Starbucks Capstone Challenge

Table of Contents

Project Overview

The Starbucks project is coming out from customer marketing domain. Traditional marketing analytics or scoreboards are essential for evaluating the success or failure of organization's past marketing activities. But today's marketers or organizations can leverage advanced marketing techniques like predictive modeling for customer behavior, predictive lead scoring, and all sorts of strategies based on predictive analytics insights.

Various Cases for Predictive Marketing Analytics are:

- Upselling and Cross-Selling Readiness
- Understanding Product Fit
- Optimization of Marketing Campaigns

In this Project, we would deal with the case of 'Optimization of Marketing Campaigns'. Predictive techniques can make an organization marketing investment much more efficient and helps in regularly validating results. Connecting customer information to the operational data provides valuable insight into customer behavior and the health of your overall business.

Refer below links in order to have better understanding of the impact of predictive analytics for better marketing performance.

1. Predictive Analytics: What it is and why it matters
2. Marketing Analytics for Data-Rich Environments
3. How to Use Predictive Analytics for Better Marketing Performance

In this project, we would go through the predictive modeling technique for customer behavior through Starbucks dataset example. Starbucks provided 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. Not all users receive the same offer, and this data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

In this project, Starbucks wants to connect offer data, customer data and transaction data (operational data) to gain insights about customer behavior and overall effectiveness of offers as a value for business.

With the above key statement in mind, we can determine the main objective or problem motivation or problem statement:

Problem statement

Identifying which groups of people are most responsive to each type of offer, and how best to present each type of offer.

Above problem is classification problem to determine the type of offer that each customer will be most responsive to. To solve this problem Starbucks give simulated data about one product

Also we can answer specific questions for example

1 – how much should customer pay in each offer

And here is a description of the data we will use

Dataset

- 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.
- 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.
- The basic task is to use the data to identify which groups of people are most responsive to each type of offer, and how best to present each type of offer.

Data Dictionary

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)

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)

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

Solution Statement

We will approach this problem using supervised learning technique (classification techniques) to merged portfolio and transcript data after data cleaning and statistical analysis.

Steps will thoroughly explained in details in project design section.

Benchmark models

We will compare our results to other similar projects to compare performance of model of determining the most suitable offer for a customer to other models
And here is work of others on the same problem and the same data to compare our results to
1 -https://towardsdatascience.com/starbucks-app-sending-the-right-offer-to-the-right-user-9d8d0e24e65c

2- https://towardsdatascience.com/using-starbucks-app-user-data-to-predict-effective-offers-20b799f3a6d5

Evaluation metrics

We will use F1 score as harmonic combination of precision and recall.
For this project recall is more important than precision as we want to send offer for customers that has an adequate chance to complete the offer not I must be very sure that

Project Design

We will follow these steps to solve the problem in the problem statement

1- Loading the 3 data files
2- Assessing data quality
3- Cleaning the data for instance removing duplicates, missing values, outlies, and resolving any inconsistencies in feature naming or representation of values

4- Statistical analysis of the 3 files
4.1- Profile data
4.1.1- Histogram of age , income
4.1.2- Bar plot of gender
4.1.3- Boxplot for income by gender

4.2- Transcript
4.2.1- Bar plot for number of transactions per offer type
4.2.2- Double bar plot for number of transactions per offer type per channel

5- Preparing data for classification models
6- Trying different classification algorithms Decision tree SVM and logistic regression

7- Evaluating the models on precision, recall and F1-score