Starbucks Capstone Project Proposal

Muhammet Çelik - 26.06.2021

1. Domain Background

Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. As the world's largest coffeehouse chain, Starbucks is seen to be the main representation of the United States' second wave of coffee culture. As of September 2020, the company had 32,660 stores in 83 countries, including 16,637 company operated stores and 16,023 licensed stores. Of these 32,660 stores, 18,354 were in the United States, Canada, and Latin America [1].

The rapid spread of mobile applications have changed not only lives of people, but also the way of doing business. Advertisement business has altered its shape into online advertisement. On contrary to conventional methods, online advertisements offer more targeted customers to the companies. This phenomenon is also valid in the scope of this project. The aim of this project, by using the data gathered by Starbucks App, to develop a smart model making promotion offers to customers according to their demographic information and coffee drinking habits.

2. Problem Statement

Starbucks demands to develop an intelligent model deciding whether to show the online offer or not to each individual customer. Moreover, the company also wants to show the best offer to them, increasing the offer acceptance probability. To deal with this problem, we are going to use the customers historical data and build a model. Performance evaluation of the model can be made by observing the past offers and their acceptance and rejection states. The solution will be based on the demographic features, the behavioral features and also the promotion features, which are going to be explained in Solution Statement, Section 4.

3. Datasets & Inputs

This dataset contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (Buy-One-Get-One free). Some users might not receive any offer during certain weeks. Not all users receive the same offer, and that is the challenge to solve with this dataset [2].

There are three available data sources: **portfolio.json, profile.json, transcript.json.** The detailed information has been given in Table 1.

Data Source	Source Explanation	Variable Name	Variable Type	Variable Explanation
portfolio.json	**10** records, contains offer ids and metadata about each offer (duration, type, etc.)	id	string	offer id
		offer_type	string	type of offer
		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	channels for showing offers
profile.json	**17000** records, contains demographic data for each customer	age	int	age of the customer
		became_member_on	int	date when customer created an app account
		gender	string	gender of the customer
		id	string	customer id
		income	float	customer's income
transcript.json	**306533** records, contains records for transactions, offers received, offers viewed, and offers completed	event	string	record description (ie transaction, offer received, offer viewed)
		person	string	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

Table 1: Data sources and their explanations

4. Solution Statement

There are two stages of the solution: Data Analysis & Modelling. In Data Analysis Stage, explatory findings will be used for Feature Engineering for the model. In profile dataset, there are NULL values and also irrational age values (118 – born in 1900). Also, when a customer became a user of app will enlighten us about which customers are more inclined to technology. In transcript dataset will lead us to understand the responses of the customer. An RFM analysis will help us for bucketing the behaviors. By using portfolio dataset, we will go through the attractiveness of the promotion types. Therefore, we will transform the customers' properties into a mathematical feature set.

For modeling, each customer's demographic features, behavioral features and offer type features will create the base table. Avoiding from unnecessary promotion offers, I will also consider not sending offer decision in the label set. Due to the fact that model dataset will be tabular, I would consider to build a tree-based algorithm such as Decision Tree, Random Forest. In the case of poor performance, I will take XGBoost into account.

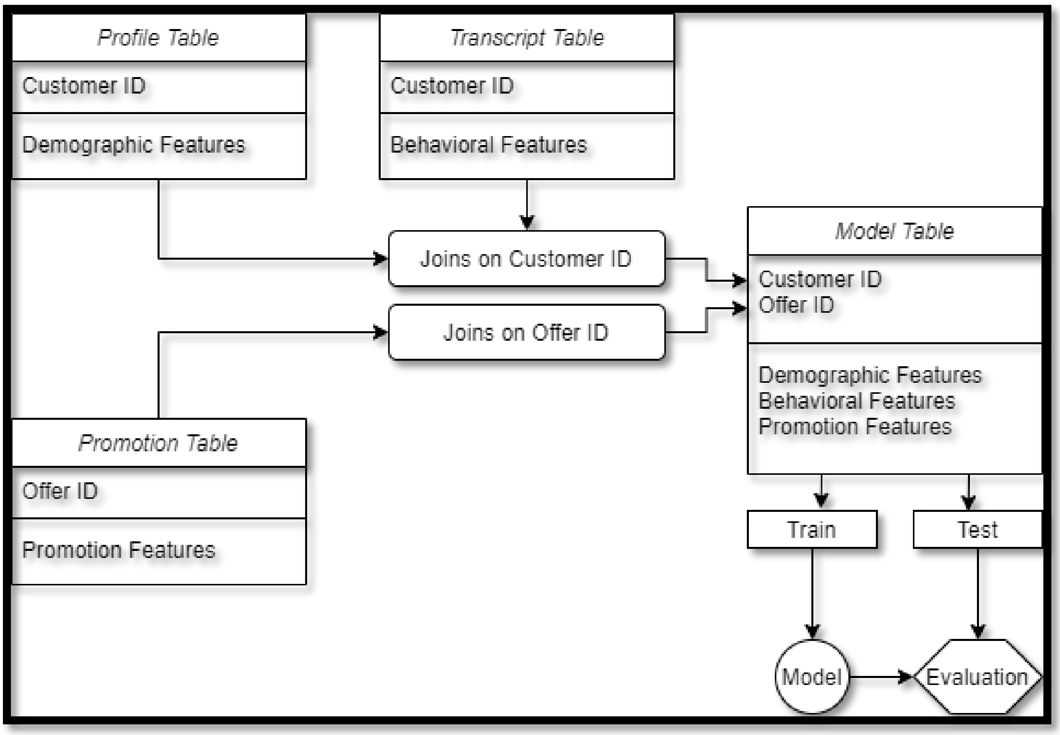

Figure 1: General Flow of the Project

5. Benchmark Model

As a benchmark model, I will use a dummy model, a KNN model and a Logistic Regression model. I will build these models by using the default parameters of them and I will not spend time on hyperparameter tuning for these models. In this project, as abovementined, both false negatives and false positives are important, I will evaluate the model performance by the help of F1 score metric.

6. Evaluation Metrics

F1 score will be the evaluation metric. F1 score is the harmonic mean of precision and recall. Since I consider the number of both false positives and false negatives to a bare minimum, this metric will provide the most confidential results. Besides, ROC AUC score can be the second alternative to compare the models in more graphical way.

7. Project Design

The following steps will shape the general flow of the project:

1. The project environment will be on Jupyter.
2. Deep-dive Explatory Data Analysis will be performed.
3. Features will be generated.
4. Tree-based algorithms will be built and optimized on AWS SageMaker.
5. A demonstration of model comparison will be presented.
6. Project work will be reported.

8. References

[1] https://en.wikipedia.org/wiki/Starbucks

[2] Udacity ML Engineer Nanodegree – Starbucks Project Workspace