Starbucks Challenge - Capstone Proposal

Illán Lois Bermejo

Index

1) Project Motivation

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

The end goal of our project is to improve the accuracy of the current promotional campaigns, by improving the rate of conversion of the offers, measured as those converted divided by the total, as a means of increasing the revenue.

In order to do so, we will evaluate the information provided in three files, related to offer campaigns, the customers such offers were sent to, and the characteristics of each offer, as to determine whether the offers sent to their customer base are being used, and which characteristics both in the offer and the customer are relevant for the determination of the success of the offer.

Once we have determined such characteristics, we will use the associated features to feed a model that will improve both the current success ratio (measured as an improved accuracy rate), and that of a less complex model.

2) Problem Statement

Sending offers and promotions to customers and potential customers of a business in order to increase revenue, get market share or publicize brands or products is a commercial strategy used for a long time.

However, sending these offers indiscriminately to the entire customer base of a business is not the most efficient way to tackle this task.

This is what has traditionally been done by businesses, due to the lack of the necessary information about their clients to be able to apply optimization strategies.

This implies loss of effectiveness of these campaigns, due to various reasons:

- sending offers to users who have purchased the products or services of these companies without the existence of the offer
- spamming: sending a large number of emails to a customer can cause them to not pay attention to these offers, or apply filters to eliminate them without being read
- generalized shipping prevents us from sending offers selectively to certain customers and at certain times, which would allow us the possibility of a possible purchase

For all these reasons, for years all businesses have been dedicated to gathering information about their clients, through, for example, loyalty programs, with which they can focus their commercial and marketing campaigns on the right people.

However, it has been with the introduction of machine learning techniques that these strategies based on user and customer data have been able to unleash the full power of this information.

Therefore, in our case, the problem that we intend to solve is the optimization of the sending of offers to Starbucks customers, to avoid the loss of effectiveness associated with the problems described above.

3) Domain Background

As described in the previous chapter, companies have since the 70s gather information about their clients and their behaviour, through direct imput, fidelization programs and other means.

That information has been being stored in databases, queried mainly through SQL.

Models as the CRM have been employed to determine the value of each customer, for instance.

Nevertheless, since the late nineties, a revolution has broken out in this field, brought by the advent of internet, on the one hand, that has permitted access to a new degree of data acquisition, and the advent of machine learning techniques, that have permitted to effectively treat these huge amounts of data.

4) Rationale

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

And so, machine learning and deep learning techniques are being used for solving these business problems and providing insight.

One of the fields in which machine learning methods have been used with great success is to solve clustering problems, for instance, for finding different kind of profiles within a customer base, or for providing useful insight on their behavior.

In this regard, here are some examples of these techniques put into use for solving such problems:

"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

Other machine learning techniques, such as Bayesian algorithms, are being used for other kind of classification problems, as it is their use in email filtering:

https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering#:~:text=Naive%20Bayes%20classifiers%20are%20a,commonly%20used%20in%20text%20classification.&text=It%20is%20one%20of%20the,with%20roots%20in%20the%201990s.

For these reasons, this kind of algorithms seem particularly suitable for the achievement of our goal, and this is the reason we have selected the use of machine learning techniques to undertake our problem.

During the present project, we will employ segmentation techniques based on machine learning algorithms to improve the impact of promotional campaigns.

We will also use other several machine learning models that using several features, some extracted directly from the data, some derived from it, will improve the success of the offering campaigns.

5) Data glossary - Datasets and Inputs

The data we have gathered to solve our problem is presented beneath. It consists of three json files:

- portfolio.json

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 duplicated values.

- profile.json

A bigger file, containing demographic information about 17.000 customers. Some missing values for all users declared with age of 118 years, in regards of gender and income. No duplicated values.

- transcript.json

A massive file, 306534 rows containing information about all offers and transactions. The *'value'* column contains dictionaries that hold information on whether it is an offer or a purchase, and, in case it is an offer and has been completed, the reward associated.

In these dictionaries, both *'offer id'* and *'offer_id'* exist, we will have to unify those values.

This table also seems at first glance to have duplicated records. A further analysis on these duplicated records will be required.

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.

In the following page we further describe the content of each file, detailing every variable.

FILE COLUMNS DESCRIPTIONS:

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. Both *'offer id'* and *'offer_id'* exist, we will have to unify those values. This table also seems at first glance to have duplicated records. A further analysis on these duplicated records will be required.

6) Proposed solution

The purpose of our project is to improve the accuracy of the current promotional campaigns, by improving the rate of conversion of the offers, measured as those converted divided by the total, as a means of increasing the revenue.

In a posterior phase, our objective would be to measure the derived gain in revenue from our promotional campaigns, compared to the previous.

To achieve our goal, we will divide the problem in two subproblems.

- The first will be finding a way to categorize customers as to decide which of them send an offer to. Non supervised algorithms of clustering will be used for this purpose.

This will allow us, for any new customer that we have information on, to determine if such customer is susceptible of redeeming an offer.

- The second problem will be broader: customer segmentation is only the first issue. Offer success relies in many other parameters. The customer susceptibility to an offer will be used as one of them, but will not be the only one.
  Other additional variables will be, for instance, the means the offer has been sent (using one-hot encoding for that purpose) and the difficulty of the offer to be redeemed.

We will try to improve the accuracy of current emailing campaigns by at least 50%, and by 10% when compared to a simpler model (linear regressor, with only readily-available features).

7) Benchmark model

We will consider that the current the massive bombardment method(i.e. sending emails to a lot of different users, considering all the same) is the benchmark model, as it seems to be the method currently being used.

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

A simple linear regressor with only the readily available information (that is, the information deriving from the existing files, treated so it can work, but with no additional features) will be used to set a second benchmark and see the improvement from the original scenario.

We will later define our model, operating on convenient features created along the data wrangling phase. Our model will be consider successful if it can improve the accuracy results derived from the simple linear regressor at least by 10%, considering that we want also to get an improvement of at least 50% from the massive bombardment method.

8) Evaluation 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.

As stated above, in later phases we would like to include a variation percentage in revenue as a metric. Currently that is not possible, since we cannot evaluate the behavior of our model in this regard, and thus we cannot include it.

9) Project design

Developing the solution a step further, 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.

We will get explore and decide on null values, unpack the values contained in both in the *channels* variable from the portfolio file, and the *value* variable from the transcript file.

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 creating new features

Useful insight and new features (as, for instance, revenue from redeemed offers per user) 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 to feed our model.

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).

We will also create our target variables: the user susceptibility to offers value, and whether an offer has or not been successful.

Finally, we will also set our benchmarks: that coming from the original strategy, and the one coming from the simpler model.

4) Variable selection

From all the collection of features created, we will then select the most suitable ones to create our model. That will be done by selecting those that explain better the behavior of our target variables, taking into consideration also the correlations existing among our features, in order to avoid those that are similar.

5) Customer segmentation

After preparing our data, we will try out several machine learning techniques, comparing their results with our benchmarks.

At end, we will select the best model and present our conclussions.

Now that we have presented our project design, we will move on to cover the tools devised to cover these steps.

10) Tools employed

For all these purposes the Python programming language will be particularly suitable.

This is because, appart from the wide range of tools provided for the analysis and treatment of data, the language offers also many machine learning libraries.

Another reason is that, for this project, we will require the use of Amazon Sagemaker for the treatment of data.

This makes compulsory the use of Python, as Sagemaker provides a library to interact with the service.

The following libraries will be required:

- pandas
- numpy
- math
- json
- datetime
- matplotlib
- seaborn
- scikit-learn
- sagemaker

The manner in which we will perform our analysis with Python will be through the use of Jupyter notebooks, which are a convenient tool for our purposes.

In later stages, the procedures contained in these notebooks will be conveniently transformed into Python scripts.

11) Conclusions

Along this proposal, we have presented the problem we are facing, the rationale for our proposed solution, we have presented our proposal for that solution, explained our benchmark models, the metrics of comparison with such models, and our project design, for which we have detailed both the tools and data to employ.

Summing up, we pretend to undertake the task of improving the success ratio of offer campaigns by employing machine learning techniques that are indeed suitable for the task.

With these techniques we will try to construct a model that we will measure against a benchmark, in order to determine our success.

Before getting to construct our model, heavy data manipulation will be necessary.

Due to this, we will use the Python programming language, since the wide range of tools that it makes available for this task makes it especially appropriate.

At the end of the project, we will determine if this proposed approach was correct, by comparing our results with those yielded by our benchmark.