

Project #1 – Blibli Future Program

E-Commerce Dataset

Overview

You will find a zip file containing 7 csv files of e-commerce data which come from 7 different systems as well. Its features allow viewing an order from multiple dimensions: from order status, price and payment to customer location, product attributes and finally reviews submitted by customers.

Data Explanation

1. user_dataset.csv

- user_name : unique identifier of a user.
- customer_zip_code : zip code
- customer_city : city
- customer_state : state / province

2. order_dataset.csv

- order_id : unique identifier of the order.
- user_name : key to the user dataset.
- order_status : Order status (processing, delivered, shipped, etc)
- order_date : Purchase timestamp
- order_approved_date : Shows the payment approval timestamp
- pickup_date : Timestamp when it was handled to the logistic partner.
- delivered_date : actual order delivery date to the customer.
- estimated_time_delivery : estimated delivery date that was informed to customer at the purchase moment

3. order_item_dataset.csv

- order_id : unique identifier of the order
- order_item_id : unique identifier of the order item
- product_id : product unique identifier
- seller_id : seller unique identifier
- pickup_limit_date : Shows the seller limit date for handling the order over to the logistic partner.

- Price : item price
- shipping_cost : shipping cost (if an order has more than one item the freight value is splitted between items)

4. payment_dataset.csv

- order_id : unique identifier of the order
- payment_id : unique identifier of the payment
- payment_type : method of payment chosen by the customer
- payment_installments : number of installments chosen by the customer
- payment_value : total order amount

5. products_dataset.csv

- product_id : unique product identifier
- product_category : root category of product
- product_name_lenght : number of characters of the product name
- product_description_lenght : number of characters of the product description
- product_photos_qty : number of product photos
- product_weight_g : product weight measure in grams
- product_length_cm : product length measure in centimeters
- product_height_cm : product height measure in centimeters
- product_width_cm : product width measure in centimeters

6. seller_dataset.csv

- seller_id : seller unique identifier
- seller_zip_code : zip code
- seller_city : city
- seller_state : state / province

7. feedback_dataset.csv

- feedback_id : feedback form unique identifier
- order_id : unique order identifier
- feedback_score : rating
- feedback_form_sent_date : timestamp that survey was sent to the customer.
- feedback_answer_date : feedback answered timestamp

Task

You have to create an analysis and apply Machine Learning algorithm to the datasets.

1. Load data to Database server (PostgreSQL)
 - Design database, create the relations
 - Create Datawarehouse / datamart
2. Understand the Data
 - Query with SQL
 - Basic EDA (null, handle missing values, distribution, etc)
 - Ask 5 Business Questions (if possible, apply different methods/statistics)
 1. Question
 2. Why those questions?
 1. Background
 2. Expectation / Assumption / In General
 3. Answer + simple visualization + explanation
3. Machine Learning Opportunities (supervised and unsupervised)

Identify what problems that can be solved with those data using machine learning

- Problem statement
- Define objectives
- EDA related to target
- Design possible solution
- Provide justification using the most suitable performance evaluation criteria

Note:

- Please find 1 Supervised Learning and 1 Unsupervised Learning case
- Try at least 3 models for each case (start from the simplest one)
 - Show the model comparison
 - Give your thoughts of that comparison
- Push to git at least once every two weeks and provide any changes you have worked on, and what will you do / explore in the next two weeks.

Deliverables

1. Git – Notebook(s), README, requirements.txt
2. Blog post (e.g. <https://medium.com/analytics-vidhya/exploring-the-seattle-airbnb-dataset-3cb0df7699d3>)
3. Final Presentation