

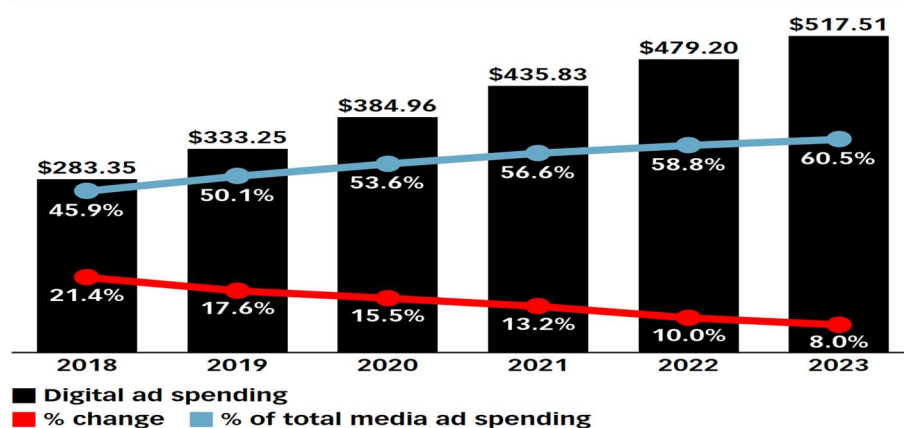
## Capstone proposal – Starbucks

2/5/2020

### Domain Background

Marketing is all about getting 4 things right – reaching out to the right **customer** with the right **product** at the right **time** through the right **channel**. While the change in information consumption has resulted in digital marketing taking over the traditional marketing, Data Science has revolutionized and reshaped the digital advertisement business to meet those objectives.

**Digital Ad Spending Worldwide, 2018-2023**  
billions, % change and % of total media ad spending



Note: includes advertising that appears on desktop and laptop computers as well as mobile phones, tablets and other internet-connected devices, and includes all the various formats of advertising on those platforms; excludes SMS, MMS and P2P messaging-based advertising  
Source: eMarketer, February 2019

T10016

www.eMarketer.com

Reference: <https://www.emarketer.com/content/global-digital-ad-spending-2019>

The use cases range from capitalizing on unexpected insights, boosting contextual relevance to generate response from target audience, defining optimum pricing based on demographics, enhancing ads creativity based on various segments to generate response, and much more.

### Problem Statement

In this project we will explore the data provided by Starbucks. This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. 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). Some users might not receive any offer during certain weeks.

**The goal in this project is to build a model that predicts whether someone will respond to an offer or not.** This should help in defining which offer should be sent to a certain customer triggering more revenue. Also, various demographics groups and offer type relation will be identified.

## The datasets and inputs

The data is contained in 3 files:

- **portfolio.json** - containing offer ids and meta data about each offer (duration, type, etc.)
  - id (string) - offer id
  - offer\_type (string) - type of offer ie BOGO, discount, informational
  - difficulty (int) - minimum required spend to complete an offer
  - reward (int) - reward given for completing an offer
  - duration (int) - time for offer to be open, in days
  - channels (list of strings)
- **profile.json** - demographic data for each customer
  - age (int) - age of the customer
  - became\_member\_on (int) - date when customer created an app account
  - gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
  - id (str) - customer id
  - income (float) - customer's income
- **transcript.json** - records for transactions, offers received, offers viewed, and offers completed
  - event (str) - record description (i.e. transaction, offer received, offer viewed, etc.)
  - person (str) - customer id
  - time (int) - time in hours since start of test. The data begins at time t=0
  - value - (dict of strings) - either an offer id or transaction amount depending on the record

## Solution statement

I will be leveraging dimensionality reduction along with unsupervised learning methods amongst our data to determine potential strategies for adjusting the Starbucks Rewards program given our customer insights. I will be using XGBoost (XGBClassifier) and deep neural network as my primary models.

## Benchmark model

Since there is no proven published method which can be used for this problem set, I will be simpler models used widely like Naïve Bayes and/or K-means clustering as the benchmark model. In addition, threshold value will be defined for comparing across performances obtained by your solution

## Set of evaluation metrics

In order to evaluate performance of my model, I will use using accuracy, precision and recall.

The company would rather **provide an offer to a customer** rather than not providing an offer that he will not opt for. Hence Starbucks may weight more on **recall**, to reduce the number of false negatives and allow the model to produce more of false positives (**lower precision**).

The company might not be too worried about accuracy as such as these offers will not result in any negative experience to the customer. In fact, post deployment, that will provide more inference and model optimization for future.

## **Outline of the project design**

- Project Overview
- Problem Statement
- Define the metrics or calculations you will use to measure performance of a model
- Get the Data and create test set
- Exploratory data analysis
  - Data Visualization
  - Correlation Analysis
  - Algorithms and Techniques intended for solving problem
  - Benchmark or threshold for comparing across performances obtained by the solution.
- Data Pre-processing
  - Data Cleaning
  - Handling Text and Categorical Attributes
  - Feature Scaling
  - Drop data records subject to
  - Data encoding
  - Feature Engineering
- Implementation
  - Split Data into training, validation and test
  - Select and train the model.
  - Fine Tune model set.
- Results
  - Model Evaluation and Validation.
  - Model deployment to a web API endpoint
  - Justification
  - Conclusion
- Reflection
- Improvement