MONDAY, 15 JUN 2563 9:59 PM GMT+7 • SOI PHAHON YOTHIN 30, BANGKOK

Machine Learning Engineer Nanodegree, Udacity

Starbuck Campaign Challenge Proposal

Wasurat Soontronchai

16 June 2020

Proposal

1.) Domain Background

Starbuck is worldwide coffee company from America. Currently the company have more than 24,000 stores across more than 75 countries.
On every day, Starbucks promote their offer to customer via mobile application platform.

Their offer can be advertiment of either a discount, BOGO (Buy One Get One Free) or do not receive any offer in some period. Each offers has period of acceptance for customer before its expire. For example, 7 days of offer validity.

So, how these data can be valuable to the company ?
If company can send the offers to the target customer, it would help them to generate more benefit from higher offer accepted rate !

This come to my problem statement for this capstone project. Can we predict whick kind of customer tend to use or accept the certain type of offer ?

I select this problem because Starbuck is my every morning coffee. Discount and BOGO also are campaigns where I normally receive from email and sometimes in message this seem to be more realistic problem for me as a premium member of the Starbuck application.

2.) Problem Statement

As mentioned in domain background, the problem statement is to predict wheter customer will accept to an offer from given demographics informations and type of campaign (BOGO, discount etc.).
The problem will be **Binary Classification**

> Postive class : customer accept the offer (1)
> Negative class : customer do not accept the offer (0)

3.) Datasets and inputs

Total data available come in .json format 3 files:

1. portfolio.json : offer ids and detail of offer (durations, type of offer, etc.)

id (str) : offer id (This columns will merge with value column in trasnscript.json)
offer_type (str) : BOGO, discount, intormational
difficulty (int) : minimum effort spent to complete offer
reward (int) : reward for completing offer

duration (int) : limit time for offer before expire (days)
channels (str list) : channel of communication in offer
Total 10 rows x 6 columns

2. profile.json : demograhic of customer
age (int) : age of customer (Spot that age is 118 when income is NaN)
became_member_on (int) : start date of account creation in application
(Year - Not uniform distributed with the timeframe)
(Month - Almost uniform distributed with the timeframe)
gender (str) : gender of customer (M, F with 1.5% amount of 'O' for rather M or F)
id (std) : customer id (This column will merge with person column in transcript.json)
income (float) : estimate customer's income
Total 17,000 users x 5 columns

3. transcript.json: transactions record, offer recieved , viewed, offer completed
event (str) : transacton description
(2 groups of event 'transaction' and 'offer' in offer there are 'recieved', 'viewed' and 'completed')
person (str) : customer id (match on profile.json)
time (str) : hours of start the campaign test
value (str dict) : amount of offer id or transaction
Total 306,648 rows x 4 columns

4.) Solution Statement

Firstly, combine the 3 files to make it easier for analysis in later stage. Table will have all provided information about customer (demographic), offer details, transcation data and the last one is target prediction which is offer acceptance from customer (0 or 1)
Then use all feature execpt offer acceptance from customer as input features of binary classifiers model.

With multiple features (expect to be with categorical combine with numeric features, below is the list of classifier models : 2 groups with 4 models
Tree-Base classifier : To get benefit of automatic feature selection and performance onimbalanced datasets.

- DecisionTreeClassifiers
- RandomForestClassifier

Boosting algorithms classifier : Both of them have good model performance but need to ware of overfitting character.

- AdaBoostClassifier
- XGBoostClassifier

5.) Benchmark Model

To benchmark the model performanace, I would select the simple classification model as **logistic regression** as a benchmark model with their simple to set up, speed in learning and a chance of providing decent results, this should be better base line model than the naive model (assume all users accept the

offer). By comparing the evaluation metrics which will be mentioned below) of logisitic regression against the list of complex classifer model, this would help to lead the model development and model selection go in the right way.

6.) Evaluation Metrics

As the problem is binary classification, I choose

- Accuracy : To measure over model correction
- F1-score : To consider both precision and recall for model accuracy.

Training time also consider as additional evaluartion metrics.

7.) Project Design

Start with fundamental process of data science project as below.

- Cleaning the data

1. portfolio.json :

Perform on multilabel binarizer on '**channels**' columns.
Perform one-hot-encoder on '**offer_type**' columns.

2. profile.json :

Change the '**became_member_on'** to be date time formatting
Detect the missing value on '**gender**' alogn with '**income**' and drop it out.
Extract '**year**' from "**became_member_on**" and perform one-hot encoding
Perform label encoder on '**gender**' after remove the 'O' value.

3.transcript.json :
Drop out all 'person' value that do not exist in profile.json dataframe
Perform one-hot-encoding on 'event'
Seperate value columns into ' offerid
Segregate offer and transction data
Extract only event without 'transaction' value
**For the target prediction, we will count 'completed' as an positive class the other in event will considered as negative class. The remaining data such as details of offer, demographic of customer and so on will be used as input data of the model.*

- Perform data exploratory analysis to get more insight from the data.

Start on single cleaned dataframe for EDA. Check all statistic value of income, age and starting member date time (year & month). Then pefrom bivariate analysis on pair of features (eg . Income by year, Income by age range, Duration of member ship with offer acceptance rate)
Then consider to combine 3 dataframe into single dataframe by merge on offerid and customerid together with the details of offer for further model development in later stage.

- Feature Engineering

Transform skewed continuous features
Nomalizaing on numerical features

Split the data into train and test (80:20) and might consider to do the cross validate for 20 folds (selectively)
Drop out all duplicate data for offerid.
Feature Importance might be selected on this step by using tree-based model.

- Model development (Trial all binary classifier and tuning)

Perform model training and check the validation results on the splitted data. Start with default setting of all selected 4 model then compare against the baseline model (Logistic Regression). Then compare the result on all of them for selection. After got the selected model, model tuning shall be perform with random search CV. Then the selected model with all hyperparameters after perform the model tuning will be achived.
All of these process shall be iterable process.

reference

1. Pros and cons of various Machine Learning algorithms used in Classification : https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6
2. Start with the baseline model : https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa
3. SKlearn model selection library : https://scikit-learn.org/stable/model_selection.html#model-selection
4. EDA practical guide : https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9
5. Starbuck Company Profile : https://en.wikipedia.org/wiki/Starbucks
6. Ensemble Method : https://scikit-learn.org/stable/modules/ensemble.html
7. Decision Tree : https://scikit-learn.org/stable/modules/tree.html