Starbucks Capstone Challenge

Luis Soares Evangelista Neto
August 10, 2021

1 - Capstone Proposal

Luis Soares Evangelista Neto
August 10, 2021

2 - Definition

2.1 - Project Overview

This document has the purpose of offering a solution for the Starbucks Capstone Challenge. Starbucks is an American multinational company, with the largest chain of coffee shops in the world. For this project, was given a set of customer information that simulates the customer behavior on the Starbucks rewards mobile app. The mobile app has a feature that sends promotional offers or just advertisements to the clients, but they are categorized in 3 topics:

- Informational offer;
- Discount offer;
- Buy one get one free (BOGO) offer.

2.2 - Problem Statement/Metrics

Sending several offers to clients can be expensive when we don't know how our public behave. Basically we need to improve the way we communicate with our clients, by doing that, we can improve ROI (Return of Investment). Using information about demographics and transactions we can explore how the clients that respond to our campaigns usually behave and we can formulate better offers for the clients. Using a clustering model, we can identify patterns in the clients that convert to an offer, patterns that we can't identify just by observing in exploratory analysis. In this project, we are going to use the K-Prototypes model and use the elbow plot method to identify the best number of clusters (in our case, personas).

3 - Analysis

In this section, we are going to explore the data given to us, to get a better understanding. For this project, was given 3 datasets containing information about offers, clients and transactional data from the Starbucks mobile app.

- **portfolio.json** — Contains information about Starbucks campaigns (offers);
- **profile.json** — Personal information about the clients;
- **transcript.json** — Transactional information about the behavior of the clients.

We are going to understand more about the datasets in the data analysis section.

3.1 - Data Analysis (Part 1)

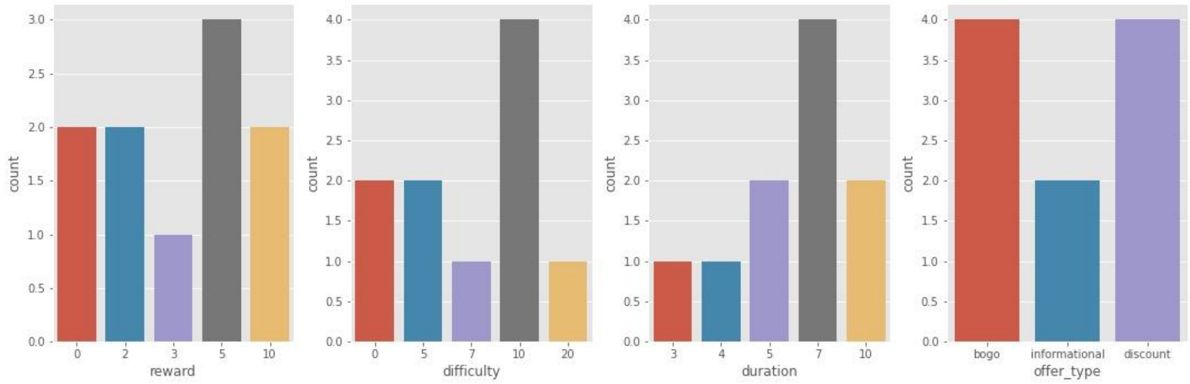
We begin our analysis exploring individually all the datasets to get a better understanding of how they are structured.

3.1.1 - Portfolio dataset

As said before, this dataset contains information about the campaigns that Starbucks offers to the clients. Below we have all the columns in the portfolio.json:

reward	*integer*
channels	*string*
difficulty	*integer*
duration	*integer*
offer_type	*string*
id	*string*

Now, let's observe how the data is distributed in some of the variables:



From the figure above, we can observe that we have a total of 10 campaigns, from that we observe:

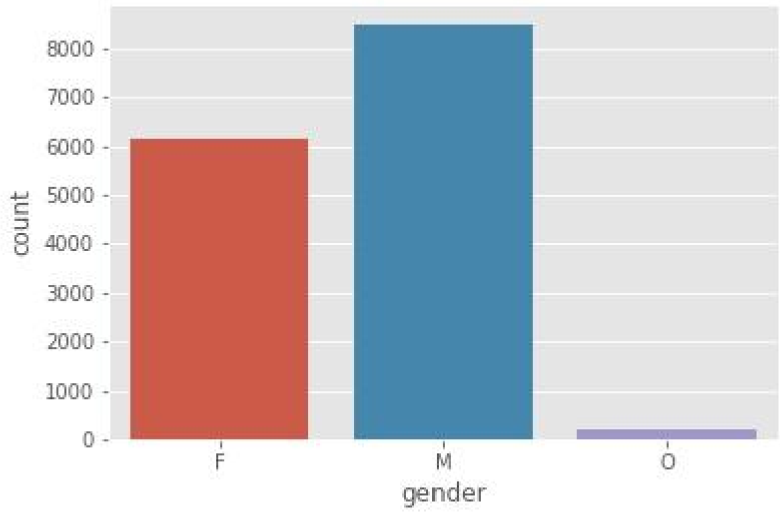
- The 3 offer types mentioned before (Bogo, Informational and Discount).
- 4 offers with high difficulty, that means the minimum required to spend to complete an offer is high.
- There are 4 channels: email, mobile, social and web, and some of the offers are limited by the number of channels.

3.1.2 - Profile dataset

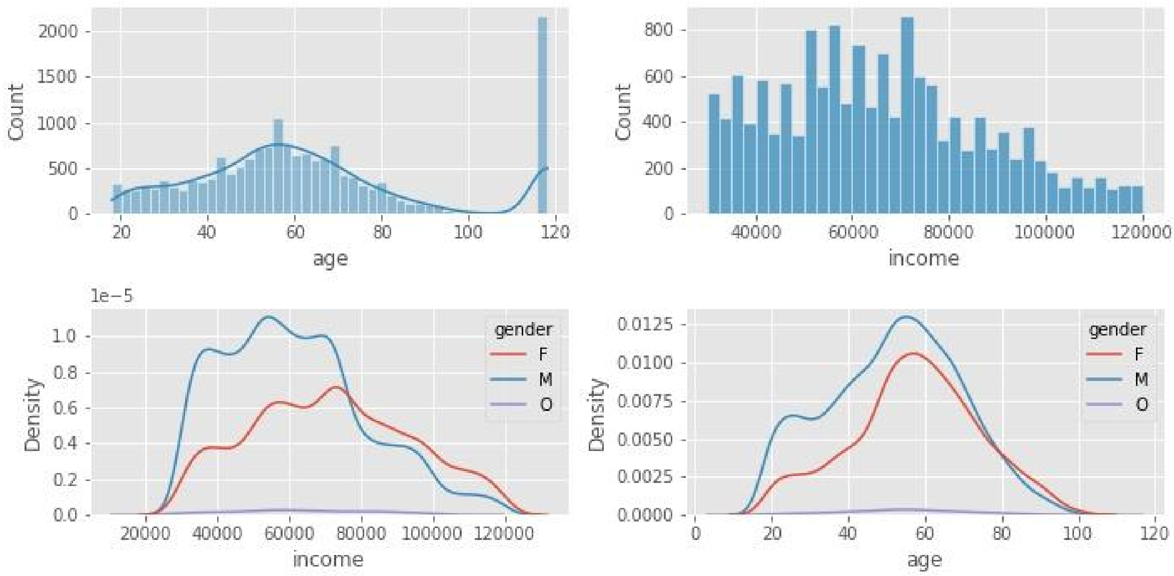
The second dataset is the profile.json, in this dataset we have some personal information about 17000 clients. Below we show all the variables:

gender	*string*
age	*integer*
id	*string*
became_member_on	*int (date format)*
income	*float*

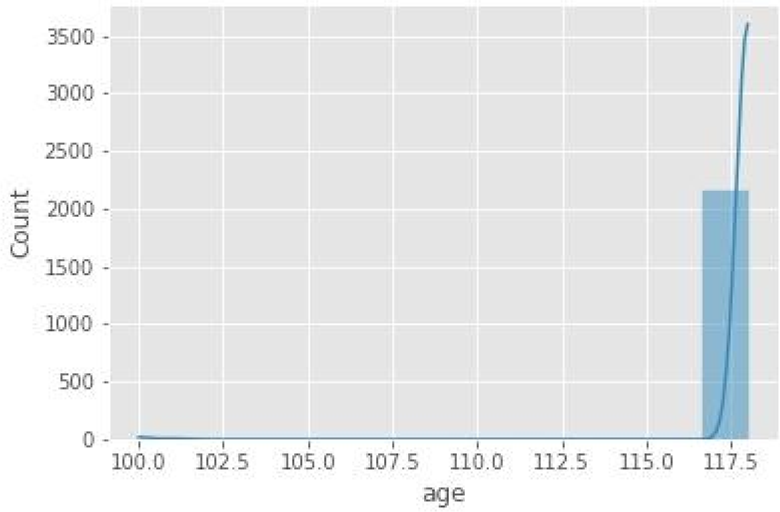
In the figure below we have the distribution from one of the variables, in this case, gender.



We can see that the majority of our clients are males. When we stratify the *income* and *age* variables by the *gender* we have as shown in the figure below:



There's not much information about how these 2 variables behave together. But we observe that the variable *age* has some clients over 100 years old as shown below:



In total we have 2175 clients with the age of 118, we choose to handle these variables later in the preprocessing step. Another point to mention is that 2175 missing values exist for each variable, *gender* and *income*, so we assume that the 2175 unusual values in *age* come from the same problem of *gender* and *income*.

3.1.3 - Profile transcript

The last dataset given is the transcript.json, in this dataset we have all the transactions from our customers, the variables are:

person	*string*
event	*string*
value	*string*
time	*integer*

From this dataset we can observe that in the *event* variable we have 4 categories: 'offer received', 'offer viewed', 'transaction', 'offer completed'. In the *value* variable we have a dictionary that contains the related transaction based on a certain event for a specific client.

4 - Data Preprocessing

In this section we are going to do some preprocessing in our datasets to correct some errors and create new features that may give us some insights. Below we have the steps we are going to make in each dataset:

Preprocessing Part 1

1. Portfolio
    - Creating unique columns for each *channel*.
    - Creating unique (dummies) columns for each *offer_type*.
    - Renaming the *id* to *offer_id*.
    - Renaming all the offers to a simpler label (i.e 1, 2,..)
2. Profile
    - Renaming the id to *customer_id*.
    - First, let's handle the null values in the *gender* variable creating a new category "N". Then we can create unique columns (dummies) of gender.
    - Fill all the 118 in *age* with *NaN* at this moment.
    - Since we have the variable became_member_on, we can create a variable that defines the membership age.
    - Fill the null values from *age* and *income* with mean.
    - Drop the *became_member_on* variable.
3. Transcript
    - Renaming the id to *customer_id*.
    - Extract *reward*, *offet_type* and *amount* from the dictionary variable *value*.
    - Drop the variable *value*.

Preprocessing Part 2

From the 3 variables extracted before in the transcript dataset, we can create a table with the information if the client successfully completes an offer. With that, we merge all the datasets for a more rich analysis.

5 - Data Analysis (Part 2)

With the dataset created before in the preprocessing part 2, we need to do some exploratory analysis to investigate with the perspective of successful offers. First, let's see what is the successful rate of each offer:

offer_id	successful rate (%)
5	70,0
7	68,18
4	58,37
9	44,63
10	44,60
8	39,40
1	38,42
6	23,09
2	0
3	0

From the table above we see that some of the offers have a high successful rate, for example, the offers 4, 5 and 7 have above of 50% successful rate, on the other hand, we see that the offers 2 and 3 have 0% impact on the public. For another point of view, let's see how the successful rate is based on some of the offers attributes.

reward	successful rate (%)
0	0
2	54,27
3	68,19
5	40,18
10	44,61

difficulty	successful rate (%)
0	0
5	48,81

7	68,18
10	49,42
20	23,09

duration	successful rate (%)
3	0
4	0
5	51,46
7	47,65
10	46,47

From the variables above, we see that some of the attributes alone don't have an impact on the successful rate.

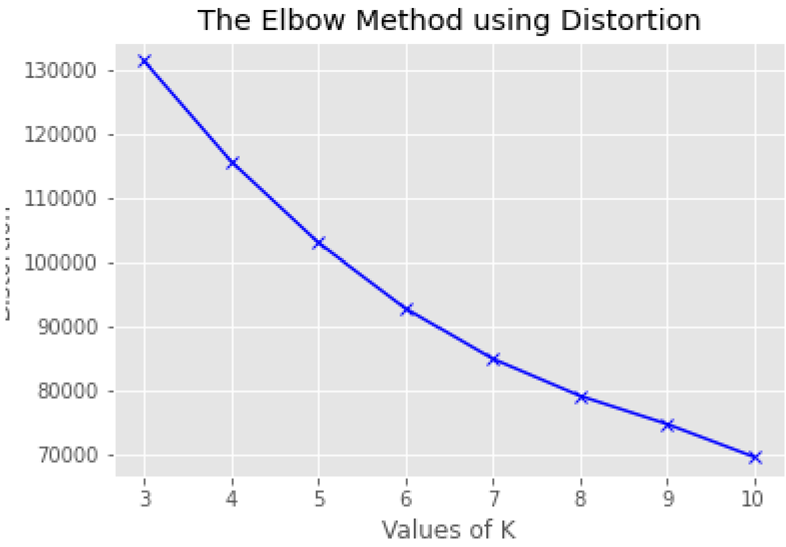
6 - Modelling (Customer Clustering)

First of all, before we go to a more advanced technique, we need to understand our customers, especially the clients who successfully convert to an offer, filtering only the successful offers from the dataset obtained from the preprocessing part 2, we have 24460 observations. Now, we need to understand how this type of client behaves. In this notebook, we are going to try to cluster all the clients that convert for an offer. We are going to use a model called **K-Prototypes** to do that. The reason we choose this model, is because we can handle mix features on our dataset, this model is a combination of K-Means and K-Modes. The variables we are going to use are:

$numericals = 'reward', 'membership\_age', 'difficulty', 'age', 'income', 'duration'$

$categoricals = 'channels', 'offer\_type\_cat', 'gender\_cat', 'offer\_id'$

After training the model for 8 iterations (from 3 to 10 clusters), we have the *elbowplot* below.



From the *elbowplot* we conclude that the model doesn't have an explicit inflection in any of the values of K (number of clusters). For the first analysis, let's choose 5 clusters just for simplicity and analyse how is the description of these clusters.

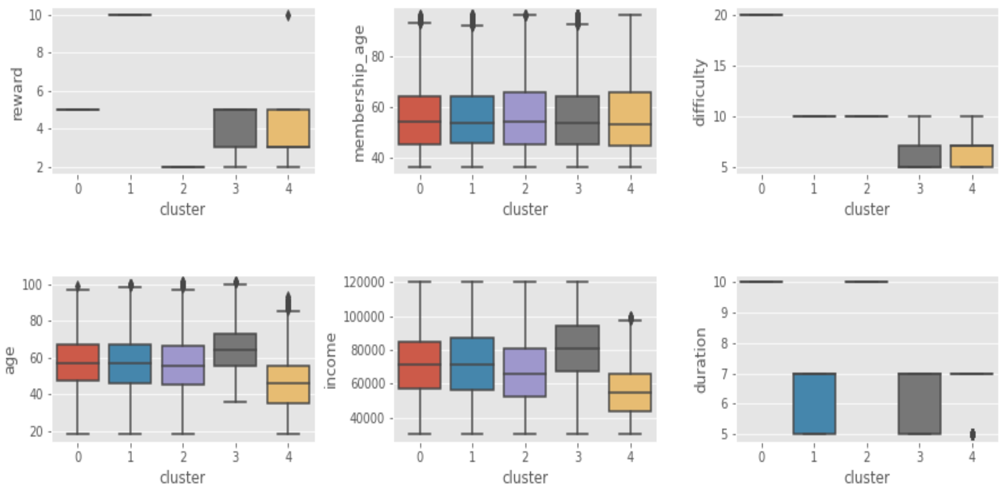
After training the model for 5 clusters, we have the following results:

Cluster	Population (%)
0	1472 (6,91%)
1	5667 (23,16%)
2	4433 (18,12%)
3	6521 (26,66%)
4	6367 (26,03%)

From the table above, we see that our model distributed well our observations, let's see how the variables in each cluster are distributed.

<u>*Numericals*</u>

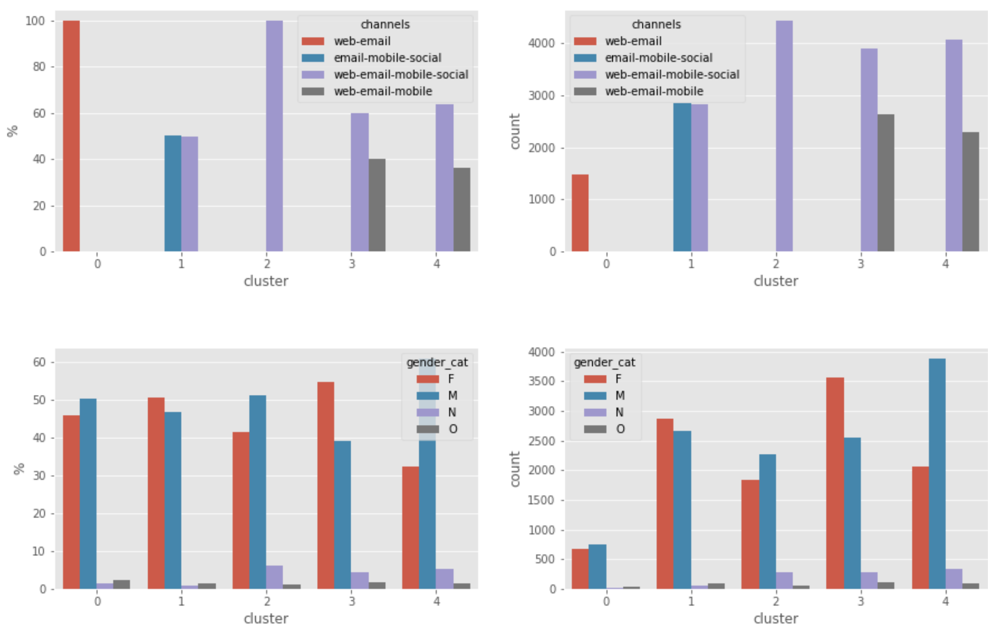
Adding the clusters into the dataset, we can explore how the variables behave when we segment by cluster, this helps us define the persona (description) of each cluster, based on the key differences in each variable. Below we have a *boxplot* showing how each variable behaves by cluster.

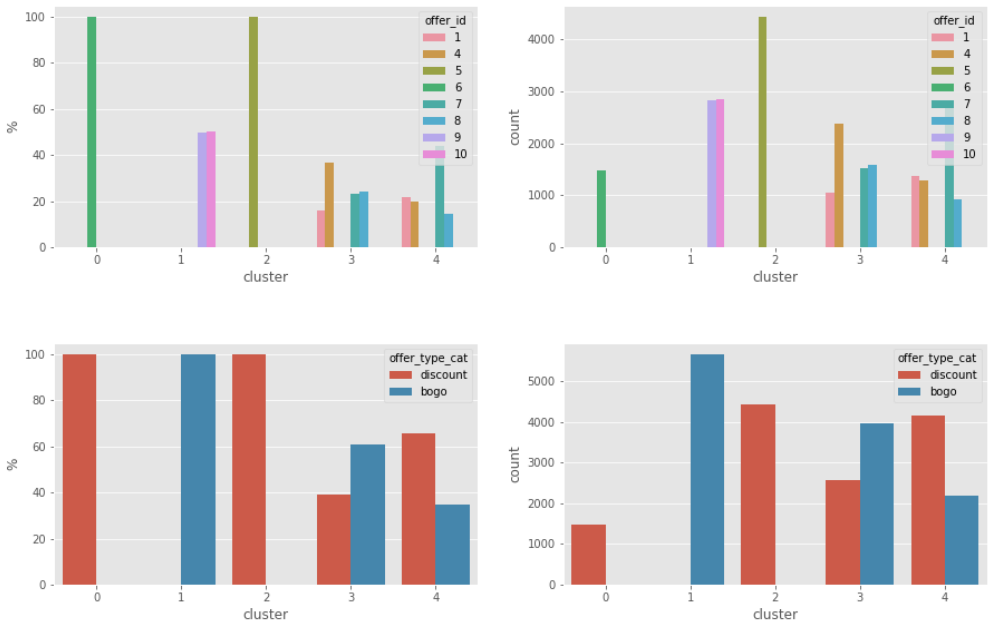


We see that some of the variables in certain clusters are basically the same, that doesn’t interest us because we want a “unique” characteristic from a variable in the cluster. So for example, cluster 0 in the *reward* variable has only the value 5, and no other cluster has the same value, so this is an unique feature from this cluster.

<u>Categoricals</u>

Exploring the categorical variables for each cluster , we have the following distributions:





From the categorical variables we have too some unique features in some of the clusters.

7 - Cluster Analysis

<u>Defining the personas</u>
From the graphics above, we can aggregate all this information and define some unique characteristics of each cluster that may help us to take some action in how we communicate with our customers. The personas are:

Cluster 0 - 1472 (~6%) clients:

- Clients with **reward** equal to 5, **difficulty** equal to 20 and **duration** equal to 10;
- All these clients convert from web and email;
- The offer type they receive is **discount**;
- And all of them convert from the **offer number 6**.

Cluster 1 - 5667 (~23.2%) clients:

- Clients with **reward** equal to 10, **difficulty** equal to 10 and **duration** falls between 5 and 7;
- All the clients in this cluster convert from all channels;
- They convert from **bogo** offers;
- And all of them convert from the **offers number 9 and 10**.

Cluster 2 - 4433 (~18.1%) clients:

- Clients with **reward** equal to 2, **difficulty** equal to 10 and **duration** equals to 10;
- All the clients in this cluster convert from all channels;
- They convert from **discount** offers;
- And all of them convert from the **offer number 5**.

Cluster 3 - 6521 (~26.6%) clients:

- Clients with **reward** that falls between to 3 and 5, **difficulty** falls between to 5 and 7 and **duration** falls between to 5 and 7;
- They are on average the oldest clients and with the highest average income.
- All the clients in this cluster convert from all channels;
- They convert from **discount** (40%) and **bogo** (60%) offers;
- About 60% of the clients are males;
- They convert from multiple offers (1, 4, 7 e 8).

Cluster 4 - 6367 (~26%) clients:

- Clients with **reward** that falls between to 3 and 5, **difficulty** falls between to 5 and 7 and **duration** equals to 7;
- They are on average the youngest clients and with the lowest average income.
- All the clients in this cluster convert from all channels;
- They convert from **discount** (60%) and **bogo** (40%) offers;
- About 60% of the clients are females;
- They convert from multiple offers (1, 4, 7, 8).

8 - Conclusion and future work

From the clusters, we observe that they separate our clients into a few characteristics that we can work with the marketing team to boost sales.

Another important observation we can make is that not all the offers we send have a good impact on the conversion. The offer that converts very few people has to be rethought.

Another way to observe the behavior of the clients is to cluster all the clients, the ones who convert and the ones who don't. this will give us more information about all the clients we have and investigate some bias in our first clusters.

If some of our clusters successfully change of conversion (A/B testing), we can go further and use some predictive machine learning model to infer what is the probability of a customer converting to a offer (y cluster). We can use a **lift chart** to verify the performance of our model and work together with the CRM teams.