Assignment 1 – Uncovering Marketing insights

Digital Marketing Analytics

Summer 2020

Prof.Sri Krishnamurthy

Northeastern University & QuantUniversity

Analyzing Digital Marketing datasets

Goals

- To work with datasets using xsv(xcsv) & Trifacta(https://community.trifacta.com/s/lesson-1-introduction-to-trifacta-wrangler),
- Stage datasets in Snowflake
- To be able to analyze marketing data using Salesforce Einstein analytics studio
- Derive insights from the datasets
- Crisply communicate and document your findings

Case

Marketa analytics has hired you to as an Algorithmic marketing analyst. Marketa is a consulting organization specializing in Marketing analytical solutions. Your client (see allocations by team number below) has provided you a sample dataset and asked you to analyze and build an analytical dashboard as a **Proof-of-concept** to illustrate the value of data driven analytics. The themes to be considered could include:

- Pricing
- Promotion
- Search

- Recommendations etc.

Marketa wants you to analyze the data using tools (xsv, Trifacta, Snowflake) and build a dashboard using Einstein analytics. They also want you to build a codelabs document to crisply illustrate the value analytical solutions would bring to the company. You are also asked to discuss what additional datasets and methodologies could be used. The company has a challenge using large scale datasets and are considering using Trifacta and xsv as data tools to work with. You are expected to illustrate how you would:

- Use the tools for joining datasets
- Filtering
- Aggregating
- Missing value handling
- Deriving additional columns from existing datasets
- Cleaning (for example removing blank spaces, formatting dates, Capitalizing etc.)

In order to do that you are asked to illustrate the strengths and weakness of each tool/package

Dashboards:

Once you clean the data, import the data into Snowflake and illustrate how to use the Einstein analytical dashboard to illustrate various aspects of analysis. (https://salesforce-trailblazer.com/snowflake-einstein-analytics/).

Questions to consider:

- Which columns are dimensions, which columns are measures?
- How would you generate new dimensions? What will you do to summarize measures?
- Who would use this dashboard?
- What value would generated using this dashboard ?

Deliverables:

- How to work with the large datasets using xsv and Trifacta
- Schemas for working with Snowflake for your chosen dataset
- Analytics Dashboard using Salesforce Einstein Analytics
- A Google Codelabs document summarizing the insights

Team allocations: (See the google sheet)

Instacart - Market Basket Analysis

The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users.

- Data: https://www.instacart.com/datasets/grocery-shopping-2017 or https://www.kaggle.com/c/instacart-market-basket-analysis/data
- Description: https://gist.github.com/jeremystan/c3b39d947d9b88b3ccff3147dbcf6c6b
- Backup copy (data and description): https://drive.google.com/drive/folders/1JC-D3vtYI6iOSGaZ9DoSXDQ4GrvMzqLL

Criteo - Attribution Modeling for Bidding

This dataset represents a sample of 30 days of Criteo live traffic data. Each line corresponds to one impression (a banner) that was displayed to a user. For each banner we have detailed information about the context, if it was clicked, if it led to a conversion and if it led to a conversion that was attributed to Criteo or not. Data has been sub-sampled and anonymized so as not to disclose proprietary elements.

- Data: https://s3-eu-west-1.amazonaws.com/attribution-dataset/criteo_attribution_dataset.zip
- Description: http://ailab.criteo.com/criteo-attribution-modeling-bidding-dataset/

- Backup copy (data and description): https://drive.google.com/open?id=1WY6DdbbL6nzcxLA3z3vWYAbqNXeCg9Qu

Dunnhumby - The Complete Journey

Household level transactions over two years from a group of 2,500 households who are frequent shoppers at a retailer All of a household's purchases within the store, not just those from a limited number of categories Demographics and direct marketing contact history for select households

- Data: https://www.dunnhumby.com/careers/engineering/sourcefiles
- Description: https://www.dunnhumby.com/careers/engineering/sourcefiles
- Backup copy (data and description): https://drive.google.com/drive/folders/1PAe62y3fgxPSgzvkMph3295Ah9WCMrhR

Yoochoose - RecSys Challenge 2015

The data represents six months of activities of a big e-commerce businesses in Europe selling all kinds of stuff such as garden tools, toys, clothes, electronics and much more.

- Data: https://recsys.yoochoose.net/challenge.html
- Description: https://recsys.yoochoose.net/challenge.html
- Backup copy (data and description): https://drive.google.com/drive/folders/1pQXY_Pl6UaLYcfvN92pqbDyk2auvyibA

Kaggle - Give Me Some Credit

Historical data are provided on 250,000 borrowers and the prize pool is $5,000 ($3,000 for first, $1,500 for second and $500 for third).

- Data: https://www.kaggle.com/c/GiveMeSomeCredit/data

- Description: https://www.kaggle.com/c/GiveMeSomeCredit/data
- Backup copy (data and description): https://drive.google.com/drive/folders/14Ss_wSOHP8L7KmHxZelttTR6OadA2ELU

MovieLens - 25M Dataset

MovieLens 25M movie ratings. Stable benchmark dataset. 25 million ratings and one million tag applications applied to 62,000 movies by 162,000 users. Includes tag genome data with 15 million relevance scores across 1,129 tags. Released 12/2019

- Data: https://grouplens.org/datasets/movielens/25m/
- Description: https://grouplens.org/datasets/movielens/25m/
- Backup copy (data and description): https://drive.google.com/drive/folders/1GhJGkFAwNb95Jnah6OEKJH2oDg0ls25g

Elo - Merchant Category Recommendation

This dataset is created by Elo, one of the largest payment brands in Brazil. The datset contain contains up to 3 months' worth of transactions for every card.

- Data: https://www.kaggle.com/c/elo-merchant-category-recommendation/data
- Description: https://www.kaggle.com/c/elo-merchant-category-recommendation/overview
- Backup copy (data and description): https://drive.google.com/drive/folders/1HmrVX4nAT3AVD9jHIe_z-pTKn7Jh-J-h?usp=sharing

Reference: https://github.com/ikatsov/tensor-house/blob/master/resources/datasets.md