Starbucks Capstone Project Proposal

By

Oluwanifesimi Ademoye

Udacity Machine Learning Nanodegree Program

Domain Background

The Starbucks Capstone Project intends to use an informed analysis of customer data to know more about the customer's behaviour with the data and metadata gleaned from transactions .The project uses unsupervised machine learning algorithms to get insights from the data in order to send personalised offers to their customers and also to find out more about the behaviour of certain groups their customers are segmented into.

Problem statement

This project aims at identifying which groups of people/individuals are most responsive to each type of offer, and how best to present such offer.

The goal of the project is to analyze historical data about the customer's previous transaction and offers previously made to the customer to develop a model to present the best package to each customer.
Starbucks would like to know which groups of people are most responsive to each of type of offers, and how best to present the types of offer

Datasets and Inputs

All the datasets used for this project are provided by Starbucks. There are three datasets, portfolio.json that contains offer ids and metadata about each offer (duration, type, etc.), profile.json that contains demographic data for each customer and transcript.json that contains records for transactions, offers received, offers viewed, and offers completed. It's in the JSON format, we will need to transform it into DataFrame for further analysis.
The datasets used to model, train and validate are:

1. Profile.json: This contains demographic data about the rewards program users. There are 17000 users and 5 fields which include:
    - 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: This contains the offers sent during the 30-days simulation period. There are 10 offers and each has 6 fields which are:
    - 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: This is the transactional data that shows the events such as when a user views, receives or completes an offer. It contains 306648 events and 4 fields which are:
    - 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 the start of test

Solution statement

The purpose of this project is to build a machine learning model to determine the best type of offer to send to each customer based on their behaviour.
The solution will be divided into 4 sub-parts:

1. Data Pre-processing – This will involve removing or processing any missing values or outliers present in the data.

2. Data Exploration – This involves a deep dive into data which shows the underlying existing relationships in the data, its distrubition and bias . We will also find interdependencies within the data.

3. Building the model – We will build the machine learning model on the understanding of the above data and since we were provided no ground truth labels for the data , I will be making use of unsupervised learning algorithms to see what clusters the customers fall into to make more inference about their specific behaviours

4. Analysis of their Groupings – This step will involve making informed analysis about the customers groupings or clusters they fall into.

Benchmark Model

For this project I will be using insights I gleaned from this Medium article which showed how to use the k-means algorithm to get insights from clusters from customers data. I will use the model employed in this article as a baseline for the model I will employ in the project.

`

Evaluation Metrics

Contrary to supervised learning where we have the ground truth to evaluate the model's performance, k-means does not have such evaluation metric, therefore most times people levelrage on domain knowledge and intuition if available. In this project the metric used will be:

- The Silhouette Score The Silhouette score would be used as an evaluation metric to determine how well the points were assigned to clusters. It returns a value between -1 and +1; the best value is +1 and the worst is -1. Values close to 0 indicate overlapping clusters. It can also be used to select the optimum number of clusters K to be used.

- The Elbow Method: Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow. We'll use the geyser dataset and evaluate SSE for different values of k and see where the curve might form an elbow and flatten out.

Project Design

The project workflow is as follows:

1. Data Loading and Exploration
2. Data Cleaning and Preprocessing
3. Dimensionality reduction
4. Feature engineering and data transformation
5. Clustering transformed data with K-means or an alternative of GMM or DBScan
6. Extracting trained model attributes and visualizing k clusters
7. Summarizing our findings and work in a detailed blog post