Udacity - Machine Learning Engineer Nanodegree

Starbucks Capstone Project Proposal

Amita Pandey
Oct 24 2020

Domain Background

Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. As the world's largest coffeehouse chain, Starbucks is seen to be the main representation of the United States' second wave of coffee culture.

Starbucks Mobile **App** Now Counts 16.8M **Users**.

This phone app for Starbucks reward program sends personalized offers to its customers. These offers can be in the form of buy 1 get 1 free (BOGO), an advertisement for a drink or an actual offer. Not all users receive the same offers and some users might not receive any offers during certain weeks.

The goal for this project is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type.

Thereby customizing such promotional offers for customers based on their previous responses.

Problem Statement

The goal of this project is to **Build a model that predicts how the customer will respond to an offer based on their response to previously sent offers.**

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 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

Solution Statement

To solve this problem we will be following the below process

1. Fetch Data : Fetch the given data in csv to utilize in the next steps.
2. Clean Data : Remove duplicates, correct errors, handle missing value cases, normalize and type conversion.
3. Data Visualization and analysis : Explore the given data further to find relationships between different variables
4. Model Training : Train the model
5. Evaluate the model : Measure the model's performance using some metric. Hyperparameter tuning and finally use the test data to check the output

During Data Visualization and analysis I plan to explore what offer was most responded to, response to an offer, age and gender which are greatly interested in the offers.

RandomForestClassifier and DecisionTreeClassifier models could be good ones to determine which model best represents the given data.

Benchmark Model

KNeighborsClassifier is a fast and standard method for binary classification machine learning problems.

I plan to use KNeighborsClassifier to build a benchmark, and evaluate the model result using F1 score as an evaluation matrix.

Evaluation Metrics

I plan on using F1 score as the metric to evaluate the quality of the approach. F1 score can be interpreted as the weighted average of the precision and recall. The traditional or balanced F-score is the harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

Project Design

I plan on following the below steps in completing this project.

1. Establishing workspace in AWS Sagemaker
2. Fetching data
3. Cleaning data as required for modeling purposes
4. Exploring the data to understand it better.
5. Building and exploring different models to determine the most suitable one for the data.
6. Using benchmark model and F1 score metric to ensure sanity
7. Summarize the findings and project work in a blog post