**Starbucks Capstone Project Report**

Overview

This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks.

Not all users receive the same offer, and that is the challenge to solve with this data set.

The task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. You'll see in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.

Dataset

The data is contained in three files:

* portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
* profile.json - demographic data for each customer
* transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

**portfolio.json**

* id (string) - offer id
* offer\_type (string) - type of offer ie BOGO, discount, informational
* difficulty (int) - minimum required spend to complete an offer
* reward (int) - reward given for completing an offer
* duration (int) - time for offer to be open, in days
* channels (list of strings)

**profile.json**

* age (int) - age of the customer
* became\_member\_on (int) - date when customer created an app account
* gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
* id (str) - customer id
* income (float) - customer's income

**transcript.json**

* event (str) - record description (ie transaction, offer received, offer viewed, etc.)
* person (str) - customer id
* time (int) - time in hours since start of test. The data begins at time t=0
* value - (dict of strings) - either an offer id or transaction amount depending on the record

Business Problem

Starbucks want to personalise their marketing more by understanding which types of customers respond best to which types of offers. So what we need to understand is, which off could be sent to a particular person which is likely to increase engagement and activity.

The key challenge is that not all users react in the same way. For example, so customers may ignore offers, while others will be influenced by them. Some users may complete an offer but might have never viewed it – so just completed it because they wanted to make an order and didn’t notice the offer. For this reason and to others, the data requires cleaning. My aim is to understand how each demographic group across, gender, age group and income bracket react towards different offer type. I want to be able to analyse to understand each groups, build a model and share suggestions on what Starbucks could do moving forward.

Data set Analysis

Portfolio

As mentioned we have x columns

A screenshot of a black screen

AI-generated content may be incorrect.

The data set consists of no null values which will help massively with the cleaning and merging process

A screen shot of a computer

AI-generated content may be incorrect.

Offers distributed are higher for BOGO and Discount

A graph of different types of data

AI-generated content may be incorrect.

The duration for Discount is the highert on average

A graph of blue bars

AI-generated content may be incorrect.

Channel type is most likely to be email or web

A graph of blue bars

AI-generated content may be incorrect.

Profile

Upon analysis there appears to be an anomly with customers listed as age 118. This value occurs with unusually high frequency then expected, particulary when compared to 101. Aditionally, the count for age 118 matches with the count of missing values for gender and income. As such, these records are likely invalid and may be best to remove from the dataset.

A graph of age distribution

AI-generated content may be incorrect.

When looking at the missing values, all the missing values come from age 118 for gender and income. This futher highlights that it is best to remove this age from the dataset to continue with analysis and recommendations.

A black screen with colorful text

AI-generated content may be incorrect.

We have more males then female es

A pie chart with numbers and text

AI-generated content may be incorrect.