Proposal

An application of Random Forest Classifier to Starbucks' direct marketing system

Saverio Pertosa

October 2020

1 Domain background

This project approaches a field that is at the heart of modern retails: direct marketing.
With the boom of data availability, the chances of applying machine learning techniques to improve the return of marketing campaigns are increasing.
The data for the project is provided by Starbucks, one of the largest and most famous coffee-house franchising, whose direct marketing campaigns consist of sending customized offers to its customers through various communication channels.

2 Problem statement

The objective of the project is to predict accurately what kind of offer have a higher chance to be completed.
Once an offer has been sent, customers behavior can be clustered according to their subsequent actions:

- Offer viewed and completed
- Offer viewed but not completed
- Offer not viewed but completed
- Offer not viewed and not completed

The most pressing problem for Starbucks is of course identifying the offers triggering the first kind of reaction.

3 Dataset and inputs

Starbucks provided 3 .json files, containing simulated data mimicking customer behaviour.

- **profile.json** - Customers' data *(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)

  - 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

4 Solution statement

My proposal consists of the application of a Machine Learning technique called "Random Forest Classifier" to predict the likelihood that an offer, characterized by a given set of features and customer's data, will be completed a certain offer sent to them digitally.
In the phase of "feature engineering", a new variable will be computed, called *Purchase* - a count of the total number of transactions that a customer completed outside the context of an offer. This will be used as a proxy of the "demand" for Starbucks' products of a user.

5 Benchmark model

The "Random Forest Classifier" will be compared to a "Decision tree" model, that will not use the *Purchase* variable and whose hyperparameters will not be optimized.

6 Evaluation metrics

For each model, a few performance metrics have been computed on the test set. They will be described in more details inside the project.

- **Accuracy**
- **Balanced accuracy**
- **Recall**
- **Precision**
- **Confusion matrix**
- **ROC-AUC**

7 Project design

The project will be developed on a Jupyter Notebook, that will follow this outline

- **Data loading and exploration** - Loading the provided files and doing some first exploratory analyses;
- **Data cleaning, feature engineering, and first analyses** - Handling of missing values, JOINs between data sources, Dataframe transformation, and feature engineering;
- **Model** - Training of main models and benchmark;
- **Final remarks** - Presenting the final results, along with discussions on some insights derived from the analyses.