Low Level Design (LLD)

Market Basket Project on

E-Commerce

Mohammed Ashfaq Ali

Document Version Control

Description	Author
First Draft	Ashfaq Ali
Added Workflow chart	Ashfaq Ali
Added Exception Scenarios Overall, Constraints	Ashfaq Ali
Data Schema	Ashfaq Ali
Architecture	Ashfaq Ali

Contents

1. Problem Statement:

- The goal of this project is to give company an estimate of how much sales they have done, who is the top seller, which category product gets more revenue, which seller gets most negative feedbacks, average delivery time of each seller and RFM.

1.1 Overview:

Title: Market Basket Project on E-Commerce.

2. Domain Knowledge:

- The E-Commerce word means online shopping. So we must have basic knowledge on how the e-commerce platform works.
- Basically we have sellers and buyers in the platform where seller uses seller account and sells the products and buyers use buyer account and buys the products.
- In order to grow the e-commerce business the company should have a track of the feedbacks that buyer gives to the seller.
- The company should check who the top and regular customer is and who the top rated seller is.
- Regarding the online purchase, during the COVID, people became more aware about e-commerce sites and started purchasing online with safety measures.

2.1 Business Problem:

- E-Commerce sector faces lot of issues with fraud listing, misbehavior of seller, default item
- Delivery etc.
- Sometimes they will be payment issues, site issues and late delivery issues.
- An order might have multiple items.
- Each item might be fulfilled by a distinct seller.
- All text identifying stores and partners where replaced by the names of Game of Thrones great houses.
- Hence trying to resolve the above problems using Machine Learning Algorithms.
- Trying to build a user friendly ML model which can save time and effort for company to understand core problem and how to resolve it.

3. Data Requirements:

For the Problem statement data collected via Kaggle Platform (https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce)

- This is a Brazilian ecommerce public dataset of orders made at Olist Store.
- The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil.

- Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.
- We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.
- This is real commercial data, it has been anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

4. Expected Solution:

- Find customer generating most revenue
- Top customer.
- Top category products.
- Cities with highest revenue generation.
- Top rated sellers.
- Relationship between delivery time and review score.
- Seller's cities with highest and lowest delivery time.
- States with highest and lowest delivery time.
- Average delivery time varies across time.
- RFM (**Recency, Frequency, Monetary).**

5. ML formulation of the business problem:
First Cut Approach

- Importing the necessary libraries in jupyter notebook.
- Importing the datasets and merging them and renaming the columns.
- Checking for the null values and removing them.
- Performing the EDA and descriptive data analysis.
- Performing the RFM
- Then lastly performing the K-Means clustering model.

6. Business Constraints:

What is RFM analysis?

- RFM stands for recency, frequency, monetary value. In business analytics, we often use this concept to divide customers into different segments, like high-value customers, medium value customers or low-value customers, and similarly many others.
- Let's assume we are a company, our company name is geek, let's perform the RFM analysis on our customers

Recency: How recently has the customer made a transaction with us?
Frequency: How frequent is the customer in ordering/buying some product from us?
Monetary: How much does the customer spend on purchasing products from us?

7. Tools & Technology Requirements:

Tools & Technology:
Python | Data-Preprocessing | EDA | Feature Engineering | Machine Learning | Github.

IDE: Jupyter Notebook.

8. Conclusion:

Market Basket Project on E-Commerce is a Machine Learning Algorithms based model. For the Problem statement data collected via Kaggle Platform and built an end-to-end deployment ML model.

Why this Low-Level Design Document?

- The purpose of this document is to present a detailed description of the E-commerce business. It will explain the purpose and features of the system, the interfaces of the system.
- Basically we have sellers and buyers in the platform where seller uses seller account and sells the products and buyers use buyer account and buys the products.
- In order to grow the e-commerce business the company should have a track of the feedbacks that buyer gives to the seller.
- The company should check who the top and regular customer is and who the top rated seller is.
- Regarding the online purchase, during the COVID, people became more aware about e-commerce sites and started purchasing online with safety measures.

E-Commerce	Finalized	Source
Olist	yes	https://github.com/ashfaq1828/Brazilian-E-Commerce-Public-Dataset-by-Olist

Logging

We should be able to log every activity done by the user.

- The System identifies at what step logging required
- The System should be able to log each and every system flow.

- Developers can choose logging methods. You can choose database logging/ File logging as well.
- System should not be hung even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

Database

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

- The User chooses the disease.
- The User gives required information.
- The system stores each and every data given by the user or received on request to the database. Database you can choose your own choice whether MySQL.

Technology stack

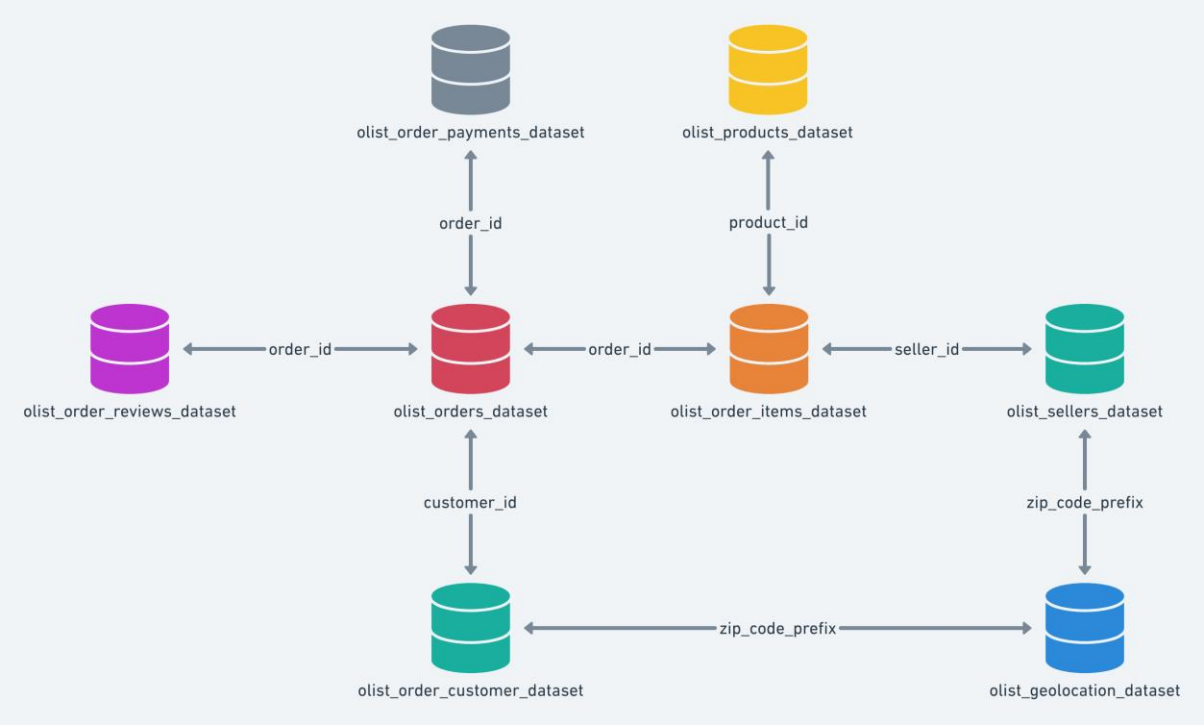
Front End	NA.
Backend	Python, ML.
Database	MySQL.
Deployment	NA.

Proposed Solution

Why making a baseline model important? Well, to compare the performance of our actual model, let say LSTM in this case, is very important to ascertain that we are in the right direction as if performance of LSTM is not better than the baseline model then there is no point of using LSTM.

1. Baseline Model: Logistic Regression, since this is a classification problem.
2. Actual model: LSTMs

Data Schema


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
olist_order_reviews_dataset
order_id
olist_orders_dataset
order_id
olist_order_items_dataset
seller_id
olist_sellers_dataset
customer_id
zip_code_prefix
olist_order_customer_dataset
zip_code_prefix
olist_geolocation_dataset

Architecture: