Machine Learning Engineer Nanodegree

Udacity

Capstone Project Proposal

Omar Hegazy

Domain Background

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Machine Learning algorithms use statistics to find patterns in massive amount of data, and data here encompasses a lot of things from numbers, words, images, clicks. If it can be digitally stored it can be fed into machine learning algorithms.

Machine learning has become increasingly important over the years. The need to get insights about the behavior of a customer based on the data collected over the years is one of the most trending uses of machine learning nowadays. My project is about Starbucks, Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. As the largest coffeehouse in the world, Starbucks is seen to be the main representation of the United States' second wave of coffee culture. Since the 2000s, third wave coffee makers have targeted quality-minded coffee drinkers with hand-made coffee based on lighter roasts, while Starbucks nowadays uses automatic espresso machines for efficiency. The company operates 30,000 locations worldwide in over 77 countries.

Starbucks means to attract and retain customers, Starbucks leverages a rewards program that honors regular customers with special offers not available to the standard customer. For this project, we'll be combing through some fabricated customer and

offer data provided by Starbucks and Udacity to understand how Starbucks may choose to alter its rewards program to better suit specific customer segments.

Problem Statement

Starbucks wants to find a way to give to each customer the right in app special offer. There are three types of offers

1- Buy One Get One (BOGO)
2- Classic Discount
3- Informational (no real offer)

Our goal is to analyze historical data about app usage and offers orders to develop an algorithm that associates each customer to the right offer type We can measure the performance of the algorithm by measuring its accuracy.

Dataset

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.

There are three datasets:

profile.json

Rewards program users (17000 users x 5 fields)

- gender: (categorical) M, F, O, or null
- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

portfolio.json

Offers sent during 30-day test period (10 offers x 6 fields)

- reward: (numeric) money awarded for the amount spent
- channels: (list) web, email, mobile, social
- difficulty: (numeric) money required to be spent to receive reward
- duration: (numeric) time for offer to be open, in days
- offer_type: (string) bogo, discount, informational
- id: (string/hash)

transcript.json

Event log (306648 events x 4 fields)

- person: (string/hash)
- event: (string) offer received, offer viewed, transaction, offer completed
- value: (dictionary) different values depending on event type
- offer id: (string/hash) not associated with any "transaction"
- amount: (numeric) money spent in "transaction"
- reward: (numeric) money gained from "offer completed"
- time: (numeric) hours after start of test

Solution statement

What I'm going to do is to use machine learning techniques to build a model to predict which offer is the best for customers. What that mean is to is to predict the type of offer the customer most likely to use.

I will develop a model for each offer and combine the results.

V. Benchmark Model

As the benchmark result, we can deduce the current Conversion Rate of the offer received. Leaving out the informational offers, which have no real “conversion”

VI. Evaluation Metrics

I will use the accuracy of the model to the performance of the algorithm. By using the same metric, we are able to quantify and compare both the benchmark and the final models.

$$accuracy = \sum_{i=1}^{n} \frac{p_i}{n}$$

p_i^c = correctly classified instance
n = total number of offers sent

VII. Project Design

The idea of workflow for approaching the solution includes several machine learning techniques, following the guideline sections below.

a. Data loading and exploration

Load files and do some data visualization to get better understanding for the distribution and characteristics of the data

b. Data cleaning, pre-processing , Feature engineering and data transformation

Having analyzed the data, handle data to fix possible issues found. Preparing the datasets to deal with the problem stated and feed the algorithms. The transcription data must be structured and labeled as appropriate offer or not.

c. Split the dataset into training, validation, and testing sets

Prepared three datasets containing different record. The largest dataset is employed to train the model, while the validation set, to evaluate the models during the training. The testing set contains That never used on the model before, so the results of the model can be considered reliable by using this dataset, it will be possible to measure the final performance and compare the results of the trained models.

d. Defining and training a machine learning model using Decision tree, KNN, Logistic regression and random forest

e. Evaluating and comparing models performances the comparison is based on the accuracy of each model to get the most suitable one for the job.