MIT 805 – Big Data

Semester Assignment – Part 1

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

School of Information Technology

Department of Computer Science

Date: 16 August 2023

By

Murendi Rampai

Student Number: u14169691

Overview of dataset on eCommerce multi-category store customer behaviour

1. Introduction

This report aims to provide an overview of a dataset that holds within it, insightful information about customer behaviour on an online retail platform. The main focus of this report is on the technical aspects of the dataset as well as how the characteristics of the dataset align with the definition of big data in terms of the V's: volume, velocity, variety and veracity. The dataset has been collected from an open source customer data platform, (REES46 Open CDP, n.d.) and is also publicly available and downloadable on the website, Kaggle (Kechinov, 2019). The data platform, REES46 Open CDP, provides data that has been collected from a variety of sources and processes and pre-packages the data to allow businesses to be able to further process the data to obtain valuable insights that can contribute to business growth. The customers and the businesses where this data is collected from remains anonymous in order to protect the privacy of persons involved.

2. Technical Aspects of Dataset

The dataset contains information about customer activity on a multi-category online store in the time span of 1 month – October 2019. The dataset is in the format of a CSV file. The size of the file is 5.5GB and it comprises of more than 42 million rows and 9 columns that describe user activities on the online platform (Kechinov, 2019). The 9 columns in the dataset are described as follows:

- **User ID** – a unique identifier of the users on the platform
- **User Session** – groups the activities that the user completed before logging off or pausing for a long time
- **Event type** – identifies the type of activity undertaken by the user such as purchasing, adding to cart, removing from cart and viewing
- **Product ID** – a unique identifier of the products
- **Brand** – names of brands of the different products
- **Category ID** – a collective group to which similar products belong
- **Price** – The price of each product
- **Event time** – the time at which the activity occurred

The data can be described in terms of the V's of big data as follows:

Volume: The data is large in volume. The csv file contains more than 42 million rows of data and adds up to approximately 5.5GB of data. This amount of data cannot be processed with traditional means in Microsoft excel sheets (for example) as it would require a lot of processing memory and take a lot of time to process which would render the analysis impractical. Big data techniques/technologies such as Python coding have to be used to process the data.

Velocity: The data is not real-time data thus the velocity does not contribute much to the description of this dataset as big data. There is however a timestamp column in the data that shows the time when an activity was started. This timestamp shows multiple activities within a single second in some cases which indicates that the data is generated at high velocities and also explains why there is such a high volume of data for a one-month period.

Variety: The dataset is a structured dataset with a variety of data types. The data types contained in the dataset are dates, times, text, alpha-numeric data and currency. Data types such as date-time require more memory to process than simple numerical data due to their structure and behaviour and therefore add to the complexity of the dataset. The dataset also contains a mixture of quantitative and qualitative data.

Veracity: In the collection phase of the data, the Customer Data Platform collects the data from different data sources and consolidates the data into one data set by matching and removing duplicates. Certain transformations have to be done to combine the data. For example, the user id is obtained by matching different identifiers from the different sources such as email address, phone number, loyalty ID etc. to one user and giving them a single user ID. Data quality issues are likely to be introduced in the consolidation phase of the data

Value: The data gives insight into customer behaviour and product performance on eCommerce platforms and can help retailers increase sales and optimise stock. The data can be analysed to determine which products and product categories are most likely to be purchased and therefore inform decisions with regards to warehouse stocking. Customer behaviour can be analysed to give insight into the needs of customers and inform the decisions on how to market certain products. There is thus a lot of value that can be derived from the data.

3. Data Processing

The data was pre-processed with Python in order to get an understanding of the data and its quality and make it usable for future processing and visualisation. The steps that were taken in pre-processing the data include changing/assigning correct data types, ensuring there are no missing values, transforming columns.

4. Expected Relationships, Insights and Predictions

There is a wealth of information that can be extracted from this dataset. It is expected to extract information such as:

- **Sales trends**: The product prices can be used to analysis sales trend and analyse the relationship between time (time of day/month) and sales. This can help to identify peak times and help to manage web resources accordingly.
- S**imilar product predictions**: Prediction of other products a user is likely to buy can be made based on categorization of products in cart and also based on the recurrence of products in the same carts/sessions

- **Customer conversion rate**: The likelihood that a customer will purchase a product based on their activity history can be modelled to make future predictions
- Popularity of certain products and product categories – The popularity of certain product and product categories can be analysed to allow businesses to make better business decisions
- **Pricing strategies**: The relationship between pricing and purchase performance of different brands of the same product can give insight into pricing strategies

5. Conclusions

The eCommerce customer dataset described in this report hols a wealth of information that can be leveraged to provide valuable insights that can help businesses in making important business decisions to increase sales and improve customer experience by tailoring products according to customer needs.

References

Kechinov, M. (2019) *ECommerce behavior data from Multi Category Store*, *Kaggle*. Available at: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store (Accessed: 15 August 2023).

Rees46 Open CDP (no date) *Open Source Customer Data Platform. REES46*. Available at: https://rees46.com/en/open-cdp (Accessed: 15 August 2023).