# Machine Learning Engineer Nanodegree

## Capstone Proposal

John David LEQUEUX
November 23rd 2020

## Starbucks Capstone Challenge Proposal

## Domain Background

To provide the next best offer to customers is a main topic in marketing. The goal is to predict the optimal personalized product offer to each individualized customers. As a consequence it will improve the chance that a customer uses the offer and then it will increase the promotion response rate. Another benefit is to improve the customer satisfaction by avoiding spamming.

With a high number of data and features to analyze, it requires a real time automated self-learning decision. Using machine learning algorithms is then necessary to create such model.

## Problem Statement

The aim of this project is to predict if an offer will be well received by different customers. In other words the goal is to determine if a customer will use an offer or not.

This binary problem can be solved by using a machine learning algorithm based on a binary classification. This is the solution we will investigate in this project.

## Data sets and Inputs

The inputs are obtain using simulated data that mimics customer behavior on the Starbucks rewards mobile app. It contains three files with different information. Here below a description of the content with the features which can be used in the model.

- *portfolio.json*: This file contains a description of 10 different offers with features related to the channels used, the difficulty, the duration, the type of offer, and the reward.

  Main characteristics:
  - An offer can be a BOGO, a discount or it can be informational only.
  - An offer can use up to four channel: email, social, mobile and/or web.
  - The difficulty, the reward and the duration are integers between 0 and 20. We can notice that the difficulty and the reward are null for the informational offers.

- *profile.json*: This file contains a list of customers with information related to the age, the gender, the customer id, the date when the customer created an app account and the customer's income.

  Main characteristics:
  - 2175 rows don't have any information concerning the gender and the income. Moreover the age associated for these rows is always 118. We can take the assumption that this a default age for customers who haven't populate their age, gender and income in the application.
  - Each customers have a different customer id.

- transcript.json: This file contains a list of record transaction with the customer id, the time in hours since start of test, and values depending on the record (either an offer id or transaction amount).

  Main characteristics:
  - There is no empty data in the file
  - A recorded transaction can be either offer received, transaction or offer completed
  - A value can be either an offer id or transaction amount

After a step of exploration and analysis, these data will be combined to create a train and test input data for the model.

## Solution Statement

Our solution is to create a model combining features related to the offers, the customers and the transactions. Then the model will be trained and tested using a binary classification. In a last part we will use it to predict if an offer is appropriate or not to a customer.
Since we already know if an offer has been completed or not thanks to the file *transcript*, we will use a supervised learning model.

## Benchmark Model

Since we are dealing with a binary problem (the customer will use the offer/ the customer will not use the offer), we will use a logistic regression as the benchmark for this model. This model is widely used in machine learning to handle binary classification and has multiple applications (refer to [4]).

## Evaluation Metrics

We are taking the hypothesis that the company is focusing on the precision of the model. So that if an offer has been predicted as appropriate for a customer, this customer would use it.

Consequently we will considerate the precision as the evaluation metric of the model.

# Project Design

We will divide the project into four parts:

- exploring the data
- splitting the data into train/ test sets
- defining and training a binary classifier
- evaluating and comparing model test performance

## Exploring data

We will need to clean, organize and normalize the features from the files in order to have an input suitable for the model.

## Splitting the data into train/ test sets

The input created during the first phase will then be shuffled and split into a train set and a test set.

## Defining and training a binary classifier

The input will be then given to a binary classification model. We will considerate LinearLearner from Sagemaker for this project.

## Evaluating and comparing model test performance

In this last part we will evaluate the model through the precision of the model.

# References

[1]: https://marketingland.com/machine-learning-for-next-best-offers-272374
[2]: https://medium.com/swlh/next-best-offer-when-you-have-few-products-but-lots-of-data-521349035a9d
[3]: https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html
[4]: https://en.wikipedia.org/wiki/Logistic_regression