**I. Definition**

### Project Overview

In a world where advertising is becoming increasingly competetive, it is becoming also more and more important to target your audience accurately in order to ensure your business can continue to thrive.

With the recent advancements in computing, machine learning algorithms are now largely available in a number of frameworks that makes it easy for everybody to train and use. Leveraging this new technology to target your customers accurately is nowadays paramount.

Starbucks is a coffee company that was founded in 1971 in Seattle and despite his strong presence is the market, the recent pandemic hit the sector quite strongly.

Starbucks reported his first quarterly loss in 7 years (source: <https://www.ft.com/content/4c876c5a-a03a-3f57-8a04-c364b12ad4b7>) and many bars and restaurants had to close down as their profitability was strongly undermined by the social distancing guidelines put in place by world governments.

### Problem Statement

Starbucks wants to improve the way it proposes deals and offers through the app to its customers.

The aim of this project is to leverage machine learning algorithms to increase revenue by targeting the right audience and drawing them to the stores.

The goal of this project is to analyse historical data to draw insights and create a model to better target customers.

### Metrics

For the issue at hand we are going to use accuracy as metric to evaluate the machine learning algorithm.

The problem at hand is a binary classification problem, and the dataset at hand is balanced.

Furthermore, the business does not incur any severe consequence in case of a false positive or a false negative.

## II. Analysis

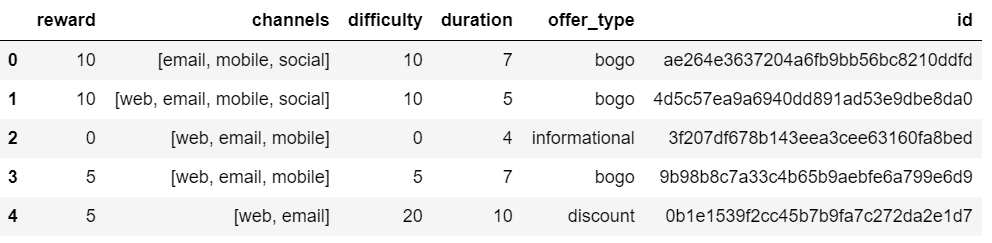
### Data Exploration

In this first section we are going to explore the 3 datasets at hand.

We will also use plots to get a sense of what data is avaiable and how we might want to use it.

The goal of this phase is to get an insight on the datasets at hands and identify potential issues to tackle.

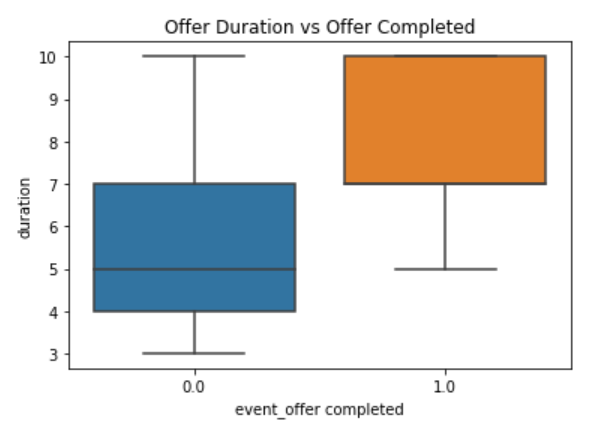
*Portfolio Dataset*



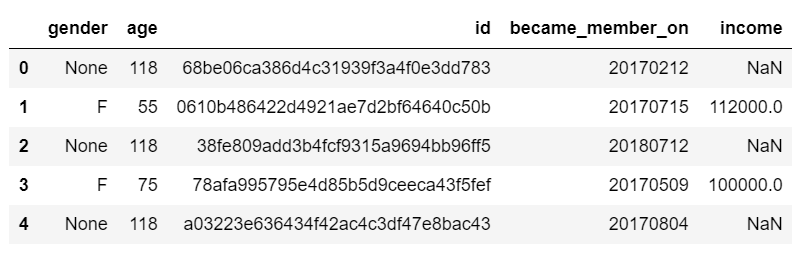
The portfolio has information about the offers Starbucks put forward. We see that 2 columns need to be one-hot encoded: Channels and offer\_type.

The following are the offer types: email, mobile, social, web.

Some of the features inside the Portfolio dataset are highly informative for the problem we are trying to tackle, for instance the higher the duration, the more likely the offer is completed.



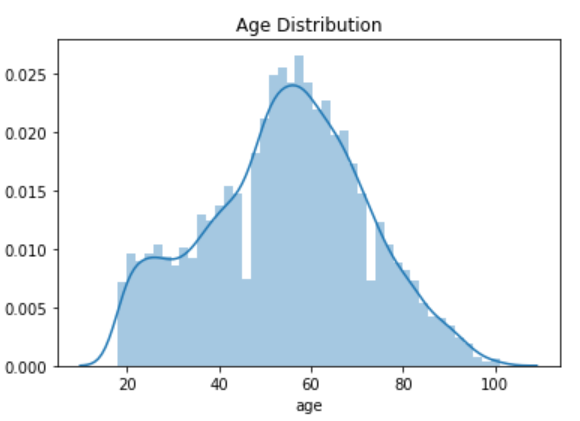
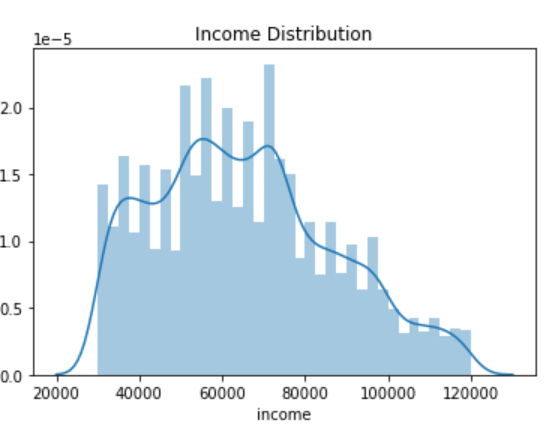
*Profile Dataset*



We have a number of missing values and all of them coincide with age 118. We have removed those.

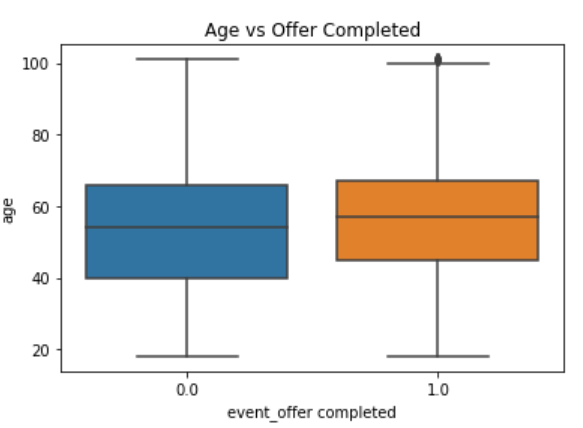
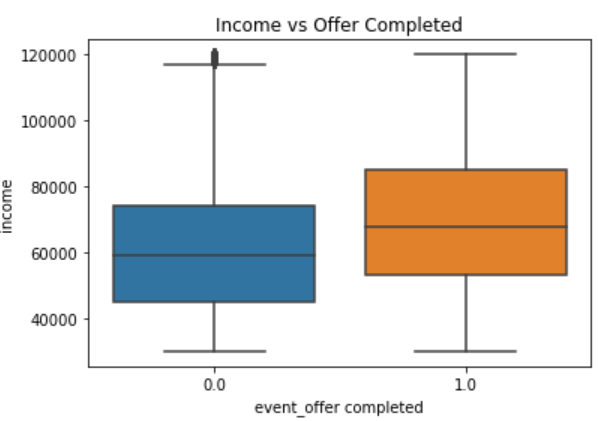
We will need to one hot encode gender andtransform the variable "became\_member\_on" to something that tells us the customer tenure, such as number of days since customer joined.

The following are the distributions of age and income with median values 55 and 64k respectively:



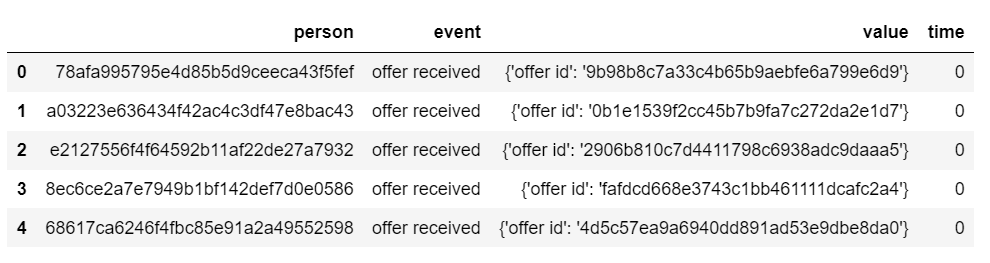
Income appears to be more influential than age in separating whether `an offer is completed.

Higher Income increases the chances of completing the order.



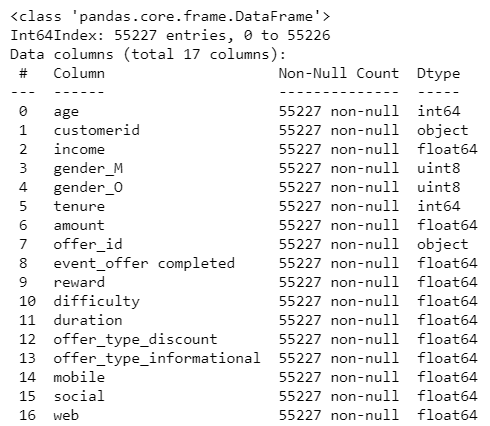
*Transcript Dataset*

The transcript dataset is essentially a mapping between the event and the offer



*Combined dataset*

This is what the dataset looks like once we have preprocessed them and combined them:

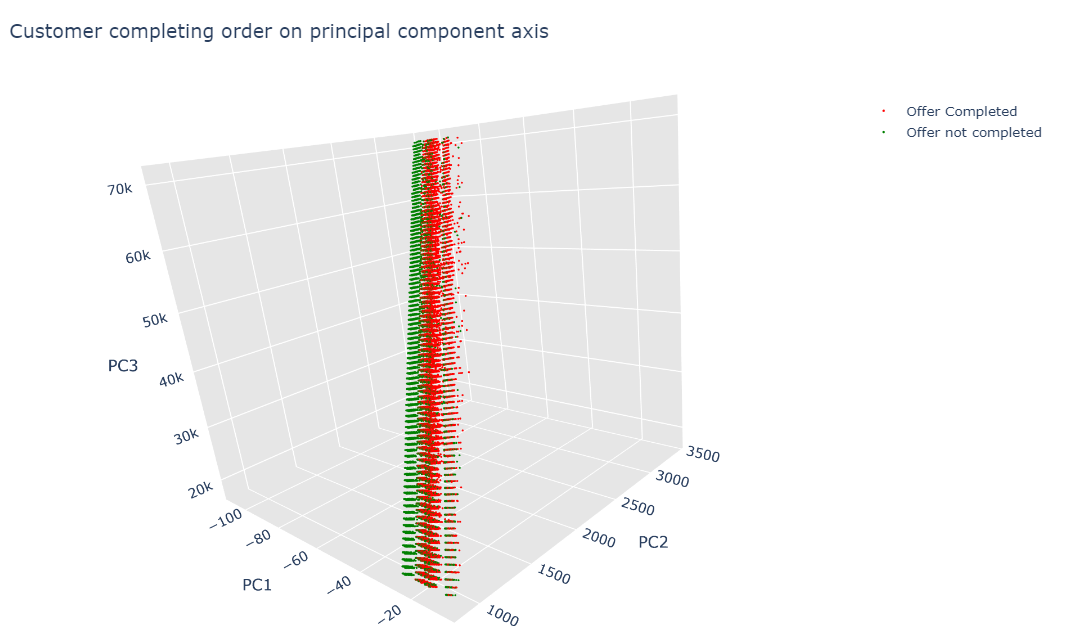


The dataset has now many dimensions and to further explore the separation between orders completed or not across all dimensions we are going to leverage a dimension reduction technique, Principal Component Analysis.

Using this technique, we can reduce the dimensionality of the data to 3 for the purpose of the visualization, retaining most of the variability in the original dataset.

Due to the high presence of categorical features, the points are sort of clustered towards the central axis.

But we do see a clear separation between the green and the red points, but also an area where they strongly overlap. We will need a strong classifier in a hyperdimensional space to achieve high accuracy.



## III. Methodology

### Data Preprocessing