****

****

**人工智能技术专题论文**

**题    目     用户行为分析**

**学   院       自动化学院**

**专 业 物联网工程**

**年级班别 20级(1)班**

**学 号 3120001384**

**学生姓名 陈铿任**

**指导老师 刘冬宁**

**2021年 6 月26 日**

**Content**

[1 Introduction 1](#_Toc70949431)

[1.1 Understanding user behavior analysis 1](#_Toc70949432)

[1.2 Purpose of user behavior analysis 1](#_Toc70949433)

[2 The specific content of user behavior analysis 1](#_Toc70949440)

[2.1 2](#_Toc70949441)

[2.2 User behavior analysis 2](#_Toc70949442)

[2.2.1 Funnel model analysis 2](#_Toc70949451)

[2.2.2 User Rentention analysis](#_Toc70949452) 2

[2.2.3 Analysis of user consumption behavior 3](#_Toc70949452)

[2.2.3.1 Box diagram 3](#_Toc70949452)

[2.2.3.2 Daily ARPPU analysis 4](#_Toc70949452)

[2.2.3.3 Analysis of daily ARPU](#_Toc70949452) 4

[2.2.3.4RFM model analysis 4](#_Toc70949452)

[2.2.4 Behavior Event Analysis 5](#_Toc70949452)

[3 Example analysis – Taobao user behavior analysis 5](#_Toc70949449)

[3.1 Introdution of the data set 5](#_Toc70949450)

[3.2 Data pre-processing – data cleaning 5](#_Toc70949453)

[3.2.1 Handling of missing values 6](#_Toc70949454)

[3.2.2 Handling of data labels 6](#_Toc70949455)

[3.3 Behavior Analysis 7](#_Toc70949453)

[3.3.1 User Rentention Analysis 7](#_Toc70949454)

[3.2.2 Data comparsion of user behavior 8](#_Toc70949455)

[3.2.3 Model anlysis 10](#_Toc70949455)

[3.2.3.1 Funnel model analysis 10](#_Toc70949454)

[3.2.3.2 Box diagram 11](#_Toc70949455)

[3.2.3.3 Daily ARPPU analysis 12](#_Toc70949455)

[3.2.3.4 Analysis of daily ARPU 12](#_Toc70949455)

[3.2.3.5 Repurchase rate 13](#_Toc70949455)

[3.2.3.6 RFM model analysis 13](#_Toc70949455)

[4 Work summary and insights 14](#_Toc70949469)

[4.1 Work summary 14](#_Toc70949470)

[4.2 Insights 15](#_Toc70949471)

**1. Introduction**

**1.1** **Understanding user behavior analysis**

With the continuous development of big data in recent years, big data has become a popular industry with applications in various fields.

User behavior analysis is one of the application directions. With the rapid development of smart devices, people's life has become inseparable from smart devices. People spend most of their time in smart devices on shopping platforms, and we browse information on shopping platforms will generate different information to reflect our behavior.

User behavior information is very important in e-commerce platform, through user behavior analysis, we can understand the user's situation, such as: their different living habits, cultural background, life location and so on these information will affect their behavior. So what is user behavior analysis?

User behavior information refers to the case of obtaining the basic data of website or APP and other platform visits, the statistics and analysis of the data therein, from which the laws of user visits to the website or APP and other platforms are found to establish, further amend, or re-designate online marketing strategies and so on to achieve refined operations.

**1.2** **Purpose of user behavior analysis**

User behavior data becomes very important for products in these industries such as Internet finance, e-commerce platforms, online education, securities, and short videos. The purpose of user behavior is to customize retention analysis, refine channel quality assessment, product analysis, and precision marketing.

1. Custom retention analysis: Based on user behavior, we can do refined retention assessment. According to the product characteristics, we can customize user retention and further evaluate and optimize the product according to the active situation of users.
2. Refined channel quality assessment: With users brought through different channels, we need to understand which channel has the richest and most valuable user behavior data, so as to reduce channel costs and improve channel conversion.
3. Product analysis: Through funnel analysis or path conversion, we can find key points for product improvement and relevant factors that promote core conversion, thus making product design more humane.。
4. 精准营销：Precise marketing: locate users' interests through their behavior of clicking on product details, search behavior, browsing history, and purchased things, so as to develop better personalized services.

**2.** **The specific content of user behavior analysis**

**2.1Metrics for user behavior analysis**

The indicators of user behavior analysis can be broadly classified into three categories based on user behavior performance, namely stickiness indicators, activity indicators and output indicators.

1. Stickiness metrics: mainly focus on the user in a certain period of time to continue to visit he use of the site, a continuous state. For example: the number and proportion of new users, the number and proportion of active users, user retention rate, user churn rate, user access rate, etc.

[2] Activity metrics: point to the behavior that occurs during each user visit, mainly refers to the engagement of user visits, such as: active users, new users, lost users, frequency of use, etc.

[3] Output metrics: measure the value output created by users, such as: number of page views (PV), number of page visitors (UV), number of clicks, etc.

**2.2** **User behavior analysis model**

Qualitative and quantitative analysis of data (user-specific behavior) is performed with the help of a number of models.

Commonly used analysis models are：

1. Funnel model analysis
2. User Retention Analysis

Analysis of user consumption behavior 1. Box chart

2. Daily ARPPU analysis

3. Analysis of daily ARPU

4. RFM model

4) Behavioral event analysis

**2.2.1** **Funnel model analysis**

Funnel model analysis is to describe the user conversion and churn rate of key links in each stage in the multi-city such as users using APP or visiting website, which can scientifically reflect the state of user behavior and the conversion rate of users in each stage from the starting point to the end point.

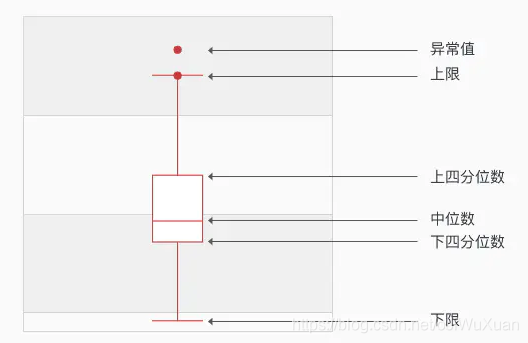
The funnel model can analyze and verify whether the design and decision of the whole process is reasonable and satisfies the needs of customers. By analyzing the conversion rate related to the user behavior process, we can find out which link has a relatively low conversion rate and needs to be improved, so as to discover and solve the problems. **2.2.2** **User Retention Analysis**

User retention analysis is a model used to analyze the level of user activity. The retention volume and retention rate (PV, UV and other indicators) are used to understand the actual behavior of current users so as to know the retention and churn of users, to find out the key factors affecting the sustainable growth of the product, to know the market decision, product improvement, to enhance user value, etc.

**2.2.3** **Analysis of user consumption behavior**

User consumption behavior is a way to understand the user's consumption frequency and consumption amount and the user's income under the access to the user's consumption information, so as to predict the user's repurchase rate (the number of users of repurchase behavior / total number of users of consumption behavior), the classification of the user's value level, so as to further make good decisions and timely adjustments.

**2.2.3.1** **Box diagram**

A box plot is one of the analytical charts used as a statistical chart of the dispersion distribution of a set of data, and is used to handle abnormal data and data filtering. (It is not affected by outliers and can accurately depict the discrete distribution of data in the first place)

Where the box plot contains 5 important factors.

1. Upper quartile U: indicates that only 1/4 of the values in all samples are greater than U, i.e., U is at 25% when sorted from largest to smallest.

2. Lower quartile L: indicates that only 1/4 of the values in all samples are smaller than L, i.e., at 75% when sorted from largest to smallest.

3. Median Q: indicates that a group of numbers arranged from small to large is in the middle of the number, if the number of sequences is even, the median of the group is the average of the middle two numbers.

4. Upper limit: indicates the maximum value of the non-abnormal range, the quartile distance is IQR=U-L, then the upper limit is U+1.5IQR.

5. Lower limit: indicates the minimum value in the non-abnormal range, i.e., the lower limit is L-1.5IQR.

**2.2.3.2** **Daily ARPPU Analysis**

ARPPU (average revenue per paying user) is the revenue earned from each spending user, i.e. the average spending on the product by the spending user within the statistical date.

ARPPU = total revenue consumed / number of users consumed **2.2.3.3** **Analysis of daily ARPU**

ARPU (average revenue per user) is the average revenue per user. ARPU focuses on the salesman situation in a time period, the higher the ARPU value then the higher the profit, the better the efficiency, which can measure the profitability and development of the product vitality.

ARPU = total consumption revenue / total number of users **2.2.3.4RFM model analysis**

RFM model is a method of customer value segmentation based on the contribution of customer activity and transaction value. Identifying quality customers, it can specify personalized communication and marketing services, and can measure customer value and customer profit generating ability.

1.R(Recency) - the time interval of the most recent transaction, based on the date of the most recent transaction to calculate the score, for example, the closer the current date, the higher the score.

2.F(Frequency) - the number of times the customer has traded in the recent period, based on the frequency of transactions, the higher the frequency of transactions, the higher the score.

3.M(Monetray) - the amount of transactions made by the customer in the recent period, the higher the transaction amount, the higher the score.

RFM The model's assumptions are based on the following premises.

1. Customers who have recently traded > those who have not traded recently

2. Customers with high transaction frequency > those with low transaction frequency

3. Customers with a high transaction amount > those with a low transaction amount

Customized scoring rules are used to classify different types of customers, which ones to focus on and maintain, which ones to explore the reasons for churn, etc.

**2.2.3** **Behavioral Event Analysis**

Behavioral event analysis refers to the behavioral analysis of users based on the acquired index data, by tracking or recording user behavioral events to understand the trend direction of the event and the user's situation, so as to dig into the impact factors of the event.

It mainly studies the 5W behaviors of users, i.e. who the user is, from what place, what event happened, how it was generated, when it happened, so as to explore the impact of which on products and enterprises, etc.

Behavioral indicators：

1) Click-through rate CTR = clicks / exposure (products, consulting content, videos) \* 100%

2) Exit Rate

3) Consumption frequency

**3.** **Example Analysis**

The following is an analysis of Taobao user behavior to illustrate the use of the process.

**3.1** **Introduction of the data set**

The dataset is sourced from AliTianchi (which provides a complete IDE as well as rich computing resources) and contains 1048576 user data for the period 2014-11-18 to 2014-12-18.

|  |  |
| --- | --- |
| Features | Instructions |
| user\_id | The user name |
| item\_id | The commodity name |
| behavior\_type | Behavior(Click,Favorite,Add to cart,Buy corresponds to 1,2,3,4) |
| user\_geohash | The geographical position of user |
| item\_categort | Commodity categories |
| time | The time that the user behavior takes place |

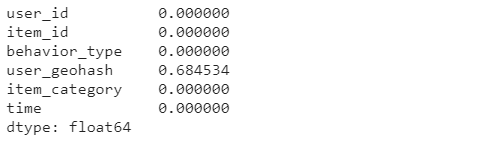
**3.2** **Data pre-processing - data cleaning**

**3.2.1** **Handling of missing values**

A custom function is used to calculate the ratio of the number of missing values under each feature to the total number of that feature.

Code Implementation：

|  |
| --- |
| DataFrame.apply(lambda x:sum(x.isnull()/len(x))) |

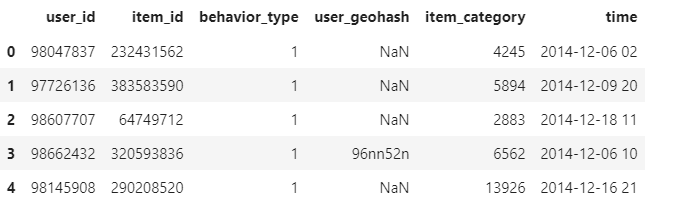


Through the observation of user\_geohash (The location of user while behavior take palce) this feature is also too many missing values, then this feature can be removed.

Code Implementation：

|  |
| --- |
| DataFrame.drop('user\_geohash',axis=1,inplace=True) |

**3.2.2** **Handling of date labels**

The data of the tag <time> contains both date and time parts as observed by DataFrame.head(). To facilitate the analysis of daily and at the same time user behavior, <time> is divided into two tags <time>, <hour>.

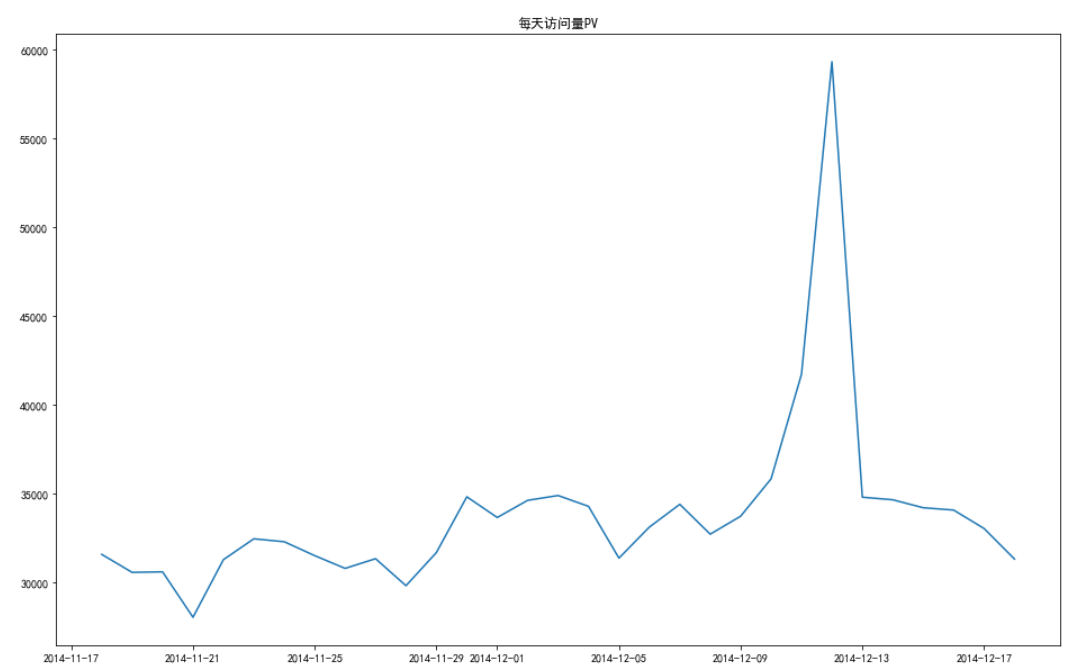
Code Implementation：

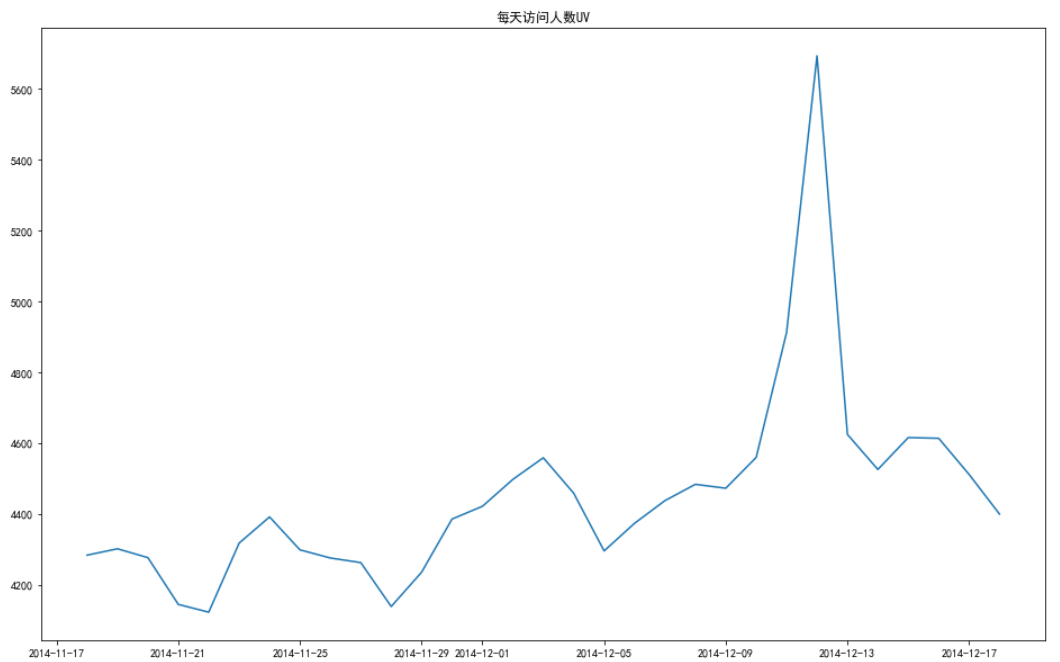
|  |
| --- |
| df['time'] = pd.to\_datetime(df['time'],format="%Y-%m-%d")  df['hour'] = df.time.dt.hour# Extract the specific time (hours) part of time  df['time'] = df.time.dt.normalize() # Extracting the date part of time |

**3.3 Behavior Analysis**

**3.3.1 User Rentention Analysis**

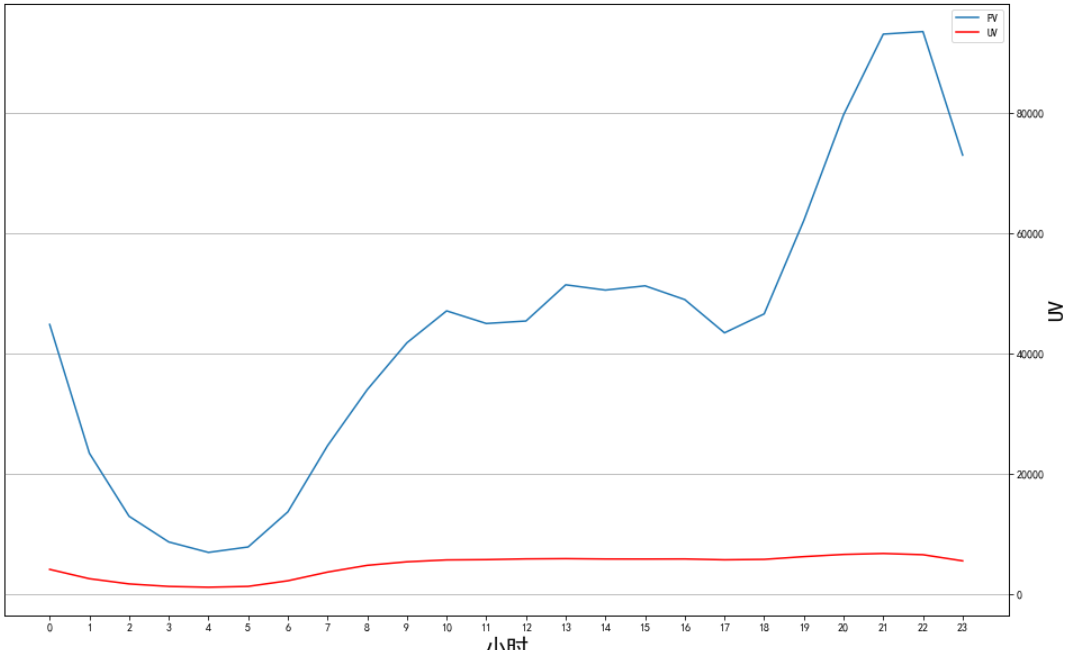
First, the data dimension is analyzed：





It was found that the number of visitors and the number of visits peaked on Double 12, and both PV and UV were similar in image shape.

Then, analysis of the hourly dimension ：



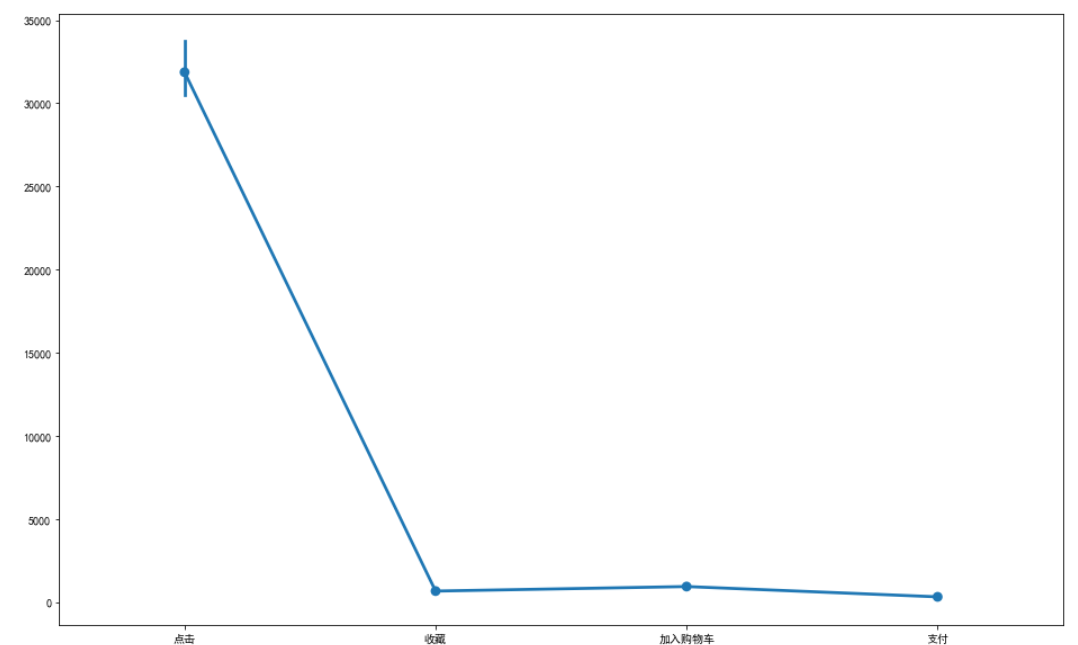
Through the chart observation, UV overall no very obvious fluctuations but in the morning 0-6 points the number of people will relatively decline;PV at 20-22 o’clock will appear to reach the peak of the visit, the data is more normal than UV.

**3.3.2 Data comparison of user behavior**

Comparison of data totals for the four behaviors：

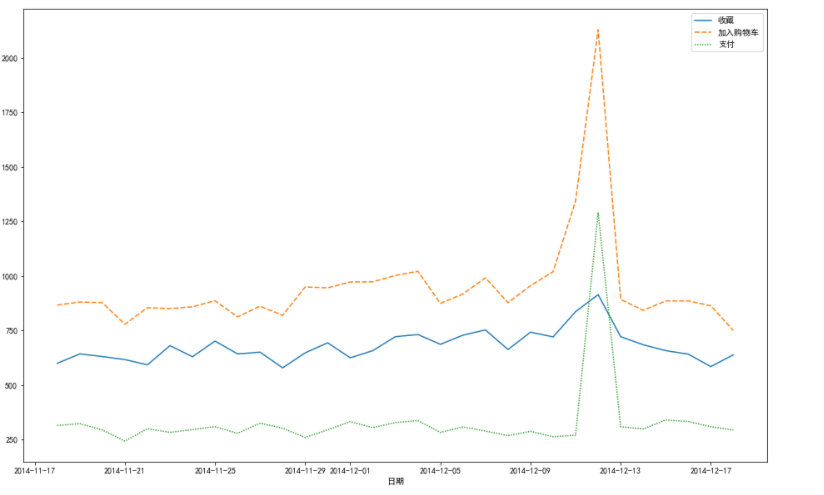
Code implementation：

|  |
| --- |
| behavior\_data=pd.pivot\_table(df,index='time',columns='behavior',values='user\_id',aggfunc=np.size)  behavior\_data.columns = ["click","favorite","add to cart","pay"]  plt.figure(figsize=(16,10))  sns.pointplot(data=behavior\_data[["click","favorite","add to cart","pay"]])  plt.show() |

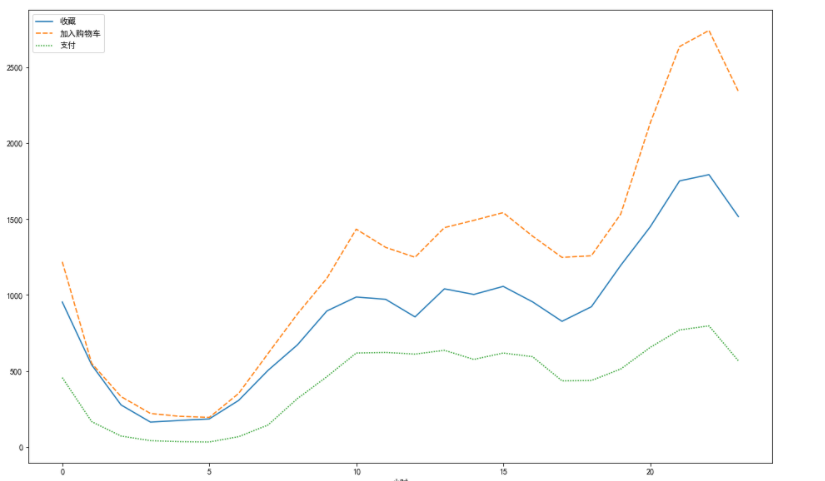


Conclusion: The amount of data clicked is much higher than the other three, so the following in the date and hour down to study the behavior and compare only three behaviors.

Analyze three types of user behavior by date dimension：



Conclusion: Again, in the double 12 shopping festival can cause the user's desire for shopping , are in the double 12 period to reach a peak

Analyze three types of user behavior in the hourly dimension：

**3.3.3 Model analysis**

**3.3.3.1 Funnel model analysis**

Order of behavior: Click->Collect->Add to Cart->Purchase

Calculation of single conversion rate: It is the number of people in the previous act divided by the number of people in the next act \* 100% (retain two decimal points), and the conversion rate of the first act needs to be set to 100%. (as a reference system)

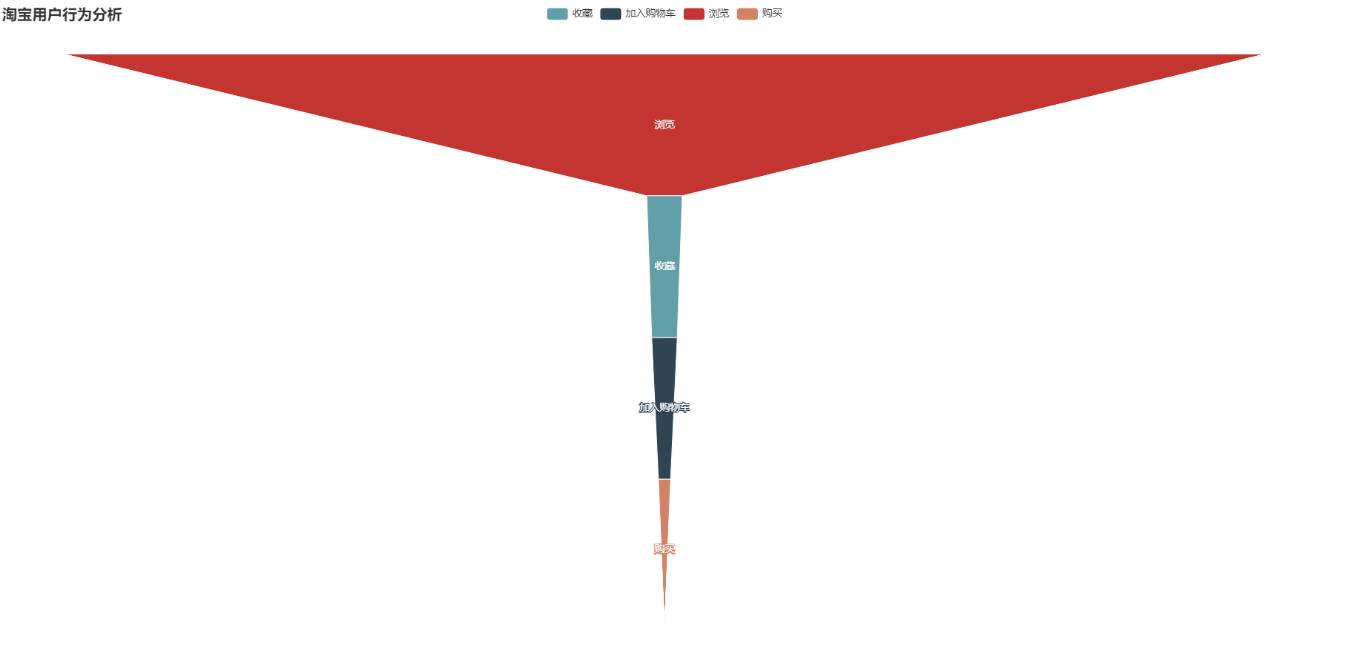
Code implementation：

|  |
| --- |
| tnum = np.array(df\_behavior["number\_of\_people"])[:-1]  num = np.array(df\_behavior["number\_of\_people "])[1:]  conversion\_rate = num/tnum  conversion\_rate.insert(0,1)  print(conversion\_rate)  conversion\_rate = [round(x\*100,2)for x in conversion\_rate] |

Calculation of the overall conversion rate: it is the number of people for each action divided by the number of people for the first action, respectively.

Code implementation：

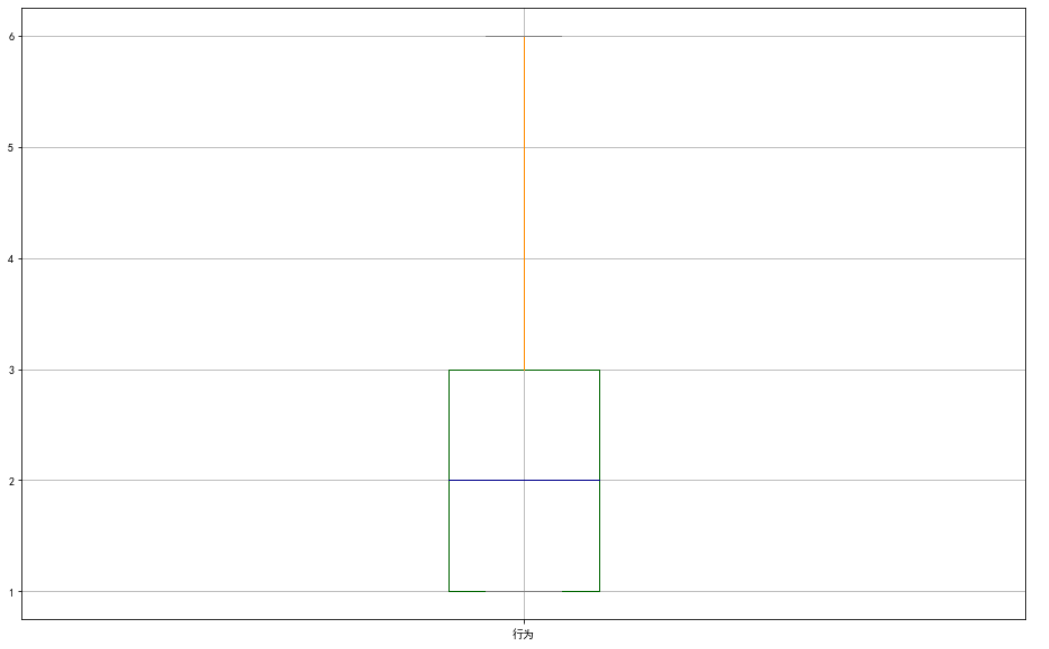
|  |
| --- |
| total\_conversion = df\_behavior["number\_of\_people "]/df\_behavior["number\_of\_people "][0]  total\_conversion = [round(x\*100,2)for x in total\_conversion] |



By looking at the funnel chart, we see that the conversion rate from browsing to buying is very low, and it is time to explore the reasons for this.

The company's conjecture is that this means that it may be due to the user not finding the product they want during the browsing process, improving the personalized recommendations of the product or the product not being suitable for most consumers resulting in the product not being sold or something going wrong when paying or seeing a more satisfying product at the time of purchase.

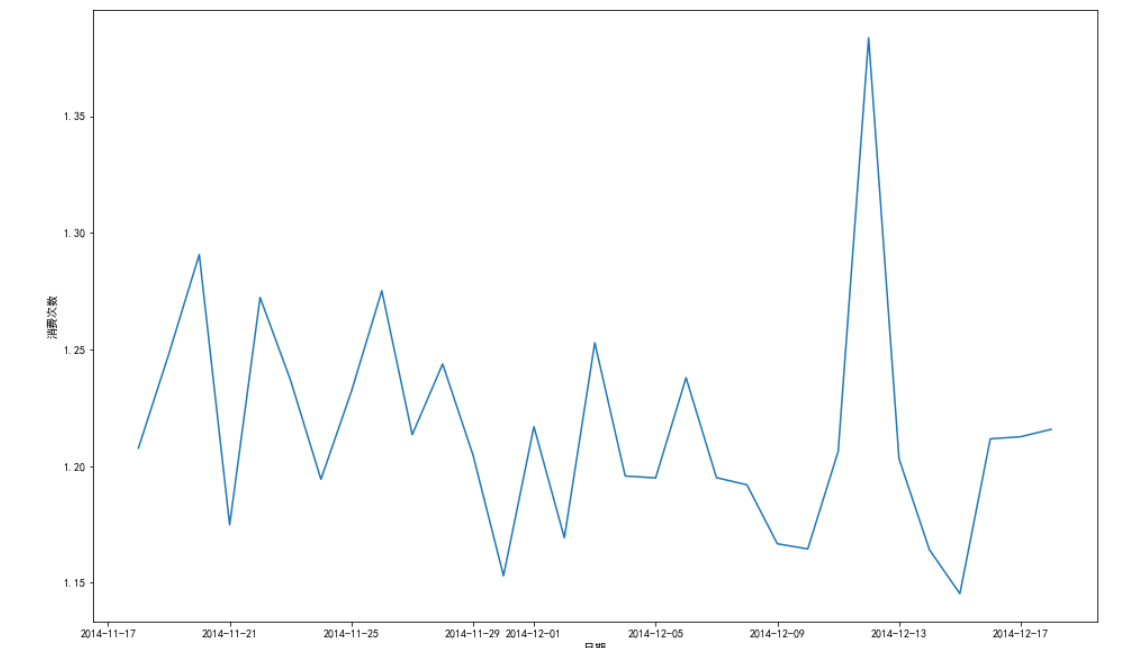
**3.3.3.2 Box diagram**



Trough the observation found that 25% to 50% and 50% to 75% box weight accounted for the same percentage, which indicates that most users are in 1-3 times consumption of this number; the number of users' consumption in general is on the low side.

**3.3.3.3 Daily ARPPU analysis**

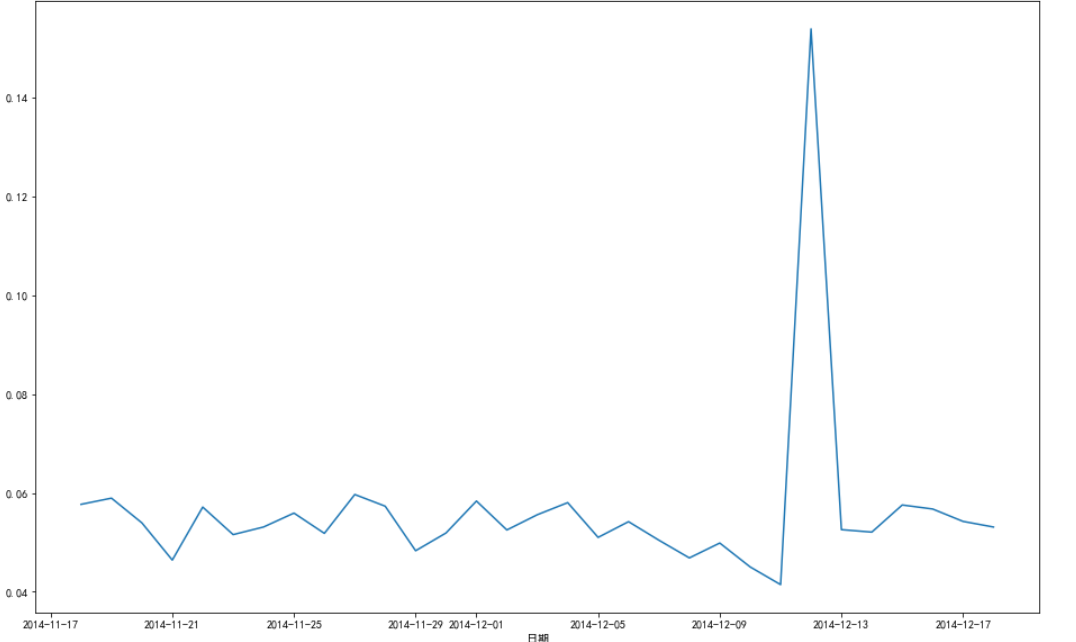
In this dataset there is no user's spending amount, we replace the spending amount with the number of spending times.



Through the ARPPU-day statistics chart, we found that the number of user consumptions fluctuates mainly between 1.20 and 1.30 on some usual days, but increases dramatically during the Double 12 shopping holiday, with a peak value of around 1.40. This indicates that some festivals will promote users to do shopping.

**3.3.3.4 Analysis of daily ARPU**

Again, since there is no consumption income data in this dataset, we use the number of consumptions instead of income.



Through the ARPU\_day statistics chart we find that the number of users' spending in this dataset is very small, the average number of spending per user is less than 0.2.

**3.3.3.5 Repurchase rate**

Code implementation：

|  |
| --- |
| Re\_buy = df[df.behavior==4].groupby('user\_id')['time'].apply(lambda x:len(x.unique()))  print(round(Re\_buy[Re\_buy>=2].count()/Re\_buy.count(),2)) |

By calculating the repurchase rate, the result is 0.47, which means that the product is quite unstable and does not meet people's needs, resulting in a rather low consumption and repurchase rate, and the company should make timely treatment to stop the loss urgently and make the next marketing strategy.

**3.3.3.6 RFM model analysis**

Since there is no spending amount of users in this dataset, here only R (last spending time) and F (spending frequency) are used for scoring to classify users.

Customized scoring rules：

Code implementation：

|  |
| --- |
| def Recency(day):  if day < 5:  return 5  elif day >=5 & day < 10:  return 4  elif day >=10 & day < 15:  return 3  elif day >= 15 & day < 20:  return 2  elif day >= 20 & day <25:  return 1  elif day >= 25:  return 0    def Frequency(time):  if time < 15:  return 0  elif time >= 15 & time < 30:  return 1  elif time >= 30 & time <45:  return 2  elif time >= 45 & time < 60:  return 3  elif time >= 60 & time < 75:  return 4  elif time >= 75:  return 5 |



Conclusion: Dividing customers with different values helps to give different personalized services for different customers and can create better value.

**4.** **Work summary and insights**

**4.1Work summary**

This course on the topic of artificial intelligence technology has given me a certain understanding of many areas of artificial intelligence, such as medical imaging, intelligent optimization, natural language processing, and text mining, etc. Among them, I have become very interested in the field of data mining and analysis. Secondly, the topic of this time is an application of data mining and analysis mentioned in the teacher's class, which combined with my own ideas to determine the topic of this thesis.

In this work, I learned to use pyecharts, numpy, pandas, seaborn, matplotlib library and understand the whole process of user behavior analysis. Through a Taobao user behavior analysis example, learn and master some methods of user behavior analysis, from data pre-processing (data cleaning, feature selection) to behavior data comparison analysis to model analysis, step by step, there are a lot of gains.

**4.2 Insights**

Although I encountered a lot of twists and turns during the learning process, I also made some gains and realized that I have a lot of shortcomings in many places, my own shortcomings in technology, and I still have a lot to learn for myself who just started to contact data mining and analysis. At the same time, I would like to thank the artificial intelligence technology course for helping me find the direction I love.

Since I love data mining and analysis, I will keep going. Starting from a person who knows nothing, reading blogs, checking documents on the official website, and doing hands-on practice. Every time I learn a point of knowledge, it will involve other new points of knowledge, although the process may be a bit complicated, but after learning a new thing, it will be applied to the example above.

The knowledge I have learned now is only the skin of data mining and analysis, and there is still a long way to go to learn data mining and analysis more deeply, so I should not slacken off ，hopeing myself will keeping learning so that i will have achievements in this area.

# 参考文献

1. [Python异常数据处理——箱型图分析-CSDN博客](https://blog.csdn.net/oJiWuXuan/article/details/107815803?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522162466751716780274113693%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=162466751716780274113693&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_positive~default-1-107815803.pc_search_result_before_js&utm_term=%E7%AE%B1%E5%9E%8B%E5%9B%BE&spm=1018.2226.3001.4187)

1. [用户行为分析模型——RFM模型-CSDN博客](https://blog.csdn.net/weixin_41725746/article/details/95210428?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522162466781816780255267496%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=162466781816780255267496&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~top_positive~default-1-95210428.pc_search_result_before_js&utm_term=RFM%E6%A8%A1%E5%9E%8B&spm=1018.2226.3001.4187)

1. [漫谈用户行为分析的价值和意义 - 简书](https://www.jianshu.com/p/fbb6897dee90?open_source=weibo_search)

1. [pyecharts图表库学习:Funnel（漏斗图）-CSDN博客](https://blog.csdn.net/HHG20171226/article/details/103114757?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522162466814616780269896653%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=162466814616780269896653&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~baidu_landing_v2~default-7-103114757.pc_search_result_before_js&utm_term=pyecharts%E6%BC%8F%E6%96%97%E5%9B%BE&spm=1018.2226.3001.4187)

1. <https://seaborn.pydata.org/generated/seaborn.pointplot.html?highlight=pointplot#seaborn.pointplot>

1. [用户消费行为分析-CSDN博客\_](https://blog.csdn.net/liu_ziyue/article/details/105645613?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522162466788316780264068757%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=162466788316780264068757&biz_id=0&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduend~default-2-105645613.pc_search_result_before_js&utm_term=%E6%B6%88%E8%B4%B9%E8%A1%8C%E4%B8%BA%E5%88%86%E6%9E%90&spm=1018.2226.3001.4187)

1. [python中的datatime.dt.day一系列-CSDN博客\_dt.days](https://blog.csdn.net/starmoth/article/details/91469115?ops_request_misc=&request_id=&biz_id=102&utm_term=pythondt&utm_medium=distribute.pc_search_result.none-task-blog-2~all~sobaiduweb~default-1-.pc_search_result_before_js&spm=1018.2226.3001.4187).