GBA 464: Assignment 3

Yufeng Huang

September 29, 2020

1 Objective

The general goal of this task is that we try to distinguish potential high-value consumers and separate them from low-value consumers. Recency, frequency and monetary value (RFM) are 3 factors that we can calculate and use to do the targeting. In this assignment, we will work with a sample dataset from a company called CDNOW, to try and figure out the potential value of a consumer in a given month, using only historical data prior to this month. We will then classify the sample by the "RFM index" we generated and see how much it is related to actual consumer spending.

2 Loading the data [20%]

You can obtain the dataset via the following link: https://dl.dropboxusercontent.com/s/xxfloksp0968mgu/CDNOW_sample.txt
or directly on Blackboard. Note that the data is stored in fixed format, meaning that each variable starts at a fixed column in the text file. I've written the code for reading the data. Basically, I use the function 'read.fwf()' to read the data.

In the raw data, the first two variables are individual consumer identifiers. The second one is a re-coded version of the first one. For simplicity, we drop the first variable and only use the recoded ID as identifiers. I've already written this part as well. After dropping the first column in

the original data, the remaining columns are individual ID ($id), date of the trip ($date), purchase quantity (i.e. number of CDs purchased, $qty) and total expenditure (in dollar values, $expd).

Our next step is to aggregate the data into individual-month level, so keys should be $id, $year, and $month. During this aggregation process, we should sum up quantity and expenditure for each consumer in each month. You also need how many trips (construct $trips) the individual has been to the shop. Assign the collapsed data (again, on the key of $id, $year and $month) to a new data frame.

Of course, most people will not go to the shop and buy something every month. But we need an RFM prediction for each individual in every month (between January 1997 and June 1998, 18 months in total). When there is no trip in a given month, replace trip, expenditure and quantity to zero. Now, you should be ready to compute recency, frequency and monetary value separately.

3 Computing the RFM measures [80%]

3.1 Recency [20%]

Keep in mind that for any measure in RFM, we can only use historical data, i.e. data in the months before the current month. In this note, we define recency as the number of months since the last month with purchase. In the example below, if an individual has been to the store in month 1, 2 and 5, her recency is NA in month 1 (because we do not know anything before the data starts), 1 in month 2, 1 in month 3, 2 in month 4, 3 in month 5, and 1 in month 6.

We talked about an example in class that is similar to this recency measure. However, the way we constructed that measure was not optimal. Optionally, try to optimize your algorithm when you construct the recency measure.

3.2 Frequency [20%]

We define frequency as the total number of trips a given individual made in the previous *quarter*. A quarter is defined as one of Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec. If the observation is in the very

first of this individual, we assign frequency to NA.

3.3 Monetary value [20%]

Monetary value is defined as –still using historical data– the average monthly expenditure for a consumer, in the previous months when she purchased something. For example, in month 1, the consumer came to the store and spent in total 15 dollars. Then, in month 2, her monetary value is 15. In month 2, the consumer came again and spent a total of 30 dollars. Then her monetary value in month 3 is the average, i.e. $(15+30)/2 = 22.5$. She did not come in month 3 and 4, so her monetary value did not change. Finally, she came in month 5 and spent 20, and thus her monetary value is $(15+30+20)/3 = 21.7$.

3.4 Example

Let's create an artificial example with individual "0". Note that we've already organized the data into individual-year-month level, and included the months with no trip. Note that we have computed trips by counting the number of trips in a month with purchase. The last row shows that these statistics should be re-calculated for the next consumer.

id	year	month	trips	qty	expd	quarter	recency	frequency	monval
0	1997	1	1	1	15	1	NA	NA	NA
0	1997	2	2	2	30	1	1	NA	15
0	1997	3	0	0	0	1	1	NA	22.5
0	1997	4	0	0	0	2	2	3	22.5
0	1997	5	1	3	20	2	3	3	22.5
0	1997	6	0	0	0	2	1	3	21.7
1	1997	1	1	2	29	1	NA	NA	NA

Following this example, please calculate recency, frequency and monetary value measures as defined.

4 Targeting [20%]

4.1 RFM index

An RFM index is an weighted sum of the 3 measures, for each individual i in month t:

$$RFM_{it} = b_1 R_{it} + b_2 F_{it} + b_3 M_{it}$$

For now, let's say it is your marketing team's responsibility to tell you what are the factor loadings.[1] For now let's take $b_1 = -0.05$, $b_2 = 3.5$ and $b_3 = 0.05$. Note that if a consumer is considered "high value" if she has low recency, or high frequency, or high onetary value. I have coded this section already but you need to run it or change the corresponding variable names.

4.2 Validation [20%]

When you have computed this measure, sort your sample according to the RFM index and split it into 10 (roughly) even-sized portions. One way is to use quantile() to generate the cut-offs. The high RFM parts refer to individuals (in particular months) that are more valuable than the low RFM parts of your sample. Examine the average monthly expenditure for each bin of sample, defined by the deciles of the RFM index.[2] Plot the average spending (and potentially some other measures if you want to) by group and confirm that the result is more or less monotonic. Which groups of consumers do you want to target?

For example, you might produce something similar to this. You might not get the exact figure because the result depends on how you segment the market.

[1] In principle, we should estimate these loadings and we should use another sample to validate our results.
[2] Deciles: 10% quantiles.

Figure 1: Average expenditure by deciles in the RFM index