

Capstone Project

Machine Learning Nanodegree Program

Dipan Sutradhar

Dated: May 20, 2020

Definition

Project Overview

Starbucks Capstone Challenge is a capstone project of the Udacity Machine Learning Nanodegree program. This project addresses a classification challenge that predicts the customers' demographics influence on responding to a Starbucks mobile reward program. The data set includes Starbucks mobile app reward data with simulated customer demographics.

- Portfolio: Currently available offer details
- Profile: Customer demographic details
- Transcript: Event details on customers' action

Problem Statement

Like any other campaigning program, the success of the Starbucks reward program depends on customers' participation in the program. Starbucks wants to analyze and identify if there is any influence on customers' demographics. Eventually, develop a model to predict the customers' responsiveness based on their demographic and offer details. This will help Starbucks to back their advertisement decision by data.

Data Exploration

Starbucks datasets separated into three such as portfolio, profile, and transcript. The exploration of these data set is as below:

- The Portfolio dataset represents the offer details including type of offer, the reward amount, duration of the validity, difficulty to earn the reward, and the communication channel. Starbucks reward offers are three types such as BOGO (buy one get one free), discount, and informational. Informational offer is mere information of Starbucks products though there is validity.

	channels	difficulty	duration	id	offer_type	reward
0	[email, mobile, social]	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10
1	[web, email, mobile, social]	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10
2	[web, email, mobile]	0	4	3f207df678b143eea3cee63160fa8bed	informational	0
3	[web, email, mobile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5
4	[web, email]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5
5	[web, email, mobile, social]	7	7	2298d6c36e964ae4a3e7e9706d1fb8c2	discount	3
6	[web, email, mobile, social]	10	10	fafdc668e3743c1bb461111dcafc2a4	discount	2
7	[email, mobile, social]	0	3	5a8bc65990b245e5a138643cd4eb9837	informational	0
8	[web, email, mobile, social]	5	5	f19421c1d4aa40978ebb69ca19b0e20d	bogo	5
9	[web, email, mobile]	10	7	2906b810c7d4411798c6938adc9daaa5	discount	2

- The Profile dataset represents the customer's gender, income, age, and membership anniversary date. The age value is encoded as 118 for a missing value.

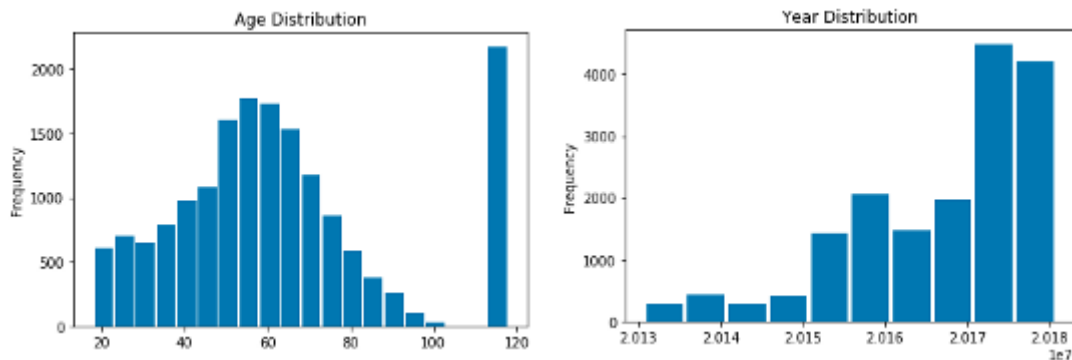
	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43	NaN

- The Transcript dataset represents the event logs when a customer performs an online action on the mobile app. There are four types of events such as offer received, offer viewed, offer completed, and a general transaction detail.

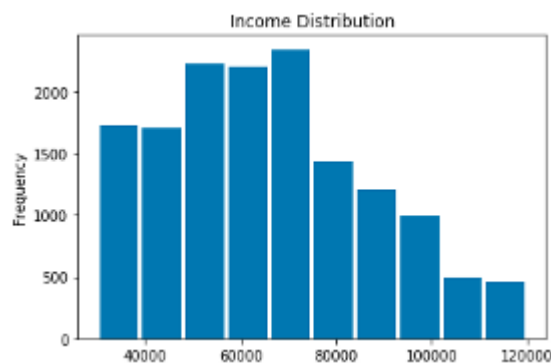
	event	person	time	value
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}
2	offer received	e2127556f4f64592b11af22de27a7932	0	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}
3	offer received	8ec6ce2a7e7949b1bf142def7d0e0586	0	{'offer id': 'fafdc668e3743c1bb461111dcafc2a4'}
4	offer received	68617ca6246f4fbc85e91a2a49552598	0	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}

Data Visualization

On visualizing customers' age, membership anniversary, and income distribution, it is found that most numbers of customers enrolled in the reward program are between 50–70 years of age. Also, there some record with incomplete data such as age is encoded 118 with no income information.



It is noticeable that Starbucks reward membership increased over the time since year 2013. But most customers signed up for reward program from 2017 to 2018.



From income visualization, we can see that majority of Starbucks reward program members have income range from 50,000 to 70,000 per annum.

Data Cleaning

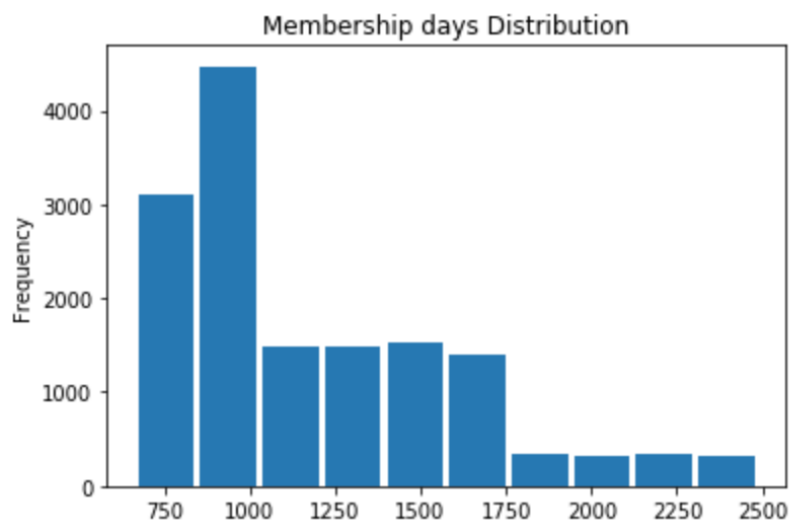
The data cleanup is done in notebook includes following steps for each dataset.

❖ Portfolio dataset

- Applying one-hot encoding to *channels* column.
- Applying one-hot encoding to *offer_type* column.
- Dropping *offer_type* column from.

❖ Profile dataset

- Dropping record with missing age. In this case, missing *age* is encoded as 118. This in turn cleans missing income records.
- Calculating membership days from *become_member_on* and dropping *become_member_on* column.
- Applying one-hot encoding to gender column. In this case, genders are *male*, *female* and *other*.
- Dropping *gender* column.



Above figure visualizes the membership days distribution. It shows the trend of customer enrolling to the reward program is increased recent days. That justifies the correctness of the data conversion showing similarity with *become_member_on* year distribution.

❖ Transcript dataset

- Extracting *offer_id* parsing dictionary string in value column.
- Extracting *amount* parsing dictionary string in value column for event type transaction.
- Creating new dataframe with *event_type*, *person_id*, *offer_id*, *amount*, and *reward*.

Training Data Preprocessing

The training dataset includes the features that I want the predictive model to be trained on. For this experiment, both customers' demographics and offer types of details are considered. So, the training data set is a new dataframe merged from Portfolio, Profile, and Transcript dataset. For true positive, a subset of the Transcript dataset with event type *offer_completed* is used. But for true negative, we will consider Transcript records with event type *offer_viewed* which customers do not participate in any transaction and there is no *offer_completed* event. Customers those who view offer but do not respond to the offer instead purchase Starbucks product are ignored from the training set.

After determining the Transcript subset applying above assumption, and then merging with Portfolio dataframe and Profile dataframe, the final training dataset looks like below:

```
Index(['label', 'age', 'income', 'membership_days', 'gender_male',  
      'gender_female', 'gender_other', 'difficulty', 'duration', 'reward',  
      'channel_web', 'channel_email', 'channel_mobile', 'channel_social',  
      'bogo_offer', 'discount_offer', 'informational_offer'],  
      dtype='object')
```

Dataset is then shuffled and split in training and testing dataset with 70% training data and 30% test data. I used scikit-learn convenient utility for this process.

Model Define and Result

This is a binary classification problem where for a given customer demographic and offer features model tries to predict if the customer will accept the offer or not.

I use DecisionTreeClassifier and SVC models from scikit-learn package. Models are trained and are used to predict test dataset.

Accuracy of the prediction is measured with the help of scikit-learn *accuracy* util.

Below is the accuracy of both the model.

- DecisionTreeClassifier: 0.98 (approx.)
- SVC: 0.97 (approx.)

Models should be trained enough to be able to predict if a customer will accept for a given offer. Model inputs are as below:

❖ Customer Details:

- Age
- Income
- Number of days member of the reward program
- Gender - male, female, or other

❖ Offer Details:

- Difficulty - money required to be spent to receive reward
- Duration – how long the reward valid
- Reward – monetary amount
- Channel – either email, web, mobile, or social
- Offer Type – either bogo, discount, or informational

Conclusion

This project experiments the influence of customers' demographic and offers details on the acceptance of Starbucks rewards. The process start with analyzing the data. The data analysis process shows us the points where data needs cleaning. After cleaning data, I prepare a training dataset and preprocess to prepare the final training and test data.

I train two classifier model with the training dataset. I use both the trained model to infer the likelihood of a customer's response on a given reward offer feature.

There are few things can be improved. I think the model can be improved with more training data. After all the analysis and preprocessing, the training dataset

seems to be inadequate. In addition, there might be room to improve training data quality by analyzing further details and scale data before training.