E-Commerce Recommendation Engine DMP

Ahmad Hawasheen
Abid Qasem
Marian Khoury
Yousef Abdalqader
Yousef Fawadleh

Introduction:

The purpose of this project is to build a recommendation engine to offer preferred products based on user history and other parameters, while designing personalized offers and campaigns. The recommendation engine's core entails collecting and analyzing historical and current data to predict and control the user journey.

1. Types of Data Produced

At the beginning, we will use Amazon Electronics dataset to build/learn our engine. The Electronics dataset contains about 21 million reviews for amazon electronic products. We will have two types of data (numerical and textual data):

- Reviews data: (review time, reviewer ID, review text, Rating of the products and the votes.)
- Products data: (Product ID, product description, product price, related products, brand, categories).

When we have our own engine, we plan to generate more data and features/attributes about the costumers. The data will be collected from the online websites when we install our service there. Raw data files will contain continuously reviews data, costumers' data, and products data. The generated data will be placed in comma-separated-values in plain ASCII format, which are readable over long time periods.

2. Data and Metadata Standards

Metadata will be comprised of two formats—contextual information about the data in a text-based document and CERIF - Common European Research Information Format standard metadata in an xml file. These two formats for metadata were chosen to provide a full explanation of the data (text format) and to ensure compatibility with international standards (xml format). The standard XML file will be more complete; the document file will be a human--readable summary of the XML file.

3. **Policies for Access and Sharing**

The raw data will be collected and maintained on a cloud server (S3 Service), and it will be accessed by the administrators and the data owners (retails) by creating accounts for these users with two factors of authentication for security purposes. Insights from the raw data will be shared with the data owners regularly.

4. **Policies for Reuse, Distribution**

Due to the normal of the data (costumers and sales data are sensitive for retails), we can't share or distribute the data with other parties. After long period, the data may be un sensitive and can be shared for research and educational purposes, if the data owners approve to share it. Access to these old data will be provided using web-based applications.

5. **Plans for Archiving and Preservation**

The generated data will be maintained in data repository (data warehouse) for archiving, future analyzing, and reporting. For security issues, we will develop comprehensive access rules to allow only authorized users with a legitimate business need to access, modify, or transmit data.

The date of the archived data files will be included and will be part of the data file name. also, metadata will be included in the archived files which is necessary for quality data analysis and reporting.