CAPSTONE PROJECT PROPOSAL: STARBUCKS PROJECT

DOMAIN BACKGROUND

Starbucks, one of the largest global coffeehouse chains, relies heavily on customer engagement strategies to maintain its competitive advantage. A key aspect of their marketing involves the use of personalized promotional offers(Jacinta Dias, 2024) to attract and retain customers. These offers—such as discounts and Buy-One-Get-One (BOGO) deals—are typically delivered through the Starbucks Rewards program and digital channels. However, predicting whether a customer will complete an offer remains a challenge due to variations in customer demographics and offer types.

Personalized marketing has become increasingly relevant in the past decade, as companies recognize the value of tailoring offers to specific customer segments.(Yixuan Zhu, 2024) For Starbucks, understanding how customer attributes such as age, income, and gender influence the likelihood of completing an offer can optimize resource allocation and improve campaign effectiveness. Previous research has highlighted the importance of customer segmentation and behavioral analysis in predicting responses to promotions, yet there is still a gap in accurately modeling these interactions in real-world scenarios.

This project aims to build a machine learning model that predicts the probability of a customer completing a Starbucks offer based on demographic profiles and offer characteristics. By addressing this problem, the model can provide actionable insights to enhance personalization, increase offer completion rates, and ultimately drive customer satisfaction and loyalty.

PROBLEM STATEMENT

The primary problem is determining which factors influence whether a customer will complete a Starbucks promotional offer. Despite the abundance of customer demographic data and detailed information on the nature of the offer, it isn’t always clear which combinations lead to higher offer completion rates. This confusion results in inefficient marketing strategies and missed opportunities for optimizing customer engagement. A potential solution is to build an ML model that predicts the likelihood of offer completion based on customer demographics and characteristics. The success of the solution can be measured using an AREA UNDER THE CURVE score.

DATASETS AND INPUTS

I was given three datasets to complete this project. All three datasets are json files. The first one is called “portfolio.json”. Each record of this file contains details about each type of offer. For each offer, the following information is given: the channels in which the offers were given, the difficulty (how much needed to spend) of each offer, the duration of each offer, the offer id, the offer type and the reward. The second dataset is named “profile.json”. Each record of this dataset contains demographic information of a particular customer. For each customer the following is collected: customer age, when the customer became a member, customer gender, customer id and income. The third dataset is named “transcript.json”. This dataset records every transaction from the moment data collection starts. Each transaction is described by 4 features. The first is the event. This states whether an offer was received, viewed, or completed. It can also state if a normal transaction took place. The next feature is the customer id. The last two features are time and value. “Time” denotes how many hours since data collection started that the transaction occurred. “Value” denotes the offer_id if the “event” feature has something to do with offer. The datasets are not actual customer data. Instead, Starbucks made simulations mimicking their actual customer data. These three datasets will be wrangled and joined together to form one dataset. Some information will be removed e.g informational offers will be removed because you don’t respond to these offers. Also transactions that don’t have anything to do with offers will be excluded. Only offers that were viewed will be dealt with. Together with the type of offer and demographics, the data will train the model to make predictions.

SOLUTION STATEMENT

The entire end-to-end solution to the problem will be solved using AWS Sagemaker. Sagemaker Studio notebooks will be instantiated and used to preprocess the data and send to S3 buckets. To solve the problem of predicting whether a customer will complete a Starbucks promotional offer, a machine learning classification model, such a XGBoost, will be developed. This model will be created using a Sagemaker training job. The model will use customer demographic data (e.g., age, income, gender) and offer characteristics (e.g., offer type, duration) as input features. The model’s performance will be evaluated using AUC to measure its effectiveness in distinguishing between customers who complete offers and those who don’t. By accurately predicting offer completion, the solution can help optimize marketing strategies and improve customer engagement.

BENCHMARK MODEL

For this project, a random classifier is used as a benchmark model. A random classifier is expected to give an AUC OF 0.5 which will serve as the baseline for comparison. The proposed ML model must achieve an AUC significantly higher than 0.5 to demonstrate its effectiveness in predicting offer completion.

EVALUATION METRICS

The primary evaluation metric used in this project is the Area Under the ROC Curve (AUC), which measures the model's ability to distinguish between customers who complete offers and those who do not. AUC is chosen because it is robust to class imbalance and provides a comprehensive measure of model performance. The benchmark model is expected to have an AUC of 0.5, while the solution model must exceed this to demonstrate effectiveness.

PROJECT DESIGN

The solution workflow begins with preprocessing and cleaning the data, where three separate datasets will be joined to create a unified dataset. Feature engineering will be applied to create new variables that better capture customer behavior and offer characteristics. Next, the dataset will be used to train a model using SageMaker's XGBoost algorithm. Hyperparameter tuning will be performed to identify the best model configuration. Once the optimal model is selected, it will be tested on a separate test dataset to evaluate its performance using AUC as the primary metric. Finally, further analysis will be conducted to interpret the model's results and gain insights into which factors influence offer completion.

REFERENCES

1. Jacinta Dias (2024). 6 Successful Marketing Strategies of Starbucks. Retrieved from https://iimskills.com/marketing-strategy-of-starbucks/
2. Yixuan Zhu (2024). Leveraging Social Media Marketing: A Case Study of Starbucks Digital Success. Retrieved from https://www.researchgate.net/publication/382586535_Leveraging_Social_Media_Marketing_A_Case_Study_of_Starbucks_Digital_Success