

Machine Learning Engineer Nanodegree

Luis Soares Evangelista Neto

August 10, 2021

1 - Capstone Proposal

Luis Soares Evangelista Neto

August 10, 2021

2 - Proposal

This document has the purpose of offering a solution for the Starbucks Capstone Challenge.

2.1 - Domain Background

The project is a study for the Starbucks company to identify the best way to communicate with clients and increase revenue by optimizing marketing campaigns.

For a company like Starbucks, it is important to identify the key characteristics of his clients and send personalized offers to them. With this, it is possible to reduce costs and track our offers portfolio and change our strategies.

My personal motivation for this project is because I recently joined a company that has projects with marketing advertisements, so it's a good way to explore, try new techniques and propose ways to solve similar problems in my company.

2.2 - Problem Statement

The main idea of this capstone project is to explore ways that can explain how the Starbucks customers respond to their offers and improve the way the company approaches them. We can use exploratory data analysis (EDA), heuristic or modelling to do so.

2.3 - Datasets and Inputs

The data given are separated into 3 datasets:

- **portfolio.json** — Contains information about Starbucks campaigns (offers);
- **profile.json** — Personal information about the clients;
- **transcript.json** — Transactional information about the behavior of the clients.

2.4 - Solution Statement

For this challenge, there are several approaches, from EDA to advanced modelling. We are going to use a clustering mixed with exploratory analysis to identify some hidden profiles in the customer that **convert to an offer**. This approach gives us some important information about these clients and how well the offers are effective.

After that, we can use a supervised model to link the profiles we have discovered with the probability of converting to an offer.

For this challenge, we are going to focus on the hidden profiles (clustering)

2.5 - Benchmark Model

A way to do a benchmark for a clustering model is to evaluate the percentage of the distribution together with the percentages in each variable.

2.6 - Evaluation Metrics

Usually, there are two known ways to evaluate a clustering model, using the **Elbow plot** and the **Silhouette plot**. These metrics help us choose how many clusters our data can be separated.

For this project we are going to use the K-Prototypes model and Elbow method to determine the optimal number of clusters for K-Prototype clusters. Instead of calculating the within the sum of squares errors (WSSE) with Euclidean distance, K-Prototype provides the cost function that combines the calculation for numerical and categorical variables.

2.7 - Project Design

The workflow for this project is described below in steps:

1. Exploratory Analysis (Part 1)

In this section we are going to explore the 3 datasets separated to get a better understanding of the data.

2. Preprocessing (Step 1)

In this section, we are going to transform our data into a unique dataframe that contains all the information in the entities (datasets) given. And create some potential useful variables.

3. Preprocessing (Step 2)

After a first preprocessing phase, let's create a variable that indicates if a customer for a specific **offer was successful or not**.

4. Exploratory Analysis (Part 1)

More exploratory analysis, but using the variable ***offer_result*** created before in step 3.

5. Modelling (Customer Clustering)

In this section we are going to create our model, applying all the transformations required for the K-Prototypes model.

6. Cluster Analysis

In this section, we are going to explore our variables and see potential insights to describe the **persona** of each cluster.

7. Observations and Future work.

This section has all the observations that we get in our analysis and potential future works together with this solution.