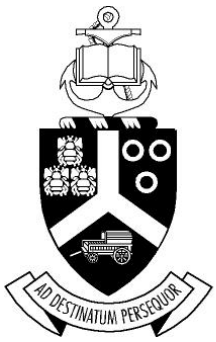

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

MIT 805

Assignment

Author:
P.T Mudau

Student number:
u22783352

September 12, 2022

Introduction

Data generation has been significantly accelerated by technological advancement, and as a result, data has grown in value and become a valuable resource for many enterprises and organizations. Traditional technologies are unable to store and process the amount of data since it is so massive. Big data is a term used to describe "collections of datasets whose velocity, variety, or volume is so great that it is difficult to store, manage, process, and analyze the data using traditional databases and data processing tools" (Snijders et al., 2012; Bahga and Madisetti, 2019). Volume refers to the size of the data, velocity refers to the rate at which data is generated, and variety refers to the different types of data. Recently, the three Vs that defined big data have been expanded to even more Vs, such as value (referring to cost) and veracity (related to accuracy), according to Prabhu et al (2019). According to Bahga and Madisetti (2019), there have been three major changes in business and technology over the past 20 years that have contributed to the challenges with big data today. The first is due to the fact that digital storage has become the most popular and cost-effective method of storing humanly consumable contents such as photographs, audio, video, and documents. Second, the increased rate of data creation and consumption via the web using single or connected devices, and finally, the ever-increasing need for businesses to monitor small and large scale operations or activities to keep up with growing market pressures and outsmart competitors. Numerous big data tools and platforms have had to be developed in the landscape of data management software and architectures as a result of all these changes (Edosio, 2014).
In this assignment, we focus on big data use case based on the many-to-many relation between products and users. With the use of these data, we may more readily comprehend the vast majority of our society and identify patterns that can be used to generate ideas and give them more substance and veracity. This platform contain consumer behavior and is important for any manufacturer since it promotes growth and financial success. The dataset is made available to all interested in customer behaviour through the eCommerce platforms and it is available at no cost for those interested in customer behavior. The dataset has also been used in other databases hosting data science competitions as a result of the popularity of this type of market. In this work, we take into account the dataset of users records from the start of the year 2019 in October and ending in April of the following year. In this assignment, we consider the data of ecommerce behaviour data from multi-category store for October 2019. As with the issuing of every purchase in online retailler, the the data capture nine features of each customer. Detail about the characteristics are provided below. For instance, the Kaggle website has a sizable volume of this dataset covering several years. The Kaggle website has access to the dataset. We give a brief description of the dataset in terms of big data features, as well as some background information on the platform and the reasons behind the data's collection and public release.

About the company and dataset

Ecommernce market are growing at noticeable rates. The online market has to growned by 56% in 2020. eCommerce businesses can gather more detailed consumer information and give these clients a more specialized online experience by using behavioral data anal-

ysis (Sachdeva and Raheja, 2013). Customer behavior analysis, sometimes referred to as consumer behavior analysis, is a data-driven observation of online consumers and their interactions with your business.
These associations engage an understanding of consumer behaviours on their site and instrument compaigns in accordance with increase consumer memory and amount adaptation (Pavithra et al, 2016). An buying store is an connected to the internet store place you can buy merchandise and seruceces. Purchasing from an ecommerce or an connected to the internet store is much smooth than buying from a common store home. An ecommerce store is an connected to the internet gate place you can enjoy purchasing or sale merchandise resorces. Though ecommerce websites, you can place orders, form fees, path your transportation and likewise also state what different costomers should mention about their purchase in conditions of "review" (Edosio, 2014). However, e-commerce store function in a similar fashion and it is classified into 3 categories, which is receiving order, payment and shipping which are as follow

- Following the site checkout process, the patron produces the unavoidable fee and shipping news.
- The ecommerce site transmit the payment news through a fee seller that validates the fee and collects the resources.
- The seller then ships the order to the customer or send the digital product immediately via email.

E-commerce has grown into a meaningful component of corporate policy and a powerful motorist of financial growth in the still-progressing worldwide saving. The resumed expansion of eCommerce commit bring about raised contest, cost savings, and changes in sellers' valuing approaches, all of that would exercise downward pressure on swelling (Edosio, 2014). Many trades, from start-boosts to limited and medium-sized trades to important brands, can benefit from bearing their own connected to the internet store where they can advertise their own merchandise and duties. Consumers of all ages immediately want an smooth-to-use, affiliated occurrence that fits in with their often customs on account of intensely rude answer of new sciences in the sell manufacturing (Sachdeva and Raheja, 2013). This is an open source dataset from Ecommerce practice dossier from multicategory store accompanying act of the consumer. The dataset holds 5,535,755 eCommerce conduct that have happened circulated where each eCommerce action has 9 attributes. The dataset is 5,27 GB and this manage hard to process in a appliance accompanying depressed estimating capacity. An occurrence is represented by each row in the file. Every event has something to do with users and products. The data is available to download on the following website *https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store.*

Characteristics of the data

0.1 Volume

As already mentioned above, data of 2019 October contain a feature of each of thecustomer. All of this information is recorded is a CSV file amounting to the size equal

to 5,27 GB. The dataset recording 2019 November contains a feature of each of the-customers. Also this information is recorded in a CSV file which amounted to 8,38 GB.

0.2 Velocity

The data is collected from October 2019 to April 2020; nevertheless, there is no appropriate determinant that can estimate the velocity of the data. The dataset lacks a time stamp that would indicate how frequently it was generated or recorded.

0.3 Variety

Among the 9 columns of features for the 2019 October, ecommerce, 1 of them is float data types and 8 of them is objects (character strings). Similarly, with the 2019 November, there is 1 columns with float data types and 8 columns with character strings.

0.4 Veracity

Although it may not be easy to assess the degree of accuracy, since the dataset is intended to reveal how e-commerce in multi-category stores behaves, it is anticipated that the behavior of numerous vehicle stores would give customers more accurate information. Although the data is anonymized to conceal the buyer's true identity, the other qualities should still be as accurate as feasible.

Explanatory Data Analysis (EDA)

The 5,535,755 GB of structured data with 9 features that were collected between October 2019 and April 2020 are the data that were used for this assignment. Continual, category, and ordinal data types make up the features. Since it was not possible to drop some other variables, domain expertise was found to be helpful when cleaning and imputing data. Finding relevant information about users and events revealed by the Ecommerce behavior data for multicategory store platform was the main goal of the exploratory data analysis. As a result, the analysis will be helpful in illuminating both the distribution of events and the user behavior patterns. Due to the impossibility of using all 9 features to create predictive models, the analysis will enable us to better understand which factors are crucial for assessing how effective users are. The exploratory data analysis using big data processing framework to find the business insights from the behaviour data. The visualisation is then utilisation to explore the characteristics of the behaviours. Big data analytics can provide users with more insight into the market itself and the type of event who come to this platform. The raw data can be summarised in the form of chats and dashboards to provide users with more interactive visualisation tools to understand this market place.

submission/Assignment MIT805/correlelation.png

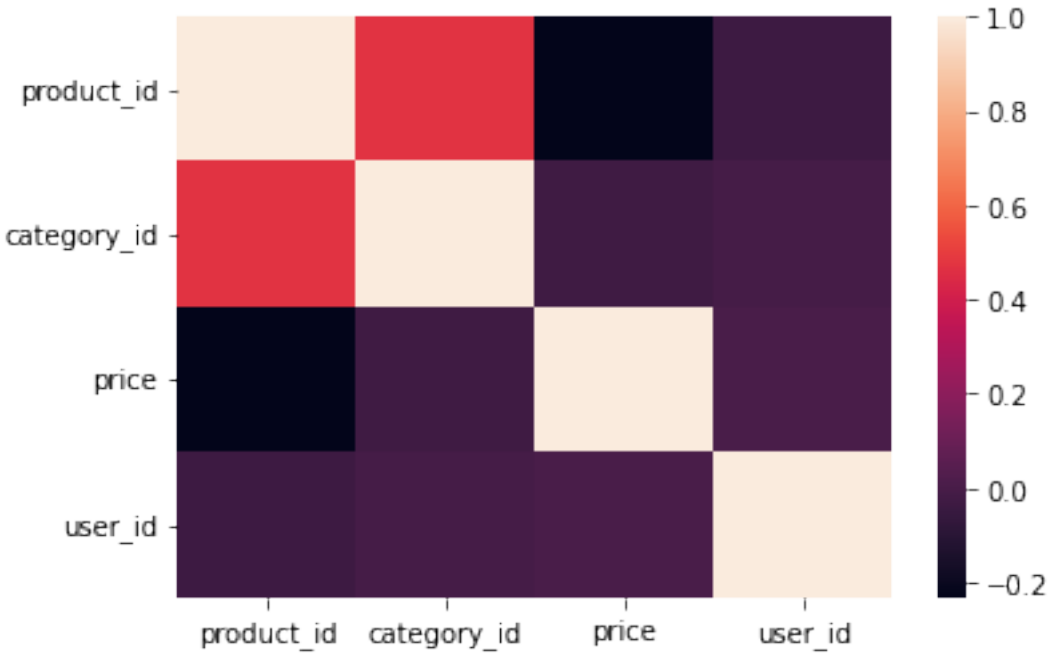

Figure 1.1: Correlation matrix

As a result, different feature variables must be investigated in order to identify those that are closely related to the target variable. We expected event type to be closely related to product id, category id, category code, user id, and price based on preliminary data analysis. Figure 1 depicts the correlation matrix of features ranging from positive 1 to negative 1, with positive 1 indicating perfectly positive linear correlation and negative 1 indicating perfectly negative correlation. Because each variable is correlated with itself, the correlation coefficient along the diagonal is equal to one in the figure above.

References

Edosio, U. Z, (2014), Big Data Analytics and its Application in E-Commerce, *Proceedings E-Commerce Technologies*, University of Bradford.

Pavithra, B., Niranjanmurthy, M., Kamal Shaker, J., Martien Sylvester Mani, F, (2016), The Study of Big Data Analytics in E-Commerce, *International Journal of Advanced Re-search in Computer and Communication Engineering*, 5(2), 126-131.

Bahga, A. and Madisetti, V. (2019), *Big Data Analytics: A Hands-On Approach*, Hand-son Book Series.

Prabhu, C. S. R., Chivukula, A. S., Mogadala, A., Ghosh, R. and Livingston, L. M. J. (2019), Big Data Analytics: Systems, Algorithms, Applications, Springer Singapore

Pte. Limited.

Snijders, C., Matzat, U. and Reips, U.-D. (2012), 'Big Data: Big gaps of Knowledge in the Field of Internet Science', *International Journal of Internet Science*, 7(1), 1–5.
Sachdeva, S. and Raheja, S., (2013), Analysis of e-commerce behavior in Multi-Category Store.

DECLARATION OF ORIGINALITY

UNIVERSITY OF PRETORIA

The University of Pretoria places great emphasis upon integrity and ethical conduct in the preparation of all written work submitted for academic evaluation.

While academic staff teach you about referencing techniques and how to avoid plagiarism, you too have a responsibility in this regard. If you are at any stage uncertain as to what is required, you should speak to your lecturer before any written work is submitted.

You are guilty of plagiarism if you copy something from another author's work (e.g. a book, an article or a website) without acknowledging the source and pass it off as your own. In effect you are stealing something that belongs to someone else. This is not only the case when you copy work word-for-word (verbatim), but also when you submit someone else's work in a slightly altered form (paraphrase) or use a line of argument without acknowledging it. You are not allowed to use work previously produced by another student. You are also not allowed to let anybody copy your work with the intention of passing if off as his/her work.

Students who commit plagiarism will not be given any credit for plagiarised work. The matter may also be referred to the Disciplinary Committee (Students) for a ruling. Plagiarism is regarded as a serious contravention of the University's rules and can lead to expulsion from the University.

The declaration which follows must accompany all written work submitted while you are a student of the University of Pretoria. No written work will be accepted unless the declaration has been completed and attached.

Full names of student: _____

Student number: _____

Declaration

1. I understand what plagiarism is and am aware of the University's policy in this regard.
2. I declare that this assignment report is my own original work. Where other people's work has been used (either from a printed source, Internet or any other source), this has been properly acknowledged and referenced in accordance with departmental requirements.
3. I have not used work previously produced by another student or any other person to hand in as my own.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

SIGNATURE: _____ DATE: _____