# Capstone Project – Machine Learning Engineer

An analysis on Starbucks simulated customer data

Christoph Deserno

October 27th 2020

## The project's domain background

I have chosen the Udacity Starbucks data project which holds simulated customer data for transaction and advertisement offers regarding the usage of a Starbucks mobile application. The data from this app is divided into three tables explained further below.

## A problem statement

There are different type of advertisements – in this context they will be called offers – which affect different customers in different ways. Some may work well on one costumer while they work poor on another customer. Since the offers come with a cost for the company – meaning discounts they are offering in the advertisement – it would be best if the offers fit well to the customer to have a high conversion rate – meaning the offers are accepted and turned into company income. Here I will look at how to determine which offer type might fit best to which demographic group of the customers.

## The datasets and inputs

The given dataset consists of three tables

- Portfolio
  The portfolio holds information about what kind of offers are used for advertising. Three different offer types are present. A “buy one get one” (BOGO), a discount offer and just pure informational offers. The first two offers come with a reward measured in dollars while the informational offer is just not rewarded and pure advertisement. Each offer is described by a unique id, a duration for how long the offer is valid, a difficulty meaning the value a customer must spend to be able to receive the award of the offer, the offer type (BOGO, discount, informational), the channels how the offers are broadcasted (email, mobile, social, web) and a reward.

| | reward | channels | difficulty | duration | offer_type | id |
|---|---|---|---|---|---|---|
| **0** | 10 | [email, mobile, social] | 10 | 168 | bogo | ae264e3637204a6fb9bb56bc8210ddfd |
| **1** | 10 | [web, email, mobile, social] | 10 | 120 | bogo | 4d5c57ea9a6940dd891ad53e9dbe8da0 |
| **2** | 0 | [web, email, mobile] | 0 | 96 | informational | 3f207df678b143eea3cee63160fa8bed |
| **3** | 5 | [web, email, mobile] | 5 | 168 | bogo | 9b98b8c7a33c4b65b9aebfe6a799e6d9 |
| **4** | 5 | [web, email] | 20 | 240 | discount | 0b1e1539f2cc45b7b9fa7c272da2e1d7 |
| **5** | 3 | [web, email, mobile, social] | 7 | 168 | discount | 2298d6c36e964ae4a3e7e9706d1fb8c2 |
| **6** | 2 | [web, email, mobile, social] | 10 | 240 | discount | fafdcd668e3743c1bb461111dcafc2a4 |
| **7** | 0 | [email, mobile, social] | 0 | 72 | informational | 5a8bc65990b245e5a138643cd4eb9837 |
| **8** | 5 | [web, email, mobile, social] | 5 | 120 | bogo | f19421c1d4aa40978ebb69ca19b0e20d |
| **9** | 2 | [web, email, mobile] | 10 | 168 | discount | 2906b810c7d4411798c6938adc9daaa5 |

- Profile
  The profile table holds information about the customers. Mainly demographic data can be found as gender, age and income. Also an information about when a person became

customer of the app.

| | gender | age | id | became_member_on | income |
|---|---|---|---|---|---|
| **1** | F | 55 | 0610b486422d4921ae7d2bf64640c50b | 20170715 | 112000.0 |
| **3** | F | 75 | 78afa995795e4d85b5d9ceeca43f5fef | 20170509 | 100000.0 |
| **5** | M | 68 | e2127556f4f64592b11af22de27a7932 | 20180426 | 70000.0 |
| **8** | M | 65 | 389bc3fa690240e798340f5a15918d5c | 20180209 | 53000.0 |
| **12** | M | 58 | 2eeac8d8feae4a8cad5a6af0499a211d | 20171111 | 51000.0 |
| **...** | ... | ... | ... | ... | ... |
| **16995** | F | 45 | 6d5f3a774f3d4714ab0c092238f3a1d7 | 20180604 | 54000.0 |
| **16996** | M | 61 | 2cb4f97358b841b9a9773a7aa05a9d77 | 20180713 | 72000.0 |
| **16997** | M | 49 | 01d26f638c274aa0b965d24cefe3183f | 20170126 | 73000.0 |
| **16998** | F | 83 | 9dc1421481194dcd9400aec7c9ae6366 | 20160307 | 50000.0 |
| **16999** | F | 62 | e4052622e5ba45a8b96b59aba68cf068 | 20170722 | 82000.0 |

- Transcript

  The transcript table contains data about different types of events. These events can be transactions where it is listed, how much a customer spent at what time or they can hold information on offer status information. This information can either be whether an offer is received, whether it is viewed or completed.

| | person | event | value | time |
|---|---|---|---|---|
| **0** | 78afa995795e4d85b5d9ceeca43f5fef | offer received | {'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'} | 0 |
| **1** | a03223e636434f42ac4c3df47e8bac43 | offer received | {'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'} | 0 |
| **2** | e2127556f4f64592b11af22de27a7932 | offer received | {'offer id': '2906b810c7d4411798c6938adc9daaa5'} | 0 |
| **3** | 8ec6ce2a7e7949b1bf142def7d0e0586 | offer received | {'offer id': 'fafdcd668e3743c1bb461111dcafc2a4'} | 0 |
| **4** | 68617ca6246f4fbc85e91a2a49552598 | offer received | {'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'} | 0 |
| **...** | ... | ... | ... | ... |
| **306529** | b3a1272bc9904337b331bf348c3e8c17 | transaction | {'amount': 1.5899999999999999} | 714 |
| **306530** | 68213b08d99a4ae1b0dcb72aebd9aa35 | transaction | {'amount': 9.53} | 714 |
| **306531** | a00058cf10334a308c68e7631c529907 | transaction | {'amount': 3.61} | 714 |
| **306532** | 76ddbd6576844afe811f1a3c0fbb5bec | transaction | {'amount': 3.5300000000000002} | 714 |
| **306533** | c02b10e8752c4d8e9b73f918558531f7 | transaction | {'amount': 4.05} | 714 |

## A solution statement

The described task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type.

For this task I plan to bring the tables into a structure that makes them easier to process. Given nested information as in the value column of the transcript data is difficult to take as is and must be pre processed. After cleansing the data must be combined and special situations must be taken care of for example that completion information on offers is available for offer types BOGO and discount but not for informational offers.

My approach for this will be a linear model that finds correlation between the different features of a person and the amount spent by a person. Depending on the influence of a demographic criteria I will try to find recommendation to play certain offers to certain demographic groups more or less often.

## A benchmark model

Since the development of the model will help to explain the dependencies a benchmark does not exist so far. Having the wish to explain dependencies the model should be of a simple type. Increasing model complexity will usually lead to less ability to interpret what the model does.

## A set of evaluation metrics

The linear model will result an r squared score for how the data correlates and determine coefficients of correlation. These correlations coefficients should be plausible

## An outline of the project design

First I will take a look at the raw data, how should it be preprocessed. Matching informational offers to a completion date will be tricky. I will focus on modelling with customer data. I assume that features of customers will explain the variance of which offer fits best to which customer. Simple models can be computed on local machines but I will deploy the model to Amazon Sagemaker for later use.