Capstone Proposal
Machine Learning Engineer Nanodegree

Richter Lau
Feb 12,2020

Starbucks Project Proposal

Domain background

Customers make multiple transactions every day if we can gather enough information from the customer, we can use machine learning algorithms to find out the customer behavior behind it. This Starbucks project is very similar to the situation we mentioned before, we can use simulated data that mimics customer behavior on the Starbucks rewards mobile app to solve our problem.

Problem Statement

In this project, we will focus on two major problems. The first problem is what demographic categories lead the customer to finish the offer. The second problem is what kind of offer is more attractive, Buy one get one free, discount or informational offer. We can analyze and preprocess the data first then we can apply machine-learning algorithms[1] to it to find out the solutions based on the preprocessed data.

Datasets and Inputs

All the datasets 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.

These are the sample from the dataset.

```
Portfolio has 10 rows and 6 columns.
```

	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

[1] Sidana, Mandy. "Types of Classification Algorithms in Machine Learning." Medium, Medium, 5 July 2019, medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14.

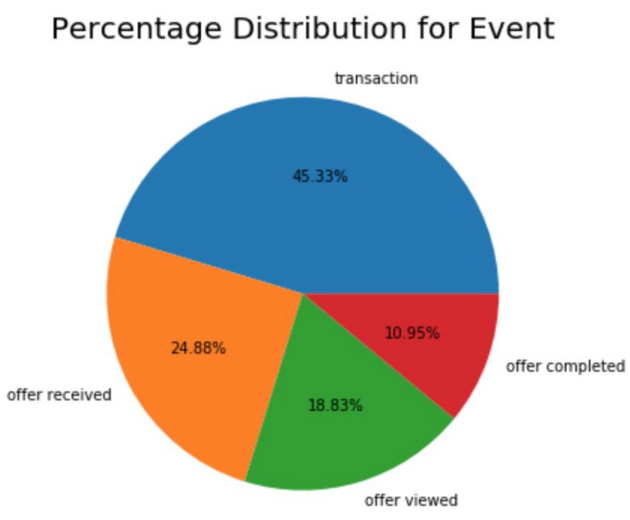
```
Profile has 17000 rows and 5 columns.
```

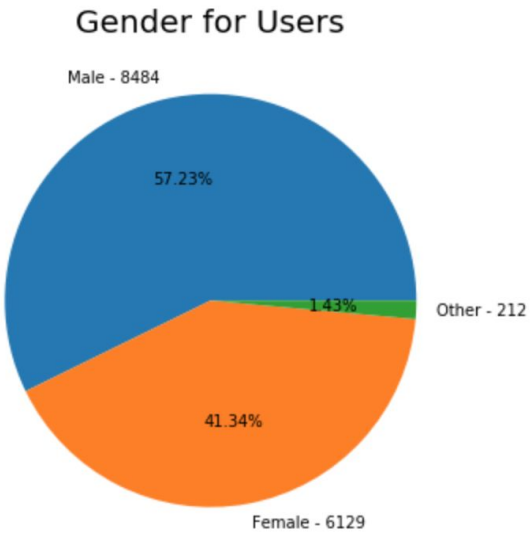
	age	became_member_on	gender	id	income
0	118	20170212	None	68be06ca386d4c31939f3a4f0e3dd783	NaN
1	55	20170715	F	0610b486422d4921ae7d2bf64640c50b	112000.0
2	118	20180712	None	38fe809add3b4fcf9315a9694bb96ff5	NaN
3	75	20170509	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0
4	118	20170804	None	a03223e636434f42ac4c3df47e8bac43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

```
Transcript has 306534 rows and 4 columns.
```

	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': 'fafdcd668e3743c1bb461111dcafc2a4'}
4	offer received	68617ca6246f4fbc85e91a2a49552598	0	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}

After some basic graph plotting we discover the datasets we have are unbalanced datasets.





Solution Statement

We will need to preprocessing the datasets first in order to apply a machine learning algorithm. After preprocessing the datasets and comine it into one dataset we will scale it into usable data. I suggest using Logistic Regression[2] for this project, because The goal of logistic regression is to find the best fitting model to describe the relationship between the dependent variable and a set of independent variables, which is very close to our goal in this research.

Benchmark Model[3]

Since there is no other obvious methodology against which you can benchmark, we will compare our solution with the KNN model[4]. To apply the KNN model we will need to find the optimal K-value for the model, and we can use the elbow method[5] to do that.

Evaluation Metrics

Since our model has an unbalanced dataset, to evaluate the solution model with the benchmark model. We would use f1-score as a main metric and accuracy as a supporting metric to compare them.

$$\text{F1-score} = \left(\frac{\text{Recall}^{-1} + \text{Precision}^{-1}}{2}\right)^{-1} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})}$$

Project Design

In this project, we will need to go over all the datasets first, remove duplicates and change the column name to more consistent throughout the datasets. Then we can draw some of the graphs related to the dataset to see the basics statistic of the data. After that, we can change the categories from text into numerical and scale it for better calculations. After we combine all the datasets into one dataset, we can process the data and put it into the algorithums.

---

[2] Brownlee, Jason. “Logistic Regression for Machine Learning.” Machine Learning Mastery, 12 Aug. 2019, machinelearningmastery.com/logistic-regression-for-machine-learning/.
[3] Jamesmf, “What is a benchmark model?” stackexchange, stackexchange, 10 November 2015 https://datascience.stackexchange.com/questions/8785/what-is-a-benchmark-model
[4] Umamaheswaran, Venkatesh. “Comprehending K-Means and KNN Algorithms.” Medium, Becoming Human: Artificial Intelligence Magazine, 14 Nov. 2018, becominghuman.ai/comprehending-k-means-and-knn-algorithms-c791be90883d.
[5] Doumbia, Moussa. “Elbow Method in Supervised Learning(Optimal K Value).” Medium, Medium, 24 Aug. 2019, medium.com/@moussadoumbia_90919/elbow-method-in-supervised-learning-optimal-k-value-99d425f229e7.

Before we put it into the algorithums we will need to split the data randomly for test and training. After everything works fine, we will run our algorithums. From the accuracy and F1-score we can find out which algorithum is better.