Machine Learning Engineer Nanodegree
Starbucks Capstone Project

by Arwa Alamoudi

November 27, 2021

1 Domain Background

In 2020, there were 32,660 Starbucks stores worldwide and the number is slightly increased in 2021 despite the coronavirus pandemic [Jer21]. Also, a study was conducted to understand the popularity of restaurants loyalty applications. It found that Starbucks has the most regularly used loyalty rewards application amongst other major restaurants. This is due to the fact that Starbucks allows its customers to actively engage with the application and thus receive rewards [Pan18] [Man18].

With this active engagement, it would be very beneficial to leverage the data produced by Starbucks' digital customers to better understand their behaviour and thus provide the needed services.

2 Problem Statement

In this project, we want to know which groups of people are influenced by each type of offer. So, the goal is to better understand the customers hence send appropriate offers that increase Starbucks profits. We will focus on the customers who really get influenced by the offers. So, we send them appropriate offers.

By influence we mean the customers who completely respond to the offers, that is, receive the offer, view it, process it, and complete it. The following are the offers proposed by Starbucks in its reward application: BOGO (Buy One Get One Free), discount, and informational.
This is a classification problem where the input is a type of customer and the output is appropriate type of offer so Starbucks marketers can better understand their customers.

3 Datasets and Inputs

3.1 Dataset Overview

The dataset used in this project is taken from Udacity as part of Machine Learning Nanodegree program. The data simulates how customer make purchasing decisions and how those decisions are influenced by promotional offers.

3.2 Data Dictionary

There are 3 files:

1. portfolio: which contains information about the offer sent during the last 30 days.
2. Profile: which contains information about 17000 customers.
3. transcript: which contains 306534 offer reaction records.

The following are the schema and description of each file:

portfolio.json: (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)

profile.json: (17000 users x 5 field)

- 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)

transcript.json: (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

3.3 Initial Exploration

After simple and basic analysis, it seems that roughly 47 of the records are related to "BOGO" offer and the remaining are related to "Discount" offer. No record at all for informational offer as it would be difficult to track.
The distribution of both classes are almost the same, so the data set is balanced.

4 Solution Statement

The plan is to build a classification machine learning algorithm that learns more about the customers who purchase Starbucks products (complete the offer cycle). Then try to predict which offers are desired by which types of customers.
By combining the three data sets, we can analyse and study customers' demographics, then try to find relationships and patterns among customers' demographics, their recorded events, and their response to offers. Next, using these findings, we will build a model to achieve this goal.

5 Benchmark Model

This is a classification problem where the model predicts the offer that influences a customer. Logistic Regression will be used as a benchmark model. It is a supervised machine learning algorithm which is mostly used in a classification problems [Wol20].

6 Evaluation Metrics

Accuracy, Precision, Recall, F1, and Confusion Matrix will be calculated to evaluate and compare between the built models (benchmark model and the others).
Accuracy will help understand how a model performs in predicting offer type for a customer. Precision, Recall, F1, and Confusion Matrix will help understand the performance measurement on classifying various classes and show the discrepancies between predicted and actual labels.

7 Project Design

1. **Data Loading:** load the JSON files and convert them to Dataframes.
2. **Exploratory Data Analysis:** explore the three data frames and perform some analysis and visualization to better understand the data.
3. **Data Cleaning and Pre-processing:** clean and pre-process the data frames based on the results obtained in step 2; dealing with missing values, removing unneeded words from the data set such as "offer_id:" in value in transcript file, dropping unwanted columns - the ones are not related to the problem being solved such as time.
4. **Feature Engineering and Data Transformation:** transform the categorical features and prepare the data for modelling step.
5. **Train-Test Data:** split the data sets to training and testing sets.
6. **Benchmark Model:** build the benchmark model, fit it on training data set, deploy it, then evaluate it on testing data set.
7. **Proposed Model:** build several models such as XGBoost, LightGBM, SVM, KNN, and decision tree, fit them on training data set, deploy them, and evaluate them on testing data set. Compare between the models and benchmark model, then choose the best model.

References

[Jer21] Nikolina Jeric. 30+ starbucks statistics and facts with pumpkin spice, Apr 2021.

[Man18] The Manifest. The success of starbucks app: A case study, June 2018.

[Pan18] Riley Panko. How customers use food delivery and restaurant loyalty apps, May 2018.

[Wol20] Rachel Wolff. 5 types of classification algorithms in machine learning, Aug 2020.