Proposal of Starbucks Project

1. Background

The project focus on the customers response for the different types of Starbuck's offers. The datasets include the customer's profile, offer's information and the actions of the customers. The objective is to establish a model to predict a specific customer's response to a type of offer.

2. Statement

In this project, I cleaned and transformed the datasets firstly. Then classified and analysed the customer clusters. Finally, a set of machine learning models were developed and compared. The specific steps are following:

1. Analysing and cleaning the datasets of portfolio, profile and transcript;
2. Transforming the fields of datasets if necessary;
3. Clustering the customers and analysing the response of different customer groups;
4. Developing benchmarking model to predict the possibility of customers using an offer;
5. Developing new models to solve the problem and comparing.

3. Datasets

The datasets include the offer's information (portfolio.json), customer's information (profile.json) and events (transcript.json). They are provided by Udacity in workspace. The data schema of them are following:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)

profile.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account

- gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
- id (str) - customer id
- income (float) - customer's income

transcript.json

- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

4. Project Solution

The project's solution including four main parts:

- Data Exploration: inspecting the original datasets;
- Data Clean and Transformation: cleaning the datasets and transforming their fields;
- Modelling: developing machine learning models;
- Evaluation: evaluating models' performance.

5. Benchmark Model

A regression model will be developed before the classification models as a benchmark model. The regression mode is used to predict the possibility of a customer view or use an offer. Then we will demonstrate how the classification models can solve the problem in more effective way.

For the classification models, the Logistic Regression models are set as benchmark models for each type of offer.

6. Evaluation Metrics

To compare the models' performance, three metrics will be used, which are Accuracy, Precision and Recall. They are calculated by following formulas:

$$Accuracy = \frac{\text{True Positives } + \text{ True Negatives}}{\text{sample size}}$$

$$Precision = \frac{\text{True Positives}}{\text{True Positives } + \text{ False Positives}}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives } + \text{ False Negatives}}$$

Finally, the models will be evaluated by the metrics above in different scenarios and the best one will be recommended.

7. Outline of Project Design

The core part of solution is developing a set of machine learning models to predict customer's response to offers. Because customers may have different actions response to different type of offer, the models will be developed for each offer types separately.

I will try to transfer the problem of project to a classification problem, then clean and transform original datasets before modelling. The models of machine learning applied in this project include Random Forest Classifier, Logistic Regression and SVM.