Student: Katharina Mayr
Date: 2022-09-16
Email: katharina.mayr@adidas.com

Capstone Project Proposal – Starbucks Capstone Challenge

Domain Background

I will be working on the Starbucks Capstone project. Starbucks is interested in retaining customers and strengthening customer relationships by providing the best possible service and offers to customers. Starbucks wants to identify the best offer for each customer on an individual personalized offer that at the same time maximizes revenue and provide an optimal customer experience. Some customers might respond best to discount or bogo (“Buy one get one free”) offers while offers prefer to receive purely informational offers or do not want to receive offers at all.

Problem Statement

The challenge in the Starbucks Capstone project is to model or quantify the between and the person’s demographics and his or her response to a specific offer type.
For this, I will build a machine learning model that predicts how much someone will spend based on demographics and offer type.

The datasets and inputs

Three different data sources will be utilized to solve the presented challenge.
First, transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase is used. This data includes records of the user receiving, viewing, and completing offers as well as records for transactions that were made without an offer.
Second, information containing offer ids and meta data about each offer (duration, type, difficulty, reward, channels) is given. Lastly, demographic data is given for each customer (age, gender income, date of the creation of the app account).

Solution Statement

I machine learning model will be built that predicts how much someone will spend based on demographics and offer type. More precisely, the XGBoost (eXtreme Gradient Boosting)[1] algorithm with regression objective will be utilized to predict the monetary spend of customers.

Benchmark model

The results of the XGBoost model will be compared to a simple linear regression model which takes the same input features.

Evaluation metrics

To evaluate the performance of the proposed model the R^2 score as well as the RSME (root-mean-square error), two popular evaluation metrics for regression use cases, will be used.

Outline of the project design

Several steps will be necessary to solve the presented problem. First, the described data needs to be processed, i.e. it needs to be cleaned and formatted. This includes checking and excluding / imputing missing values, excluding outliers, and bringing the data in a shape to be able to analyze it further. The last point includes generating one row per transaction which holds information on the offer (if one was viewed and completed) and the consumer. Second the data needs to be analyzed by generating descriptive statistics and new features need to be generated. Potentially, the data will also be scaled. Thirdly, an XGBoost model will be trained to make predictions for the amount spent by each customer given the available data. For this, hyperparameters need to be tuned to obtain the model with the best prediction accuracy. Lastly, the results of the XGBoost model will be compared to those of the benchmark model, the linear regression model, to see whether it gives a superior performance.

---

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785