AWS Machine Learning Engineer Capstone Proposal

Hieu Nguyen Minh

August 2024

1 Domain Background

This project stems from Starbucks' direct marketing system, which maintains customer engagement by sending personalized offers through various channels like email, social media, the web, or its app. These offers include buy-one-get-one (BOGO) deals, discounts, and informational messages, each designed to incentivize purchases or provide product information. However, in order for marketing campaigns to be successful, they must generate profits. Companies need to carefully target customers likely to respond to offers while attracting new consumers and rewarding loyal ones without unnecessarily cutting into profits. Some customers only respond to rewards, while others dislike marketing entirely, highlighting the complexity of marketing decisions. With the rise of machine learning and the availability of large datasets, intelligent systems can enhance marketing campaigns by analyzing consumer behavior patterns.

2 Problem Statement

Starbucks invests in marketing campaigns with the goal of generating profits, making it crucial to deliver the most relevant offers to the right customers. However, some customers never see the offers, which may indicate an issue with the chosen communication channel, while others see the offers but don't make purchases, suggesting a mismatch in offer type or targeting. On the other hand, some customers engage with the offer, trying new products or spending more, which is the desired outcome. This project aims to solve the problem of identifying the most appropriate offer for each customer, which is the one that leads to a purchase influenced by the offer. If a customer doesn't see the offer or doesn't act on it, or if they make a purchase without being influenced by the offer, it is not considered effective.

3 Datasets and Inputs

The dataset used in this project is provided by Udacity and Starbucks, consisting of simulated data that replicates customer behavior on the Starbucks rewards mobile app. The data generation program models how individuals make purchasing decisions and how promotional offers affect those decisions. Each simulated person has hidden traits that impact their buying patterns and are linked to their visible characteristics. The simulation tracks events such as receiving, opening, and responding to offers, as well as making purchases. However, the dataset does not track specific products, only the transaction or offer amounts.

Data dictionary

1. `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)
2. `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)
3. `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

4 Solution Statement

To address the problem, this project proposes using machine learning techniques to analyze customer behavior based on their interactions with Starbucks. Specifically, a neural network will be trained to predict how customers will respond to different offers, determining whether they will complete the offer cycle. Considering that consumer behavior is influenced by past experiences, the project will employ a Recurrent Neural Network (RNN) to account for time-dependency in decision-making.

5 Benchmark Model

A Feedforward Neural Network (FNN) will be trained on the same dataset as the RNN to enable comparison of their results. While an FNN analyzes static input without considering customer history, an RNN can make predictions based on past events. A conventional FNN model might repeatedly suggest an offer that once yielded good results, without recognizing when it becomes irrelevant, such as when a customer no longer wants to make the same purchase or is already bought without incentives. The RNN, on the other hand, is designed to capture the relationship between past experiences and future behavior.

6 Evaluation Metrics

The accuracy of will be used to evaluate the performance of both the FNN and RNN.

$$accuracy = \frac{TP}{N}, \tag{1}$$

where TP is the number of correctly classified samples, and N is the total number of samples.

7 Project Design

I follow the guideline to design the project workflow.

1. Exploratory Data Analysis (EDA) Read data files and produce data visualization to understand the data distribution and characteristics.

2. Data cleaning and engineering Process the data, like filling the N/A values, remove duplicates, and create new features from existing features; Prepare the data to be ready to feed the neural networks; Label the record as appropriate offer or not.

3. Split the dataset into training, validation, and test sets. The training set is for training the networks, while the validation set is for evaluating the models during the training phase. The test set contains data never seen before by the network, so that we can evaluate how well the model performs on new data.

4. Build and train the FNN and RNN.

5. Evaluate and compare model performances