eCommerce behavior data from multi category store

Mzoxolo Mbini

MIT805 Big data, Department of Computer Science, Univrsity of Pretoria, Hatfield, Pretoria, South Africa
Email: u16350244@tuks.co.za

Keywords: Stream data; Big Data V's; E-Commerce

Received 31 September 2021; revised 21 August 2021

1. INTRODUCTION

The trend of shopping online with e-shops goes hand in hand with the digital age. Shopping online seems to be a success story.

However, several vendors support developing retail shops to support face-to-face contact with customers as a unique selling point. Each way of shopping has its own perks, but also has downfalls the customer who chooses to buy.

Using commercial data to drive eCommerce business has become the gold standard. An eCommerce strategy must be based on data just like any other business. Without reliable data, there's no way to know the traffic on the website or even how the business is doing.

Understanding what drives customers' behaviour is the holy grail of the eCommerce industry. Knowledge can be used to improve the shopping process and ultimately result in higher sales and customer satisfaction.

REES46 Technologies built an eCommerce platform Open Online CDP for the United States-based retail stores. The Open Online CDP listens to the user activities (business processes) such requests, product view, move product(s) cart and update user details. All these processes are actioned and recorded in real-time.

The paper focuses on the technical aspects of the output of the recorded business processes. The paper continues to describe why and how the dataset is generated, its metadata, and insights that can be drawn from it.

2. DATA-SET ORIGINS

In one convenient interface, REES46 provides marketing professionals and online store owners with the intelligence and technology they need to develop their online businesses. A multinational team of aspiring marketers and developers with eCommerce backgrounds founded REES46 in 2013.

The founders introduced Progressive Penalisation and have been building a database of virtual customer profiles ever since. It holds over 199 million customer profiles, each with its detailed digital footprint; from gender and age to interests, location, travel statistics and even information about kids and pets. That brings targeting to an entirely new level.

The founders optimized the recommending algorithms in 2014 to run all calculations in under 30 milliseconds and used machine learning to improve the cold start issue. The founders added triggering emails with dynamic product recommendations personalized for each viewer in 2014, and location-based marketing services based on iBeacon technology offered promising results.

In 2016, the founders launched many new features: web push notifications, a re-marketing tool to re-target abandoned visitors, a feedback a system incorporating customer reviews and seller reputations, and new segmentation algorithms with smarter segmentation.

The founders operate worldwide and build strong connections in Europe and the United States. A user-friendly unified solution is being developed, combining powerful marketing tools that can be used separately or as part of a bundle, and calculating every key performance indicator based on real-time analytics. A single interface to manage every step

3. DATASET

The eCommerce dataset from REES46 contains transactions generated from 285 millions of users. The file contains behaviour data from a large online multi-category store, collected by the Open Online CDP project that is owned by the REES46 for a month (November 2019). The file consists of rows where each row represents events. Each event recorded is a user and product interaction. Therefore, there is no event without a product or user recorded.

3.1. Metadata

The dataset description and how each of the columns can be interpreted as follows.

1. **event_time**
   The time the event occurred
2. **event_type**
   Types of events that took place that includes view, cart, remove from cart, purchase
3. **product_id**
   Product unique identifier
4. **category_id**
   Product category unique identifier
5. **category_code**
   Business name for a category the product belongs to
6. **brand**
   Name of the product brand or manufacturer
7. **price**
   Product purchase cost
8. **user_id**
   Permanent user unique identifier
9. **user_session**
   User session unique identifier

4. V'S ON THE DATASET

4.1. Volume

Data volume is measured by the amount of data being processed and stored. Using the eCommerce behaviour data from the multi-category store dataset, there are 285 million events relating to eCommerce websites. The total time spent on the website in November 2019 was 18750.55 hours. The transactions are stored as a .csv file format and it has a capacity of 9.01 Gigabytes with 68 million records.

4.2. Velocity

Velocity is the speed at which the data is created, gushed and collected from the source. Open CPD uses streaming data, caching and long background tasks are not used. Only real-time data is processed. The dataset is processed and recorded under a sub-second. The records time difference between records varies from sub-second to minutes.

4.3. Veracity

Veracity is a degree measurement used for accuracy, reliability and trust. The collected data from Open CDP (technology platform owned by REESE46) on large eCommerce stores serving 15M monthly visitors. Certain data details were removed, including sensitive and personal information. Broken data, such as NULL prices or products without categories was also removed.

4.4. Variety

Variety refers to the observed data types contained within the dataset of interest. The dataset is not variable as it only contains structured retail transactions, and are stored in a .csv format. Hence the data is structured as it comes from a relational database.

5. OBSERVATIONS

The data set contains time, product information, and price. They are few businesses process observations that are contained on the dataset.

5.1. Daily traffic

The total count of customer visits in November is 3,696,117, but it's unlikely they came from all platform directions. The spike around November 16th and 17th is probably an on-sell event. This can be further proven if the daily price trend for the price of the product i.e. product_id = 1003461, Xiaomi or 1005115, Apple, and the price is lowered during the 16th and 17th.

5.2. Top five viewed categories

The top five categories that are viewed by customers are: electronic.smartphone, electronic.video.tv, computer.notebook, electronic.clocks and apparel.shoes. Samsung, Apple, Xiaomi, and other brands of electronics are the most popular. The majority of customers come to view or buy electronic gadgets.

It would thus be a strategic decision for the store's upper echelons to decide whether they should focus on this specific category instead of being a multi-category store, or whether other categories would be beneficial?

5.3. Customer journey

Once the item is placed in the cart, only 30% of people add it to the cart (1.36% divided by 4.49%).

6. MAPREDUCE ALGORITHM

6.1. Repository

The code base with the algorithms results are stored on the below repository
url: https://github.com/mzoxolombini/MIT805-Part2

6.2. Attributes chosen

6.2.1. Total sales per brand
A mapper was used to extract the product brand value of the sale for each transaction. A reducer was then used to aggregate the value of these transactions by brand.

Reason for choice of algorithm

The algorithm is fast and not computationally expensive

Results

The total value of sales per brand was found to be **126.61525**

6.2.2. *Total sales per category*

A mapper was used to extract the product category and the value of the sale for each transaction. A reducer was then used to aggregate the value of these transactions by product category.

Reason for choice of algorithm

The algorithm is fast and not computationally expensive

Results

The total value of sales per product category was found to be **140.63878**

6.2.3. *Total number of product brand removed from the cart*

A mapper was used to extract the brand product and the count value removed from the cart. A reducer was then used to aggregate the count value of these transactions by product brand.

Reason for choice of algorithm

The algorithm is fast and not computationally expensive

Results

The total value of the product brand removed from the cart was found to be **242.84**

ACKNOWLEDGEMENTS

The dataset was posted by Michael Kechinov is made available on Kaggle url: https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store

REFERENCES

[1] Khalid Adam, Ismail Hammad, Mohammed Adam, Ibrahim Fakharaldien, and Mazlina Abdul Majid. Big Data Analysis and Storage. *Proceedings of the 2015 International Conference on Operations Excellence and Service Engineering*, (September 2015):648–659, 2015.

[2] Kaggle.com. eCommerce behavior data from multi category store, 2020.

[3] REES46 Technologies. REES46 Open CDP - open source customer data platform.

[4] Andy Yu. Will Customers Buy the Products in their Cart?, 2020.