

**Business Understanding**

* 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.
* 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 questions:
  1. Which offer should be sent to a particular customer to let the customer buy more?
  2. What is the impact of the customer demographic on the offer completion?
  3. What is the impact of the membership duration on the offer completion?
  4. Which are the best channels for that leads to the most offer completion?

**STRATEGY**

* First, I will wrangle and combine the data from offer portfolio, customer profile, and transaction. Each row of this combined dataset will describe the customer demographic data, offer's attributes, and whether the offer was successful. In this, I will take into account the possibility that a person may have completed the offer without even actually viewing the offer. Such outliers will have to be taken care and only those transaction will be considered where the user have actually viewed the offer and then completed the offer.
* Second, I will create a model (I am thinking Random Forest) that will be able to predict the offer success based on the provided customer demographics and the offer attributes.
* Third, I will obtain the important feature columns that influences the success of an offer and use the visualization of the data to answer the questions that were framed above.

**Metrics**

I will assess the accuracy and F1-score of the model. Accuracy measures how well a model correctly predicts whether an offer is successful. However, if the percentage of successful or unsuccessful offers is very low, accuracy is not a good measure of model performance. For this situation, evaluating a models' precision and recall provides better insight to its performance. I chose the F1-score metric because it is "a weighted average of the precision and recall metrics".

**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

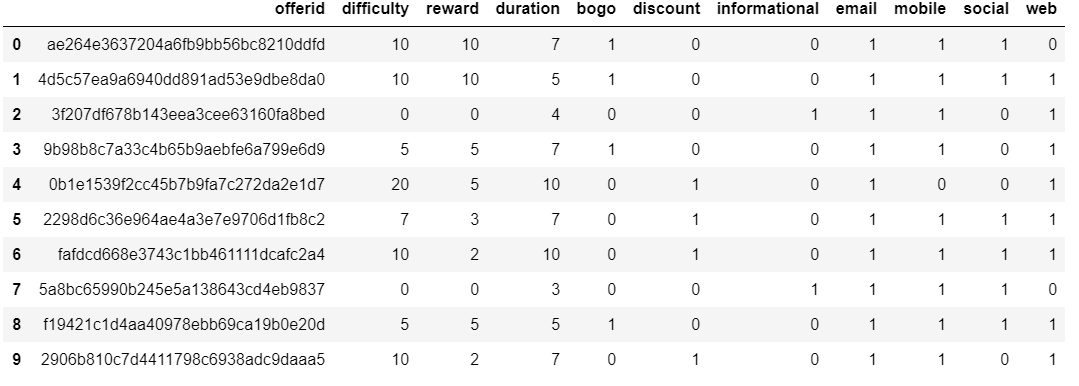
**transcript.json**

* 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

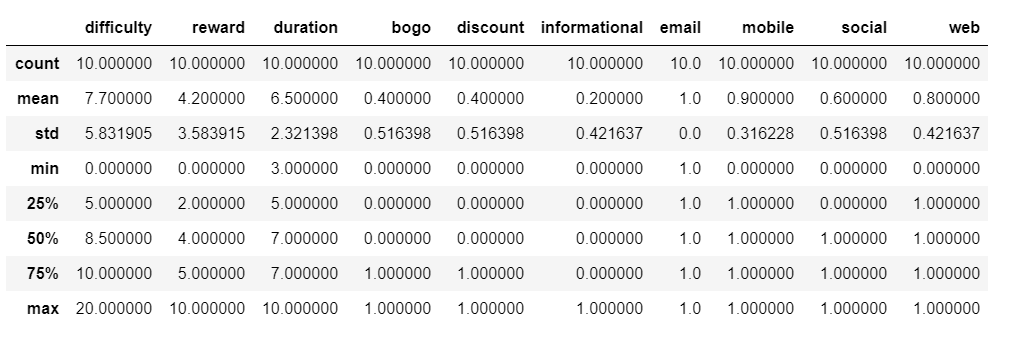
**Data Wrangling**

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)

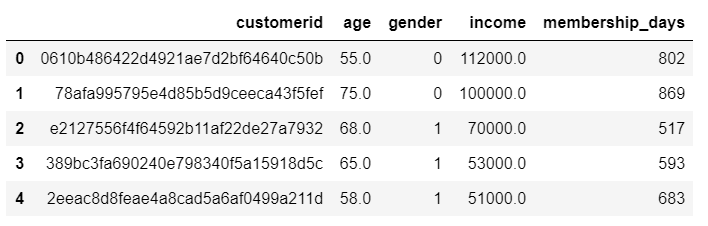


The following image shows important statistics about the dataset:

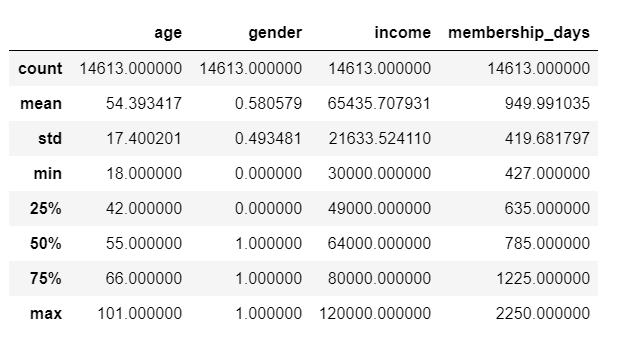


PROFILE:

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



The following image shows important statistics about the dataset:



TRANSCRIPT:

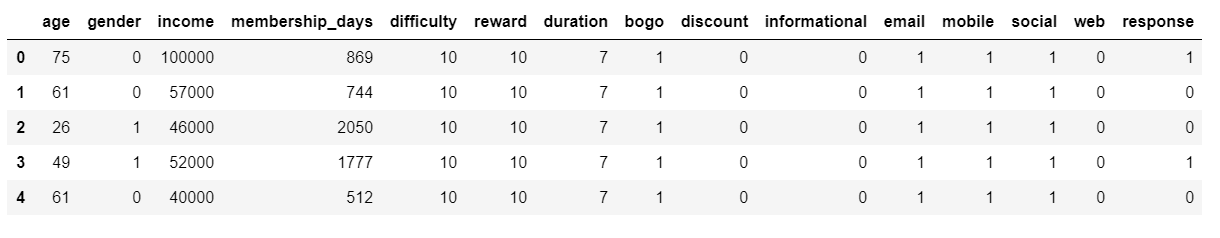
1. Change the name of the 'person' column to 'customerid'
2. Remove customer id's that are not in the customer profile DataFrame
3. Extract the offerid from the value column into a separate column
4. Only keep the following events: offer viewed, offer completed
5. Change column ordering

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 AND ANALYSIS**

MODEL: RANDOM FOREST

ACCURACY: 0.665487768936

F1 SCORE: 0.720305569246

BEST PARAMETERS: {'n\_estimators': 300, 'min\_samples\_split': 10, 'min\_samples\_leaf': 4}

We will be validating the robustness of the model’s solution by running the model with multiple different random states and then checking the mean / variance of the results.

The following results was obtained:

Mean Model Accuracy: 0.664161508989

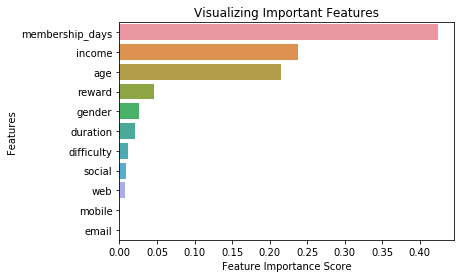
Variance of Model Accuracy: 1.75896544688e-06

Mean Model F1 Score: 0.719837831379

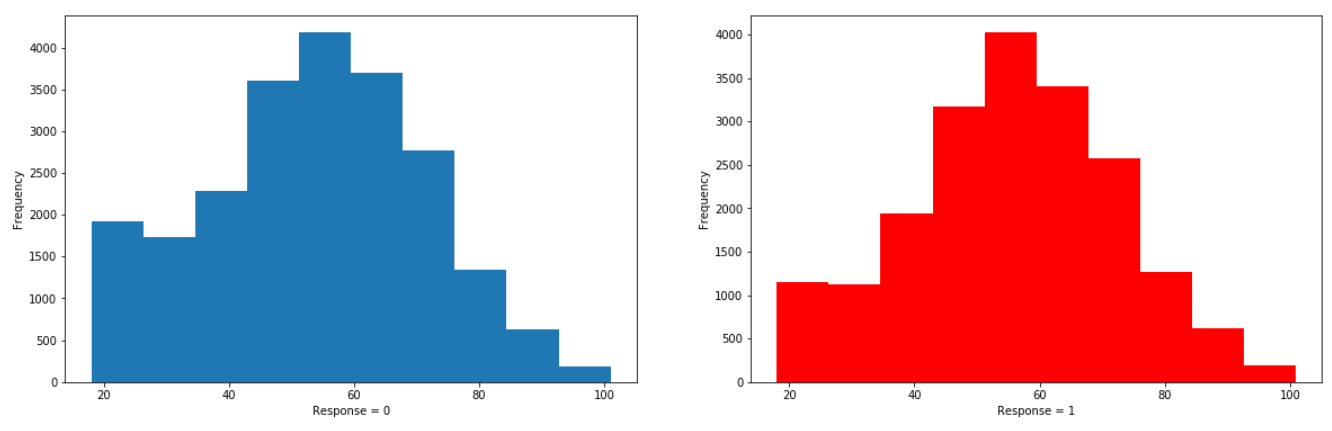
Variance of Model F1 Score: 1.75896544688e-06

By observing the above values, we can confirm that the model that was developed is very robust to data changes!

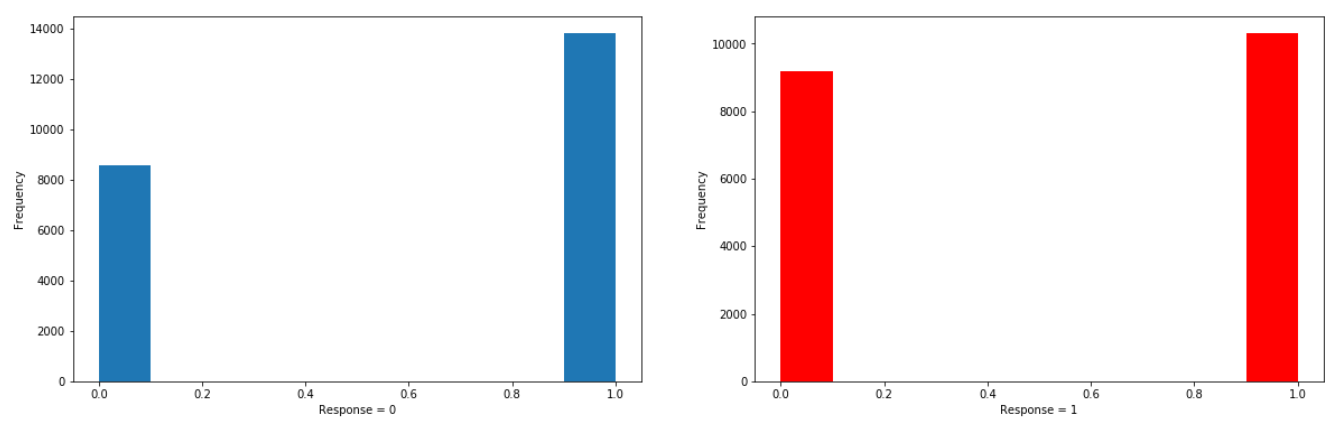
Important Features:



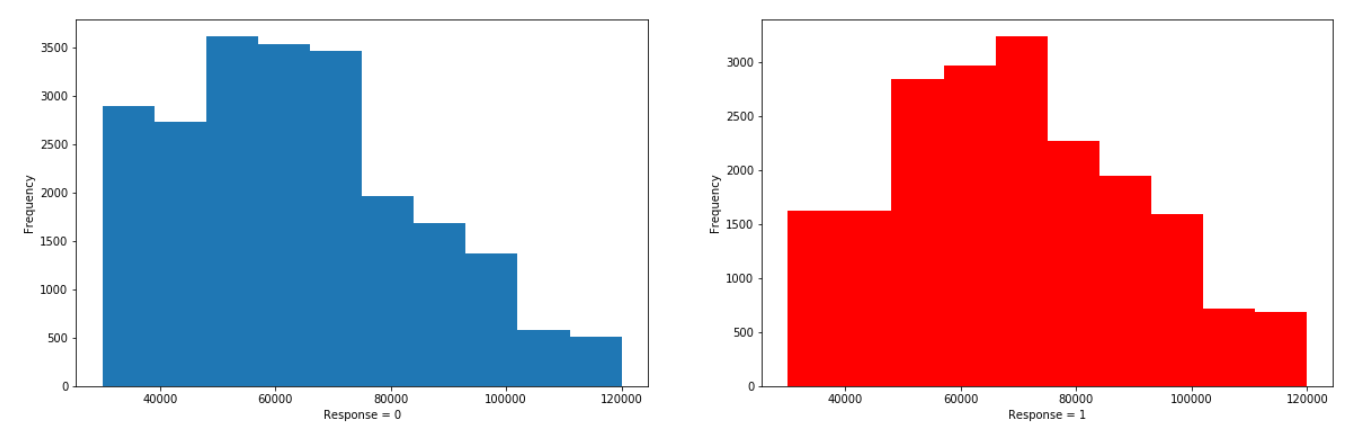
Age Distribution:



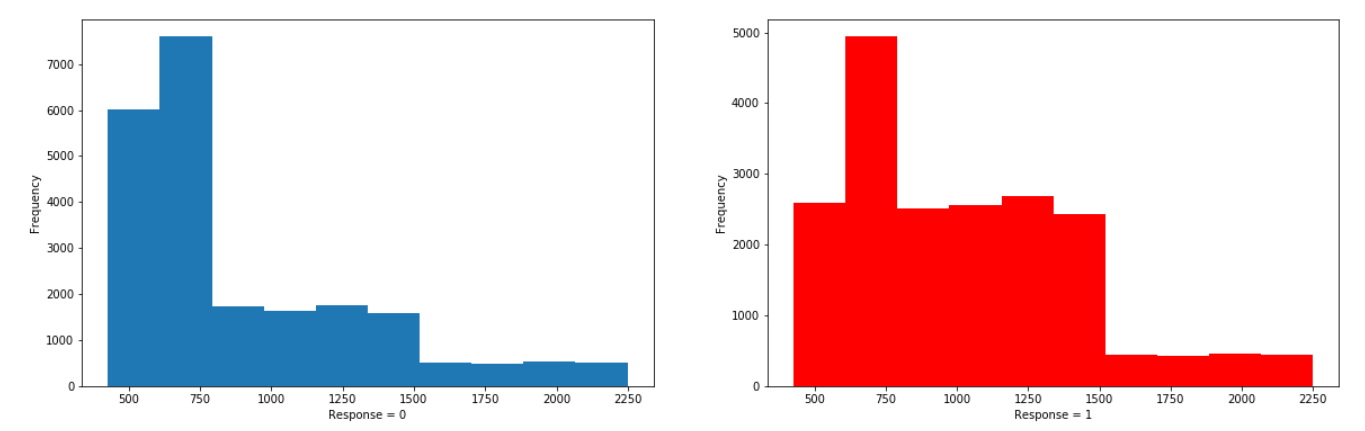
Gender Distribution:



Income Distribution:



Membership Days Distribution:



Now the best tuned model has been created which gave us the important features. We have also performed data visualization about the important columns. Armed with this information, we will be able to answer the questions we had proposed initially.

1. Which offer should be sent to a particular customer to let the customer buy more?

We were able to achieve an accuracy value of 66.54% which shows that our model will be able to predict the offer response based on the customer demographics and the offer details very nicely. We also were able to get a high value of F1 Score (72.03%) which signifies that the model is not biased towards one particular result. By observation of the important features result, we can say that offers having a higher reward have higher offer completion rate. Simply put, the more the reward, the better the chance of an individual responding to the offers. The duration of the offer also plays an important role and thus offers having a longer duration tends to have a getter completion rate. The understanding that the people will have more time to complete the offer as compared to offers whose duration is very less. Imagine you getting the offer on a Monday, but the offer expires in 5 days. Let’s say you may have a habit of going to Starbucks on weekends. But since the offer will expire in 5 days (on Friday), you will not be able to take benefit from the same. Now if the offer duration was more, let’s say 7 days then you will be able to take the benefit of the offer on weekends more easily.

2. What is the impact of the customer demographic on the offer completion?

It can be observed from the important features graph (Feature Importance: refers to a numerical value that describes a feature's contribution to building a model that maximizes its evaluation metric) that the following parameters have the most influence on the offer completion rate related to the customer demographics:

* 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.

* Income

Reasoning: People who have comparatively high income are more likely to respond to the offers.

* Age

Reasoning: Age plays an important factor in deciding as to how likely a person will respond to the offers.

3. What is the impact of the membership duration on the offer completion?

People who are Starbucks member for very long are more loyal and more likely to respond to the offers.

4. Which are the best channels for that leads to the most offer completion?

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. We can also see that social media have a greater influence and impact on the offer completion as compared to other channels!

**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.
* Make a web app