

We will be exploring the Starbuck’s Dataset which simulates how people make purchasing decisions and how those decisions are influenced by promotional offers. We want to make a recommendation engine that recommends Starbucks which offer should be sent to a particular customer.

**Business Understanding**

There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount spent. In an informational offer, there is no reward, but neither is there a required amount that the user is expected to spend. Offers can be delivered via multiple channels.

We are interested to answer the following two questions:

1. Which offer should be sent to a particular customer to let the customer buy more?
2. Which demographic groups respond best to which offer type?

**Dataset Description**

The data is contained in three files:

* portfolio.json — containing offer ids and metadata about each offer (duration, type, etc.)
* profile.json — demographic data for each customer
* transcript.json — records for “transactions”, “offers received”, “offers viewed”, and “offers completed”

Here is the schema and explanation of each variable in the files:

**portfolio.json**

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

**profile.json**

* age (int) — age of the customer
* became\_member\_on (int) — the 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

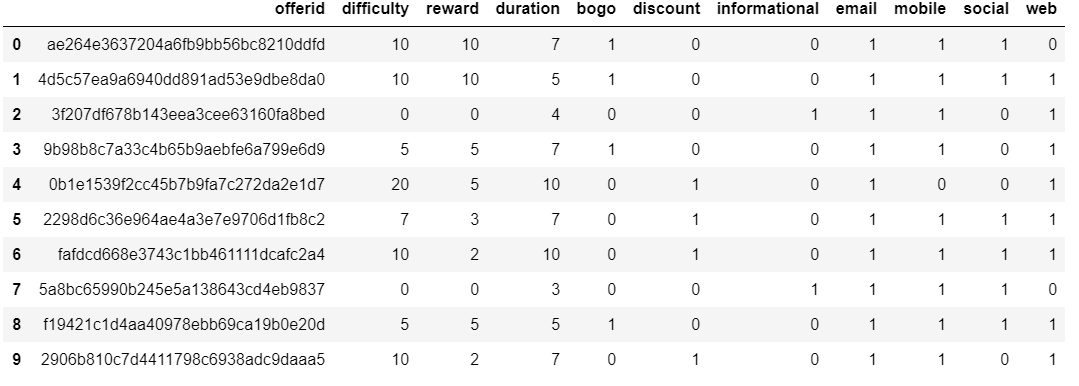
**transcript.json**

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

**Data Wrangling**

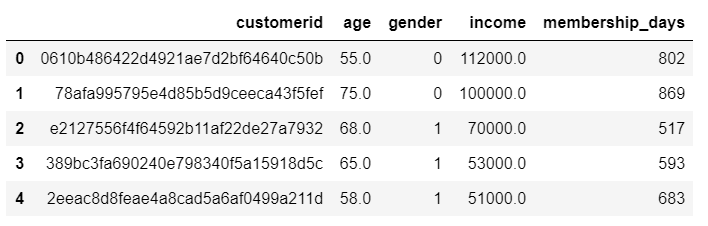
PORTFOLIO:

1. Change column ordering
2. Change the name of the 'id' column to 'offerid'
3. Generate One-Hot encoded columns from 'offertype' column (and replace original)
4. Generate One-Hot encoded columns from 'channels' column [with multiple labels] (and replace original)



PROFILE:

1. Change column ordering
2. Change the name of the 'id' column to 'customerid'
3. Convert missing value encoded as 118 to N/A in the "age" column
4. Remove customers with N/A income data, N/A gender data and unspecified gender
5. Transform the 'became\_member\_on' column to a datetime object
6. Perform Encoding on the "gender" column



TRANSCRIPT:

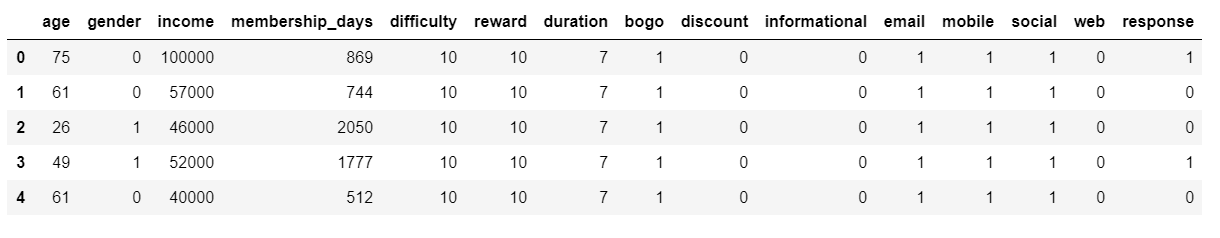
1. Change the name of the 'person' column to 'customerid'
2. Remove customer id's that are not in the customer profile DataFrame
3. Extract the offerid from the value column into a separate column
4. Only keep the following events: offer viewed, offer completed
5. Change column ordering

Now what we want to do is to create a column named as a response. For a particular customer, if a particular offer was viewed and then the offer was completed, the value in the response column for that particular offer should be one. If a particular offer was only viewed and not completed, then the value in the response column for that particular offer should be zero. This will signify the response of an individual towards different offers. Note that the sequence of viewing the offer and then completing the offer should be maintained.

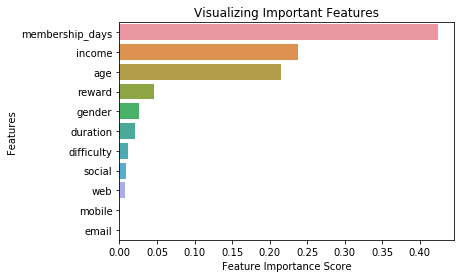


**COMBINED DATASET**

Starbucks Dataset = Profile + Portfolio + Transcript



**RESULTS**



**CONCLUSION**

The problem that I chose to solve was to build a model that predicts whether a customer will respond to an offer.

After performing the exploratory data analysis and data wrangling, I ran a Machine Learning algorithm, specifically Random Forest to give prediction if a customer will respond to the offer (complete the offer) or not. After performing hyperparameter tuning, I was able to achieve an accuract of 67.28%.

I was also able to extract the important features (Feature Importance: refers to a numerical value that describes a feature's contribution to building a model that maximizes its evaluation metric). The analysis suggests the following top five features based on their importance:

1. Membership Days Reasoning: People who are starbucks member for very long are more loyal and more likely to respond to the offers.
2. Income Reasoning: People who have a comparatively high income are more likely to respond to the offers.
3. Age Reasoning: Age plays an important factor in deciding as to how likely a person will respond to the offers.
4. Reward Reasoning: The more the reward, the better the chance of an individual responding to the offers.
5. Gender Reasoning: In the group of people who responds positively to the offers, the contribution of female members is more as compared to the group of people who do not respond to the offers.

I will also like to add that the column mobile and email have a negligible contribution for the simple reason that the above two options are present for all kind of promotions (offers) and thereby are not providing any additional information.

**FUTURE IMPROVEMENTS**

* It is possible to build a machine learning model that predicts how much someone will spend based on demographics and offer type.
* Perform normalization/scaling.
* Test additional machine learning models.
* Make a web app