41%	72.73%	65.65%
NB	81.60%	76.71%	73.61%	70.63%
DT	75.82%	71.22%	68.21%	65.09%
SVM	83.03%	77.43%	74.99%	70.55%
RF	83.30%	77.00%	73.88%	69.47%
LR	84.05%	78.38%	75.34%	70.68%
Algorithms	Recall score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	54.20%	54.90%	55.46%	56.14%
NB	**64.48%**	**63.28%**	**62.19%**	**62.84%**
DT	59.00%	59.39%	59.25%	59.82%
SVM	50.80%	54.81%	57.71%	59.30%
RF	51.77%	53.09%	54.13%	56.00%
LR	58.37%	59.94%	60.15%	60.60%
Algorithms	F1 score for different time interval			
	One week	Two weeks	Three weeks	One month
k-NN	54.08%	53.74%	54.17%	55.57%
NB	**65.41%**	**64.47%**	**63.13%**	**63.40%**
DT	58.72%	59.63%	59.48%	60.02%
SVM	47.03%	52.99%	57.02%	58.40%
RF	49.00%	49.62%	50.51%	52.39%
LR	60.17%	61.09%	60.74%	60.47%

users' daily and periodic purchasing behaviors of takeout food. Then we combine the demographic information and location information with the purchasing activities to find the most potential purchasing groups of takeout food. Based on the features extracted from historical purchasing records, demographic information and location information, we use several popular machine learning methods to predict the future purchasing activities within a specific time. The experiments show that our extracted features can be well used for the takeout food purchasing prediction problem.

We used various machine learning methods to predict the repeat buyer of takeout food and the experimental results demonstrate the predictability of takeout food purchasers. The experimental results show that NB model have better performance when predicting short-term repeat purchasing actions while DT outperforms the other algorithms for long-term prediction.

As for future research, we will try to have more detailed data of more users to further improve the prediction performance of takeout food purchasing. We will also try to analyze and predict the consumer behaviors across different takeout food ordering platforms to find consumers' loyalty to different platforms. On the basis, we can give more personalized recommendation to consumers to further improve the quality of service to benefit both platform providers and consumers. We will also try to apply some deep learning and reinforcement learning methods to further improve the prediction performance.

Chapter 5

Repeat Buyer Prediction in E-commerce

Merchants sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, "Black Friday" or "Double 11 (Nov 11th)", in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI). It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. With the long-term user behavior log accumulated by Tmall.com, we get a set of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day. Our goal is to predict which new buyers for given merchants will become loyal customers in the future. In other words, we need to predict the probability that these new buyers would purchase items from the same merchants again within 6 months. A data set containing around 200k users is given for training, while the other of similar size for testing. We extracted as many features as possible and find the key features to train our models. We proposed an ensemble model based on different classification models and an ensemble lightGBM model using different parameter sets. The experimental results show that our ensemble models can bring about great performance improvements comparing with the original models.

5.1 Introduction

In the past few years mobile phones have witnessed a remarkable evolution and explosive popularization[46]. Meanwhile, e-commerce also has a prosperous development and drastically changed traditional commercial relationships, as well as the shopping process for the fast-growing online shoppers[5]. With a smart phone at hand, the consumer can check the details of products, compare the prices across various e-commerce platforms, save items into charts and enjoy a great many benefits such as personalization from merchants and recommendation from social networks[9, 21, 32, 69].

To attract more attention and clean up inventory, many large e-commerce platforms tend to carry out special promotion events several times per year, such as "Black Friday" of Amazon and "11.11" of Alibaba. Because of the large discounts and other preferential policy, the volume of business of most merchants in e-commerce platforms will reach the peak in the year. However, most of the new buyers during special shopping events tend to be one-deal hunter and will not buy products again from the merchant in the future. The cost will be very high if a merchant give advertisements to all the new buyers. Therefore, it is necessary to identify the potential repeat buyer for each merchant to give more precised and personalized service to its potential customers. Fig 5.5 shows the historical records of a two consumers in e-commerce as an example. Empirically, even though user 2 has more successful purchases from the merchant than that of user1, he still tends to be a one-deal hunter while user 1 is a potential repeat buyer of the merchant. Our work here is to identify the possible repeat buyers in the future based on their profiles, historical records and other features related to the components of a successful purchase.

Generally 2% of shoppers make a purchase on the first visit to an online store while the other 98% enjoys only window-shopping. To bring people back to the store and close the deal, "retargeting" has been a vital online advertising strategy that leads to "conversion" of window-shoppers into buyers[66]. Further more, we care about the other question: whether a buyer will buy again from one store in the future? The repeat buyer prediction problem can be formulated as a typical classification problem and the consumer can be divided into two groups, repeat buyer and not[34]. Model training of this task is similar with that of other classification tasks for which feature engineering is the main component that distinguishes this task from others. Researchers have paid plenty of attention to the classification algorithms design and optimization in the research community while not much work has been reported on model merging for prediction tasks in e-commerce. Consequently, in this paper we focus on feature engineering at first. We extract various types of features from user activity log data test these features via extensive experiments. For the model training, we train several typical classification model alone at first and have a performance ranking about the models. Then we propose an ensemble model based on the classical machine learning models, lightGBM and XGBoost. The ensemble model can bring great performance improvement comparing with the original single models. We hope that our work can make sense for data science practitioners, who need to develop solutions for prediction tasks in e-commerce.

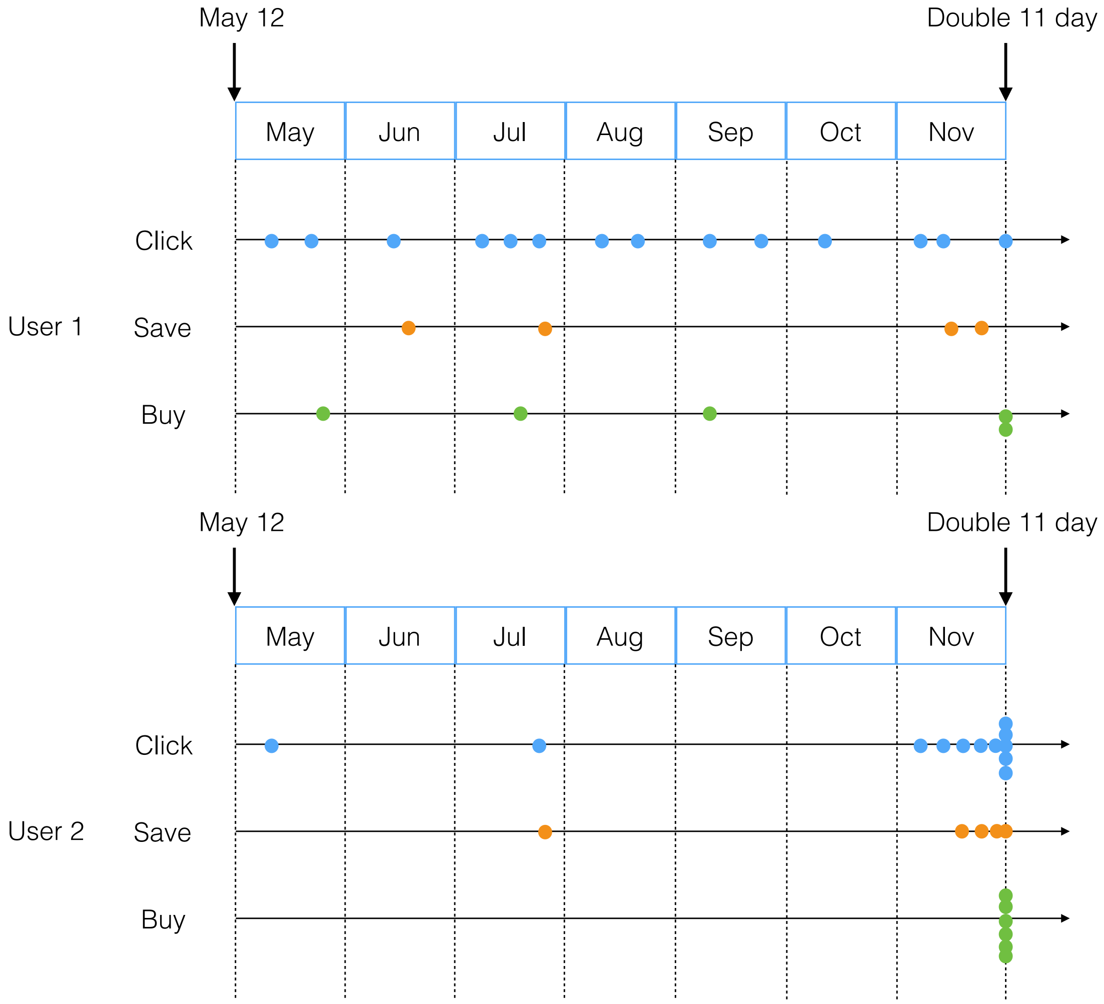


Figure 5.1: Example of a user's historical records and possible consumption sequences

5.2 Data Description

The data set contains anonymized users' shopping logs in the past 6 months before and on the "Double 11" day,and the label information indicating whether they are repeated buyers. Due to privacy issue, data is sampled in a biased way, so the statistical result on this data set would deviate from the actual of Tmall.com. But it will not affect the applicability of the solution. We have four files in our dataset. Details of the data format can be found in the table below.

Table 5.1 shows the user logs on and before the double 11 day of 2014 with a duration period of 6 months. The data fields contain user_id, item_id, category_id, brand_id, merchant_id, time_stamp and action_type. The ids are unique and have no overlap. Products sold in different merchants are assigned different item_ids even if the products are exactly the same. There are four typical actions in e-commerce, namely click, add to card, add to favorite and buy. Click is the first step for a possible purchase progress and buy is the final step for a successful purchase. We use binary coding for the four action types: 0 for click, 1 for add-to-cart, 2 for purchase and 3 for add-to-favourite.

Table 5.2 shows the statistics of the user activity log data. According to the statistic results, we can see that many merchants in the log data do not have new buyers in the training or testing data actually. They are included in the log data because some new buyers may have click actions on them. The activities of the new buyers at these merchants are valuable information for inferring the preferences and habits of the new buyers, which can be used to calculate the similarity of different consumers.

Table 5.3 shows the counts of four action types. Most of the actions are clicks and only a very small ratio of the actions can lead to a successful purchase. The small ratio of add-to-cart action indicates that most consumers only buy one item directly from a merchant, which has a possible influence on the consumer loyalty to the merchants. Add-to-favourite can be seen as a symbol that a consumer will probably come back again to a merchant in the future. We will analyze their influence in our experiments in detail in the following part.

Table 5.6 shows the statistics of the training and testing data. The set of merchants in training data and that in testing data are the same except for a single merchant. Users in the training and testing data have no overlap. The second last column is the number of positive <new buyer, merchant> pairs such that the new buyer bought items from the merchant again within six months. The last column is the percentage of such positive pairs. The percentage of positive pairs is around 6%, which indicates that most of the new buyers are indeed one-time deal hunters.

Table 5.1: User Behavior Logs

Data Fields	Definition
user_id	A unique id for the shopper.
item_id	A unique id for the item.
cat_id	A unique id for the category that the item belongs to.
merchant_id	A unique id for the merchant.
brand_id	A unique id for the brand of the item.
time_stamp	Date the action took place (format: mmdd)
action_type	It is an enumerated type 0, 1, 2, 3, where 0 is for click, 1 is for add-to-cart, 2 is for purchase and 3 is for add-to-favourite.

Table 5.2: Statistics of log activity data

#rows	#users	#merchants	#items	#categories	#brands
54,925,330	424,170	4,995	1,090,390	1,658	8,444

Table 5.7 shows user demographic data, which contains the age and gender of users. The age values are divided into seven ranges and the gender values are divided into three groups. The class label of a training <new buyer, merchant> pair is known, and it indicates whether the new buyer bought items from the merchant again within six months after the “Double 11” promotion. The class labels of testing <new buyer, merchant> pairs are hidden, as shown in Table 5.5. The task is to predict the class labels of the testing pairs.

5.3 Our Goal

Our goal is to predict whether the given user will become a repeat buyer of the given merchant. Value should be 0 or 1.

$$label < merchant, user > \in \{0, 1\}$$

where 0 indicates the user is a one-time deal hunter or non-repeat buyer while 1 indicates the user will be a repeat buyer of the merchant. In other form, we should calculate the probability of each user-merchant pair in the test data, which is a float number between 0 and 1. We use the roc-auc score to validate the performance of each machine learning algorithm.

Table 5.3: Statistics of action types

click	add-to-cart	purchase	add-to-favourite
48,550,713(88.39%)	76,750(0.14%)	3,292,144(5.99%)	3,005,723(5.47%)

Table 5.4: User Profile

Data Fields	Definition
user_id	A unique id for the shopper.

age_range	User' s age range: 1 for under 18; 2 for [18,24];
	3 for [25,29]; 4 for [30,34];
	5 for [35,39]; 6 for [40,49];
	7 and 8 for more than 50; 0 and NULL for unknown.

gender	User' s gender: 0 for female, 1 for male, 2 and NULL for unknown.

Table 5.5: Training and Testing Data

Data Fields	Definition
user_id	A unique id for the shopper.
merchant_id	A unique id for the merchant.
label	It is an enumerated type 0, 1, where 1 means repeat buyer, 0 is for non-repeat buyer. This field is empty for test data.

5.4 Feature Engineering

5.4.1 Statistic Analysis

Intuitively, male and female consumers should have different shopping style and loyalty to some specific merchants, brands, categories and items. For example, female consumers may have more interest in clothes and makeups while male consumers may pay more attention to electronics and games. Among the 260,863 users provided in the training dataset, there are 176,413 female consumers with a repeat buyer ratio of 6.45%. As for the remaining 84,450 male consumers, only 5.41% of them will buy again in the future. This statistic result indicates that gender factor can be a positive feature for the repeat buyer prediction task.

There are mainly five components of one successful purchasing, namely merchants, consumers, brands, categories and items, respectively. There are four types of actions, namely click, add-to-cart, add-to-favourite and buy respectively. In addition, we also have the gender and age information of each user as well as the time_stamp information of each log record. Since our goal is to predict the repeat buyers of a merchant. Our main attention is paid to the related two entities, user and merchant.

Table 5.6: Statistics of training and testing data

data	#users	#merchants	#pairs	#positive pairs	positive%
train	212,062	1,993	260,864	15,952	6.12%
test	212,108	1,993	261,477	16,037	6.13%

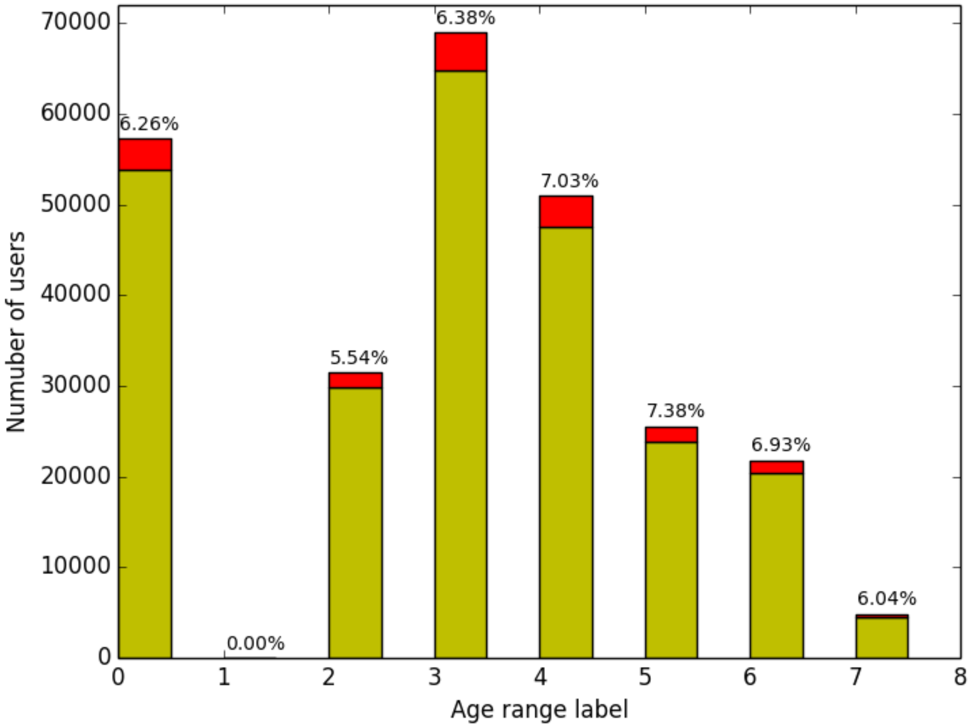


Figure 5.2: Distribution of user counts with different age range, 1 for under 18; 2 for [18,24];3 for [25,29]; 4 for [30,34];5 for [35,39]; 6 for [40,49];7 and 8 for more than 50; 0 and NULL for unknown. The number of percentage in each column shows the repeat buyer ratio of each age group.

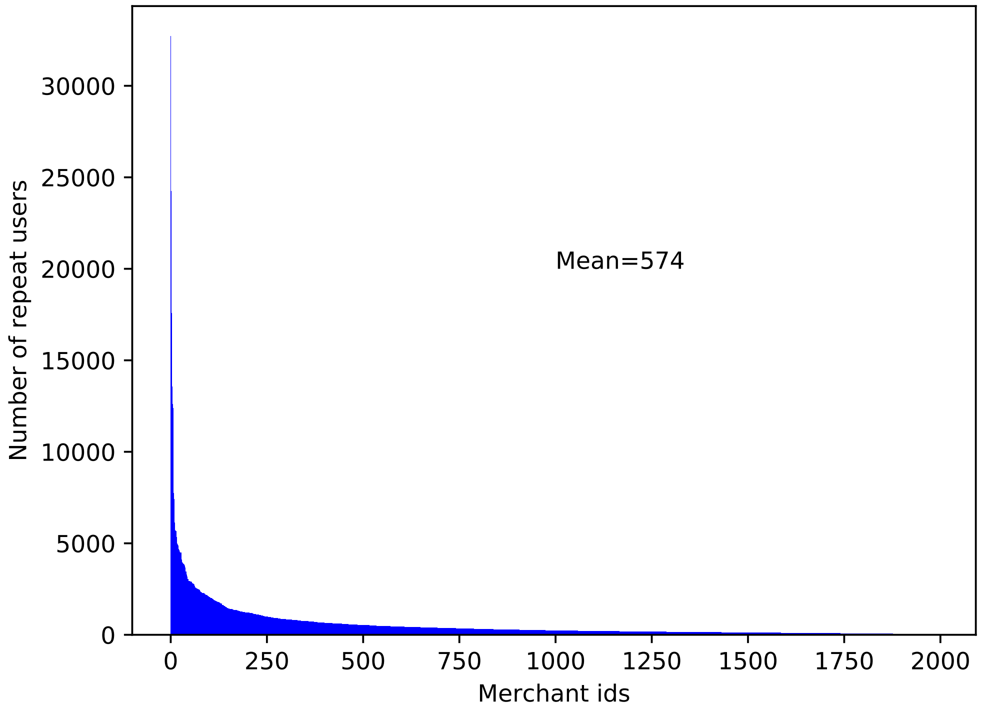


Figure 5.3: Distribution of repeat buyers of each merchant

Fig 5.2 shows the user distribution with different age range. Only 14 users have the age label '1', which means they are under 18 years old. The younger and elder consumers have lower repeat purchase ratios while the consumer group between 30 and 39 years old have higher loyalty. The reason may be that young people are more familiar with electronic device and surfing online so there will be more choice for them to select the variety of merchants instead of keeping shopping in one store. We made an interview through Wechat, the most popular social app in China. We sent out 200 questionnaires and received 156 feedback. The younger consumers care more about the price and different styles about the products. As a result, they may move across different merchants to find their ideal products. The elder consumers do online shopping not so frequently which result in a lower repeat buy ratio in the following 6 months. The consumers between 30 and 39 years old have already had their own shopping taste and economic foundation and tend to have some firm merchants which they believe reliable.

Fig 5.3 shows the distribution of repeat buyer counts of each merchant. The mean value is 574.

5.4.2 Features

To train the models and make prediction of repeated users, we should extract features from logs and personal profiles. We already make statistic analysis for the behaviors and user information in last section. In this part, the technique one-hot coding is used to avoid the additional influence caused by the value of features.

There are five components involved in an online shopping transaction, namely user, merchant, item, brand and category, respectively. Besides, there are four types of actions between users and merchants, namely click, add-to-cart, add-to-favorite and buy. In addition, the dataset also provides the demographic information of related users. As a result, we can generate various features to train our model. Since our task is to predict the repeat buyers for specific merchants, the user-merchant interaction features should have more influence on the results.

User Profile Features

Different user groups may favor different types of products. For example, clothes and cosmetics are more attractive to women while electric products are more appealing to men. As such, we generated features to describe the popularity of merchants, brands, categories, and items within different user groups, where users are grouped based on their gender or age range. These features include overall buy counts, monthly aggregation on monthly buy counts, penetration features and repeat buyer features. Only users of a particular age range or a particular gender are used to calculate these features. Besides of that, the numbers of merchants, brands, categories or items that a user has actions on them can also be seen as user profile features. Considering the time stamp, action type and action objectives, we can generate various user profile features, shown as Table 5.7.

Merchant Profile Features

Different merchants have different main brands or categories, market shares and reputations, etc.. The profile information of merchants can be seen as significant features for our repeat buyer prediction task. The detailed description of merchant profile features is shown as Table 5.8

User-Merchant Features

Since our task is to predict whether a consumer will be a repeat buyer for a specific merchant, the features between the user-merchant pair play important roles in the prediction. Considering the action types, brands and categories, we can generate various user-merchant features, shown as Table 5.9.

Table 5.7: Main features related to user profile

Feature Name	Description
u_gender	the gender type of a user
u_age	the age range of a user
u_click	the click counts of a user
u_cart	the add-to-cart counts of a user
u_fav	the add-to-favorite counts of a user
u_buy	the buy counts of a user
u_action	the action counts of a user
u_day	the day counts that a user has actions
u_click_{month_id}	the click counts of a user for each month
u_cart_{month_id}	the add-to-cart counts of a user for each month
u_fav_{month_id}	the add-to-favorite counts of a user for each month
u_buy_{month_id}	the buy counts of a user for each month
u_action_{day_id}	the one-hot coded action counts of a user for each day
u_item	the item counts on which a user has actions
u_cat	the category counts on which a user has actions
u_brand	the brand counts on which a user has actions
u_merchant	the merchant counts on which a user has actions
u_click_1111	the click counts of a user on double 11 day
u_cart_1111	the add-to-cart counts of a user on double 11 day
u_fav_1111	the add-to-favorite counts of a user on double 11 day
u_buy_1111	the buy counts of a user on double 11 day
u_day_last	the last day in which a user has actions
u_day_first	the first day in which a user has actions

User-merchant Similarity Features

User-merchant similarity features measure how similar a user and a merchant are based on brands or categories. It is derived by the merchant market share features and user preference features on brands or categories. Taking a merchant-brand pair $< M, B >$ as example, let N_{MB} be the number of purchases of the brand from the merchant and N_B be the number of purchases of the brand from all the merchants. We define the merchant's market share on the brand as $S_{MB} = N_{MB}/N_B$. Similarly, taking a user-brand pair $< U, B >$ as example, let N_{UB} be the number of purchases of the brand from the user and N_U be the number of all purchases of the user from all the brands. We define the user's preference on the brand as $P_{UB} = N_{UB}/N_U$. Based on the definition above, for a use-merchant pair $< U, M >$ and the brand list $B = \{B_1, B_2, ..., B_{8444}\}$ in our dataset, we can generate the market share vector of the merchant $S =< S_{MB_1}, S_{MB_2}, ..., S_{MB_{8444}} >$ and the preference vector of the user $P =< P_{UB_1}, P_{UB_2}, ..., P_{UB_{8444}} >$. Then the similarity of the $< user, merchant >$ pair based on brand is calculated as $S_{UM} = P \times S^T$. The similarity of the $< user, merchant >$ pair based on category can be calculated similarly. Intuitively, the more similar a user and a merchant are, the more likely the user will buy from the merchant again.

Table 5.8: Main features related to merchant profile

Feature Name	Description
m_item	the number of items merchants have
m_cat	the number of categories of the merchant
m_click	the click counts of a merchant
m_cart	the add-to-cart counts of a merchant
m_fav	the add-to-favorite counts of a merchant
m_buy	the buy counts of a merchant
m_action	the action counts of a merchant
m_click_{month_id}	the click counts of a merchant for each month
m_cart_{month_id}	the add-to-cart counts of a merchant for each month
m_fav_{month_id}	the add-to-favorite counts of a merchant for each month
m_buy_{month_id}	the buy counts of a merchant for each month
m_gender_features	the features of merchants of each gender group
m_age_features	the features of merchants of each age group
m_age_gender_features	the features of merchants of each age and gender group

Repeat Action Features

Intuitively, for the $< user, merchant >$ pair to be predicted, the repeat buyer feature for the merchant or the repeat purchasing merchant feature for the user should make great sense. Here we extracted the repeat buy features for both the user and the merchant based on the buy action. The detailed description is shown as Table 5.10.

5.4.3 Collaborative filtering based feature

There are four matrices to calculate user similarity, namely user-merchant matrix, user-category matrix, user-brand matrix and user-item matrix. Similarly, there are also four matrices to calculate merchant similarity, namely merchant-user matrix, merchant-category matrix, merchant-brand matrix and merchant-item matrix.

$U\_train = \{u_1, u_2, ..., u_M\}$

$M\_train = \{m_1, m_2, ..., m_N\}$

S_{u,u^*} is the similarity of users u and u^*.

S_{m,m^*} is the similarity of merchants m and m^*.

$P_{u,m} \in \{0, 1\}$ is the probability of whether user u is a repeat buyer of merchant m.

For $u \in U\_train$ and $m \in M\_train$, $P_{u,m} \in \{0, 1\}$.

For $u \in U\_test$ and $m \in M\_test$, $P_{u,m} \in [0, 1]$.

For a test pair (u^*, m^*) to be predicted,

Table 5.9: Main features related to user-merchant pairs

Feature Name	Description
u_m_click	the click counts of the user-merchant pair
u_m_cart	the add-to-cart counts of the user-merchant pair
u_m_fav	the add-to-favorite counts of the user-merchant pair
u_m_buy	the buy counts of the user-merchant pair
u_m_action	the action counts of the user-merchant pair
u_m_{action type}_{items ,categories,brands}	the size, mean, max and min of actions for a user on items, categories and brands under one merchant
u_m_ratio_{action type}	the ratios of different actions for the user-merchant pair
user_merchant_time_delta	the time delta of user actions for merchants

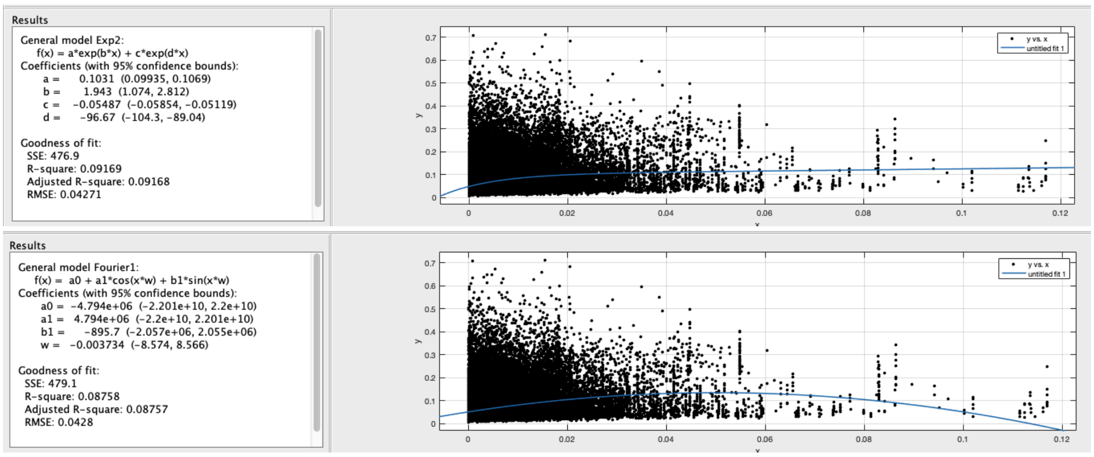
$$S_u = \sum_{u \in U\_train} S_{u,u*}$$

$$S_m = \sum_{m \in M\_train} S_{m,m*}$$

$$P_{u*,m*} = \sum_{u \in U\_train} \sum_{m \in M\_train} \frac{S_{u,u*} \times S_{m,m*} \times P_{u,m}}{S_u \times S_m}$$

$$= \sum_{u \in U\_train} \frac{S_{u,u*} \times P_{u,m*}}{S_u} + \sum_{u \in U\_train} \sum_{m \in M\_train - \{m*\}} \frac{S_{u,u*} \times S_{m,m*} \times P_{u,m}}{S_u \times S_m}$$

However, the filtering results between CF based model and our training model show that the CF based model works not so well, as shown in Fig. 5.4. A possible reason is that our dataset is too small and the similarity matrices are too sparse to get the precise similarity results. If we can get the whole dataset of the e-commerce platform,



Results

```
General model Exp2:
    f(x) = a*exp(b*x) + c*exp(d*x)
Coefficients (with 95% confidence bounds):
      a =     0.1031  (0.09935, 0.1069)
      b =      1.943  (1.074, 2.812)
      c =   -0.05487  (-0.05854, -0.05119)
      d =     -96.67  (-104.3, -89.04)

Goodness of fit:
  SSE: 476.9
  R-square: 0.09169
  Adjusted R-square: 0.09168
  RMSE: 0.04271
```

Results

```
General model Fourier1:
    f(x) =  a0 + a1*cos(x*w) + b1*sin(x*w)
Coefficients (with 95% confidence bounds):
      a0 =  -4.794e+06  (-2.201e+10, 2.2e+10)
      a1 =   4.794e+06  (-2.2e+10, 2.201e+10)
      b1 =      -895.7  (-2.057e+06, 2.055e+06)
      w =   -0.003734  (-8.574, 8.566)

Goodness of fit:
  SSE: 479.1
  R-square: 0.08758
  Adjusted R-square: 0.08757
  RMSE: 0.0428
```

Figure 5.4: Filtering results of the CF based model and the results from our training model

we should obtain the user similarity results with higher accuracy. This can be a future research direction since we only have a small dataset currently.

5.5 Model training and testing

5.5.1 Model for training and testing

There are many classification algorithms to solve our repeat buyer prediction problem. Typical algorithms, such as Factorization Machine(FM)[48], Logistic Regression(LR), GBM[19], Random Forest(RF), and XGBoost[10] as well as some deep learning models could all be considered independently or together.

Factorization Machines(FM) are a new model class that combines the advantages of Support Vector Machines (SVM) with factorization models. Like SVMs, FMs are a general predictor working with any real valued feature vector. In contrast to SVMs, FMs model all interactions between variables using factorized parameters. Thus they are able to estimate interactions even in problems with huge sparsity(like recommender systems) where SVMs fail. The model equation of FMs can be calculated in linear time and thus FMs can be optimized directly. So unlike nonlinear SVMs, a transformation in the dual form is not necessary and the model parameters can be estimated directly without the need of any support vector in the solution[47].

Logistic regression is a widely-use linear classifier. It is intrinsically simple and so is less prone to over-fitting. In fact, our experimental result shows that LR achieves the best performance among all the individual models. We use the implementation of LibLinear[18].

Gradient Boosting Decision Tree is a tree-based additive model. GBDT learns multiple decision trees iteratively, where the learning target of the current tree is defined as the loss gradient of the previous trees. The outputs of GBDT are the additive predictions of all trees to calculate the final prediction. It has a strong predictive power and naturally handles data with heterogeneous features. In this chapter, we use the implementation of lightGBM and XGBoost [10].

5.5.2 Ensemble model for repeat buyer prediction

To further improve the prediction performance, we propose an ensemble model for this task considering the classification results of each independent model. There are various integration results. Here we use the integrated result based on the two best individual models to validate the effectiveness of our solution.

For the ensemble model training and testing, we use LightGBM and XGBoost models to obtain the primary probability respectively and then get an intermediate value of

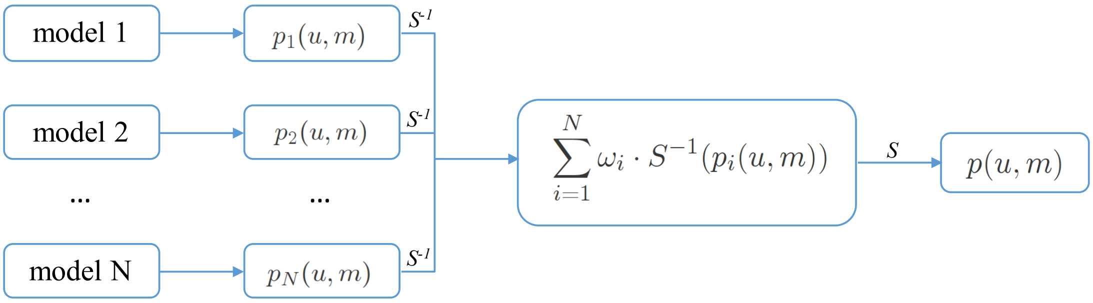


Figure 5.5: Ensemble model for the prediction task.

each probability using Sigmoid inverse function. We then calculate the mean value of the intermediate values and use Sigmoid function to get the final repeat buyer probability.

Sigmoid Function:

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

Sigmoid Inverse Function:

$$S^{-1}(x) = ln\frac{x}{1-x}$$

Repeat Buyer Probability:

$$p(u,m) = S\left(\sum_{i=1}^{k} w_i \times S^{-1}(p_i(u,m))\right)$$

where $p(u,m)$ is the final probability that a user u will make a repeated purchase from a merchant m, $p_i(u,m)$ is the probability predicted by the i-th single model, w_i is the weight assigned to the i-th single model, k is the number of single models, and $\alpha \in [0,1]$ and $1-\alpha$ are the weights assigned to different feature groups, in which vec means the feature group consisting of vector features while $base$ means the feature group without vector features.

5.6 Experimental results

At first, we tested several classical models and the results are shown as Table 5.11. We choose the two best models to validate our idea of ensemble training model. The calculation formula can be simplified as following:

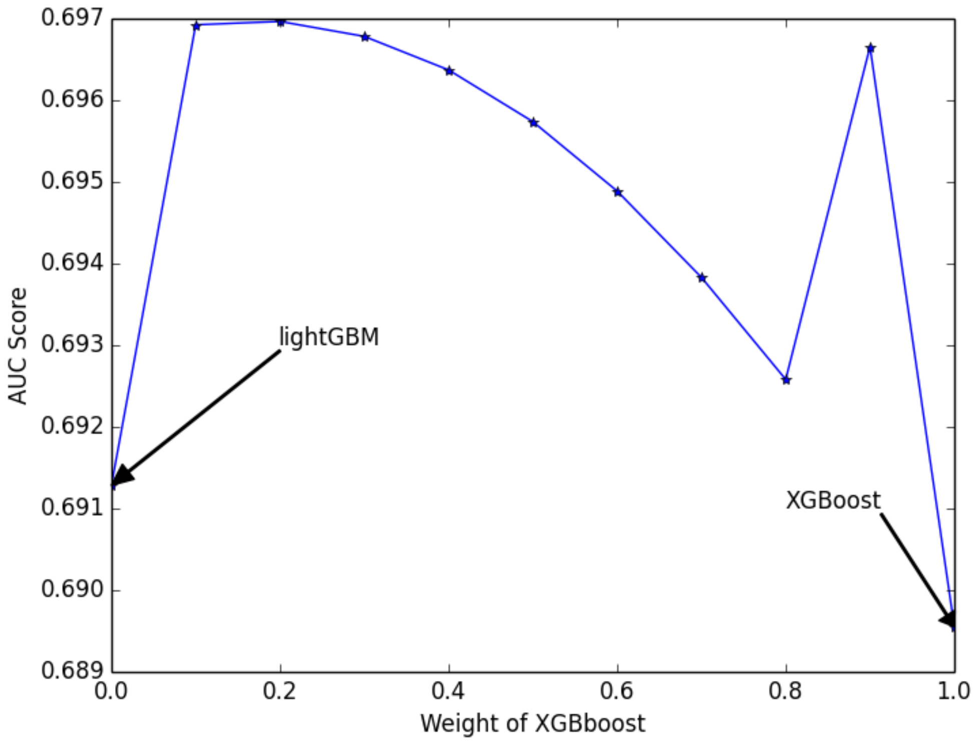


Figure 5.6: Ensemble model based on lightGBM and XGBoost

$$p(u,m) = S(\alpha \times S^{-1}(p_{xgb}(u,m)) + (1-\alpha) \times S^{-1}(p_{lgb}(u,m)))$$

where $p(u,m)$ is the final probability that a user u will make a repeated purchase from a merchant m, $p_{xgb}(u,m)$ and $p_{lgb}(u,m)$ are the probabilities predicted by the XGBoost model and the lightGBM model. $\alpha \in [0,1]$ and $1-\alpha$ are the weights assigned to the two different models.

In this part, we set the granularity of α as 0.1 and the result is shown in Fig. 5.6. The auc score of lightGBM model is 0.691 and the auc score of XGBoost is 0.689. However, when combining the results of these two models, we can get better results comparing with the single models. The best result achieved by our ensemble model is 0.697, which is obviously better than than the single lightGBM and XGBoost models. Considering the economic volume of e-commerce with yearly sales in thousands of billions U.S. dollars currently[1], even 0.1% improvement of auc score can make great sense for one of the most important part in E-commerce, repeat buyer prediction and target marketing. As for the lightGBM model, we used two different parameter sets when training the model. The details of the parameteres are shown as Table 5.12. The two models are similar to each other except for the difference of learning rate, which has influence on the

[1] https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

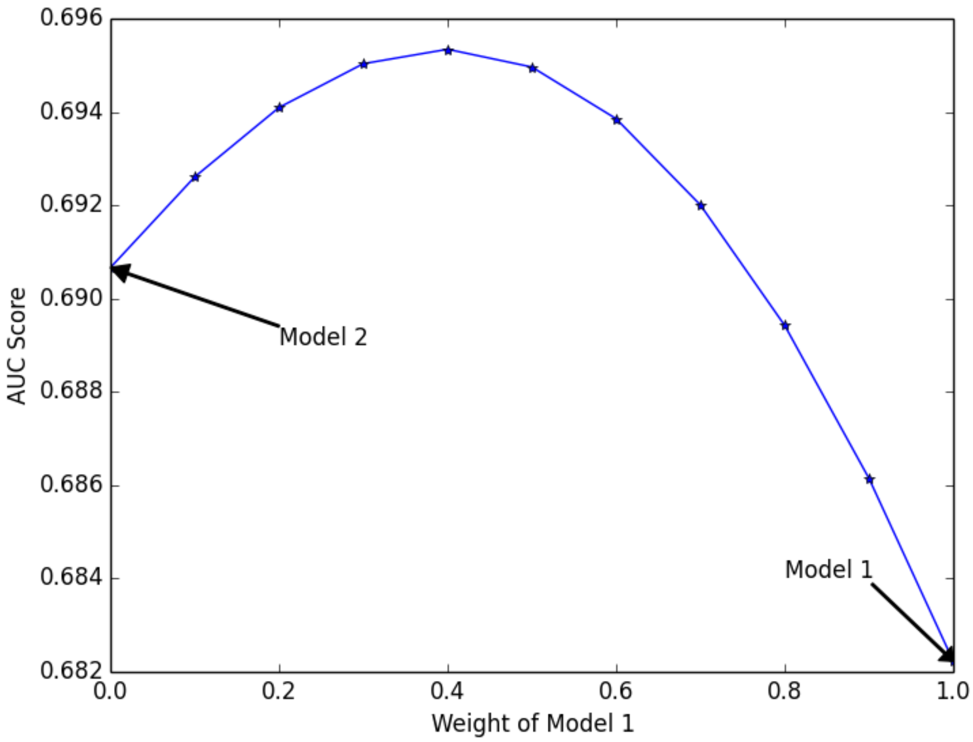


Figure 5.7: Ensemble model based on two lightGBM models with different parameter sets

training loss. The auc score of model 1 is 0.682 and the auc score of model 2 is 0.691. However, when combining the two results with weighted average, the performance will have significant improvement, as shown in Fig. 5.7. The best result is around 0.695. When the weight of the model 1 is set as 0, the result is actually that of model2. With the weight of model 1 grows, the performance of the ensemble model increases at first and reaches the peak around 0.4.

5.7 Conclusion

In this chapter, we extracted various features that may influence the repeat purchasing of consumers in the future, including user-profile features, merchant-profile features, user-merchant interaction features and repeat features, as well as some aggregation features derived from the original features. Based on the features, we used several classical machine learning models to train the dataset to perform repeat buyer prediction task. For the lightGBM model which has best performance, we used two different parameter sets to train the model and get two different auc scores. On the basis, we combined the two models with different parameter sets and get better results comparing with the original ones. For the two best models, we combined the lightGBM and XGBoost models and achieved better results again. We also tried to combine all the five models, however, no further improvements were achieved.

For the future work, we will consider to extract more related features to further improve the performance of the models. In addition, we will try to rank each feature automatically to find the most influential features that lead to repeat purchase in the future. The combination of different training models can bring about great performance improvement comparing with single ones. We will try to integrate more effective models and find the mechanism that influence the performance of the ensemble model.

Table 5.10: Main features related to repeat actions

Feature Name	Description
m_re__u	the number of users having actions on the merchant more than one day
m_re_buy_u	the number of users having buy actions on the merchant more than one day
m_buy_u	the number of users having buy actions on the merchant
ratio_m_re_buy_u	the ratio of {m_re_buy_u}/{m_buy_u}
m_re_click_u	the number of users having click actions on the merchant more than one day
m_click_u	the number of users having click actions on the merchant
ratio_m_re_click_u	the ratio of {m_re_click_u}/{m_click_u}
m_re_cart_u	the number of users having add-to-cart actions on the merchant more than one day
m_cart_u	the number of users having add-to-cart actions on the merchant
ratio_m_re_cart_u	the ratio of {m_re_cart_u}/{m_cart_u}
m_re_fav_u	the number of users having add-to-favorite actions on the merchant more than one day
m_fav_u	the number of users having add-to-favorite actions on the merchant
ratio_m_re_fav_u	the ratio of {m_re_fav_u}/{m_fav_u}
u_re_buy_m	the number of merchants on which the user has buy actions more than one day
u_buy_m	the the number of merchants on which the user has buy actions
ratio_u_re_buy_m	the ratio of {u_re_buy_m}/{u_buy_m}
u_re_click_m	the number of merchants on which the user has click actions more than one day
u_click_m	the the number of merchants on which the user has click actions
ratio_u_re_click_m	the ratio of {u_re_click_m}/{u_click_m}
u_re_cart_m	the number of merchants on which the user has add-to-cart actions more than one day
u_cart_m	the the number of merchants on which the user has add-to-cart actions
ratio_u_re_cart_m	the ratio of {u_re_cart_m}/{u_cart_m}
u_re_fav_m	the number of merchants on which the user has add-to-favorite actions more than one day
u_fav_m	the the number of merchants on which the user has add-to-favorite actions
ratio_u_re_fav_m	the ratio of {u_re_fav_m}/{u_fav_m}

Table 5.11: AUC scores of different single model

Model	RF	LR	FFM	lightGBM	XGBoost
AUC Score	0.678	0.669	0.673	0.691	0.689

Table 5.12: Two parameter sets of the lightGBM model

Model 1	Model 2

params = {
’task’: ’train’,
’boosting_type’: ’gbdt’,
’objective’: ’binary’,
’metric’: ’auc’,
’num_leaves’: 31,
’learning_rate’: 0.05,
’feature_fraction’: 0.9,
’bagging_fraction’: 0.8,
’bagging_freq’: 5,
’verbose’: 0
}
	params = {
	’task’: ’train’,
	’boosting_type’: ’gbdt’,
	’objective’: ’binary’,
	’metric’: ’auc’,
	’num_leaves’: 31,
	’learning_rate’: 0.03,
	’feature_fraction’: 0.9,
	’bagging_fraction’: 0.8,
	’bagging_freq’: 5,
	’verbose’: 0
	}

Chapter 6

Conclusion

6.1 Conclusion

This thesis focuses on consumer behavior analysis in E-commerce and online food delivery system. On the basis of consumer behavior analysis, we make repeat buyer prediction using feature engineering method.

For the consumer behavior analysis in E-commerce, we introduce the significance of our research and make a comprehensive analysis about consumer behavior in e-commerce. For the preliminary analysis, we mainly analyze the different consumption patterns and trends in different time frames, such as hour-level in a day, day-level in a week and the whole consumption trend through several months. Further more, we analyze the influence of demographic factors, social status and special shopping events. In addition, we study the loyalty of consumers to specific E-commerce platforms based on across-platform analysis. Based on the analysis, we make a clustering to divide the customers into different groups. The consumers in each cluster have similar shopping patterns and can be used for personalized recommendation and precise advertisements.

For the consumer behavior analysis in online food delivery system, we try to have a detailed understanding about consumer behavior in terms of daily food purchasing. Different from online shopping of e-commerce, the food consumption is more related to our daily life. One guy may not do shopping for a month while he has to take food everyday. We analyze the consumer behavior of online food consumption and find some different behavior patterns. Based on the analysis, we further validate the predictability of takeout food purchasing.

For the repeat buyer prediction, we firstly extract various features that may have influence on consumers' future buy actions. The features include user-profile features, merchant-profile features, user-merchant interaction features, repeat action features and aggregation features derived from original features. Based on the various features, we firstly use several classical machine learning models to do the repeat buyer prediction task. Then we merged some of the models to get better results. The experimental results show that our merged model can bring about great performance improvement comparing with original single models.

6.2 Future Work

Here we introduce and discuss some possible research directions in the future based on our current progress.

For the consumer behavior analysis issue, we will make a comprehensive cross-cultural analysis to find the behavior difference of consumers with different society, economic and culture backgrounds. This can be helpful for the cross-border e-commerce platforms to better make their market strategies to provide more personalized service to different online shopping groups.

For the repeat buyer prediction issue, we will explore how to automate the feature generation and selection process for e-commerce prediction tasks. Currently, we mainly tune our models and parameters manually to achieve higher performance. However, it is very time-consuming and fallible. An automatic process system is helpful and necessary consequently. Besides of repeat buyer prediction, we will also try to make repeat products purchasing prediction in the future. On the basis, we can have better understanding on consumers' interest and behavior the combination of these two prediction tasks should be helpful to improve the prediction performance of each task.

Socioeconomic status prediction based on consumer behavior analysis is our another research interest in the future. As the social standing or class of an individual or group, socioeconomic status is a complex assessment measured in a variety of ways that account for a person's work experience and economic and social position in relation to others, based on income, education, and occupation. It can be used to many aspects, such as credit reference systems for banks, policy making for governments, health prediction for the public, etc. As a direct representation of socioeconomic status, consumer behavior analysis in E-commerce and food purchasing should be an important direction to predict the consumers' socioeconomic status.

Bibliography

[1] Gediminas Adomavicius and Alexander Tuzhilin. “User profiling in personalization applications through rule discovery and validation”. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 1999, pp. 377–381.

[2] Neha Sahu Barkha Agrawal. “THE REVOLUTION IN ONLINE FOOD SERVICES: A PERCEPTUAL STUDY”. In: *Journal Current Science* 20.1 (2019).

[3] Barış Akbaş, Dilek Karahoca, Adem Karahoca, and Ali Güngör. “Predicting newspaper sales by using data mining techniques”. In: *Proceedings of the 15th International Conference on Computer Systems and Technologies*. ACM. 2014, pp. 158–165.

[4] Nilanjan Banerjee, Dipanjan Chakraborty, Anupam Joshi, et al. “Towards analyzing microblogs for detection and classification of real-time intentions”. In: *Sixth International AAAI Conference on Weblogs and Social Media*. 2012.

[5] Steven Bellman, Gerald Lohse, and Eric J Johnson. “Predictors of online buying behavior”. In: (2009).

[6] Roger Bennett. “Ticket sales forecasting methods and performance of UK theatre companies”. In: *International Journal of Arts Management* (2002), pp. 36–49.

[7] Smriti Bhagat, Amit Goyal, and Laks VS Lakshmanan. “Maximizing product adoption in social networks”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pp. 603–612.

[8] Janice Boyce, Charles C Broz, and Margaret Binkley. “Consumer perspectives: take-out packaging and food safety”. In: *British Food Journal* 110.8 (2008), pp. 819–828.

[9] Mark Brown, Nigel Pope, and Kevin Voges. “Buying or browsing? An exploration of shopping orientations and online purchase intention”. In: *European Journal of Marketing* 37.11/12 (2003), pp. 1666–1684.

[10] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. “Xgboost: extreme gradient boosting”. In: *R package version 0.4-2* (2015), pp. 1–4.

[11] Lijing Cheng, Yongquan Fan, Chun Yu, and Yajun Du. “An improved trust-aware recommender system for personalized user recommendation in Tmall”. In: *DEStech Transactions on Engineering and Technology Research* ICMITE2016 (2016).

[12] Jan Christiaens and Greet Vanden Berghe. “A fresh ruin & recreate implementation for the capacitated vehicle routing problem”. In: (2016).

[13] Geng Cui, Hon-Kwong Lui, and Xiaoning Guo. “The effect of online consumer reviews on new product sales”. In: *International Journal of Electronic Commerce* 17.1 (2012), pp. 39–58.

[14] Honghua Kathy Dai, Lingzhi Zhao, Zaiqing Nie, et al. “Detecting online commercial intention (OCI)”. In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 829–837.

[15] Suvodip Dey, Pabitra Mitra, and Kratika Gupta. “Recommending repeat purchases using product segment statistics”. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM. 2016, pp. 357–360.

[16] Xiao Ding, Ting Liu, Junwen Duan, and Jian-Yun Nie. “Mining user consumption intention from social media using domain adaptive convolutional neural network”. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[18] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. “LIBLINEAR: A library for large linear classification”. In: *Journal of machine learning research* 9.Aug (2008), pp. 1871–1874.

[19] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[20] Qi Guo and Eugene Agichtein. “Ready to buy or just browsing?: detecting web searcher goals from interaction data”. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2010, pp. 130–137.

[21] Stephen Guo, Mengqiu Wang, and Jure Leskovec. “The role of social networks in online shopping: information passing, price of trust, and consumer choice”. In: *Proceedings of the 12th ACM conference on Electronic commerce*. ACM. 2011, pp. 157–166.

[22] Vineet Gupta, Devesh Varshney, Harsh Jhamtani, Deepam Kedia, and Shweta Karwa. “Identifying purchase intent from social posts”. In: *Eighth International AAAI Conference on Weblogs and Social Media*. 2014.

[23] Blanca Hernández, Julio Jiménez, and M José Martín. “Age, gender and income: do they really moderate online shopping behaviour?” In: *Online information review* 35.1 (2011), pp. 113–133.

[24] Bernd Hollerit, Mark Kröll, and Markus Strohmaier. “Towards linking buyers and sellers: detecting commercial intent on twitter”. In: *Proceedings of the 22nd International Conference on World Wide Web*. ACM. 2013, pp. 629–632.

[25] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[26] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. *Recommender systems: an introduction*. Cambridge University Press, 2010.

[27] Liu Juan. “From festive ceremony culture to marketing: Case study of T-Mall “Double Eleven” online shopping event”. In: *Journal of Advertising Study (Academic Edition)* 2 (2013), p. 2013.

[28] Gabriella Kazai, Iskander Yusof, and Daoud Clarke. “Personalised news and blog recommendations based on user location, Facebook and Twitter user profiling”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM. 2016, pp. 1129–1132.

[29] Eunju Kim, Wooju Kim, and Yillbyung Lee. “Combination of multiple classifiers for the customer’s purchase behavior prediction”. In: *Decision Support Systems* 34.2 (2003), pp. 167–175.

[30] Niels Buus Lassen, Rene Madsen, and Ravi Vatrapu. “Predicting iphone sales from iphone tweets”. In: *Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International*. IEEE. 2014, pp. 81–90.

[31] Uichin Lee, Zhenyu Liu, and Junghoo Cho. “Automatic identification of user goals in web search”. In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 391–400.

[32] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. “The dynamics of viral marketing”. In: *ACM Transactions on the Web (TWEB)* 1.1 (2007), p. 5.

[33] Vladimir I Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[34] Guimei Liu, Tam T Nguyen, Gang Zhao, et al. “Repeat buyer prediction for e-commerce”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 155–164.

[35] Yan Liu, Bin Guo, Chao Chen, et al. “FooDNet: Toward an Optimized Food Delivery Network based on Spatial Crowdsourcing”. In: *IEEE Transactions on Mobile Computing* (2018).

[36] Caroline Lo, Dan Frankowski, and Jure Leskovec. “Understanding behaviors that lead to purchasing: A case study of pinterest”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. 2016, pp. 531–540.

[37] Rajan Lukose, Jiye Li, Jing Zhou, and Satyanarayana Raju P Venkata. *Learning user purchase intent from user-centric data*. US Patent App. 12/263,176. 2010.

[38] Patrick Meulstee and Mykola Pechenizkiy. “Food sales prediction:" If only it knew what we know"”. In: *Data Mining Workshops, 2008. ICDMW’08. IEEE International Conference on*. IEEE. 2008, pp. 134–143.

[39] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. “Ontological user profiling in recommender systems”. In: *ACM Transactions on Information Systems (TOIS)* 22.1 (2004), pp. 54–88.

[40] Vicki G Morwitz and David Schmittlein. “Using segmentation to improve sales forecasts based on purchase intent: Which “intenders” actually buy?” In: *Journal of marketing research* 29.4 (1992), pp. 391–405.

[41] Gagandeep Nagra and R Gopal. “An study of factors affecting on online shopping behavior of consumers”. In: *International journal of scientific and research publications* 3.6 (2013), pp. 1–4.

[42] Richard L Oliver. “Whence consumer loyalty?” In: *Journal of marketing* 63.4_suppl1 (1999), pp. 33–44.

[43] Aron O’cass and Tino Fenech. “Web retailing adoption: exploring the nature of internet users Web retailing behaviour”. In: *Journal of Retailing and Consumer services* 10.2 (2003), pp. 81–94.

[44] Chanyoung Park, Donghyun Kim, Jinoh Oh, and Hwanjo Yu. “Predicting user purchase in E-commerce by comprehensive feature engineering and decision boundary focused under-sampling”. In: *Proceedings of the 2015 International ACM Recommender Systems Challenge*. ACM. 2015, p. 8.

[45] Dirk Van den Poel and Wouter Buckinx. “Predicting online-purchasing behaviour”. In: *European journal of operational research* 166.2 (2005), pp. 557–575.

[46] Jacob Poushter et al. “Smartphone ownership and internet usage continues to climb in emerging economies”. In: *Pew Research Center* 22 (2016), pp. 1–44.

[47] Steffen Rendle. “Factorization machines”. In: *2010 IEEE International Conference on Data Mining*. IEEE. 2010, pp. 995–1000.

[48] Steffen Rendle. “Factorization machines with libfm”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.3 (2012), p. 57.

[49]Paul Resnick and Hal R Varian. “Recommender systems”. In: *Communications of the ACM* 40.3 (1997), pp. 56–59.

[50]Francesco Ricci, Lior Rokach, and Bracha Shapira. “Introduction to recommender systems handbook”. In: *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[51]Francesco Ricci, Lior Rokach, and Bracha Shapira. “Recommender systems: introduction and challenges”. In: *Recommender systems handbook*. Springer, 2015, pp. 1–34.

[52]Ruby Roy Dholakia. “Going shopping: key determinants of shopping behaviors and motivations”. In: *International Journal of Retail & Distribution Management* 27.4 (1999), pp. 154–165.

[53]Bracha Shapira. *Recommender systems handbook*. Springer-verlag New York Incorporated, 2015.

[54]Ahu Sieg, Bamshad Mobasher, and Robin Burke. “Web search personalization with ontological user profiles”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 525–534.

[55]Catarina Sismeiro and Randolph E Bucklin. “Modeling purchase behavior at an e-commerce web site: A task-completion approach”. In: *Journal of marketing research* 41.3 (2004), pp. 306–323.

[56]Michael R Solomon. *Consumer behaviour: buying, having and being (6th eds)*. 2004.

[57]Michael R Solomon, Darren William Dahl, Katherine White, Judith L Zaichkowsky, and Rosemary Polegato. *Consumer behavior: Buying, having, and being*. Vol. 10. Pearson Toronto, Canada, 2014.

[58]Euiho Suh, Seungjae Lim, Hyunseok Hwang, and Suyeon Kim. “A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study”. In: *Expert Systems with Applications* 27.2 (2004), pp. 245–255.

[59]Esther Swilley and Ronald E Goldsmith. “Black Friday and Cyber Monday: Understanding consumer intentions on two major shopping days”. In: *Journal of retailing and consumer services* 20.1 (2013), pp. 43–50.

[60]Paolo Toth and Daniele Vigo. *Vehicle routing: problems, methods, and applications*. SIAM, 2014.

[61]Efraim Turban, David King, Jae Kyu Lee, Ting-Peng Liang, and Deborrah C Turban. “E-commerce: mechanisms, platforms, and tools”. In: *Electronic Commerce*. Springer, 2015, pp. 51–99.

[62]Marilyn A Winkleby, Darius E Jatulis, Erica Frank, and Stephen P Fortmann. “Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease.” In: *American journal of public health* 82.6 (1992), pp. 816–820.

[63]Ming-ming Xi and Li-meng Zhu. “Consumer Behavior and Bandwagon Effect: Evidence from “Double Eleven” Online Shopping”. In: *Contemporary Finance & Economics* 7 (2016), p. 001.

[64]Yandi Xia, Giuseppe Di Fabbrizio, Shikhar Vaibhav, and Ankur Datta. “A Content-based Recommender System for E-commerce Offers and Coupons”. In: (2017).

[65]Rui Yao and Jianhua Chen. “Predicting movie sales revenue using online reviews”. In: *Granular Computing (GrC), 2013 IEEE International Conference on*. IEEE. 2013, pp. 396–401.

[66]Jinyoung Yeo, Sungchul Kim, Eunyee Koh, Seung-won Hwang, and Nedim Lipka. “Predicting Online Purchase Conversion for Retargeting”. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM. 2017, pp. 591–600.

[67] Vincent Cheow Sern Yeo, See-Kwong Goh, and Sajad Rezaei. “Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services”. In: *Journal of Retailing and Consumer Services* 35 (2017), pp. 150–162.

[68] Nicholas Jing Yuan, Yu Zheng, Xing Xie, et al. “Discovering urban functional zones using latent activity trajectories”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.3 (2014), pp. 712–725.

[69] Yongzheng Zhang and Marco Pennacchiotti. “Predicting purchase behaviors from social media”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 1521–1532.

[70] Bo Zhao, Hong Huang, Jar-Der Luo, et al. “A Preliminary Study of E-Commerce User Behavior Based on Mobile Big Data-Invited Paper”. In: *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE. 2018, pp. 1–5.

[71] Han Zhu, Xiang Li, Pengye Zhang, et al. “Learning Tree-based Deep Model for Recommender Systems”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM. 2018, pp. 1079–1088.

[72] Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin, Cheng-Wei Wu, and Vincent S Tseng. “EFIM: a highly efficient algorithm for high-utility itemset mining”. In: *Mexican International Conference on Artificial Intelligence*. Springer. 2015, pp. 530–546.

[73] Indre Žliobaite, Jorn Bakker, and Mykola Pechenizkiy. “Towards context aware food sales prediction”. In: *Data Mining Workshops, 2009. ICDMW’09. IEEE International Conference on*. IEEE. 2009, pp. 94–99.

List of Figures

List of Tables