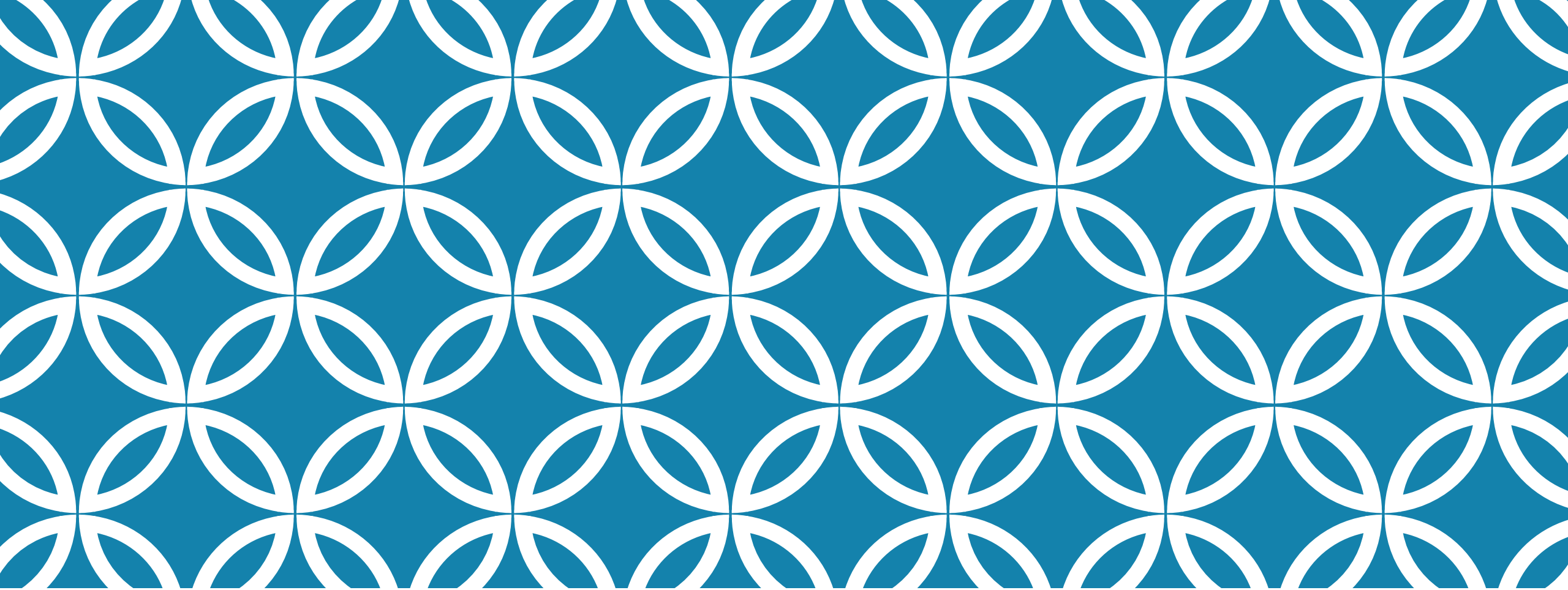

CAPSTONE PROJECT PROPOSAL

Machine Learning Engineer
Nanodegree

INFO

Toni Magdy

15th August, 2020

Starbucks Capstone Challenge

DOMAIN BACKGROUND

- This project from the marketing system, An application by Starbucks to keep in touch with its costumers and to make orders online.
- Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free).
- There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount spent. In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend. Offers can be delivered via multiple channels.

PROBLEM STATEMENT

- Starbucks needs a way to send to each costumer the right offer.
- Our goal is to analyze historical app data to find most appropriate offer for each one of the costumers.
- The appropriate offer when the costumer sees the offer and buy the products under the offer influence.
- The user can receive an offer, never actually view the offer, and still complete the offer, there will be an offer completion record in the data set, however, the customer was not influenced by the offer because the customer never viewed the offer.

PROBLEM STATEMENT (CONT.)

- The offer lifecycle

- It's a classification problem, we aim to know if the user will complete the offer or not, we will build a model that predicts whether or not the user will complete the offer,

DATASETS

The data is contained in three files:

- **portfolio.json:** containing offer ids and meta data about each offer (duration, type, etc.)
- **profile.json:** demographic data for each customer.
- **transcript.json:** records for transactions, offers received, offers viewed, and offers completed.

DATASETS (CONT.)

- The dataset is provided by Udacity and Starbucks.
- The program used to create the data simulates how people make purchasing decisions and how those decisions are influenced by promotional offers.
- Each person in the simulation has some hidden traits that influence their purchasing patterns and are associated with their observable traits. People produce various events, including receiving offers, opening offers, and making purchases.
- As a simplification, there are no explicit products to track. Only the amounts of each transaction or offer are recorded.

DATASETS (CONT.)

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)

DATASETS (CONT.)

profile.json

```
profile rows: 17000, profile columns: 5
```

Out[13]:

	age	became_member_on	gender	id	income
10412	85	20180404	F	503053089f114898b546bc6740d8e978	84000.0
14449	45	20180607	F	f2e49f5002c540eb92ca320fea990319	73000.0
16016	68	20171009	M	ab68c87257344ba7963064dd8b4b9350	33000.0
9449	118	20161031	None	c99a06c81f8540b49cb6a66719ea62dc	NaN
3302	60	20170302	M	7366bef4c288476dab78b09a33d0e692	52000.0
12484	118	20160727	None	b04385001db14fdf87829c6163ae9ddd	NaN
14313	98	20150403	M	75225655a1c44546a18f100f7c864f98	37000.0
15713	26	20180117	F	28416b56bdc94890a4996dd2dcc598b4	45000.0
5257	56	20180711	M	2d3c956111ad434786e39ed79354dd5a	66000.0
15232	118	20180525	None	270e7fd65f7e45c58b79d0d8ad2c72ab	NaN

DATASETS (CONT.)

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

DATASETS (CONT.)

transcript.json

```
transcript rows: 306534, transcript columns: 4
```

Out[11]:

	event	person	time	value
39728	transaction	9b9bd320b3b34859abfee2109a0b4831	90	{'amount': 3.6}
50487	transaction	06b1031271174d8596c1996478f07ede	150	{'amount': 0.31}
246979	offer received	489f08a011894421991b8cc0e6e0a946	576	{'offer id': '2298d6c36e964ae4a3e7e9706d1fb8c2'}
146286	transaction	991386e4c20041428093919ed3c8f2ba	390	{'amount': 0.43}
15223	offer viewed	48225ea573e545e0b704ce3fcca8bb9e	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
54706	offer received	055640cd12d04eb4b8a51ec67d451fc7	168	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}
51845	transaction	ae0e47bc419940d68686ae364e73212b	156	{'amount': 2.24}
232039	transaction	796cc7c1e8534e78bdff45f9e11494d6	534	{'amount': 1.97}
22257	offer completed	e110e63527c24ad1b482f76acde24a42	18	{'offer_id': 'f19421c1d4aa40978ebb69ca19b0e20d...
7390	offer received	8cc0db430879405898d8390ca74ad13a	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}

DATASETS (CONT.)

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)

DATASETS (CONT.)

portfolio.json

```
portfolio rows: 10, portfolio columns: 6
```

Out[12]:

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

SOLUTION STATEMENT

- In order to solve this problem, we will build a machine learning model to study costumers behaviors.
- we'll try to train several classification models to predict whether or not the user will complete the offer, we'll use Random Forest, Sagemaker XGBoost, RNN, SageMaker LinearLearner.
- The output of the model should be 0 if the user will not complete the offer, and 1 if the user will complete the offer.

BENCHMARK MODEL

- We will use Logistic Regression as a simple machine learning algorithm to compare the results with.
- Logistic Regression is simple and easy to implement.

EVALUATION METRICS

It's a classification problem, so that below evaluation metrics should be able to determine the model performance.

- Accuracy $= \frac{True\ Positives + True\ Negatives}{Total}$

  Accuracy is how many points did we classify correctly.

- Precision $= \frac{True\ Positives}{True\ Positives + False\ Positives}$

  Precision is proportion of positive cases that we classify correctly.

- Recall $= \frac{True\ Positives}{True\ Positives + False\ Negatives}$

  Recall is proportion of actual positive cases the we classify correctly.

PROJECT DESIGN

- Data loading and exploration: load the dataset and present some data visualization in order to understand the data.
- Data Cleaning: clean the dataset and fix any issues.
- Feature Engineering: prepare the data to be suitable for the model.
- Split Data: split the data into training and test sets.
- Train the model: train the machine learning model.
- Train the benchmark model: train the benchmark model to compare the results with.
- Evaluate the models: Test the models and compare the results.

THANK YOU!