**Capstone Project Proposal**

Udacity Machine Learning Engineer Nanodegree

Starbucks Project

By Jing Li

July 23, 2020

**Background Information**

Starbucks is the largest international coffeehouse chain that based in Seattle, Washington. Introduced in April 2008, Starbucks launched its first ever rewards program. Over the course of past several years, the company had replaced its one-star-per-transaction model in favor of a revenue-based loyalty program. Early last year, 2019, Starbucks redesigned the program to include tiers, opening more ways to redeem rewards.

Closely related, this project is dedicated to providing actionable insights that help drive the company’s marketing campaign efforts to a rather optimal state which in return, more effectively reaching out the right customers with individually tailored rewards offerings.

Intuitively, the benefit of this analytical process, which described below, is to enable healthy marketing promotions that could be more smoothly embedded within multiple channels such as email, mobile, and web, etc.. Thus, achieving higher revenue.

**Dataset Overview**

Starbucks has provided three datasets in the format of Json. The three date files are names as the following:

* profile.json (1700 users, 5 features)

This table contains rewards program’s user demographic information such as gender, income and age.

* + gender: (categorical) M, F, O, or null
  + age: (numeric) missing values encoded as 118
  + id: (string)
  + became\_member\_on: (date) format YYYYMMDD
  + income: (numeric)
* portfolio.json

This table provides overall information about total of 10 offers designed to be distributed.

* + 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 offers to be open, in days
  + offer\_type: (string) bogo, discount, informational
  + id: (string)
* transcript.json

This table provides log details including which customer reacted to which offer they received, and what is the result.

* + person: (string)
  + event: (string) offer received, offer viewed, transaction, offer completed
  + value: (dictionary) different values depending on event type
    - offer id: (string) not associated with any “transaction”
    - amount: (numeric) money spent in “transaction”
    - reward: (numeric) money gained from “offer completed”
  + time: (numeric) hours after start of the test

**Problem Statement**

This analysis aims to provide business users an end-to-end solution with the goal of answering the following key business questions:

1. Should the marketing team send offer to a particular rewards program customer?
2. If so, then which offer would help business achieve highest revenue and lowest cost?

There are two key pivotal points that are worth mentioning when addressing the above questions. They are “customers” and “offers”. These two important aspects set up the main theme of the entire analysis. One may notice that all solution components of this analysis focuses on answering either “customers” related question, “which person should we send offer or prioritize”, or “offers” related question, “which offer would maximize the value that it would bring to the business”.

In addition, the key business questions are further decomposed into smaller questions that are answered throughout each solution components. These decomposed questions are listed in the following:

1. What are the success rates for each offer?
2. What statements can we make when comparing the bogo and discount offers?
3. When people make certain decisions about the offers, are there any common characteristics that these people share?
4. If so, how does this group of people react towards each type of offer?
5. From looking at the duration of certain customers from they first review the offer to offer completion, what conclusion can we make?

**Solution Structure**

To address the key business questions provided above based on both “customers” and “offers” perspective, the entire analysis has been divided up to four components, they are:

1. Descriptive Analysis (perspective: customers and offers)

The descriptive analysis is designed to provide a business user with a soft understanding of how customers can be viewed as a certain group based on some common characteristics, as well as where do the current rewards offerings stand in terms of success rate, and monetary value.

Upon the successful execution of this analysis, a user would be able to make conclusions in regard to 1-3 sub questions stated in the previous problem statement section.

Methodology to be examined:

* frequency, recency, monetary segmentation
* segmentations based on common statistics
* unsupervised learning

1. Predictive Modeling (perspective: customers)

The goal of this part of the analysis is to assign a “expected value” to each rewards program customer. Therefore, the predictive modeling component is consisted of two modeling technique, classification and regression modeling. The expected value is defined as multiplying the probabilities of each customer becoming a bogo or discount promotion participant with the predicted spend amount. Please note, since informational offers does not require any monetary value to complete, we will treat this type of offer separately from the other two types of offers. However, how likely one would make a purchase on the bogo or discount offers after viewing the previous informational offers that sent to them is to be studied.

Another interesting topic to investigate from a customer perspective is to study the time that took for each customer who decided to participate from opening the offers to competition. Survival analysis is suited for further investigation into this topic.

Upon successful execution of this component, sub question 4 and 5 will be answered.

Methodology to be examined:

* classification
* regression
* survival analysis

1. Impact, Cost Analysis (perspective: offers)

To analyze offers’ impact, one needs to consider the cost which bringing in a new part of the analysis to the workflow, and that is cost analysis. Please note that certain cost assumptions will be made in the following section as it is required for this type of analysis. Specifically, a cost function will be introduced to determine the cost when a classification model wrongly predict customers’ reaction towards certain offering types.

Upon successful completion of this component, one would be able to fine tune the previously trained classification probability threshold to an optimal setting so that a particular customer group’s cost could be minimized, thus having a better chance to obtain higher expected value for that particular customer group.

Methodology to be examined:

* cost function
* fine tune probability threshold for the classification model based on the provided cost function

1. Report, Recommendation

Required as part of the deliverable of this capstone project, a report is to be created to communicate to the technical audience about the enabling process of each analytical components of this analysis.

However, since this capstone project aims to provide an end-to-end solution not only to the technical users but also to a broader business audience, the report serves as a key component to provide actionable recommendations to the business.

Upon success creation of the report, business users are provided with key findings obtained from answering previously defined sub questions. The report also further clarifies how these key findings help one deriving a conclusion to the key business questions stated in the beginning.

Methodology to be examined:

* conducting key findings
* generalize to conclusions to the key business questions
* derive specific recommendations

**Business Assumptions**

So far, we have stated two key business questions, along with two breakthrough perspectives, customers and offers. In addition, we have also decomposed the key business questions into several smaller business questions through which we will use the defined solution structures to lead us to a conclusion to the key business questions.

Nevertheless, the following general business assumptions are made:

* A customer can make a purchase decision with/without being a rewards program member, and the provided dataset only contains customers that are rewards program members.
* However, promotion offers are only available to rewards program members.
* There is a “funnel” stage to follow. For example, after viewing the bogo and discount offers, a customer can choose whether participating the event. However, in order for this customer to view these offers in the first place, he/she must first receive an informational offer and opened it.
* No one leaves the rewards program after becoming a member.
* There is a cost associated to each type of classification modeling error, and there is a fixed cost to produce each type of promotion offers.

Please note that making cost related assumptions before view the data is extremely difficult, therefore the following assumption may change during the actual analysis:

* Based on the entire population, which is unknown to us, the percentage of rewards program members who viewed their first informational offers then became participants of the bogo and discount offers is 20%.
* The bogo and discount offers cost the same, $20 dollars per customer (one-time).
* A lost customer who is potentially interested costs the business $200 dollars per customer (one-time).
* Based on the provided dataset, the bogo and discount offer will be X% effective (this “X” is to be calculated in the analysis). That is X% of the program members who are identified as potential participants then receive the bogo and discount offers decide to participate and make purchasing decisions. The remaining Y% (this “Y” is to be to be calculated in the analysis) will not participate anyway.

**Metrics**

Value is naturally important to many businesses; this concept is also the motivation of the second and third components (predictive analysis, cost analysis) of the solution structure. Therefore, “expected purchase amount” and “cost” are inevitably discussed throughout the analysis.

For a rewards program member to be valuable, it means he/she must have a high expected purchase amount and in order to achieve this, the concept of MAE, F1 score, Type I and Type II errors will be used during the regression and classification modeling which implemented in the second component.

In order for a promotion offering to be valuable to business, it must effectively help increasing the revenue while maintaining low cost. While the provided dataset does not include any revenue information, we can make assumptions, stated in the last section, about cost. This is the backbone of our third component, the cost analysis.

**Benchmark Model**

For the second component of this project, predictive analysis, linear regression and logistic regression will be used as our baseline model.

For the third component of this project, cost analysis, admittedly, it is difficult to find a metric to justify whether a cost is lower enough. However, I will provide two “what-if” scenarios as reference points in the following that describes the options of “do-nothing” and “offers-to-all”. These two scenarios take into account the probability of both Type I and Type II errors, and the probability that the promotion offer will be X% effective.

* Option 1: “do-nothing”

This option would cost $0.00 per customer in producing the promotion cost but incur a lost potential-purchasing-opportunity of 20% \* $200 = $40.

* Option 2: “offers-to-all”

This option would cost $20 per customer in producing the promotion offers as a fixed cost plus 20% \* Y% \* $200 =$40Y, totaling $20+$40Y per customer.