Domain & Background

Starbucks, as a marketing strategy, provides discounts or free services on its offerings to users. This helps in increasing sales and creating customer awareness of new offering. Starbucks has a diverse user base. Hence not all offers are rolled out to each customer. Starbucks gives different offers to different customers based on their demographics and user profile. In a simple words Starbucks' offering are personalised. Business lean towards propensity model to reduce the expenditure in advertising, as model give them more specific target customer.

https://datascience.foundation/sciencewhitepaper/propensity-modelling-for-business

Problem Statement

Since we have varied groups of users with varied preference or liking, we need to be smart in rolling out offers to them. Based on the historical data that is available to Starbucks, we can make a data backed decision whether to rollout the offer will result in a successful business value.

So, in this project we will try to build an ML model which predicts whether a user or group of users will respond to an advertisement or offer (propensity modelling). With this known beforehand with some accuracy we can be extremely specific while we do advertisement targeting.

Dataset and Inputs

The data is divided in three files:

Each file's attributes are listed as below.

portfolio.json — contains details of offer [10 records 6 attributes]

- id (string) — offer id
- offer_type (string) — Offer Type (discount, buy one get one, informational)
- difficulty (int) — minimum spend to complete the offer
- reward (int) — reward offered for completing the offer
- duration (int) — time, in days, for the offer is valid
- channels (list of strings)

profile.json — demographics of each customer [17000 records, 5 attributes]

- age (int) — age of the customer
- became_member_on (int) — date when customer became member
- gender (str) — gender of the customer male(M), female(F), others(O)
- id (str) — customer id
- income (float) — member's income

transcript.json — customer response to the offer [306534 records, 4 attributes]

- event (str) — event type like transaction, offer received, offer viewed
- person (str) — customer id
- time (int) — time in hours since the start of the test. The data begins at time t=0
- value — (dict of strings) — either an offer id or transaction amount depending on the record

Of the ~76K offers around 33k resulted in completion, which is roughly 43%. So there is marginal imbalance. But that should not be much of an issue.

Solution Statement

The task is to combine transaction, demographic, and offer data to determine which demographic groups respond best to which offer type.

Evaluation Metrics

Since the model will predict success or failure of offer consumption, it's a classification and hence we will use below metrics(F2 Score) for evaluation:

- Precision
- Recall
- F1 score
- **F2 score – Since there is class imbalance**
- **AUC**

Benchmark Model

- Logistic Regression

We will be creating a logistic regression as our benchmark model. Its accuracy is ~80%. We will try to outperform this with our final model.

Project Design

EDA on each file

Data distribution

Combine three datasets into one

EDA on aggregate data

- Target variable bias

Data cleaning

- Remove null or outliers if any
- Handling missing values Imputation vs Drop
- De duplication

Feature engineering

- One hot encoding for categorical fields
- Binning of age and income into groups
- Derived features
- Scaling the numerical columns

Split the data into test and train

Train the model on various algorithms

- Decision Tree
- Random Forest

- Xgboost
- Adaboost
- KNeighbors

Hyperparameter tuning of best model

Get parameters for the best model and find important feature

Evaluate performance of the model