UDACITY

Machine Learning Engineer Nanodegree

Capstone Project Proposal

I. Domain Background

Starbucks is a multinational chain with over 32,000 coffeehouses worldwide and 20 million members in its loyalty program. With over 10 million downloads and 600,000 ratings on the Android Google Play store (3.8 million ratings on the Apple store), the Starbucks mobile app makes it so convenient for customers to frequently order their favorites or try the newest seasonals.

To increase its profits, the marketing and data science teams would leverage data about its customers, such as their transactions and demographics, to create personalized and timed promotions that would incentivize customers to spend more or visit more often.

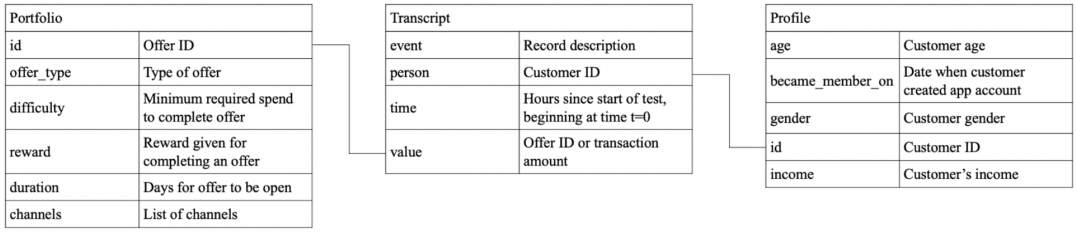
II. Problem Statement

Using data from the company's digital channels, the capstone project's objective is to use historical data to determine the best offer to send to the customer. Intuitively, the model should generate the probability of a customer responding to each offer, and the final outcome should be the offer with the highest, positive return on investment.

III. Datasets and Inputs

The data contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. The typical flow of a transaction through the rewards app would be receiving the offer, viewing the offer, making a transaction, and finally completing the offer. Customers can make transactions without seeing or using the offers. However, in order for a promotion to have a positive ROI, the customer must have seen the offer first. This assumes that the customer was influenced to make a purchase.

The simulated data contains 3 JSON files on offers, customer demographics, and transaction history that can be transformed into the ERD below.



Portfolio	
id	Offer ID
offer_type	Type of offer
difficulty	Minimum required spend to complete offer
reward	Reward given for completing an offer
duration	Days for offer to be open
channels	List of channels

Transcript	
event	Record description
person	Customer ID
time	Hours since start of test, beginning at time t=0
value	Offer ID or transaction amount

Profile	
age	Customer age
became_member_on	Date when customer created app account
gender	Customer gender
id	Customer ID
income	Customer's income

For simplicity, there are only 10 offers in the portfolio, but in reality, there are hundreds over the years. The data is simulated for 30 days, leading to 306,534 events captured with 17,000 customers. Sample data is below.

`portfolio: 10 records`

	reward	channels	difficulty	duration	offer_type	id
0	10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd
1	10	[web, email, mobile, social]	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0
2	0	[web, email, mobile]	0	4	informational	3f207df678b143eea3cee63160fa8bed

`profile: 17000 records`

	gender	age	id	became_member_on	income
0	None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	None	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN

`transcript: 306534 records`

	person	event	value	time
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	0
1	a03223e636434f42ac4c3df47e8bac43	offer received	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}	0
2	e2127556f4f64592b11af22de27a7932	offer received	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}	0

IV. Solution Statement

The return on investment (ROI) of these promotions would be equal to the incremental revenue generated by the customer minus the cost of reward (from BOGO or discounts). For simplification, the overhead costs to run the campaign and create the model are negligible. If the ROI is positive then the company is gaining profit.

Based on the model results, the final outcome should pick the offer based on the highest expected value, which is the probability the customer would use the offer multiplied by the corresponding incremental margin. With this method, the model could recommend an offer with a lower success rate but higher margin, which averages out with a higher ROI.

V. Benchmark Model

Since there is no existing model used to determine the right offer for each customer, a simple benchmark model should be created first. The project will use K-nearest neighbors, being a quick and popular multi-classification approach, and use its accuracy score as the baseline.

VI. Evaluation Metrics

The model performance will be measured against test data for accuracy, precision, and recall. If all marketing promotions generate more incremental revenue than reward costs and campaign operational costs, then it's better to have higher recall or fewer false negatives. In order words, the number of promotions given is based on the quantity rather than the quality of customers responding.

VII. Project Design

The project is broken into 3 steps: preprocessing, modeling, and evaluating.

1. Preprocessing

Exploratory analysis is done alongside preprocessing, as discovering simple trends or correlations could lead to creating or deleting features for the model. Besides exploring and cleaning the data, joining the 3 datasets into 1 flattened table is required. Finally, the cleaned data is split 80/20 (as a rule of thumb) into training and test data.

2. Modeling

Once the training and test data are preprocessed, they will be used for model selection and hyperparameter tuning. This could be a classification problem, since this situation only has 11 distinct choices (10 offers and "do nothing"). The model outcome would be a score for each choice, or probability of choosing each offer.

An additional consideration for model selection is that marketing stakeholders would want to know how groups of customers respond differently. Therefore, stakeholders would likely choose explainable models over "black box" models, such as ensemble models.

3. Evaluating

The model outcomes with test data will be compared against the benchmark. If the model increases the response rate of customers, then the proposed solution should be used to improve marketing campaigns.