Cohort's Analysis in eCommerce

Manali M Walanj

Masters in Computer Science, Somaiya University, Mumbai, Maharashtra, India

ABSTRACT

Cohort analysis treats an outcome variable as a function of cohort membership, age, and period. The linear dependency of the three temporal dimensions always creates an identification problem. Resolution of this problem requires external knowledge that is often difficult to acquire. Most satisfactory is the introduction of variables held to measure the dimensions that underlie at least one of age, period and cohort. Such measured, substantive variables can provide direct tests of cohort-based explanations. A Promising path for future technical development is a hierarchical Bayes approach, which treats appropriately defined cohort, age, and period contrasts as randomly distributed and allows for their dependence on substantive, measured variables. Models that include age, period, and cohort can also include interactions between these dimensions, but not all such interactions are identified. This extends the realism of cohort models, since many phenomena seem to require specifications that allow for interactions between two or more of age, period, and cohort. Panel studies and cross-sectional studies with retrospective information not only support cohort analyses, they engender them. These longitudinal data structures do not, however, provide the basis for a solution to the identification problem.[5]

Keywords : Cohorts Analysis, Data, Extraction, Online, Graphs

I. INTRODUCTION

In today's digital life and pandemic situation, eCommerce businesses are in high demand. This also means that the competition in the field is on the rise. Online businesses are on the constant lookout for making strategies that can help them grow as the number of competitors are increasing. One relevant way is to utilize and analyze data appropriately to understand the customer behaviors and buying patterns. This is where cohort analysis comes in. A cohort is "a group of individuals having a statistical factor (such as age or it belongs to a specific class) in common in a demographic study". In terms of eCommerce, a cohort is decided based on specific characteristics, usually, a particular action carried out within a specific frame of time. For example, everyone who bought their first product from the company in January 2020. Cohort analysis refers to the practice of studying and analyzing the behaviors, habits, and activities of a particular cohort over a fixed period.[3]

II. METHOD

The steps typically involved in the analysis process include:

1. Extracting raw data: Raw data is taken from a database using MySQL and exported into spreadsheet software, where user attributes can be joined and further segmented.
2. Creating cohort identifiers: Group user data into different parts, such as date of joining, date of first purchase, graduation year, all mobile devices at a particular place and time, etc.
3. Calculating life cycle stages: Once users have been divided into cohorts, the amount of time between events attributed to each customer is measured in order to calculate life cycle stages.
4. Creating tables and graphs: Pivot tables and graphs create visual representations of user data comparisons, and help calculate the aggregation of multiple dimensions of user data.

Example

Performing cohort analysis on online retail data

1. Performing df.head() on the data set to get the starting values out of it

	Invoice	StockCode	Description
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS
1	489434	79323P	PINK CHERRY LIGHTS
2	489434	79323W	WHITE CHERRY LIGHTS
3	489434	22041	RECORD FRAME 7" SINGLE SIZE
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX

2. df.info() getting information in table

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  datetime64[ns]
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 65.1+ MB
```
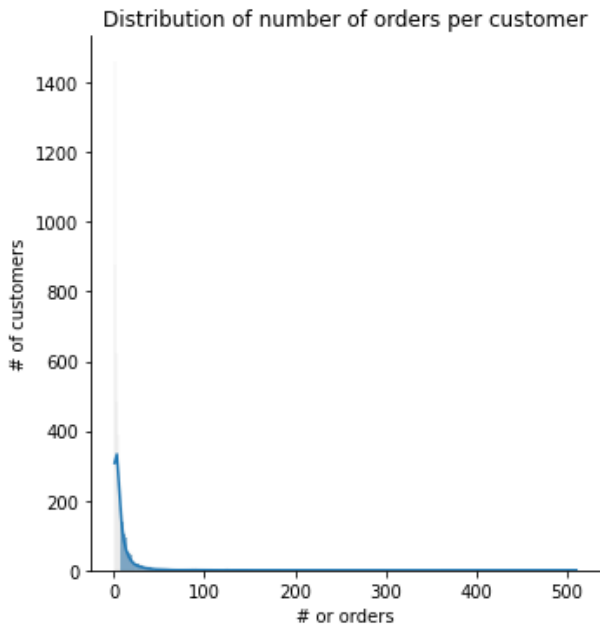
3. df.describe() describing the values in table

	Quantity	Price	Customer ID
count	1.067371e+06	1.067371e+06	824364.000000
mean	9.938898e+00	4.649388e+00	15324.638504
std	1.727058e+02	1.235531e+02	1697.464450
min	-8.099500e+04	-5.359436e+04	12346.000000
25%	1.000000e+00	1.250000e+00	13975.000000
50%	3.000000e+00	2.100000e+00	15255.000000
75%	1.000000e+01	4.150000e+00	16797.000000
max	8.099500e+04	3.897000e+04	18287.000000

4. df.isnull() checking null values from the dataset

```
Invoice           0
StockCode         0
Description    4382
Quantity          0
InvoiceDate       0
Price             0
Customer ID  243007
Country           0
dtype: int64
```

5. Plotting the graph of the values

Distribution of number of orders per customer

6. Cohort of the given data

Advantages of cohort analysis

● Improve Customer Retention

At the highest level cohort analysis can quickly provide you a clear picture of customer retention in terms of speed to second transaction and orders per customer over time. Both of these translate to Lifetime Value (LTV) and most importantly profits as you're pulling more revenue without additional acquisition costs. Based on the results of this high level analysis you'll have a better understanding of the improvement or degradation in quality of newly acquired customers.

● Predict Future Behaviors

Past customer behavior can often be your best indicator of future behavior. Being able to chart out expectations from a retention perspective can enable better planning of upcoming revenue and inventory allowing you to invest more efficiently into filling your funnel. Additionally, for seasonal businesses it can provide crucial understanding of how seasonal shoppers perform long term.[1]

Important Metrics for Cohort Analysis

Below are a few of the important metrics that are necessary for effective cohort analysis

● Repeat Rate / % to Second Order - This depicts the number of returning customers vs. one-time buyers in a particular cohort. This is a key metric for understanding your retention rate.

● Orders per Customer - As the name implies, it is the number of orders from a single customer in a cohort. Repeat customers usually have higher orders per customer.

● Time between Orders - This is a way to measure the frequency of buying of a customer. It is particularly useful for understanding when to send out those marketing emails to your customer base.

● Customer Lifetime Value (CLV) - This cohort-specific metric allows us to measure customer loyalty in terms of the money spent.

● Average Order Value (AOV) - This explains how much different customers spend when buying from us. Sometimes, it helps to boost sales by targeting marketing efforts towards buyers with higher AOV.

III. RESULTS AND DISCUSSION

The goal of cohort analysis is to identify the customer's behaviour and buying patterns which indirectly helps in boosting the sales of the company.

Cohort studies are sorts of observational studies during which a cohort, or a gaggle of people sharing some characteristic, are followed up over time, and outcomes are measured at one or more time points. Cohort studies are often classified as prospective or retrospective studies, and they have several advantages and drawbacks . The essential characteristics of cohort studies and includes recommendations on the design, statistical analysis, and reporting of cohort studies in critical care medicine. Tools are provided for researchers and reviewers.Such analyses must contend with the linear dependency between age, period, and cohort membership. The use of one or more measured variables held to underlie at least one of age, period, or cohort can break the linear dependency. So too can application of credible prior information, whether expressed as constraints on coefficients in fixed effect models, or as exchangeability assumptions in random effects or hierarchical models. Measured variables can, of course, be incorporated into both fixed effect and random effect models. This strategy is to be preferred, since it makes it possible to check ideas about substantive processes within the most direct way[2]

IV. REFERENCES

[1]. Lopez Rivera, Ibrahim "Developing Online Trust in Electronic Commerce: A Generational Cohort Study in Puerto Rico" ProQuest LLC, D.B.A. Dissertation, Universidad del Turabo (Puerto Rico) 978-0-3551-7065-8

[2]. Fridell, Gustav "IT'S IN THE DATA: A multimethod study on how SaaS-businesses can utilize cohort analysis to improve marketing decision-making" Linköping University, Department of Management and Engineering. ISRN: LIU-IEI-TEK-A--20/03665—SE

[3]. Inter-Generational Comparison of Social Media Use: Investigating the Online Behavior of Different Generational Cohorts , 10.1109/HICSS.2016.477

[4]. LawrenceL.KupperJoseph M.JanisAzzaKarmousBernard G.Greenberg "Statistical age-period-cohort analysis: A review and critique" Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC 27514, U.S.A.

[5]. David Barrett, Helen Noble "What are cohort studies?", This approach to research does bring with it some important challenges—often related to their size, complexity and longevity.