Starbucks Challenge - Capstone Proposal

Index

1) Project Motivation

I have selected the Starbucks project for this Udacity Machine Learning Capstone Project.

The project will evaluate the information provided within this dataset, as to determine whether the offers sent to their customer base have an impact on sales, and along the project we will try to perform customer segmentation and clustering, to increase the impact of the campaigns.

2) State of the art

There is a huge interest nowadays in applying machine learning algorithms to customer segmentation problems, in order to maximize revenue, minimize costs, or simply avoiding inconvenience to customers.

In this regards, there is lot of scientific research covering this and other related issues, as for instance:

“A machine learning approach to segmentation of tourists based on perceived destination sustainability and trustworthiness”

Gabriel I. Penago Londoñoa, Carla Rodríguez Sánchez, Felipe Ruiz Moreno, Eduardo Torres

https://www.sciencedirect.com/science/article/pii/S2212571X20301542

"Towards Intelligent Risk-based Customer Segmentation in Banking"

Shahabodin Khadivi Zand

https://arxiv.org/abs/2009.13929

"Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services"

Chinedu Pascal Ezenkwu , Simeon Ozuomba , Constance Kalu

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.736.3182

"Soft computing applications in customer segmentation: State-of-art review and critique"

Abdulkadir Hiziroglu

https://www.sciencedirect.com/science/article/abs/pii/S0957417413003503

During the present project, we would like to try out segmentation techniques based on machine learning algorithms to improve the impact of promotional campaigns.

3) File Descriptions

The dataset comes from Starbucks simulated data. The data is provided in three files:

- portfolio.json - contains offer ids and meta data about each offer (duration, type, etc.)

This is a small dataset, containing only 10 rows, describing the portfolio of offers that are sent to customers. The '*channels'* column contains strings and needs treatment.

No null nor duplicates.

- profile.json - demographic data for each customer

A bigger file, containing information about 17.000 customers. Some missing values for all users declared with age of 118 years, in regards of gender and income.

- transcript.json - records for transactions, offers received, offers viewed, and offers completed

A massive file, 306534 rows containing information about all offers and transactions. The *'value'* column is in dict format, and needs treatment.

These tables are related with each other through the customer id and offer id values.

The transcript.json file will need heavy treatment to relate the transactional data and the offers with each other, and also the own offer development through time.

Here is the schema and explanation of each variable of each one of the files:

portfolio.json

- id (string) - offer id

- offer_type (string) - the type of offer ie BOGO, discount, informational

- difficulty (int) - the minimum required to spend to complete an offer

- reward (int) - the reward is given for completing an offer

- duration (int) - time for the offer to be open, in days

- channels (list of strings)

profile.json

- age (int) - age of the customer

- became_member_on (int) - the 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 the start of the test. The data begins at time t=0

- value (dict of strings) - either an offer id or transaction amount depending on the record

4) Proposed solution

Broadly speaking, we will try to assess the following question: Is there a way to categorize customers as to decide which customers send an offer to, in order to increase spending?

We will try to identify which customers don't need an offer to spend, which are highly sensitive to offers sent to them and finally, which customers respond negatively to any offer sent to them.

Our objective is to send promotional emailing only to those customers that are sensitive to it.

We will try to improve the accuracy of current emailing campaigns by at least 50%.

5) Benchmark model

We will use the massive bombardment emailing method(i.e. sending emails to all users, considering all the same) as the benchmark model, as it seems to be the method currently being used.

The accuracy for our benchmark model will be defined as the number of promotions used (completed), divided per the total promotions sent.

6) Metrics

For the clustering, explained variance will be used as to determine which variables explain the most variance of our target variable.

For our segmentation, accuracy will be the variable used.

We would like to include a variation percentage in revenue as a metric, but as it is not possible, since we cannot evaluate the behavior of our model afterwards, we cannot include it.

7) Project design

The project will consist in the following phases:

1) Wrangling of data

Inspecting the data and arranging the aforementioned anomalies, until it is adequate for an exploratory data analysis

2) Exploratory Data Analysis

An EDA will be subsequently performed as to extract the first insights from the data and decide the correct approach

3) Merging of tables and transcript transformation

There is a lot of information that can be extracted from joining the tables with each other. We will in this phase undertake this task, in order to add a few more useful features

4) Customer classification and benchmark setting

In order to proceed with our model, we will classify our customer based on their expenses and susceptibility to offers and not (creating our target variable)

4) Variable selection

From all the collection of features created, we will then select the most appropriate ones to create our model

5) Customer segmentation

After preparing our data, we will try out several machine learning techniques, comparing them with the benchmark and selecting the best model