

# Machine Learning Engineer Capstone Project

## Capstone Proposal

Yemi Awosanya

April 21, 2021

### Section 1

#### 1.1 Domain background

Starbucks is passionate about improving and uplifting the lives of its customers, one beverage at a time. In doing so, they have created a mobile application that rewards and sends offers to their valued customers. The application offers promotions such as discounts or buy-one-get-one-free (BOGO), and occasionally, informational offers.

We will be using the study of consumer behavior to see how marketing campaigns can be adapted and improved to more effectively influence the consumer.

#### 1.2 Problem statement

Because of Starbucks' dedication to their customers, we will be analyzing customer data to explore ways of improving their experience and to serve them better. That said, this project aims to predict whether or not a customer will respond to an offer based on how each customer makes purchasing decisions.

#### 1.3 Datasets and inputs

We have three datasets as follows:

##### **profile.json**

##### **Rewards program users (17000 users x 5 fields)**

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became\_member\_on: (date) format YYYYMMDD
- income: (numeric)

	gender	age	id	became_member_on	income
0	None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	None	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN
3	F	75	78afa995795e4d85b5d9ceeca43f5fef	20170509	100000.0
4	None	118	a03223e636434f42ac4c3df47e8bac43	20170804	NaN

## portfolio.json

Offers sent during 30-day test period (10 offers x 6 fields)

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer\_type: (string) bogo, discount, informational
- id: (string/hash)

	reward	channels	difficulty	duration	offer_type	id	email	v
0	10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd		1.0
1	10	[web, email, mobile, social]	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0		1.0
2	0	[web, email, mobile]	0	4	informational	3f207df678b143eea3cee63160fa8bed		1.0
3	5	[web, email, mobile]	5	7	bogo	9b98b8c7a33c4b65b9aebfe6a799e6d9		1.0
4	5	[web, email]	20	10	discount	0b1e1539f2cc45b7b9fa7c272da2e1d7		1.0
5	3	[web, email, mobile, social]	7	7	discount	2298d6c36e964ae4a3e7e9706d1fb8c2		1.0

## transcript.json

Event log (306648 events x 4 fields)

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type

- offer id: (string/hash) not associated with any "transaction"
- amount: (numeric) money spent in "transaction"
- reward: (numeric) money gained from "offer completed"
- time: (numeric) hours after start of test

	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

#### 1.4 Solution statement

We will create a machine learning model to predict if a user will accept an offer or not. The process to a clean solution is listed below:

1. Data Cleanup - This includes dropping or filling null/empty cells, removing duplicates, etc
2. Data Analysis- This includes feature engineering, etc
3. Data Visualizations - Visual representation of our data, to provid more insights
4. Model Training - Train model using the algorithms discussed in the benchmark section
5. Model Analysis - Evaluate how well the model performs using the evaluation metrics

#### 1.5 Benchmark model

For the benchmark model, we will be using the GradientBoostingClassifier. Also, we will be exploring the xgboost algorithm from sklearn.

#### 1.6 Evaluation metrics

We will be using the following metrics to evaluate my model:

1. roc\_auc\_score - Area Under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores
2. accuracy\_score - The number of correct predictions made divided by the total number of predictions made
3. f1\_score - Weighted average of the precision and recall, where an F1 score reaches its best value at 1

#### 1.7 Project design

We will go over the three datasets provided; remove inavlid or null values, drop duplicates, fill columns with mean/mode/median values, etc.

In other to see the relationship(s) between customer traits that influences a customer's buying power, we will combine all cleaned datasets, and plot graphs. Usually, this step influences the generation of new features that can be used to build an excellent predictive model.

We will split the final, cleaned data into training and testing datasets. Train our model and see how well it predicts if a customer/user will respond an offer or not.