Starbucks Capstone Challenge

Project Proposal

Domain Background | Problem Statement | Datasets and Inputs | Solution Statement

Benchmark model | Evaluation metrics | Project Design

Sooyeon Won

21. January 2021

won.sooyeon@live.com

Domain Background

This project is derived from the field of Customer Relationship Management (CRM) in Starbucks. One of the main concerns for CRM field is to interact with not only current customers, but also previous and potential customers, so that a company can boost its business relationship with the customers, and eventually expect its continuous sales growth. Nowadays, there are various channels that a company can manage its customers such as using mobile application, company's website, emails, and so on. [1] Recently I am employed as a CRM data analyst at a retail industry. This could be one reason, I chose this project, which improves my future professional skills further.

Starbucks Corporation is a multinational coffee company, whose headquarter is in Seattle. It serves all kinds of coffee-related products.[2] Starbucks is often evaluated as the leader on the highly substitutable market. Its customers realise that they not solely purchase a coffee at Starbucks, but also enjoy the unique atmosphere at Starbucks store. Moreover, Starbucks is well-known for its own solid online marketing program to manage its global customers. [3]

As mentioned in the provided business case, one of its marketing campaigns is to send out an offer to customers through its mobile application. An offer can be either just an advertisement for a certain beverage, or a coupon-type offer such as a 'discount' or 'buy 1, get 1 (BOGO)'.

- Informational offer
  Its main purpose is providing an (updated) information about products to customers. There is no reward, neither a required spending that a customer is expected to spend.

- Discount offer
  Customers get monetary rewards which is equivalent to a certain proportion of the amount of spending.

- Buy 1, Get 1 (BOGO)
  Customers should spend certain amounts (threshold) so that they can receive monetary rewards, which is equal amount of the cut-off amount.

Each offer is valid for certain number of days. The validity periods are different from offers.

Problem Statement

Not all of application users obtain the offer, and the type of the offers can be different based on customer purchasing patterns. As a CRM analyst, we should maximize a return-on-investment (ROI), since all marketing campaigns have related-costs.

To be specific, if a customer segment which is likely to react regular offers, so that the offers make the customers regularly purchase drinks at our chains, then our marketing events should focus on the customers, rather than customers who show frequent purchasing patterns regardless of product offers.

Additionally, if a customer feels reluctant to receive advertisement offers, this might be able to result in losing users or lead to decrease customer retention rate in the long run. These types of customers should be removed in advance from the “offering-list”.

Therefore, the major challenges in this project are to classify the most appropriate offer for individual customers based on their individual characteristics which is represented by each purchasing pattern. In addition, we should identify the customers who feel reluctant to receiving offers or who are not influenced by offers.

Datasets and Inputs

To solve the business issues, Starbucks provides 3 datasets:

Profile

- Rewards program users
- Format: json
- Size: 17,000 users x 5 features
- Included Features:
    - 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

- Offers sent during 30-day test period
- Format: json
- Size: 10 offers x 6 features
- Included Features:
    - 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)

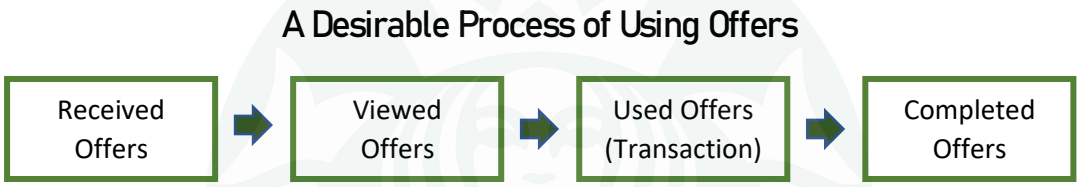
Transcript

- Event log, basically transaction records
- Format: json

- Size: 306,648 events (transactions) x 4 features
- Included Features:
    - 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 (experiment)

Solution Statement

Note that it is appropriate when each stage in the 'event' field of "Transcript" dataset occurs in consecutive order.



However, the prior stages are not necessarily required to complete the offers. For example, a customer who did not view the offer, but the customer could complete to use the offer, though. Based on this fact, I will filter such customers who are not influenced by offers, because their purchases are independent from the fact whether they received the offers.

After then, I will conduct the analysis using frequently used Classification models such as Decision Tree- , Random Forest-, XG Boost Classifiers. Additionally, CatBoost and LightGBM Classification models will be applied as complimentary supervised learning approaches. Since I create multiple supervised learning models, I try combining them all together into a custom ensemble model.

Then I will compare the performances of each model. Finally, I will select one which showed the best performance. In this analysis, 'offer type' is set as a target, and other features such as demographic features and several purchasing characteristics will be used as inputs.

Benchmark model

Although the models which show less performance than the model with best performance could be considered as benchmark models, I designated a "multinomial logistic regression" as a benchmark model for this project to compare my solution objectively.

Multinomial logistic regression is a modified form of logistic regression used to predict a target variable have more than 2 classes. Basically, it uses the SoftMax function instead of the sigmoid function.

Evaluation metrics

The performance of individual models can be measured according to various evaluation metrics. As mentioned earlier, I will solve this business case by applying machine learning classification models which is a type of supervised machine learning techniques. This is because we are trying to figure out the best type of offers for each customer.

I will firstly compute Accuracy, Precision, Recall, and F1-Score. "Accuracy" however is not appropriate to be used as an evaluation metric for predictive models when classifying in predictive analytics. As "Accuracy Paradox" indicates, a simple model might be able to achieve a high score of accuracy but be too crude to be useful. [4] On the other hand, F1-Score is computed with the prediction and recall of the test.

$$F1\ Score = 2 * \frac{Prediction * Recall}{Prediction + Recall}$$

Since the F1-Score indicates a weighted average of the prediction and recall values, it takes false positive as well as false negatives into account. F1-Score reaches its maximum (best) value at 1 and its minimum (worst) value at 0. F1-Score is often useful in comparison to accuracy, especially if each class is not evenly distributed. Thus, I will decide a model with the best performance according to its F1-Score. [5]

Evaluation Metrics	Meaning
Accuracy	The proportion of correctly predicted cases.
Precision	The number of true positive cases over the number of true positive - and false **positive** cases.
Recall	The number of true positive cases over the number of true positive - and false **negative** cases.
F1 - Score	It is the harmonic mean of precision and recall. Higher scores indicate better performance.

Project Design

My approach to the solution is as follows.

1. Introduction

- Brief explanation of current business case
- Clarifying the final goal of this project

2. Data Preparation

- Data Pre-processing
  - Deleting duplicates, fulfilling missing values
  - Missing values in the 'age' column will be replaced with the median value of ages whose incomes are identical
  - Outliers: Figure out whether the extreme values are realistic. If they are unrealistic, outliers will be removed by taking the datapoints from 5^{th} percentile to 95^{th} percentile[1]
  - Adjusting datatypes, Scaling the data values (if needed)
- Splitting data points into training, (validation,) and testing sets

3. Data exploration

- Exploring the current business situations
  - The Change of traffics during the test period
  - The Trend of Sales Amount during the test period
  - Distributions of each offer categories
- Exploring demographic factors of current customers
  - gender, age, income, number of days as a Starbucks member
- Understanding of Starbucks offers based on each offer-type
- Exploring Purchasing patterns of current customers
  - RFM Customer Segmentation Analysis
  - Investigating the customers who are not influenced from offers
  - Investigating the customers who follows the desirable purchasing process
  - Investigating the customers who shows other unexpected purchasing patterns

4. Data Analysis

- Creating a custom ensemble model by combining all models
- Tuning the supervised learning algorithms using a randomized search
- Training classification models
- Testing the fitted models with the test dataset
- Computing evaluation metrics for each model
- Evaluating and comparing the model performances

5. Conclusion

- Presenting the results of predictions
- Suggesting a business solution from my perspective based on the outputs

[1] The percentile can be changed, depending on the size of dataset.

References

[1] https://en.wikipedia.org/wiki/Customer_relationship_management

[2] https://de.wikipedia.org/wiki/Starbucks

[3] https://www.ukessays.com/essays/marketing/starbucks-the-company-philosophy-marketing-essay.php

[4] https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b

[5] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/