# Capstone Project Report

## Definition

### A. Context

Starbucks is a worlwide brand with coffe stores in multiple countries. Aiming to incentivize customer purchases, they reach out it's customers with offers by app, email or social media.

There are two main types of offers that they promote `BOGO` (buy one get another) and `discount` offers. Usually they have a limited duration, for example, two weeks and a difficulty (the amount of money to spend). However, they also involve a reward, for example in BOGO offers your reward is the same as the amount that you bought.

### B. Problem Statement

Starbucks will like to develop a Machine Learning solution to evaluate how likely is an offer to achieve a high Completition Rate or CR for short. For this, they have collected some data that simulates the real behavior of their customers. Once, this Machine Learning tool is ready they would like to use it to take the decision of going or not forward with a proposed offer.

### C. Metrics

To evaluate the model accuracy, we're considering the the Mean Absolute Percentage error or $MAPE$ for short,

$$MAPE = \frac{1}{n} \sum_{n} \frac{y_n - \hat{y_n}}{y_n},$$

where $y_n$ is the real value and $\hat{y_n}$ is the predicted value. We choose these metrics cause it allow us an easy interpretation, it is more quantifiable than the `RMSE`. For example, if the model achieves a 10% MAPE, it is telling us that on average it is missing the real value by 10%

## 2. Analysis

### A. What is the customer's purchase behavior?

As a base question we wanted to see if the customer purchase behavior has changed from year to year and among income categories. For this consider the following

- `Top Income`. Customer with 100k+ annual income.
- `High Income`. Customer with 60k+ annual income.
- `Standard Income`. Customer below 60k annual income.

The graph below help us to visualize that Top Income customers seem to have higher purchase than other income category. Additionnally, we see that in recent year this purchase behavior has been decreasing, at least in median terms. Once we saw this, a natural question that arises is what is the effect of promotion, it could that on 2015 Starbucks launched more promotions that on recent years or that the strategies for launching promotions has been changing recently.

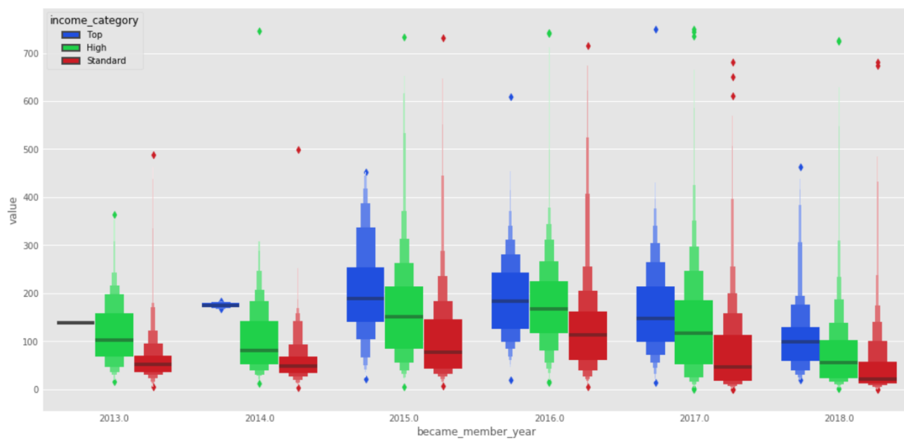


Figure 1: Customer_Purchase_Behavior

**B. What is the effect of promotions?**

We can evaluate a promotion by measuring its ***Completition Rate***, that is,

$$CR = \frac{\text{\# Offers Completed}}{\text{\# Offers received}},$$

and its ***Engaging Rate*** or Attractiveness Rate

$$ER = \frac{\text{\# Offers Viewed}}{\text{\# Offers received}},$$

once we define this, we found out that **Bogo offers are highly engaging** above 80%, think about this for a second, if Starbucks offers me another Moka Capuccino if I purchase one I will be really tempted to take that offer. However,

**discount offers incentive a purchase** nearly 60%, this could be explained by different factors like the customer adquisition capacity. For example if I only have five dollar and got two Starbucks promotions, one that I need to spent 10 dollars to receive another beverage of ones that let me buy a coffe at 6 USD with a 10% discount, then I will take the easy way and go for the discount offer.

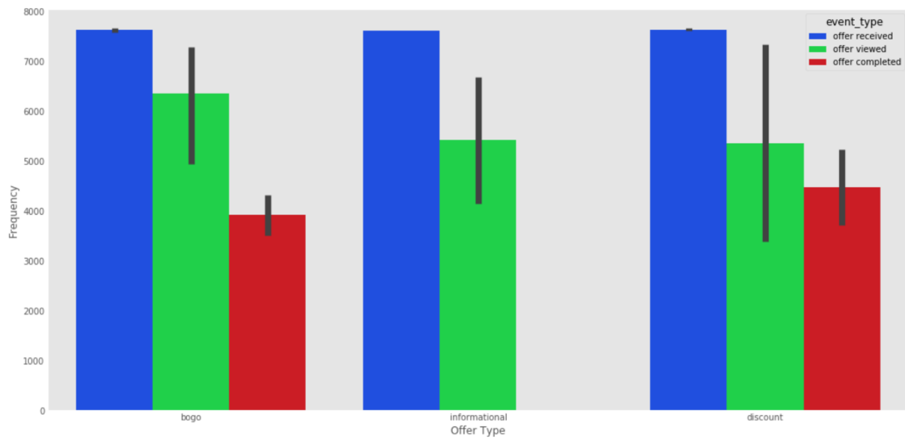


Figure 2: CR_ER

To better understand the engaging effect, let's take a look at the chart below. Here, we see two line plots, the green one with a positive slope and the blue one with a negative. In other words, **for BOGO offers the higher the difficulty, the higher its engaging effect**. By constrast, discount offers reduce its engaging effect if they are more difficult.

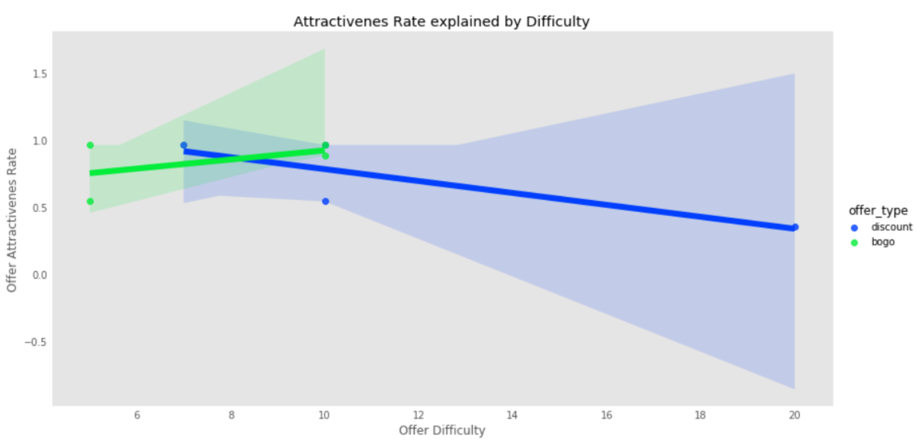


Figure 3: Engaging_Effect

Another key finding is that **Completition Rate is not affect by duration**, that is, one does not expect to have a higher completition rate if the promotion last longer. This has some further implications, for example, instead of having the same bogo offer for two weeks (1 Cappuccino for 1 Cappuccino), it will be better two have the first week the cappuccino offer and the second week a similar offer with a different Starbuck's product.

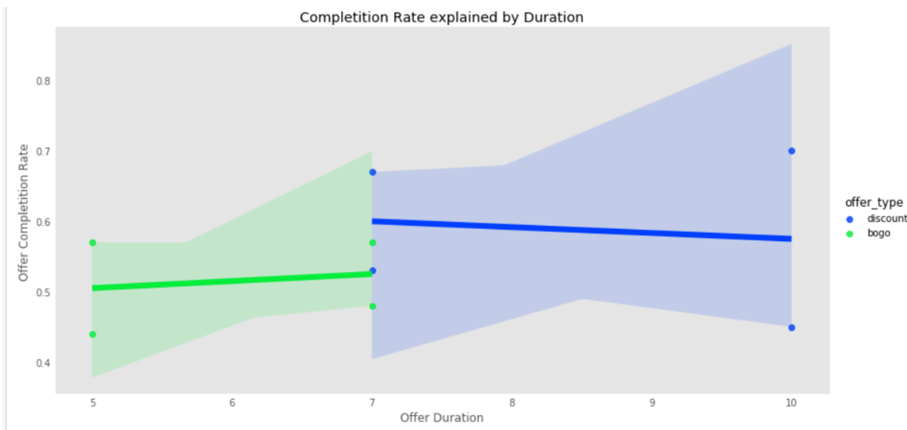


Figure 4: Duration_Effect

## 3. Methodology

### A. Imputing missing data through MICE algorithm

From the beginning, we seek to answer two key questions

- `How strong is the relationship between the variables?`.
- `Are there any hidden interaction Effects?`.

But, we should first address the issue of missing data, more specifically the `gender` and `income` information. It is natural to think that this data is missing cause it could be sensible, for example, a user that likes some privacy and don't want to gives its income or gender information.

To decide wether or not this data has relevant information, we split it in two groups, *group A*, the users that don't have information of age and gender. While, *group B* are the users that have full information. Next, we test the following hypothesis

$$H_1 : \text{Group A and B are different} H_2 : \text{Group A and B are equal} ,$$

to support $H_1$ we should statistically prove that the `mean`, `median` and `standard deviation` between the groups is different. Check out the notebook `Data_Exploration_and_Confirmatory_Data_Analysis` to see the statistical tests that we're taken based on wheter or not the data has a normal distribution.

Once we proof that missing data has relevant information, we decided to go with the Multiple Chained Equations Algorithm , or `MICE` for simplicity, to impute the missing values. MICE works basically by applying iteratively linear regression models to learn the distribution of the missing data.

### B. Engineering Features

To better explain the behavior of the promotions we decided to engineer to engineer the following features:

- `Average Time to Complete an Offer`. Considering the time when the offer was sent as time zero, what is the average time to complete the difficulty level of the offer.
- `Frequent Gender`. Among my customer, does women or men take more the given offer.
- `Frequent Income`. Of what income category does the users that completed a certain offer come from.
- `Customer Antiquity`. How long has the user been a Starbuck's customer.
- `Difficulty Rate`. We know that difficulty can take positive values, but we don't want that it increases linearly, we want to regularize its impact, thus we consider the difficulty rate as

$$DFR = \frac{\log \text{difficulty}}{\log \text{difficulty} + 1}$$

- `Reward Rate`. We want to measure the rate between reward and difficulty. As with the difficulty rate, we do not want a steady decrease, thus we consider the square root

$$RR = \sqrt{\frac{reward}{difficulty}}$$

### C. Why a Bayesian approach?

Remember that our objective is to **predict the Completition Rate** or `CR` (that is $y = CR$) for every offer. However, by framing the problem in this way, we encounter with the fact that we do only have 8 samples (we're not taking into account non-informational offers). This is one of the reasons we will consider a linear regression model.

From a frequentist approach, the linear regression model assummes that the data behaves like a Normal distribution. However, we are far even from Limit Central Theorem, cause we only have 8 observations. However, when we consider a bayesian approach we can very flexible about how do we assume that the data distributes, for example, we can take a non-informative prior if we assume we do not nothing about the data, mainly because we have only seen 8 observations of it, thus

$$f(\theta) = \frac{1}{\theta},$$

where $\theta$ is the vector of parameters that we would like to estimate.

## Results

### A. Boostraping Data

We reintroduce the evaluation metric, the Mean Absolute Percentage error or $MAPE$ for short,

$$MAPE = \frac{1}{n} \sum_{n} \frac{y_n - \hat{y_n}}{y_n},$$

where $y_n$ is the real value and $\hat{y_n}$ is the predicted value. A good way to estimate the error is by boostraping the data, that is, if our sample size is $n$ then we take $m$ desired samples with replacement of size $n$ and then apply the Bayesian Linear Regression Model to each one of this samples. For illustration, see the image below taken from this Medium article.

If you want to further see how we applied this, see the function `reg_bootstrap` under the notebook `Model_Building`. We should mention that we're considering `normalize=True` so that the `BayesianRidge` estimator normalizes the data and `compute_score=True` so that at every iteration it computes the `log-marginal likelihood`.

```
clf = BayesianRidge(compute_score=True, normalize=True)
```

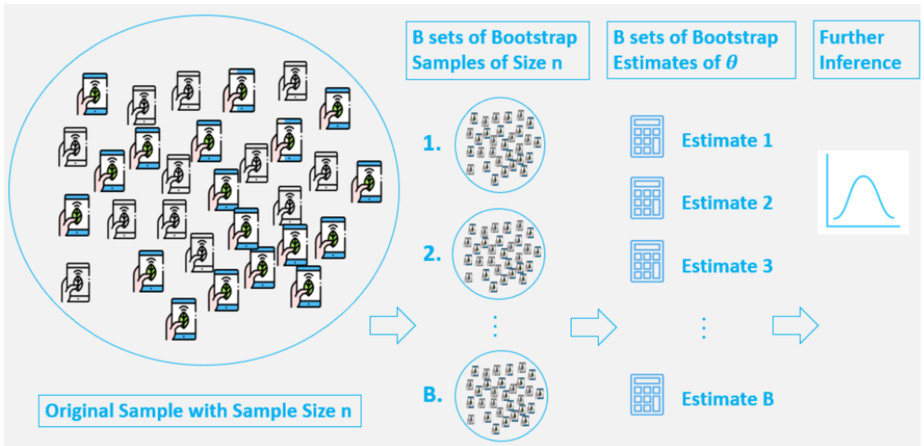


Figure 5: bootstrap_data

### B. Model Evaluation

Next, take a look at the behavior of the $MAPE$ at every iteration. Notice that most of the time it is between five and twenty, in fact, the average value is 12%. In other words, nearly 88% close to the real value on average.

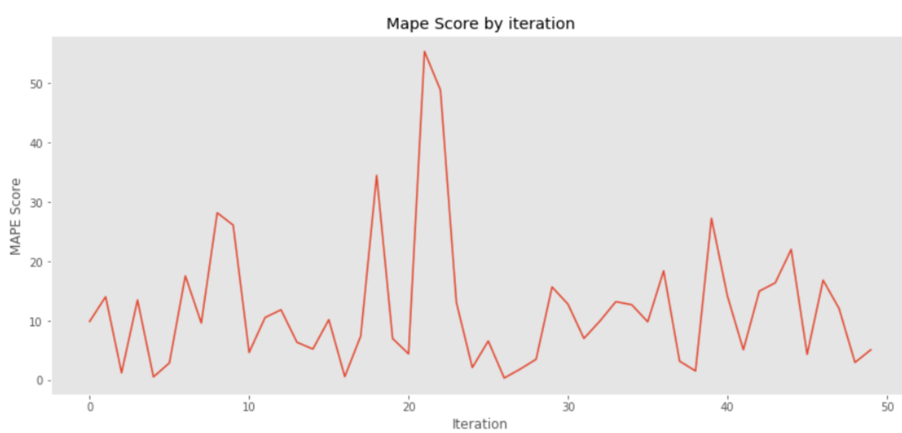


Figure 6: Mape_behavior

### C. Which offer should we recommend?

As we do not want to split an eight observation data on train-test. We assume a fake scenario. Suppose that a Starbuck's Project Manager comes to us with three different offers (that target three different customer segments) that they would like to assess their completition rate `CR`.

Based on what the PM told us about the offers, we resume each one of them in the following:

- **Bogo for Top Income and Recent Users**.
- **Bogo for Standard Income and Old Users**.
- **Bogo for High Income and all users**.

We can represent this offers as one dimensional arrays to be taken as inputs for the model, it outputs the following Completition rates:

```
Offer 1  predicted CR: 33.992025821141816
Offer 2  predicted CR: 78.07046826678777
Offer 3  predicted CR: 75.67366728846704
```

As a first sight, we could go with offer three cause it has the highest predicted `CR`. However, we should take into consideration some limitations that are inherent to offers, for example, it may be that offer 3 is cheaper and faster to deply as offer two, or that it is not valid for all the Starbucks stores. Imagine that only High Income people go to a certain Starbucks because it is in an exclusive residential zone. Thus, we should take into consideration these limitations, before giving a final recommendation.

## Conclusion

### A. Reflection (Framework Limitation)

The way we address the problem (predicting the completition rate) let us to a problem of few data. Fortunately, there are statistical techniques that we can apply to address this issue like a Bayesian Framework or bootstraping data.

By considering this approach we will expect in the future to receive more data, remember that the data consider only one month, so having at least one quarter data or one year data should be considered enough.

Additionally, this limitation let us to consider only simple models, particularly a Bayesian Linear Regression Model. In this sense, one could not try out more robust models like Random Forest Regressor or a Neural Network.

### B. Improvement

As we mentioned before, to consider a most robust model, we would need to gather more data. However, it remained to compare the linear model to a simple heuristic or a business rule. For example, based on the results of the Exploratory Data Analysis we could just split our customers based on their income and frequency to test different offers.

To deal with the lack of data, we could have considered generating synthetic data, for example, with a bayesian neural network. In theory, the bayesian framework should be robust enough to do inference with few samples, remember what we said about the prior. Thus, it would be an interesting approach to generate at least 1000 samples with a bayesian network.