ADM Assignment – Part II

Design:

Discuss who is the TPC-DS dashboard targeted towards and the use cases you will accomplish with it.

We have implemented the following use cases in the sales perspective in the TPC-DS dashboard:

- Average list price to identify the sales price of every product item on average to analyze the sales of expensive products and lesser expensive products and compare how the price of items is impacting the sales.
- Average sales quantity to identify how many items are being sold on an average in a sale and to identify the popular items amongst the customer clusters.
- Average sales per brand to analyze the sales of every brand and identify the top performing brands. This information can be used to identify the most popular products in these brand and create more promotions for these brands to target more customers and increase sales.
- Average coupon amount to calculate the coupons with more discount values on the list prices of items. These coupons can be used to propagate among the respective customer clusters to attract more sales.

All of the above use cases can be used to identify the most popular products, top performing product brands and categories, and the popular promotions used amongst respective customer clusters. All of this data can be used to generate insights about advertisements promoting the top products and promotional offers being provided, to direct the top products, promotions, and advertisements as recommendations to the particular customer clusters, and to make analytical decisions regarding the pricing of products.

Describe the design on how you would onboard the dataset.

We have included 4 datasets as part of the new dataset design:

- Customers – to store the demographic, identity, and descriptive data of the customer base of the company.
- Sales – to store the sales information of all orders of products under different categories and brands on all the sales channels of the company.
- Promotions – to store the information of the all the promotional offers provided on the products under different brands and categories of the company
- Products – to store all the descriptive information of all the products along with their type, category, description, brand, and cost.

All of these datasets can be used to design a new marketing dashboard for the company to visualize several marketing programmatic services of the company such as top customers in every region, products with highest sales under the top brands for specific customer clusters, overall sales per month and year, top promotions with discount amounts per item, most preferred sales channels etc.

Describe what tools will be used for data cleaning.

We have used Python for data cleaning using in-built functions of Python libraries including NumPy, Pandas, Seaborn, and Matplotlib. Jupyter Notebook environment is used in this case. Some of the data cleaning approaches that we have implemented are:

- Identifying missing values in the data and replacing it with appropriate values
- Identifying missing values in the data and dropping the observations with missing values
- Dropping the features not required
- Removing special characters, extra whitespaces etc. from the data in specific features
- Performing data type conversions of features
- Transforming data to consistent formats (ex. Address, date etc.)
- Removing duplicate observations with repetitive data
- Correcting incorrect data
- Skipping observations not required
- Renaming feature names

Prototype your application:

Choose a marketing dataset from Kaggle.com

We have chosen the following datasets from Kaggle.com to build our new dataset design:

- Addresses - Link
- Estee Lauder Products - Link
- Customers - Link

Show how will you upload it to Snowflake

We have completed the data cleaning process using Python on Jupyter environment and exported the Pandas dataframe to csv containing the cleaned updated datasets. Then, we have loaded the exported csv files to Snowflake using the Load Data Wizard with the following steps:

1. Create a new database and create new tables corresponding to the datasets.
2. Select a table by clicking on the table row and click on Load Data button.
3. The Load Data Wizard opens.
4. Select the warehouse on which there are USAGE privileges.
5. Click on Load files from your computer, click Select Files button, and select the csv file on the local system.
6. Select the existing input file format. If not, click on + button and create a new input with the fields matching the file format and select it.
7. Select the load options specifying what should be done in case of errors in data files. Click on Load button.
8. Click on OK button and the data wizard closes. The data loading process is completed displaying the number of rows and size in each table in the database.

What algorithms and frameworks have facilitated the development of the site?

Algorithms can be created in a variety of ways to optimize product personalisation for consumers using big data. By analyzing customer physical characteristics, beauty brands can determine the current state of a person’s skin or hair. Consumer data can be utilized to create high-quality, custom-tailored skincare, cosmetics, perfume, and haircare formulae for each customer. Chatbots employ data to give customised, relevant information along the sales funnel, and matching engine algorithms can recommend or categorize the suitability of pre-existing products.

To make these processes work, Estee Lauder require consumer input. Smart serums use data after each usage to learn more about skin issues and change the formula to keep it optimized. Estee Lauder employs Curology that uses machine learning and artificial intelligence to cure acne. Another example is the HelloAva chatbot, which uses algorithms to help customers develop a personalized skincare routine and purchase goods.

3D printing, a more contemporary technology, can also leverage customer data to alter the way beauty items are purchased and consumed. With the potential to scan and print face masks at home, print customised cosmetics at home, and even print foundation makeup right onto the consumer's face, 3D printing's capabilities, when combined with Estee Lauder products and services, could transform how customers think about purchasing patterns.

1. Response Modeling Framework

This algorithm can be used for identifying the most promising individuals for treatment in order to increase the marketing campaign's total worth. Finding the strategy that optimizes the value function is the most basic marketing optimization problem. In the instance of campaign response, we model the campaign's overall worth in terms of response probability and projected net value from a client. The set of customers who will receive the promotion, or the campaign's audience, will be the subject of our optimization.

$$U_{opt} = \underset{U \subseteq P}{\operatorname{argmax}} \quad G(U)$$

Where P is the entire population of consumers
U is the subset of consumers reached in the scope of the campaign
G(U) is the expected profit of the campaign defined as:

$$G(U) = \sum_{u \in U} \Pr(R \mid u, T) \cdot (G(u \mid R) - C) + (1 - \Pr(R \mid u, T)) \cdot (-C)$$

$Pr(R|u, T)$ - probability of a response to the treatment (promotion) T from customer u
$G(u \mid R)$ - is the response net value for customer u
C is a cost of the promotion resource

The first term corresponds to the expected gain from a responding consumer, and the second term corresponds to the expected loss of sending a promotion to which there is no response. The objective is to maximize the expected profit by finding the subset of customers that are likely to respond in the most profitable way.

2. Recency–frequency– monetary (RFM) analysis.

Recency - The number of time units that have passed since the customer last purchased.

This metric can be measured in time units (for example, months) or linked to a score. Customers can, for example, be sorted by the most recent purchase date, with the most recent 20% receiving a score of 5, the next 20% receiving a score of 4, and so on, until the last 20% receives a score of 1.

Frequency - The average number of purchases per time unit.

This metric can also be measured directly in units or scores.

Monetary - The total dollar amount spent per time unit.

This metric is typically measured by using intervals or scores.

3. Lifetime Value Modeling

Customer Lifetime Value is commonly abbreviated as LTV, CLV, or CLTV. The purpose of LTV modeling is to predict how much money a brand will make from a particular consumer over the course of their relationship.

Estimate the lifetime value of customer to sum the average expected profits for some time interval in the future:

$$LTV(u) = \sum_{t=1}^{T} (R - C) = T(R - C)$$

Here, time t is measured in some units, typically months.
R and C are the average expected revenues and costs, respectively, per customer per time unit.

T is the expected lifetime or projection horizon.

The average predicted revenues and costs are usually calculated using past data like transaction histories and campaign budgets. This estimate isn't individualized (it's based on the average of all consumers), thus it's simple. Revenues and costs for different consumer segments might vary dramatically. So, it is very common to estimate R and C for each segment separately and then calculate segment-specific LTVs if the customer segment (persona) is known.

$$\mathrm{LTV}(u) = \sum_{t=1}^{T} \frac{(R - C)r^{t-1}}{(1 + d)^{t-1}}$$

The basic LTV formula does not account for a number of significant effects. For starters, it does not take client retention into account. Although the time horizon T can be adjusted according to the average customer lifetime, it may be more convenient to add the customer retention rate r as a parameter in the formula.
For example, an annual retention rate of 0.8 means that 20% of the present customers will terminate the relationship within a year.
Second, because the LTV is commonly calculated across 2–3 year time intervals, we may need to account for the fact that money in the present is worth more than money in the future, or for a discount rate d. The expense of tying up capital for a period is reflected in the discount rate.
An annual discount rate of 0.15, for example, suggests that the present value of $1 should be regarded as the equivalent of $1.15 received in one year. The net present value of the LTV is calculated after the discount rate is taken into account. By considering these two parameters, we can calculate the customer net profit as (R-C) for the first period, (R-C) *r/(1+d) for the second period, and so on, resulting in the LTV definition.

Year	Net profit	Retention rate	Expected net profit	Discount multiplier	Discounted net profit
1	$100.00	1.00	$100.00	1.00	$100.00
2	$110.31	0.90	$99.28	0.91	$90.26
3	$118.50	0.81	$95.99	0.83	$79.33
4	$125.00	0.73	$91.13	0.75	$68.46
5	$130.16	0.66	$85.40	0.68	$58.33
LTV					**$396.38**