Udacity Machine Learning Nanodegree
Capstone Project Proposal

Mohamed Hesham Ibrahim Abdalla

1 Domain Background

Starbucks is an American coffeehouse that has been providing it's service both locally and internationally across the globe. Starbucks is known all across the globe with around 31,256 stores in 80 countries [3]. For this reason, Starbucks is trying to attract more customers by offering them rewards and offers through a mobile application. Certain rewards could trigger customer's interest and others could negatively affect the customers and hence not buying any products from the app. Therefore, offers must be personalized and efficiently used to grab customer's interest and attention.

In this project, I will develop a machine learning model to predict if a certain customer would be interested in a certain offer and analyze what type of customers could be interested in what type of offers and rewards.

2 Problem Statement

This project tackles the issue of identifying the customer type that could be interested in a certain offer. This is done by answering these two questions: if a customer has certain properties, would he be interested in seeing/using a certain offer? and what type of demographic group would use a certain offer?

I will use data visualization techniques to analyze the data and hence know what group of customers could pay attention or interest to a certain offer. In addition, I will develop a classifier to detect whether or not a certain customer would be interested in a certain offer.

3 Datasets and Inputs

This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. The data will be used in the context of Udacity Machine Learning Nanodegree capstone project.

datsets to be used in this project:
portfolio.json: containing offer ids and meta data about each offer (10 rows x 6 columns)

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational

- 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)

profile.json: demographic data for each customer (17000 rows x 5 columns)

- age (int) - age of the customer
- became_member_on (int) - 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: records for transactions, offers received, offers viewed, and offers completed (306534 rows 4 columns)

- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- 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

The mentioned datsets could be merged and combined to statically analyze the demographic groups that could be interested in a certain offer.

4 Solution Statement

For analyzing the demographic groups, I will use data visualization libraries (e.g. Seaborn [2] and matplotlib [1]). For classifying customers as interested or not in a certain offer, I will test different classifiers and measure their predictions against each other. The set of classifiers that could be used are: Decision Trees and Random Forests, Logistic Regression and Naive Bayes. In addition, grid search and random search will be used for model tuning. Since the data has not been analyzed yet, I would like to keep an open mind when it comes to choosing the classification algorithm. Therefore, the previously mentioned algorthim may or may not be used to get the final results depending on how well they perform against each other and the dataset distribution.

5 Benchmark Models

For bench-marking, I am going to compare my results to a standard Logistic Regressor.

6 Evaluation Metrics

Since evaluating the final model depends merely on how the data look, a decision on which evaluation metrics to be used is hard to take. Therefore, I will just list evaluation metrics that are most likely to be used when evaluating a classifier: Precision, Recall, Specificity, F1-Score, ROC and ROC AUC.

7 Project Design

7.1 Data Visualization 1

initial visualization of the dataset.

7.2 Data Cleanup

Cleaning the data by filling the empty rows and dropping any overly incomplete columns.

7.3 Data Visualization 2

A second visualization step after cleaning the data.

7.4 Feature Engineering

Transforming categorical columns to numerical ones and combining features if needed.

7.5 Model Selection

comparing the proposed models against each other to choose a final model for prediction.

7.6 Model Tuning

Using hyperparameters tuning techniques to tune the final model.

7.7 Inference and Testing

Using the previously mentioned evaluation matrices to test the model.

References

1. J. D. Hunter, ”Matplotlib: A 2D Graphics Environment”, Computing in Science Engineering, vol. 9, no. 3, pp. 90-95, 2007
2. Seaborn.pydata.org. 2020. Seaborn: Statistical Data Visualization — Seaborn 0.11.0 Documentation. [online] Available at: ¡https://seaborn.pydata.org/¿ [Accessed 29 September 2020].
3. Statista. 2020. Starbucks Stores In The World — Statista. [online] Available at: ¡https://www.statista.com/statistics/266465/number-of-starbucks-stores-worldwide/¿ [Accessed 29 September 2020].