Predicting the Effectiveness of Starbucks Offers Based on Demographic Groups

---

Domain Background

Starbucks, one of the largest coffeehouse chains globally, has a rewards mobile app that sends out offers to its customers periodically. These offers can vary from advertisements for a drink to discounts or buy-one-get-one-free (BOGO) deals. However, not all customers receive the same offers, and the challenge lies in determining which demographic groups respond best to each offer type. Understanding this relationship can help Starbucks tailor their offers and maximize customer satisfaction and sales. Predicting human behavior is one of the most challenging tasks, but if successful, it can yield tremendous benefits for a company's sales numbers. Therefore, there is a continuous need to improve algorithms and predictions to obtain accurate and reliable results, helping businesses make more informed decisions in their marketing strategies.

Problem Statement

The problem to be solved is determining which demographic groups respond best to different types of Starbucks offers. This involves analyzing customer transaction, demographic, and offer data to identify patterns and relationships. By solving this problem, Starbucks can improve its targeted marketing strategies and increase the effectiveness of its offers.

Solution Statement

A machine learning model will be developed to predict how customers respond to various offers based on their demographic information. This model will help identify which offers are most effective for specific demographic groups, allowing Starbucks to tailor its promotional strategies accordingly.

Datasets and Inputs

The data for this project consists of three JSON files:

1. portfolio.json: Contains offer IDs and metadata about each offer (duration, type, difficulty, reward, and channels).
2. profile.json: Contains demographic data for each customer (age, gender, income, and membership date).
3. transcript.json: Contains records for transactions, offers received, offers viewed, and offers completed.

The portfolio.json contains information about the offer types. There are three types of offers that can be sent:

1. buy-one-get-one (BOGO): if the recipient spends a certain amount the recipient gets a reward of equal amount
2. discount: the recipient receives a reward of a percentage of the amount spent
3. and informational: the recipient receives information about a product. There is no reward involved

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

These datasets will be used to analyze customer responses to different offers and the impact of demographic factors on their behavior.

Benchmark Model

The benchmark model for this project will be a simple classification model (e.g., logistic regression) that predicts customer responses to different offers based on their demographic information. This model will serve as a baseline to compare the performance of more complex models developed during the project.

Evaluation Metrics

The evaluation metrics for this project will be accuracy, precision, recall, and F1 score. These metrics will be used to assess the performance of both the benchmark model and the solution model in predicting customer responses to different offers.

Project Design

The project will follow these steps:

1. Data Cleaning: Clean the data by handling missing values, removing duplicates, and ensuring data consistency.
2. Data Exploration: Explore the data to understand the relationships between demographic factors, offers, and customer responses.
3. Feature Engineering: Create new features or transform existing ones to improve the predictive power of the model.
4. Model Selection: Test various machine learning algorithms (e.g., logistic regression, decision trees, random forests, and gradient boosting) and select the best performing model based on evaluation metrics.

5. Model Evaluation: Evaluate the selected model's performance using cross-validation and compare it to the benchmark model.
6. Model Interpretation: Analyze the model's results to identify the most important features and draw insights about the relationships between demographic factors and offer effectiveness.
7. Conclusion: Summarize the findings and discuss potential improvements and future work.