Machine Learning Engineer Nanodegree Program

Capstone Project Proposal

Mitrofanos Ntatidis

1. Domain Background

Starbucks[1] is the world's largest coffeehouse chain. That level of success cannot be achieved without some successful marketing strategies. Artificial Intelligence has been a great ally in the marketing department in the last years[2][3].

In this project we will be focusing on the offers and promotions part of marketing. Everybody knows offers can attract customers and increase sales and machine learning can be a valuable tool to choose when and which offers to send to customers.

2. Problem Statement

In this project the problem we will be solving is predicting whether a customer will respond to an offer or not. Responding to an offer means first viewing it and then making the required transactions to complete it.

3. Datasets and Inputs

The datasets were acquired on Udacity as part of the machine learning engineer nanodegree program.

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

1 https://www.starbucks.com/
2 https://www.researchgate.net/publication/333671063_Marketing_and_Artificial_Intelligence
3 https://www.researchgate.net/publication/328580914_Artificial_intelligence_Marketing

- 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) - 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

- 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

- 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

4. Solution Statement

The proposed solution to this problem is to apply the XGBoost algorithm, which is one of top performers in Kaggle. The algorithm is going to be trained and deployed in Amazon Sagemaker and its prediction parameter is going to be binary classifier because we only have two outcomes (response or no response).

5. Benchmark Model

The proposed benchmark model is a logistic regression model from SageMaker's LinearLearner class.

6. Evaluation Metrics

Some early analysis has shown that the number of negative and positive labels is relatively even, therefore a simple metric like accuracy is enough.

7. Project Design

A planned workflow for this problem is as follows

a. Data exploration and processing

Exploring the data, find patterns, delete null values and fix any other errors

b. Feature Engineering

Convert categorical features into numeric, scale features and even create new ones that might be helpful in the classification (eg total money spent)

c. Train the benchmark and solution models

d. Test the models