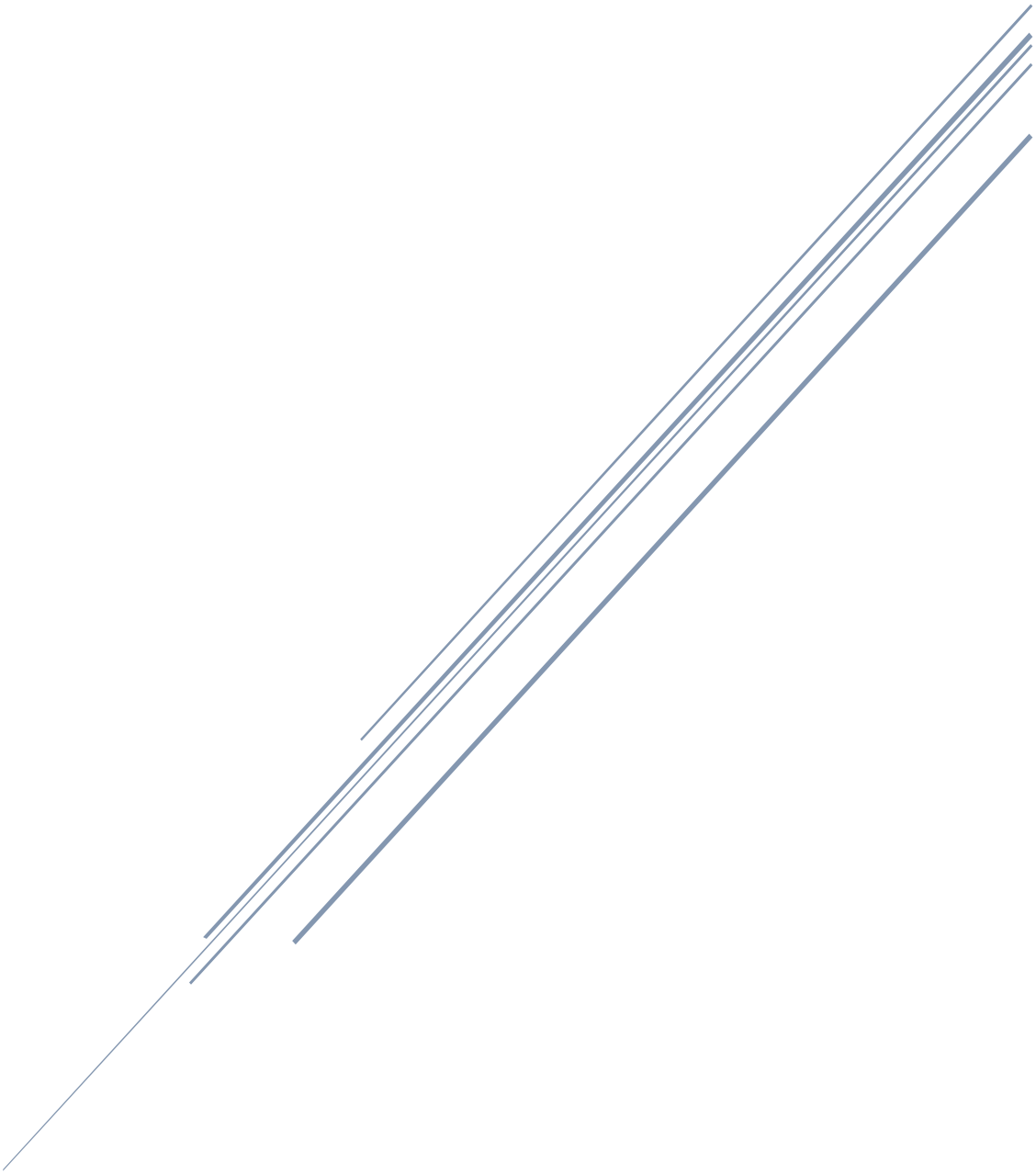
# CAPSTONE PROPOSAL - STARBUCKS

## Udacity - Machine Learning Engineer Nanodegree Program



Leonardo Simões

10/19/2020

**Summary**

## 1. Domain Background

This is one of the projects needed to complete the Nanodegree Program in Machine Learning Engineer at Udacity. The project in question consists only of the proposal and specifications for the next Machine Learning project in the course.
The chosen theme was one suggested by the course program, analyzing data from the Starbucks app and applying machine learning models to achieve a meaningful classification that can provide insights into this part of the company's business. The project aims at the practical application of theoretical and practical concepts taught in the course for an application in the real world, and not just academic ones.

## 2. Problem Statement

The problem is to make a forecast relevant to the business, using a Machine Learning model for classification or regression. This problem can be seen as the classification of the type of offer, given some characteristics of the offer and the target customers, and thus obtain relevant insights and check whether new offers will properly fulfill their objectives.

## 3. Datasets and Inputs

The data was provided by the Udacity platform in the section of the project proposal, with the objective and authorization to carry out this project. The dataset is a simplified sample of the data from the Starbucks app, and there are three separate sets, each in a different json file.

The description provided with the datasets includes that of each file and each column. The description provided by Udacity has been pasted below:

The data is contained in three files:

- portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
- profile.json - demographic data for each customer
- 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

The distributions of the labels 'bogo' and 'discount' are larger than that of 'informational', indicating that there will be a supposed superiority in the classification of the first two labels. This will be assessed after training and testing the models.

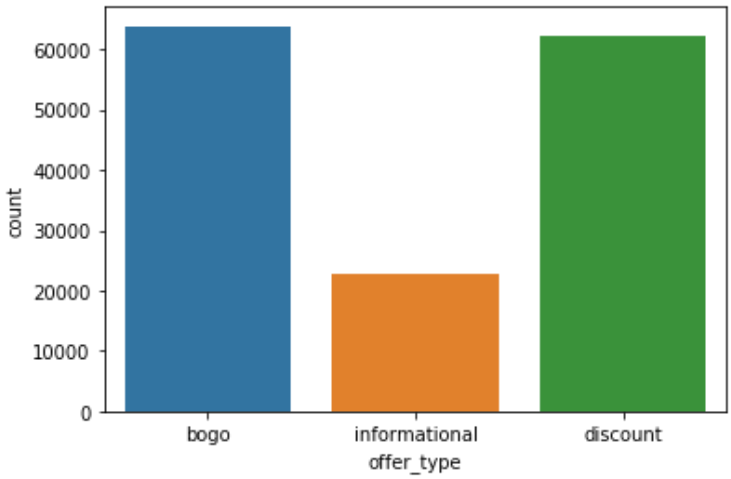



*Figure 1 – labels of the offer_type*

## 4. Solution Statement

After data collection, preparation (evaluation and cleaning), engineering and selection of features, and division into training and testing, the data set will be ready to be used in the models. The classification of the type of offer will take into account the relevant features of transcript, portfolio and profile. The effectiveness of the solution will be assessed according to a specific metric, the accuracy, and will determine the viability of the solution and the choice of model.

## 5. Benchmark Model

The models used are multilabels classifiers that receive numeric features and result in the numerical label with the highest probability of correspondence. The models are LinearLearner and XGBoost, both seen in the course and usable in the AWS SageMaker environment. After being trained, each model will be evaluated according to the chosen metric on a test set.

## 6. Evaluation Metrics

The evaluation metric that will be used to evaluate the models will be the accuracy. In the project, a multilabel classification (3 labels) will be carried out, so with the chosen approach, measuring the classification correctness for each class is adequate and sufficient.

## 7. Project Design

The project is carried out in stages, which are Presentation, Data Wrangling, Exploratory Analysis, Feature Enginner, Model 1, Model 2. Each is present in a different Jupyter notebook to increase the organization and readability of the code and information.

The Presentation stage will contain information on other stages of the project and an introduction. The Data Wrangling step will consist of analyzing the data, assessing its quality and structure, and then correcting errors and restructuring the data so that they can be better analyzed. The Feature Enginner stage will also contain the selection of features, there will be normalization of data, creation of new features, it will transform all relevant features into numeric ones and discard those that are irrelevant for certain forecasts.

There will be a stage for each of the two Machine Learning models employed. The models used are LinearLearner and XGBoost, both seen in the course and usable in the AWS SageMaker environment. At the end of this stage, each model will be evaluated according to a metric, accuracy.

# REFERENCES

**UDACITY - Machine Learning Engineer Nanodegree:**
https://www.udacity.com/course/machine-learning-engineer-nanodegree--nd009t