# Machine Learning Engineer Nanodegree

# Capstone Proposal

Mengjia Lyu June 1, 2020

## Domain Background

Starbucks is the largest coffeehouse chain in the world. Business insights are invaluable when it comes to targeted advertising. In order to maximize the effectiveness of app advertising, we need to be able to predict with reasonable accuracy what kinds of offers are most likely to lead into purchases for different customers. Machine learning algorithms can be employed to learn from transactional data showing user purchases made on the app and make relevant predictions.

## Problem Statement

The problem is to use the transaction, demographic and offer data to predict what types of offers are the most likely to be completed for a given customer. More specifically, my goal is to predict whether or not someone will respond to an offer based on the demographic data available about the customer.

## Datasets and Inputs

The dataset is provided by Starbucks and contains three *json* files:

- portfolio.json
    - contains information on offer, including
        - the specific offer type out of three possible circumstances: BOGO, informational or discount
        - the minimum amount of money required to spend in order to take advantage of the offer
        - the reward given to the customer for completing an offer
        - the shelf time for the offer to be available in days
        - offer id represented by a string
        - the channels where the offers were sent, including mail, mobile, social and web
    - has 10 rows and 6 columns
        - transcript.json
    - contains information on transactions, including
        - the event or the record description out of four scenarios: transaction, offer received, offer viewed or offer completed
        - the customer id represented by a string
        - the time in hours since start of test (t=0)
        - value incurred in the transaction represented by an offer id or transaction amount
    - has 306534 rows and 4 columns
        - profile.json
    - contains information on customers, including
        - the age of the customer
        - the date when customer created an app account
        - the gender of the customer ( 'O' for other, 'M' for male and 'F' for female)
        - customer id represented by a string
        - customer's income
    - has 17000 rows and 6 columns

The files can be accessed in the **data** folder.

## Solution Statement

As whether or not someone will make a purchase is inherently binary, binary classification models can be used to tackle the problem.

First I will preprocess and cleanse the data, then split the datasets into train and test sets. Then I'll train several possible models and pick out the one that has the most predictive power in assessing what a certain demographic group will buy when not receiving any offers.

## Evaluation Metrics

As there is not any related literature in this problem, the final model can be benchmarked against other possible models, especially the relatively *naive* models such logistic regression or naive Bayes classifier.

## Project Design

### Data Preprocessing

First and foremost comes data cleaning.
Since our data contain non-numeric values, I need to convert them into numeric features. I will also check for missing values and outliers in the datasets.

### Model Experimentation & Fine-tuning

As it is a classification problem, available options include logistic regression, decision tree, SVM, naive Bayes, neural network, random forest and AdaBoost/XGBoost.

I can then pick out the best three models by K-fold cross-validation. Using grid search, I can find the optimized hyper-parameters for the selected models.

### Evaluation

The models will be evaluated on the test sets and I will look at metrics including model accuracy, precision and recall, F1 score and AUC(Area under the ROC Curve). The best model will be selected based on the above metrics.

## References


- Starbucks
- Workflow of a Machine Learning Project
- Evaluation Metrics for Machine Learning Models