MIT 805 - Project Assignment

Part 1

Vimal Ranchhod
u21794023

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

Faculty of Engineering, Built Environment & IT
University of Pretoria

Date: 21 September 2021

1 Introduction

This specific dataset was made available from www.kaggle.com, courtesy of an ecommerce multi-category store. The source was REES46 ecommerce marketing platform. The data[2] consists of transactional items for product purchases and was for the period October 2019 and November 2019. The total size of the dataset is 14.68 GB for both months, however the analysis will be done for a single month which is approximately 5.67 GB with specific reference to October 2019. The data is in the format of a raw CSV file and contains 9 columns, which can be described below.

1.1 Dataset Feature Description[2]

1.1.1 Column1 - ”event_time”

This is a time stamp of when the specific purchase took place. The data type is a date-time object.

1.1.2 Column2 - “event_type”

These are online purchases, so a few different types of events can occur. A customer could view the product, add the product to the cart, remove the product and purchase the product. The data type is a string.

1.1.3 Column3 – “product_id”

This is a unique product identification number. Most likely format is int64

1.1.4 Column4 – “category_id”

This is a unique category number for which a few products might fall into. Most likely format is int64

1.1.5 Column5 – “category_code”

This is a code name for the category ID. The data type is a string.

1.1.6 Column6 – ”brand”

The brand name of the product such “Apple” or “Samsung”. The data type is a string.

1.1.7 Column7 – ”price”

This is the actual price of the product that was purchased. The data type is a float.

1.1.8 Column8 – ”user_id”

The unique reference to a customer. The data type is int64.

1.1.9 Column9 – ”user_session”

The unique reference to a customer session. If the customer logs in after a long pause, then a new session id is assigned. The data type is a string.

2 The Vs of Big data

2.1 Volume

This is a large dataset of approximately 5.6 GB for a single month of user transactions in a tabular format. In total, for October 2019 there are approximately 42 million lines of transactions that have occurred. Due to the size of the data, processing of this data would require some form of big data architecture.

2.2 Variety

This data is in a structured format. The data is obtained from an ecommerce site, from which a lot of useful information can be extracted. It may not have different primary sources but includes variety in features e.g., time, product, customer, price information.

2.3 Velocity

Although this data is static, one could possibly look at streaming data as real time purchases are made for real time analysis. In this case, organizations could make rather accurate and timeous data-driven decisions.

2.4 Value

In creating value, a dataset should contain key features that once analysed can inform managers on what is the best course of action to take. This dataset consists of a number of important features on customer and product information that can provide some useful information to the marketing team. A handy form of analysis can be customer or product segmentation, where a few useful measures can be identified. Three distinct measures common to the ecommerce space are recency, frequency and monetary.

Recency[1]:

This is defined as the last purchase of a certain product or even by a customer. One typically takes the date at the end of the last purchase date of the dataset as the time stamp. Every purchase with respect to a product or customer is calculated when measured in days from the time stamp. In summary, recency will be measured in days where companies usually provide a lot more attention to recent customers or recently purchased products.

Frequency[1]:

This is how often the customers made a purchase within a defined period. Again, this could be applied to the customer and the product. In the dataset, one must be cautious as to the type of product

categories that are to be looked at. Items could be less frequent based on their type. Electronic goods are slow moving goods and might in general have a low frequency.
Monetary[1]:
This is the total expenditure a customer has made within a given period. Again, one should be cautious in terms of high value and low value items as this could be misleading in interpretation. Generally, customers who have spent significantly more will inevitably spend again in the future.

2.4.1 An interpretation of RFM and business value creation

It is the combination of all three of the above that one can start proper segmentation of customers and really create business value. Since all three variables can be calculated and are continuous, one could possibly cluster the categories using an unsupervised clustering algorithm. Once some exploratory analysis has been done, then data will be aggregated either by product or customer id or both. The exploratory analysis will give some high-level insights on popular products, frequency of products, and distributions. I expect clustering allows visibility of high spend and frequent customers, which is indicative that a good customer-retention strategy needs to be in place to constantly engage with these customers. A low recency and frequency cluster could possibly indicate a better promotional strategy or maybe offering a discount option for items or to those customers.

2.5 Visualisation

The expectation is that a lot of insightful visualisation can be attained once analysis has been completed. Exploratory analysis will involve scatter plots and graphs. Clustering solutions will use scatter plots that will clearly define segmented customer groupings.

3 Conclusion

The dataset[2] in summary is a source of interesting analysis and could possibly provide guidance for effective data-driven decisions for ecommerce companies. Going forward, a map-reduce algorithm will be used to process the data of this significant size for final analysis.

References

[1] A Joy Christy, A Umamakeswari, L Priyatharsini, and A Neyaa. Rfm ranking–an effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, 2018.

[2] Michael Kechinov. ecommerce behavior data from multi category store, Dec 2019. `https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store`.