MIT805 Assignment 1

Collection & processing of a big dataset.

Contents

NO visualisation in the evaluation

7

Introduction

The report is to give details of the data to be explored as part of the MIT805 big data project. The dataset identified and chosen consists of events on an eCommerce website. The benefit expected from processing the data is to be able to predict the next action the customer will take as they browse on products with the intention to offer more value added services to the customer that is in line with their interest and also to gage how likely the customer is going to buy a product based on their views and purchases. Value added services to customers makes and organisation be more preferred by cutomers.

Dataset

The data was collected to study the eCommerce behaviour from multi category store by the OpenCDP project. It was downloaded from Kaggle[1]. It contains 285 million user events or user actions on an eCommerce website selling different categories of items. It was collected for a period of seven months (from October 2019 to April 2020) from the store website. Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users Michael Kechinov M[2].

The fields in the data set are event_time, event_type, product_id, category_id, category_code, brand, price, user_id, and user_session.

I was looking for transactions or events data with the intention to analyse people behaviours with regards to window shopping and making a decision to buy an item.

Characteristics of the data set

Volume

The volume of this data set is relatively big, making two hundred and eighty five millions of transactions stored in two comma separated files of five and three gigabytes. The file could not be opened with an ordinary excel or text application as they all have limitations, there might be need to further reduce this size to allow for processing on an ordinary machine with an Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz processor and sixteen gigabytes memory. One of the file has the following records.

```
In [13]: events.shape
Out[13]: (67501979, 1)
```

Variety

The data is structured as the collection of it is predefined with specific elements for which data needs to be collected. The variables in the datasets are broad with the fields mentioned above. There are millions of customers involved, such number gives a variety of variables as different people have different behaviours and tastes. In real life the classifications given above will be more making the events captured to have even more diversity.

Velocity

For the purpose of this project, static data will be used however this data in real life moves very fast noting from the data multiple events occurring in a second. Some days are busier than others and this trend will be explored as well.

Veracity

The qualities of this data can much be relied on as it was collected from real life events as and when they occurred. The events are captured by machines hence there is no room for capturing errors. This will give the results which are reliable, and much value can be expected from the predictions made from using the data.

Value

To derive value from the data, I will use technologies and analytics named by Ohlhorst F. J.(2012)[3] as to be used to define value of data sources which translate to actionable elements that move business forward. These technologies are Traditional business intelligence, Data mining, Statistical application, predictive analysis and data modelling. Proportions, movements, and information that leads to valuable actions will be illustrated from the data.

Some of the patterns I am expecting to uncover are as follows:

For customer service

1. Number of views a customer makes in a period of time.
2. Number of views a customer makes before buying an item.
3. The items that specific customers mostly view.

These frequencies will then be modelled into actions that will automatically trigger offers to the clients.

For internal Operations

4. The period with the most views in order to be able to allocate more resources to service customers.
5. We may need to know if events drop which will guide us to check the economic conditions or any other factors that may affect customers.
6. Check dormant customers and lock their accounts to save resources.

All the items mentioned above go a long way to assist business to make informed decisions and we can make this decision in a timely manner. The customer service is likely to improve as we can individually offer service that suits individuals. We are also able to optimise our processes and operations to the best of efficiency as we can make.

The other characteristics of the dataset

The data is valid to the simulated institution as it is real data collected real time from the source. It is volatile in that many of the customers transact daily, and the trends can change any day due to seasons or economic conditions. Due to its structure, it can be visualised easily.

Data Processing[4]

Feature Selection

The user identification was selected and the idea is to get to know what type of products the user views so that a model can be built to suggest new products that relates to the user activity. Also to check which days there are most views so that computing resources can be allocated to enhance user experience

Data Cleaning

There were no redundant samples. The key feature in the dataset is the user id and all the other fields will mostly be analysed centred on this key feature. Below is the view of the listed data.

```
events.head(5)
```

Out[1]:

	vent_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
0	2019-10-01 00:00:00 UTC	view	44600062	2103807459595387724	NaN	shiseido	35.79	541312140	72d76fde-8bb3-4e00-8c23-a032dfed738c
1	2019-10-01 00:00:00 UTC	view	3900821	2053013552326770905	appliances.environment.water_heater	aqua	33.20	554748717	9333dfbd-b87a-4708-9857-6336556b0fcc
2	2019-10-01 00:00:01 UTC	view	17200506	2053013559792632471	furniture.living_room.sofa	NaN	543.10	519107250	566511c2-e2e3-422b-b695-cf8e6e792ca8
3	2019-10-01 00:00:01 UTC	view	1307067	2053013558920217191	computers.notebook	lenovo	251.74	550050854	7c90fc70-0e80-4590-96f3-13c02c18c713
4	2019-10-01 00:00:04 UTC	view	1004237	2053013555631882655	electronics.smartphone	apple	1081.98	535871217	c6bd7419-2748-4c56-95b4-8cec9ff8b80d

There are missing categories, but these are not significant to the outcome of the project as the category codes are completed and the categories are just a description.

There are about one hundred thousand duplicate records on the file that was explored and these will be removed from the database.

```
In [8]: print(events.duplicated().value_counts())
```

```
False    67401460
True       100519
dtype: int64
```

For the purpose of this project the session field may not be necessary hence it will be removed.

Feature Selection

The user ID was selected, and the idea is to get to know what type of products the user views so that a model can be built to suggest new products that relates to the user activity. Below is a view of the user id count which will further be broken down to frequencies per period and also the items they view and buy.

```
In [3]: events['user_id'].value_counts(dropna=False)
```

```
Out[3]: 568778435    22929
        569335945    14810
        568818636     6171
        512475445     6111
        512365995     6042
                     ...
        573146032        1
        566812808        1
        573146061        1
        573145629        1
        579969851        1
        Name: user_id, Length: 3696117, dtype: int64
```

The most busy days with most views will be explored so that computing resources can be allocated to enhance user experience.

Data Transformation

As the transactions were downloaded into two comma separated files, they are going to be loaded into Hadoop distributed file system from which they will be processed. Some programs will be written in MapReduce and Java to visualise and extract insights from the data.

Conclusion

The dataset to be processed has been described and despite the limitations of the machine to be used for processing it does meet the criteria described to be a big data and it can be increased as and when needed. It has been established that the data is structured and it can be modelled. The next step is to process the data using the Hadoop framework.

References

1. Kaggle (2019) *eCommerce behaviour data from multi category store* [online] Available: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store
2. M. Kechinov(2019) Available: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store
3. F.J. Ohlhorst (2012. Nov 28) Big Data Analytics: *Turning Big Data into Big Money*, John Wiley & Sons
4. J. Brownlee, (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.