Udacity Nanodegree - Machine Learning Engineer | *September 2021* | *S. Blum*

Capstone Project Proposal - Starbucks propensity modeling

1 Introduction & Background

This Capstone project is part of the Udacity Machine Learning Engineer Nanodegree. Udacity partnered with Starbucks to provide a real-world business problem for its students to learn. Yet, the data are not real-life data but rather simulated data mimicking their customer behavior. This ensures data privacy. The data allow for propensity modeling to attempt to predict the likelihood of users of the starbucks mobile web app to react to specific offers. Overall, we want to propose marketing offers on a personalized selection of users.

2 Dataset

The data consists of 3 .json-files containing simulated data that mimic customer behavior on the Starbucks rewards mobile app.

- portfolio.json - contains information about the offers
- profile.json - contains information about the customers
- transcipt.json - contains information about customer purchases and interaction with the offers

Overall, there are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount spent. In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend. Offers can be delivered via multiple channels.

3 Problem statement

The aim is to determine what user would like to receive which kind of offer. Some customers do not want to receive offers and might be turned off by them, so we want to avoid sending offers to those customers. User are distinguished using personal data like gender, age, and income.

4 Solution Statement

A Support Vector Machine (SVM) will be used as a **benchmark model**, using the kernel method to utilize non-linearity. Additionally, a Artificial Neural Netwrk (ANN) will be implemented and compared with the benchmark model.

The SVM will be implemented using sklean, whereas the ANN will be implemented using tensorflow. The data will be split into training, validation, and testing sets using a 60-20-20 split.

5 Evaluation metrics

There are two possible errors in prediction, False Positives and False Negatives.

- A False Positive prediction will most likely result in the user ignoring the marketing effort. This results in wasted effort of the company as well as the user feeling bothered by it. *Precision* is used for high False Positives.
- False Negative predictions means there is no offer sent but user would have likely used it, resulting in an wasted business opportunity. *Recall* is used for high False Negatives.

Following those thoughts the evaluation will most likely focus on the *recall* to reduce the number of False Negatives.

6 Project design

The structure of the project consists of three separate notebooks. The naming convention states their functionality.

- 1-Data_Exploration.ipynb
- 2-Feature_Engineering.ipynb
- 3-Model-training.ipynb

The data sets for data exploration and for model training will be stored on an AWS S3 bucket and loaded accordingly. All scripts will be executed using AWS Sagemaker.