CUSTOMER SEGMENTATION FOR MARKETING CAMPAIGN OPTIMIZATION

Autor: Carlos Barros

BACKGROUND

It is no secret that we live in an increasingly hyper-connected technological era that has changed the way we relate and interact and has come to impact every aspect of our lives. The Internet has broken the communication barrier between cultures, it is possible to connect with a person in another part of the world without having to move to that place.

Business is no stranger to this, as companies can sell their products and/or services in different latitudes. However, this is a double-edged sword because in this market dynamic, competition increases, citing Forbes magazine [1]:

> "[…] In many technology-related industries, competition is intense and can often be the reason why a startup is not able to be profitable. However, competition cuts across the board and can contribute to potential business failure, regardless of the sector."

Because of the above, only half of the small businesses survive beyond five years, between 45.4% and 51%, depending on the year of creation. Beyond that, only one in three small businesses reaches the 10-year mark and lives to tell the tale [1].

Faced with this scenario, marketing has a crucial task and that is to understand how to meet each of the needs of its customers, bearing in mind that they have different tastes, customs, traditions, economic conditions, among other factors. Artificial Intelligence, thanks to the algorithms that have been developed in recent years, has contributed to apply this knowledge to marketing, discovering behavioral patterns, and detecting different homogeneous groups in a large set of customers. In this way, companies can take full advantage of it and anticipate the needs of their customers and thus be able to make them loyal to their brand, which will eventually become a greater economic benefit.

PROBLEM STATEMENT

The main objective of this project takes as a starting point those customers who completed the offer by performing a segmentation that allows discovering patterns of purchasing behavior and thus determine what type of offer is the most optimal according to the different groups of segmented customers. This will help Starbucks to focus its marketing campaigns and increase its economic benefit.

BENCHMARK MODEL

The benchmark model for the present project was developed by several researchers from the Sardar Patel Institute of Technology [2] in which they propose a systematic approach that combines clustering and recommender systems offering an improved and more novel approach to customer segmentation that will be of greater benefit to companies.

It is important to note that the authors use different metrics to those proposed for this project in choosing the optimal number of clusters such as Akaike and Bayes Information Criterion, Renyi and Shannon Entropies, so these factors will be considered when making a comparison between the different models developed in this project.

DATASETS AND INPUTS

The data is contained in three files:

1. portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
2. profile.json - demographic data for each customer
3. 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 start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

PROJECT METRICS

Since different unsupervised clustering models will be implemented the metrics to be used are:

1. **Elbow method:** used to determine the optimal number of clusters in k-means clusters by plotting the value of the cost function produced by different values of k.
2. **Silhouette value:** Measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).
3. **Davies Bouldin metric:** Average similarity measure of each cluster to its most similar cluster, where similarity is the ratio of distances within the cluster to distances between clusters. The minimum score is zero, and lower values indicate better clustering.

PROJECT DESIGN

The process to be used in this project is framed in the first place to understand the problem to be solved, which was mentioned in the ***Problem Statement*** section. Subsequently, data cleaning will be performed to obtain the customers who completed the offer with the different sources of information provided, as described in the ***Datasets and Inputs*** section. It is important to highlight that there are 3 types of events: offers received, offers viewed, and offers complete, which can be seen as a conversion channel within the campaign used by Starbucks. The following image shows the above:

Funnel of the entire data set

Note that at the top are the offers received with a total of 76.277, in the middle are the offers viewed with 57.725, representing 76% of the total offers received.

Finally, the conversion of completed bids was 33.579, representing 44% of the bids received. In other words, this project will work with only 44% of the overall data set.

Once the above is understood, statistical analysis can be considered to understand the nature of the data to prepare the data for the modeling of the algorithms. In a first approach with the data sets, it was possible to define the final data to work with, as shown below:

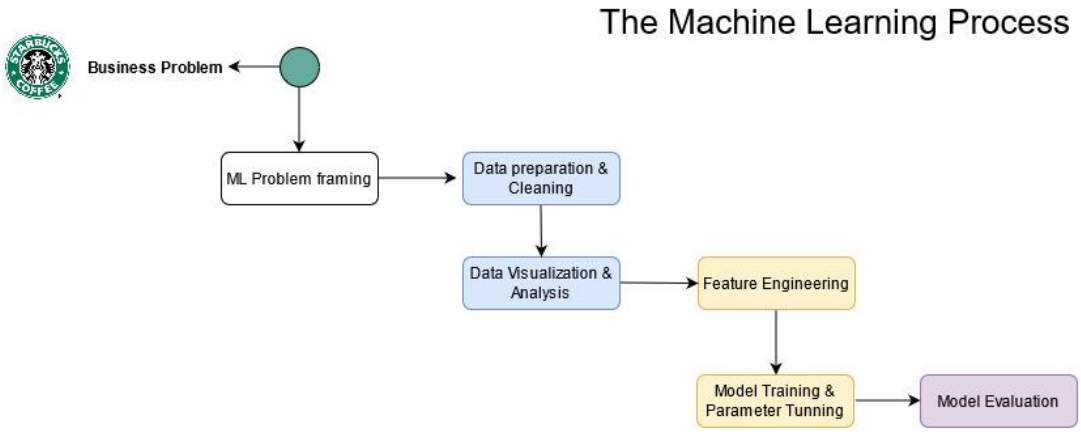
```
final_dataset.drop(columns_df_completed, axis=1, inplace=True)
final_dataset.head(5)
```

	gender	age	income	person	time	reward	channels	difficulty	duration	offer_type
0	None	118	NaN	68be06ca386d4c31939f3a4f0e3dd783	552	2	[web, email, mobile, social]	10	10	discount
1	M	68	70000.0	e2127556f4f64592b11af22de27a7932	522	2	[web, email, mobile, social]	10	10	discount
2	M	65	53000.0	389bc3fa690240e798340f5a15918d5c	600	2	[web, email, mobile, social]	10	10	discount
3	M	58	51000.0	2eeac8d8feae4a8cad5a6af0499a211d	456	2	[web, email, mobile, social]	10	10	discount
4	F	62	71000.0	31dda685af34476cad5bc968bdb01c53	258	2	[web, email, mobile, social]	10	10	discount

Post-processing data set

According to the previous image, there will be 10 variables that will be analyzed to define customer segmentation. Meanwhile, dimensionality reduction techniques such as PCA can be applied to define the important variables for data modeling.

Now, to condescend the above, the process flow of the project design can be seen in the following image:



General flow of the machine learning process

Finally, it is important to mention that a comparison will be made between 6 different segmentation algorithms to observe which model optimally segments customers according to the metrics defined in the ***Project metrics*** section. The algorithms are:

1. K-means
2. MiniBatch K-Means
3. Hierarchical Clustering
4. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
5. Gaussian Mixture Modelling (GMM)
6. MeanShift

REFERENCES

[1] Otar, C., 2018. *What Percentage Of Small Businesses Fail -- And How Can You Avoid Being One Of Them?*. [online] Forbes. Available at: <https://www.forbes.com/sites/forbesfinancecouncil/2018/10/25/what-percentage-of-small-businesses-fail-and-how-can-you-avoid-being-one-of-them/?sh=2cf9ecba43b5>

[2] K. Bhade, V. Gulalkari, N. Harwani and S. Dhage, "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization", *Ieeexplore.ieee.org*, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8494019.