Udacity Nanodegree

Machine Learning Engineer

Capstone Project Proposal

Kevin Hubert - February 2021

Table of content

Intro

I this project I will apply my learned skills from the "Machine learning Engineer" Nanodegree and will demonstrate the potential of machine learning using a dataset provided by Udacity and Starbucks.

Starbucks & Machine Learning - Introducing

Small snacks such as cookies and cakes, free Wi-Fi and many good coffees in countless varieties - there are many things for which Starbucks, 50 years after its founding, is not only known but also incredibly successful and rightfully counts as the market leader[1] in this segment.

In this project, I demonstrate how this success story is linked to digitization and what additional potential there is for Starbucks in machine learning technology.

Starbucks employs approximately 346,000 people in moren than 30,000 stores in 80 countries worldwide[2]. Due to the frequency of Starbucks stores and the associated huge number of satisfied customers every day, Starbucks generates millions or even billions of data records annually, which can be collected, analyzed and integrated into the optimization of business models with the help of modern technology in the field of artificial intelligence.

For this project, a dataset provided by Starbucks is used as part of the Udacity "Machine learning Engineer" Nanodegree. In this dataset, the following data was provided:

File	Description
portfolio.json	JSON structured information about offers done to customs. Could be rewards, information or BOGO (Buy one & get one free)
profile.json	Demographic information about customers. Containing register-date, gender, income and age for the individual customers.
transcript.json	Information about fulfilled events like receiving-offer, viewed-offer, transaction done etc.

All files were formatted as JSON

- Entries of files are related - Relation is recognizable by the given ID

Source: https://de.wikipedia.org/wiki/Starbucks

1 https://www.statista.com/statistics/250166/market-share-of-major-us-coffee-shops/
2 https://de.wikipedia.org/wiki/Starbucks

Domain Background

This project relates to the marketing model used by Starbucks to stay in contact with existing customers and to address them at regular intervals with targeted offers. For this purpose, Starbucks has an app where customers can register with personal information and from then on receive individualized messages with offers from Starbucks. These messages are addressed to customers either directly via the app, e-mail, social media or the web.

A distinction is made here on the basis of three different offer types:

- BOGO: Acronym which stands for "buy one get one." This means that when a customer buys one product, he or she receives a second one free of charge.
- Discount: Price reduction when a given threshold is reached. (e.g.) 10% discount when buy something for at least 10$.
- Informational: Just information texts about new items etc.

As with any successful marketing campaign, the goal here is to cover the marketing costs incurred and to achieve the highest possible profit. The following must be taken into account:

- Customers should not be deterred/annoyed by too many offers. The offers must therefore appeal to the individual customer
- No offers to customers who do not respond to coupons/offers.
- Address new customers with offers
- Entice existing customers to buy again with offers.
- Do not make offers that reduce profit (example: no "10% discount for a purchase value of atl least 15€" offers if the customer regularly buys for 15€ even without a discount).
- CTA (Call to action) maximization - Percentage of customers responding to the offers (100% customer interaction would be optimal).

Through targeted and individual marketing campaigns, an optimal situation can be created for customers as well as Starbucks, since a high level of customer loyalty and satisfaction can be achieved through the targeted optimized approach of customers and, on the other hand, revenue maximization on the part of Starbucks.

Problem Statement

Customers are as individual as their needs. This makes addressing customers directly through marketing campaigns an extremely complex issue. There are various reasons why customers do not respond to offers from companies, a few of which are:

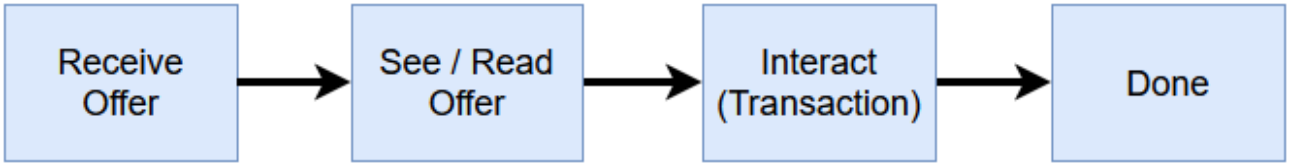
- The customer has not read/received the message - Wrong marketing channel.
- The offer does not meet the customer's needs
- The customer is not interested in the offer

- The customer was approached at the wrong time (e.g. receives money at begin of month and has not money left at the end of the month)

On the other hand, success through marketing campaigns depends on customers responding to the offers and going through the following steps to do so

1. Receive offer - Customer receives the offer
2. See / Read Offer - Customer sees/reads the offer
3. Transaction - Customer interacts with offer
4. Done - offer fulfilled

Visualized[3]:



As you can see there are several steps required until the offer is fulfilled (Done). Only if all steps are fulfilled then the custom will be happy and we will increase the companies profit

Dataset and Inputs

The data and files used by this project are provided by Udacity and Starbucks for the given nanodegree program. The given data contains simulated customers, transactions and offers similar as it is done by the Starbucks app.

3 Visualization created using https://app.diagrams.net/

portfolio.json

- JSON structured information about offers sent to customs. Could be „rewards“, „information“ or „BOGO“ (Buy one & get one (free)) |
- Shape (Rows x Columns): 10 x 6
- Column description:
    - id (string) - unique ID to identify offer referenced in other data
    - offer_type (string) - type of offer (one of: 'BOGO', 'discount', 'informational')
    - difficulty (int) - minimum required spend to complete an offer
    - reward (int) - reward given for completing that offer
    - duration (int) - time (in days) for offer to be open
    - channels (strings[]) - Channels the offer is sent out (i.e.: 'web', 'email', 'mobile', 'social')

profile.json

- Demographic information about customers. Containing register-date, gender, income and age for the individual customers.
- Shape (Rows x Columns): 17.000 x 5
- Column description:
    - id (str) - unique ID to identify customers referenced in other data
    - age (int) - age of the customer (value 118 indicates it is missing)
    - became_member_on (int) - date when customer created an app account (formatted: yyyymmdd)
    - gender (str) - gender of the customer ('O' = other, 'M' = male, 'F' = female, 'null' = missing)
    - income (float) - customer income (currency is not defined)

transcript.json

- Information about fulfilled events like receiving-offer, viewed-offer, transaction done etc.
- Shape (Rows x Columns): 306.648 x 4
- Column description:
    - event (str) - record description (ie 'transaction', 'offer received', 'offer viewed', 'offer completed')
    - 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

- All files were formatted as JSON[4]

4 https://en.wikipedia.org/wiki/JSON

Solution Statement

My solution attempts to predict the propability of a offer to be successful[5] for a given customer. Based on this prediction it should be possible to find the best matching offer for each customer.

My attempts based on machine learning. Therefor I'll transform the data into my required format, analyse the data and visualize my results and then use my data to train a deep-neural-network which will be then used to estimate the probability of a given offer to be successful or not.

To train my model I'll use the data provided by Udacity and Starbucks for this project and based on this data I should be able to build/train a neural network with a good probability if a offer will be successful for a specific customer.

Keep in mind: The outcome of a predictions could also be "Make no offer" especially for customers which may think that offers are annoying or just don't interact with them.

The model then can be used either to plan the future marketing offers completly or as a additional instrument next to the existing markething actions.

Benchmark Model

To evaluate the results of my model, I will compare these predictions of my solution with that of the current marketing strategy using the given test data. For this, of course, only the results of BOGO (buy one get one) and of discounts will be compared, since offers of the "information" type do not lead to any transaction.

For this I'll train a very simple AWS Linear Learner Model (logistic regression) to do a binary classification if a offer would be successfull or not.

5 A „successfull offers“ describes an offer which leads to a transaction (fulfillment) of the given customer.

Evaluation Metrics[6]

For the statistical evaluation of my model, I use the Confusion Matrix which divides the predictions into:

- True Positives: Correctly classified as positive
- False Positives: Incorrectly classified as positive
- True Negatives: Correctly classified as negative
- False Negatives: Incorrectly classified as negative

I'll calculate the accurancy, precision and recall which are kindly explained by the graphic (Figure 1) below based on the given predictions:

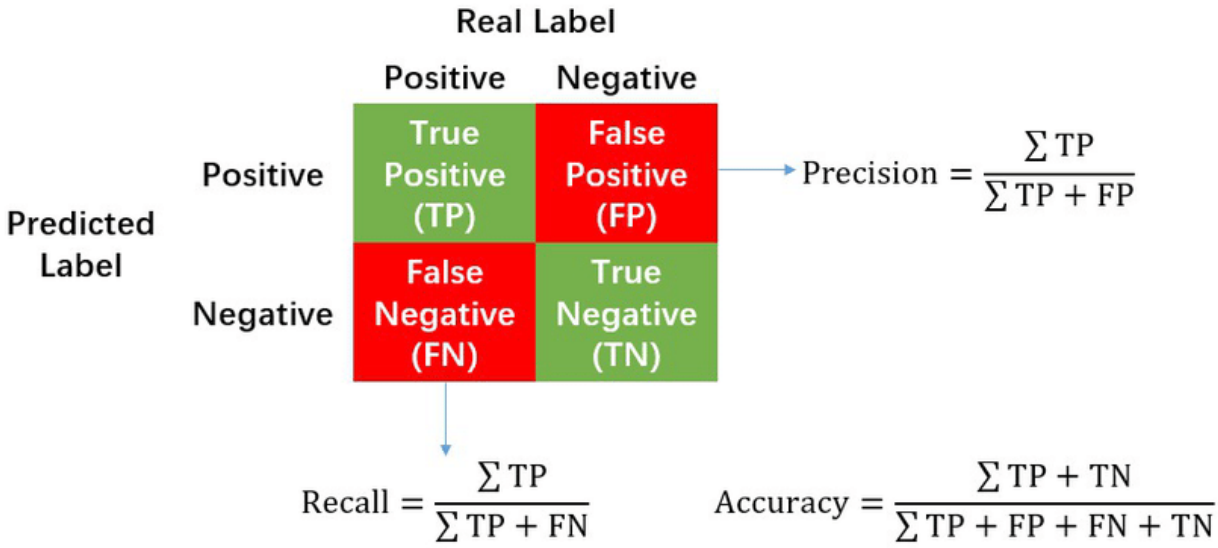


Figure 1

6 https://en.wikipedia.org/wiki/Confusion_matrix

Project Design

For my project I'll used multiple different steps and machine learning techniques to archive my goal.

I use at least the following Machine learning techniques:

- Principal component analyses
- KMean Clustering
- AWS Sagemaker
- Logistic regression for binary classification
- Deep neural network (high amount of hidden-layers/-neurons)
- AWS Custom neural network models

The following steps will be used for my solution:

1. Data loading - Load the data from files into the workspace
2. Data analyse - Explore the data given and find interesting features and structures
3. Data cleaning - Identify and remove incomplete datasets/entries which are not usable anymore
4. Visualization - Visualize the data for comparision of quantity ratio and also to make the data easier to overview in a glance
5. Preprocessing/Normalize - Convert the data into a structured and uniformed format (e.g. using MinMaxScaler)
6. Train/Test/Validation Split - Test the data given into the following sets
    - Train - Percentage largest part, which is used for training the neural network.
    - Validation - Smaller percentage part used to validate and adjust the model.
    - Test - Smaller percentage part, which was never seen by the model before - for final estimation/evaluation of the model https://datascience.stackexchange.com/questions/52632/cross-validation-vs-train-validate-test
7. Modelling - Create the machine learning models to work with
8. Training - Train created models
9. Model evaluation - Evalute the performance of the model and visualze the results

Sources:

Starbucks Logo	https://de.wikipedia.org/wiki/Starbucks	16.02.2021
Figure 1	https://www.researchgate.net/figure/Calculation-of-Precision-Recall-and-Accuracy-in-the-confusion-matrix_fig3_336402347	16.02.2021