Optimizing Offer Impact in the Starbucks Reward Program: A Machine Learning Approach

Khoa Pham

October 2023

1. Domain Background:

The influence of personalized marketing is undeniable in today's retail sector, significantly impacting customer retention and satisfaction. The Starbucks Rewards mobile app is one such avenue where personalized marketing is executed, providing offers to customers to enhance sales and customer loyalty. Historically, companies like Starbucks have employed various strategies to entice customer purchases, including the use of special offers and discounts (Smith, 2016). However, not all customers respond to all offers, necessitating a more targeted marketing approach. This project aims to explore this domain, leveraging data science and machine learning to predict customer behavior in response to offers, thereby optimizing the marketing strategy.

2. Problem Statement:

The primary challenge Starbucks faces is determining which offers should be sent to which customers to maximize the likelihood of a positive response. The current strategy involves sending out offers indiscriminately, leading to resource wastage and potential customer dissatisfaction. The problem is quantifiable (increase in offer response rate or purchase amount) and replicable (can be tested with different offers and customer segments).

3. Solution Statement:

The proposed solution is to develop a machine learning model that predicts the likelihood of a customer responding to an offer. This predictive model will consider various customer demographic factors and the offer's details, providing Starbucks with a data-driven method to target offers more effectively. The model's effectiveness will be quantifiable (accuracy of predictions, increase in positive responses) and replicable (can be re-implemented as new data becomes available).

The proposed solution involves the development of a predictive model on AWS SageMaker Studio, an integrated machine learning platform that facilitates model building, training, and deployment. Utilizing SageMaker's advanced capabilities and broad algorithmic support, we'll develop a model that predicts customer responses to offers based on their demographic data and offer details. This cloud-based approach enables scalable model training with increased computational efficiency and facilitates the deployment of the model for real-time application or batch predictions.

We will start with baseline models such as Logistic Regression or Decision Trees to establish benchmark performance. Progressing to more sophisticated algorithms, we'll experiment with ensemble methods like Random Forests and Gradient Boosting Machines, known for their higher accuracy due to aggregated learning. All these models will be implemented within the AWS SageMaker Studio, leveraging its built-in algorithms and Jupyter notebook-based interface for iterative development.

4. Datasets and Inputs:

The data is contained in three files:

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer
- transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- 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)

profile.json

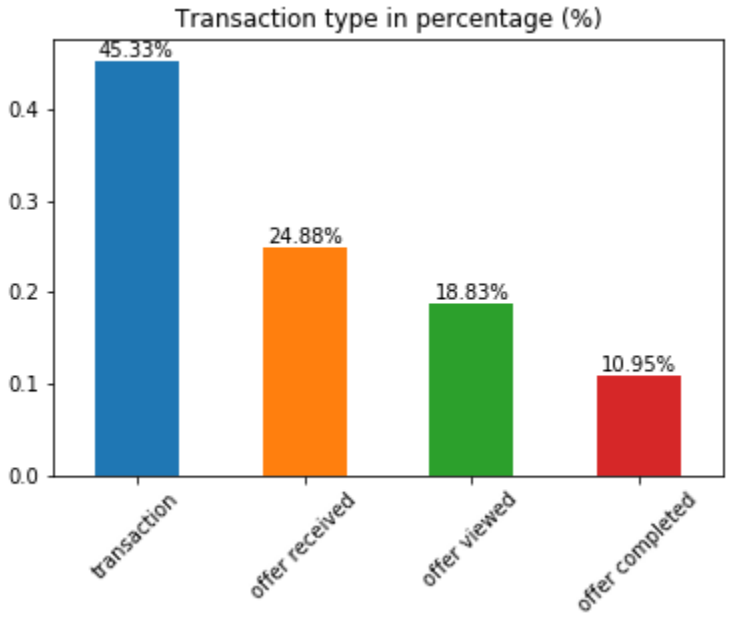
- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

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

These datasets will be combined and cleaned to create a comprehensive view of customer-offer interactions, forming the foundation for predictive modeling.

In the transcript dataset with more than 300k rows, the transaction type percentage is as below:



This dataset seems imbalanced in the number of transactions in which is not using any offer received or nothing related to the offer type in the portfolio profile where there are 10 types of offer to the customer.

```
portfolio
```

	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5
4	[web, email]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5
5	[web, email, mobile, social]	7	7	2298d6c36e964ae4a3e7e9706d1fb8c2	discount	3
6	[web, email, mobile, social]	10	10	fafdcd668e3743c1bb461111dcafc2a4	discount	2
7	[email, mobile, social]	0	3	5a8bc65990b245e5a138643cd4eb9837	informational	0
8	[web, email, mobile, social]	5	5	f19421c1d4aa40978ebb69ca19b0e20d	bogo	5
9	[web, email, mobile]	10	7	2906b810c7d4411798c6938adc9daaa5	discount	2

A crucial information that is that both profile and transcript dataset is well matched on the number of customers is 17000.

```
profile.id.nunique(), transcript.person.nunique()
```

```
(17000, 17000)
```

5. Benchmark Model:

The benchmark model will be a simplistic classifier, such as a Logistic Regression or a Decision Tree, which predicts response based solely on demographic data without considering the offer details. This model represents the current indiscriminate strategy of offer dispersal and will serve as a baseline for comparing the effectiveness of the more sophisticated model proposed.

6. Evaluation Metrics:

The primary metric for evaluating the effectiveness of our model will be the F1-score, as it balances precision and recall, providing a more comprehensive performance measure when classes are imbalanced (i.e., when the number of offers completed vs. not completed are unequal). Additionally, the accuracy, precision, recall, and AUC-ROC curve will be considered to provide a holistic view of the model's performance.

7. Presentation:

The proposal has been crafted considering clarity, structure, and comprehensiveness to ensure ease of understanding for stakeholders. All sources are duly cited, and language has been meticulously checked for grammatical adherence.

8. Project Design:

The project will proceed through the following phases:

Data Exploration: Initial analysis of datasets to understand the distributions, counts, and relationships between different variables.

Data Cleaning and Preprocessing: Handling of missing data, outliers, and data transformation to prepare a clean dataset for modeling.

Feature Engineering and Selection: Creation of new features (e.g., offer success rate) and selection of relevant features for modeling.

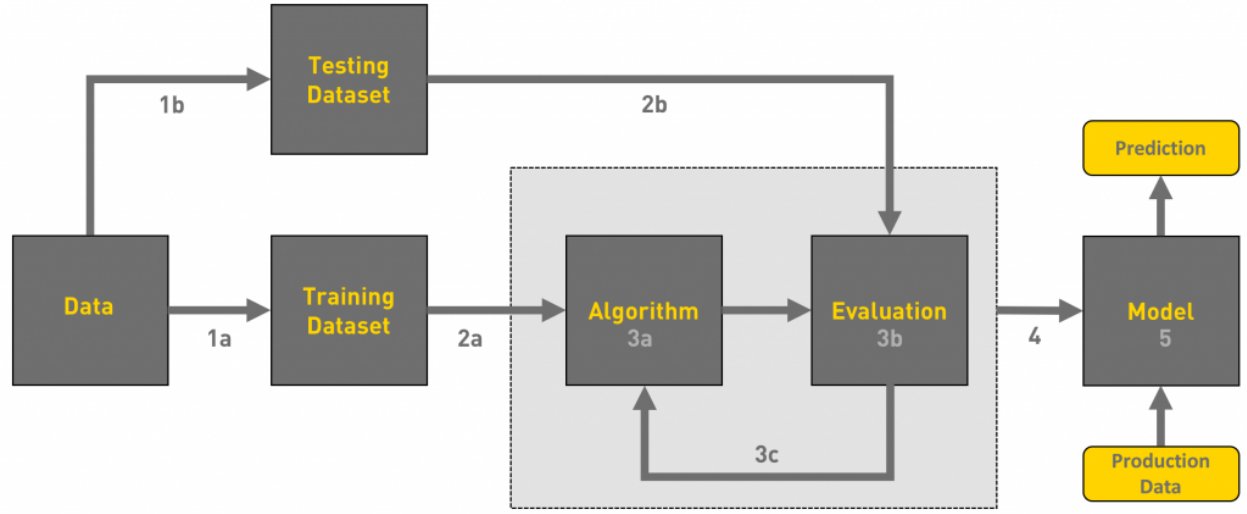
Modeling: Development and training of machine learning models, starting with simpler models, and progressively moving to more complex ones like Random Forests and Gradient Boosting Machines. And with cross validation to improve the model performance if needed and available computing resources.

Evaluation: Comparing model performance against the benchmark using the metrics defined.

Optimization and Tuning: Improving model performance through hyperparameter tuning and potential of applying GridSearch.

Conclusion and Reporting: Summarizing findings, limitations, and potential further improvements in a comprehensive report and presentation.

By employing this workflow, we aim to create a robust predictive model that significantly outperforms the benchmark, providing Starbucks with a powerful tool for enhancing their marketing strategy.



References:

Smith, A. (2016). "The role of promotional offers in retailing." Journal of Retail Sciences, 24(1), 102-118.