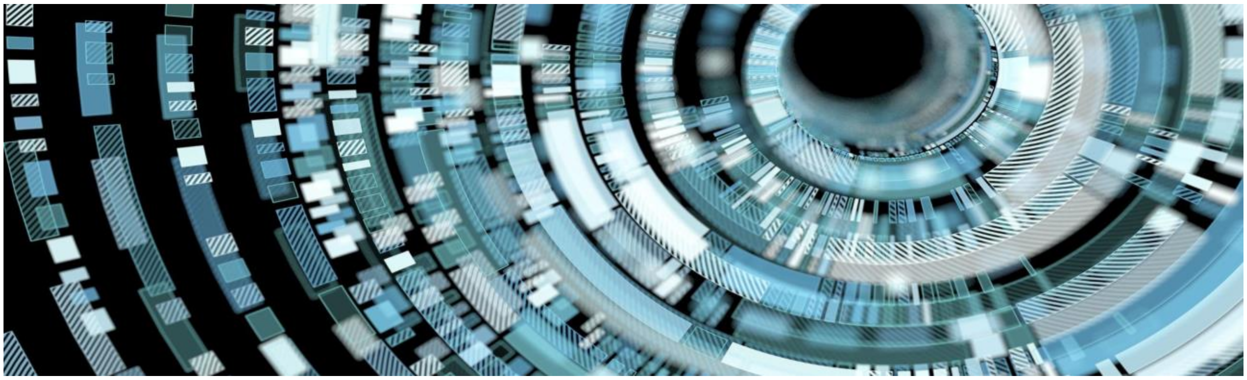


Machine Learning Engineer Nanodegree

Udacity

Douglas Wong 2022

Capstone Project Proposal

Table of Contents

Domain Background

Starbucks is by far one of the world's largest franchise coffee shops. They are known for having implemented one of the most successful information technology solution, which has allowed them to grow into industry leaders. Starbucks debuted its mobile order-ahead app feature in late 2014 and it quickly caught on with Starbucks Rewards members *(a)*. Starbucks has also become a leading mobile payment app that competes with Google Pay, Apple Pay, and Samsung Pay(*c). S*tarbuck is also well-known for its loyalty program My Starbucks® Rewards where it able to offer individualized offers to their member. The program grew by 16 percent year over year in the first quarter of 2020, reaching 18.9 million active users *(b)*. The membership increase is correlated to the increase in sales growth *(c).*

According to Starbucks, a quarter of their transactions will be completed over the phone by the end of 2020 *(a)*. This suggests that the rewards app accounts for a sizable portion of their revenue. In the recent decade, the use of artificial intelligence (AI) in marketing has grown exponentially and machine learning is providing a huge advantageous to target customers, predict product performance and customer behavior *(d)*.

In this capstone project, we want to look at how the customers used the Starbucks rewards app so that we can improve earnings through targeted offers to drive sales.

This project contains simulated data from the Starbucks reward mobile app. The Starbucks app rewards registered customers on its platform to entice them to make purchases. There are 3 main types of offers that are sent to the customer.

- Buy one get one Offer (BOGO)

- Discount Offer
- Informational Offer

In a BOGO offer, a user needs to spend a certain amount to get the reward. In the discount offer, the user receives a reward equal to the fraction of the amount spent. In an information offer, there will be no reward nor minimum amount spend.

However not all customer response to the same marketing campaign, some customers will response to campaign regardless of reward such as recurring customer, while certain customer such as new customer need to be attracted via discount.

The data that has been collected by the app reward is a data mine which offers us insights on customer base spending habits, and thus based on this valuable data, we can utilize this using Machine Learning to increase the ROI of the marketing campaign.

Problem Statement

Every company invests money in marketing campaigns expecting that it will be successful in bringing more profit as an outcome. Therefore, it is imperative that we can increase the return on investment (ROI) by identifying the most effective offer type to be offered to the different subgroups of our customer base.

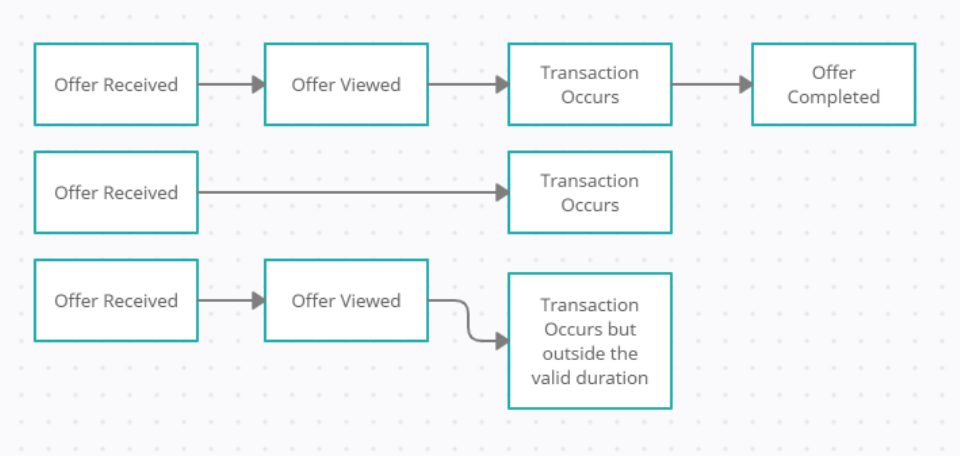
Solution Statement

Based on the dataset that we have which we obtained from the Starbucks reward mobile app, we are proposing to utilize machine learning methodology to build a model to predict the success of campaign offer type. This can allow us to determine which offer should be targeted at different subgroups of customers as well.

The proposed plan was to merge the portfolio of offer, the profile and transcript together into a big dataset where we can analyze the data. We will also determine if a particular offer is successful based on the record of the transactions that occur during the duration of an offer. Besides investigating if an offer is successful, we will also be looking into the *amount* of profit that each offer brings in as another metric that we can investigate.

We will consider a successful offer as follows. It must satisfy two conditions, offers received must be viewed, and the transaction must occur during the duration of the offer.

If an offer were received but not viewed, it would mean that the customer would have made the purchase regardless of the offer.



We will be utilizing the Amazon SageMaker platform for its integrated development environments, creating a Sagemaker notebook instance. It is a machine learning compute instance running the Jupyter Notebook App. We will use this environment to do our data processing, then we will upload our training data to Amazon S3 Cloud Object Storage.

As we have a labeled dataset, we will be using supervised learning algorithms to predict if an offer is going to be successful. We will be exploring and comparing three algorithms, scikit-learn library logistic regression, scikit-learn library random forest, and XGBoost algorithm provided by Amazon SageMaker. Logistic regression is a statistical method for predicting binary classes *(e*). Random forests classifier is an ensemble of decision trees trained on randomly selected data samples, then the best prediction from each tree and select the best solution by voting*(f)*. XGBoost is an efficient implementation of gradient boosting. Gradient boosting is an algorithm that combines many weak learning models together to create a strong predictive model *(g)*.

We will also determine the **`Feature Importance`** to estimate feature importance to describe how important that a particular feature in a model at predicting the success of an offer.

Datasets and Input

Data Overview

The dataset used in this project contains data from the Starbucks reward mobile app. It contains the event of receiving offers, opening offers, and making purchases. In this simplified dataset. Only the type of offer and the transaction and the purchase amount are available in this dataset but not the actual product contributed to the purchase.

Data Dictionary

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) - 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

Benchmark Model

We will calculate the accuracy and F1 score from a Logistic Regression Models to create a baseline model which will then be used to compare it with all other subsequent Models. This is because this model is simple to set up and provides a reasonably good result. We will also be looking at comparing the results to the contemporary work done on the same dataset that have accuracy score of: 0.80 and the f1 score is: 0.73. *(h)*

Evaluation Metric

To see how well our classification model performs, we will assess and compare the accuracy score and the F1 score (weighted average of the precision and recall value). Since the accuracy of the True positive and True negative is important in our cases, ie to see if the data model can successfully predict if an offer campaign will be successful. Depending on the class distribution of our data, we will also compare the F1 Score as it will give a better measure of the incorrectly classified cases than the Accuracy Metric (*i*).

Project Design

The planned workflow for this project is as follows.

1. **Data Preparation**
2. **Data Cleaning**
3. **Feature Engineering**
4. **Data Exploration Analysis (EDA)**
5. **Splitting Data**
    a. **Training Data** - The main data used for training our model.

b. **Validation Data**

c. **Testing Data** - Using the testing data to measure the performance of our model

6. **Data Modeling**

    a. sklearn library logistic regression

    b. sklearn library random forest

    c. XGBoost

7. **Evaluating and comparing different model performances**

    a. Accuracy and F1 Score

8. **Feature Importance**

9. **Discussion / Conclusion**

References

a. Starbucks Industry AI Case Study (Domain Background) https://info.formation.ai/rs/435-BMS-371/images/Starbucks_CaseStudy_final.pdf

b. Starbucks Industry (Domain Background) https://brainstation.io/magazine/digital-loyalty-lifts-starbucks-q1-to-record-7-1-billion

c. Starbucks Industry (Domain Background) https://www.geekwire.com/2021/quarter-starbucks-orders-u-s-now-paid-smartphone/

d. AI in marketing https://www.sciencedirect.com/science/article/pii/S2667096820300021

e. Logistic Regression Definition https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

f. RandomForest Definition https://www.datacamp.com/community/tutorials/random-forests-classifier-python

g. XGBoost https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

h. Accuracy vs. F1-Score https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

i. Baseline Model from contemporary works done on the same dataset https://medium.com/@sha821/starbucks-capstone-project-4b1eb8015bee

j. Udacity Starbucks Capstone Challenge Notebook and Dataset