Alyson Weidmann

Udacity Machine Learning Engineer – Project 3 Capstone Proposal

DOMAIN BACKGROUND

Starbucks is a global coffee chain with millions of customers worldwide. The diversity of these customers is reflected not just in their demographics, location, and coffee preferences, but in how they respond to promotions and advertisements. The Starbucks app is a platform to reach the broadest possible customer base and entice them to choose Starbucks as their caffeinator of choice. This is achieved through seamless online ordering, customer service, and special offers. Special offers, such as discounts, BOGO, and bonus points, can boost sales by encouraging a user to purchase a product that they otherwise might not have.

PROBLEM STATEMENT

The goal is to develop a model to identify which special offers and promotions might appeal to certain users based on their demographics, purchase history, and a detailed analysis of the types of offers they have responded to in the past. A successful model will be able to correctly match a promotional offer to a customer likely to respond to it, ideally resulting in a positive business outcome (such as a sale, increased customer retention, etc.).

DATASETS AND INPUTS

There are 3 main sources of data from which the model will be built:

1.) portfolio.json: Contains offer IDs and meta data about each offer (duration, type, etc.)

Contains the following fields:

- id (string) - offer ID
- offer_type (string) - type of offer (BOGO, discount, etc.)
- difficulty (int) - minimum required to spend
- duration (int) - days offer is open
- channels (list of strings)

2.) profile.json: Demographic data for each customer

Contains the following fields:

- age (int) - age of customer
- became_member_on (int) - date when customer created app account
- gender (str) - customer gender (M/F/O)
- id (str) - customer ID
- income (float) - customer's income

3.) transcript.json: Records for transactions, offers received, offers viewed, and offers completed. Contains the following fields:

- event (str) - record description (transaction, offer received, etc.)
- person (str) - customer ID
- time (int) - time in hours since start of test (at time t=0)
- value (dict of strings) - offer ID or transaction amount depending on the record

SOLUTION STATEMENT

The proposed solution to the above problem statement is to first use Exploratory Data Analysis (EDA) to clean the data and identify patterns or trends between customers and which offers result in a transaction or completion that would not have otherwise occurred. I will then build a machine learning model to predict which type of offer will be most likely to be successful with each user.

BENCHMARK MODEL

The simplest place to start would be a binary classifier to predict whether an offer will "convert" into a completion or a transaction. A "successful" offer, resulting in a sale/completion, will be class 1 while all other offers will be class 0. We can start with a single data set where offer type is included as one of the features. An ensemble of classifier models can be explored, including logistic regression, RandomForest, XGBoost, and Support Vector Classifier. Based on model performance and the results of my EDA, I will explore whether segmentation of the data (e.g. by customer attribute, offer type, income strata, etc.) yields improved predictions.

Ironically, modeling and data analysis can be unpredictable. Sometimes a model you initially thought would perform well, or reveal useful information, does not do so at all. I may find, once I dig into the data, that a classification approach does not work at all. In that case, I will explore alternative models that may tell me useful information about how promotional offers can best reach customers — a regression approach to the ideal discount amount, a survival analysis to identify the best offer durations, or a clustering approach to segment customer populations are all options I would consider as a Data Scientist at Starbucks.

EVALUATION METRICS

For a binary classifier, finding good metrics to evaluate model performance depends a lot on the class imbalance of the data set. For example, accuracy measure is not a good metric for model performance with a class imbalance of 99:1 — because the model can predict 100% negatives and still be 99% accurate! For highly imbalanced data, I would look at the Receiver Operator Characteristic area under the curve, to visualize the ratio of True Positive to False Positive predictions. I would also look at the confusion matrix, Precision and Recall (F1 score), and the prediction probability distributions to see how data is getting classified (or mis-classified). For more balanced classes, metrics such as accuracy, Mean Square Error, and Log Loss are all good metrics to determine model performance.

PROJECT DESIGN

I will complete the project according to the following steps:

1. Load the data and any necessary libraries into the workspace.

2. Clean and restructure the data so that I can perform EDA. This may include handling outliers and null values, joining tables, renaming columns, etc.

3. Perform exploratory data analysis (EDA) on the cleaned data to identify correlations between customer features, offer features, and “successful” offers.

4. Create benchmark model(s) to identify the promotional offers most likely to convert (binary classifier).

5. Evaluate model performance according to the aforementioned Evaluation Metrics (AUC, log loss, confusion matric)

6. Tune hyperparameters and refine model as needed based on the evaluation results. If multiple models are explored, they will be compared and the best one will be selected.

7. Deploy best model and write a blog post to summarize.