Starbucks Capstone

# Introduction

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.

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.

You'll be given transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

# Problem

## Domain Background

General article of advert targeting using ML algorithm, that describes how important personalized advertising is.

<https://www.sciencedirect.com/science/article/pii/S2405959520301090>

Specific article for the same data set where an unsupervised learning with clustering is used.

<https://medium.com/@jeffrisandy/investigating-starbucks-customers-segmentation-using-unsupervised-machine-learning-10b2ac0cfd3b>

The targeted control of offers to customers is a prime example of methods of machine learning and data science.

Which offer can I make to which customers. Which one is the customer most likely to look at and accept?

Is it possible from successfully made offers to develop other offers that are promising? These are questions of e-commerce that can be applied to many business areas.

## Problem Statement

There are many questions to answer in this capstone project

* Is it possible with a combination of demographic and customers features to predict the offer completed value? This is part of a supervised machine learning algorithm.
* Same question is for viewed value.
* There are customers who make up 80% of sales. Let’s call them power customer. If we label the customers as power or no power customer, is it possible to predict this label, based only on the customer demoraphic features?

# Datasets

## 2 – 3 Paragraphs

The Starbucks dataset contains three different files.

* Profile:  
  This file contains 17000 customers with income, gender, age and the date of creating an account.
* Portfolio:  
  This file contains the different offers in total 10 different offers. The columns are the type of an offer, the duration, reward the money to invest (called difficulty) and very important for the viewed classification the channels which are used to provide the offer.
* Transcripts:  
  This file is a large dataset where customers and offers comes together in events. The events can be offer received, offer viewed, offer completed and transactions. From this transcript events it is possible to extract also offers which are unviewed and uncompleted.

# Solution Statement

## 1 Paragraph

To create a machine learning model, it is necessary to get the labels for each received offer. In the transcript file are all received offers, all viewed and all completed offers. But there is no link between the received, viewed and completed person offer combinations. Therefor it is necessary to create a dataset where for each received offer the information will be added about viewed and completed. The columns for viewed and completed are the label columns for the machine learning models.

After labeling the data an explorative data analysis is performed to decide which customer offer combinations will be part of the machine learning models.

# Benchmark Model

The feature input vector for this dataset is with around 25 input features in csv file format relative less complex. I decide to use aws xgboost as the benchmark model the aws linear learner for comparison.

# Evaluation Metrics

As evaluation metrics first the accuracy score will be defined, additional the confusion matrix will be plotted, and the precision and recall will be calculated. With the linear learner it is easy possible to optimize these values.

# Project Design

## Data Cleaning

First there is a lot of data cleaning to do. Are there duplicate rows? Change the format for date time columns, normalize json columns in transcript file. Then a bunch of customers have no information provided about there income, age, gender. I decide to keep these customers because it is an own group. After cleaning the datasets, the different datasets can be merged into one full dataset.

## Labeling

As described in the dataset section, there is no clear information which offer is unviewed or uncompleted. Therefore I decided to search for each received offer the corresponding viewed and the correspond completed event, which must be in the duration time frame. If there is no corresponding event, the received offer is labeled with minus 1.

match = sub\_match.query('offer\_id == @offer\_id and @time <= time <= @validity')

## Exploratory Data Analysis

After data cleaning it is time for an explorative data analysis. Starting with univariate analysis regarding features from profile and offer. Then do some bivariate analysis regarding features in profile dataset. Finally do some analysis regarding the customers and there behave on offers. Are there customers which view or complete all offers, no offers, some offers?

How is the offer viewed rate? Are there popular offers depending on the received, viewed and completed counts?

One more interesting question is about power customers? In addition to the offer data, there are also normal purchases with amount values in the data. When I group this data by customer and sum up the amount values, I can sort it by amount values and find all customers which makes i.e. 80% of total amount or 20% of customers with highest amount summation.

Depending on the results of the data analysis it is possible that more different machine learning models could be developed.

## Feature Engineering

To create data for machine learning models some additional feature engineering must be done. There are some categorical variables, which will get one hot encoding. Continuous data will be normalized to a range from zero to one.

## Machine Learning

### Which customer offer combinations will be completed?

For the machine learning model features from customers and offers which were received and there corresponding “offer competed” label will be used to train a model and to predict if a customer offer combination will complete an offer.

As algorithms for the machine learning models I decide to use “xgboost” and the “Linear Learner” both as amazon web services implemented algorithms.

### Who is a power customer?

From the regular purchases I have created labels for power customers yes or no. Is it possible to predict a power customer only on demographic features? For this question also a machine learning model will be created.