Starbucks Capstone Challenge

Udacity Machine Learning Engineer Nanodegree Program - Capstone project
Essi Lehtola, April 2020

The project background

In this Udacity Machine Learning Nanodegree Capstone project I am going to analyse and process data that mimics customer behavior on the Starbucks rewards mobile app.

Starbucks is the biggest international coffeehouse and roastery chain in the world and has coffee shops in over 60 countries. In the app the company is sending out different types of offers (for example informational offers, discounts and ‘buy one get one’ offers) and my goal is to use the data provided to determine how the customers react to different offers.

Problem Statement

I decided to find an answer to a real business question: is a customers going to buy something in the influence of a certain offer based on past customer behaviour?

I will build a machine learning model that can answer to that problem. When customer and offer details are given to the model it will predict if the customer will make a purchase on Starbucks or not.

Datasets and Inputs

There are three simplified data tables provided by Starbucks:

1) Portfolio.json

Contains 6 columns and 10 rows of offer related data.

2) Profile.json

Contains all customer related details in 17 000 rows and 5 columns.

3) Transcript.json

Contains events that happened during the testing period in 306 534 rows and 4 columns.

This is the most important table but also hardest to preprocess. Events column contains relevant details such as if the action was a transaction or if an offer was received or viewed. Value column on the other hand provides data like offer ids and transaction amounts depending on event. In this table there are of course person id's and timestamps as well.

Solution statement

When the question is will a customer make a purchase during the influence period of an offer the solution can be found in this way:

- I familiarize myself with the data available and gather and combine as much customer and offer details as possible

- I transfer that information to a machine readable form so that it can be used as an input for a machine learning model and also separate the data to train and test sets

- I will test few models to find a good one and the model will find patterns on data. I might need to tune the model a bit we to find best possible prediction rate. The model can then be used to predict if an offer has an influence on someone or not.

Benchmark model

Since this is a classification task I will test how the gathered and preprocessed data works with a simple Decision Tree Classifier. That will tell us the percentage of the correct labels so that it is easy to compare other models to it.

Evaluation metrics

I will evaluate if the models work with percentages of correctly predicted labels. Along with that I will also use confusion matrix that shows how many of the labels were predicted correctly and how many went wrong and how (false negatives or false positives).

Project Design

The workflow will be following:

1. Gathering data from different sources

   First, I will have upload the data and analyze it a bit so that all relevant details can be found. I also need to combine the data from different sources so that I have good amount of relevant information for a model.

   We have customer ages, genders, incomes and dates when they became members but also actions they have made during the test period. We also have information like offer durations and difficulties.

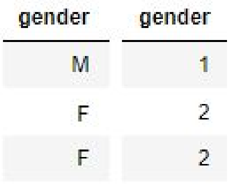
The hardest part is to collect information for example on how many offers customer has viewed and completed and how much money they have spent to Starbucks. The offers are different and the offer related customer actions are collected as records containing timestamps and event details so collecting data together might be tricky. To be able to see if an offer really had an effect on a person it requires quite a lot of preprocessing

2. Preprocessing data for machine readable form

The data collected needs to be transferred to a machine readable form. For example empty values are not really usable in a model, but I am pretty sure there is enough data to remove all rows with null values.

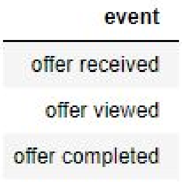
We also need to change all text to numbers. To do that we can use encoding (counting all unique values and replacing each of them with a number) or one-hot-encode them (move all options to columns and give a value 1 if the original value matches the column name).

Encoding

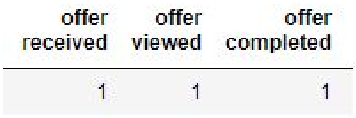


gender	gender
M	1
F	2
F	2

One-hot-encoding



event
offer received
offer viewed
offer completed



offer received	offer viewed	offer completed
1	1	1

To make the data easier for a model to use I think it will be beneficial to scale/normalize numbers as well.

The last step here is to separate the data to training and testing sets.

3. Finding a good machine learning model

As a last step of the whole process, I need to find a good model to use to predict the result as accurately as possible.

I will use a Decision Tree Classifier as a benchmark. After that we can compare the results to results of other models and even try to improve the prediction rate using hyperparameter tuning. It is easy to see if a model performs better than others by looking at the percentage of the correctly labeled rows.

In the end we should have a model that predicts the result of a combination of a customer and an offer. For example, what happens if a 60 years old lady gets a discount - is she going to use it or not.