Proposal

Evaluating Customer Engagement Programs at Starbucks

Domain Background

Many companies use incentive programs to increase sales, raise brand awareness, and encourage customer loyalty.
The design of the program often depends on the type of customer it is targeting – new (coupons), infrequent (discounts), regular (rewards proportional to purchase value, eg. Amtrak Rewards) , frequent (membership-based programs, eg. Amazon Prime). The type of offer presented also depends on the product being sold (single product, limited variety, large selection)

Problem Statement

Instead of sending all app users the same offer, we would like to target our customers carefully – by sending them offers that they are likely to respond to, else they could stop using the app. Also, some users are regulars – and they will spend anyway. They should not receive discounts.

We need to predict the offers a person should receive based on age, gender, income, and purchasing history. This prediction is quantifiable, and measurable, because it will include an expected spend for the duration of the offer. This can be compared to the true spend.

Datasets and Inputs

The data comes from usage of the Starbucks app. Three datasets are available:

1. Portfolio:

    The Portfolio table has 10 rows corresponding to 10 types of offers. The fields are:

- id (string) - offer Id
- offer_type (string) - BOGO, discount, informational (just an ad, no value)
- 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) – web, email, mobile, social

For example – a discount type offer, where, if $20 (difficulty) is spent within 10 days (duration), a reward of $5 will be given. It will be sent out by web or email channels, but not be mobile or social networks.

2. Profile:

In this table, each row corresponds to a customer and his or her demographic information – age, gender and income. There are 17,000 customers. Complete records are available for 14,825 individuals. In this project, these variables – age, gender and income - will be used to predict the offer a customer should receive.

3. Transcript:

This table contains events for all customers over a period of 30 days (720 hours). An event can be

i. a transaction, where a customer makes a purchase in dollars
ii. an offer event – offer received, offer viewed, or offer completed
The other field is "time" of the event, described above, in hours since the time of the 30-day program

Solution Statement

The model must predict the offer a customer should receive from a set of 10 offers. This is a multi-class prediction. The inputs – age, gender, and income, are structured data. So a boosted Decision Tree or Random Forest will work well. These models are explainable.
The model will be trained and deployed using Amazon SageMaker.

Benchmark Model

Standard benchmark models are focused on the overall performance of a loyalty program. Typical measures are:

- Rewards Redemption
- Loyalty Program Penetration
- Active Participation Quotient (APQ)
- Percentage of Earning Members
- Program Impact on Motivation to Spend
- Program Appeal at Enrollment

Evaluation Metrics

We will compare predicted offers against actual offers for the test cases. We will estimate the accuracy of completion of the offer, and compare spend during the offer duration.

Project Design

In this project, we're predicting offers that an individual will like. Because the customer will receive many (3-6) offers during the 30-day study period, the number of records will be a small multiple of the number of customers (17,000). The Transcript table will have to be joined on customer id, then re-shaped, so each row corresponds to a single offer period for each customer.

Then a simple multinomial logistic model will be applied, to get a sense of the model parameters. XgBoost in SageMaker or AdaBoost in Scikit Learn (decision tree models) will then be used, because they tend to perform better.