Creating an RFM Summary Using Excel

Peter S. Fader
www.petefader.com

Bruce G. S. Hardie
www.brucehardie.com†

December 2008

1. Introduction

In order to estimate the parameters of transaction-flow models such as the Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) and BG/NBD (Fader, Hardie, and Lee 2005a), as well as those of the associated models for spend per transaction (Schmittlein and Peterson 1994; Fader, Hardie, and Lee 2005b), we need an RFM (recency, frequency, monetary value) summary of each customer's purchasing behavior. In particular,

- The transaction-flow model requires three pieces of information about each customer's purchasing history: their "recency" (when their last transaction occurred), "frequency" (how many transactions they made in a specified time period), and the length of time over which we have observed their purchasing behavior. The notation used to represent this information is (x, t_x, T), where x is the number of transactions observed in the time period $(0, T]$ and t_x $(0 \leq t_x \leq T)$ is the time of the last transaction.
- The spend model requires two pieces of information about each customer's purchasing history: the average "monetary value" of each transaction (denoted by m_x) and the number of transactions over which this average is computed (i.e., frequency, x).

In this note we describe how to create such a summary from raw customer-level transaction data using pre-Office 2007 versions of Excel. (The only issue with the Office 2007 version of Excel is that some of the menu-related instructions given in this note are no longer correct. Readers familiar with Excel 2007 should be able to work out the necessary changes for themselves.)

2. Preliminaries

We will make use of the CDNOW dataset, as used in Fader et al. (2005a,b). The master dataset contains the entire purchase history up to the end of June 1998 of the cohort of 23,570 individuals who made their first-ever purchase at CDNOW in the first quarter of 1997. (See Fader and Hardie (2001) for further details about this dataset.)

The file `CDNOW_sample.txt`[1] contains purchasing data for a 1/10th systematic sample of the whole cohort (2357 customers). Each record in this file, 6919 in total, comprises five fields: the customer's ID in the master dataset, the customer's ID in the 1/10th sample dataset (ranging from 1 to 2357), the date of the transaction, the number of CDs purchased, and the dollar value of the transaction.

We start the process of creating our summary of each customer's purchasing in the following manner:[2]

- We import the text file `CDNOW_sample.txt` into an empty blank Excel workbook. We note that the associated worksheet is called `CDNOW_sample`. (We assume that the records in the raw transaction data file are grouped by customer, and sorted within customer by date of transaction. If in doubt, sort the raw dataset by customer ID and date of transaction.)

- As they are of no interest to us in this particular case (unlike, say, in Fader and Hardie (2001)), we delete the first and fourth columns (master dataset customer ID and # CDs purchased, respectively). We insert a row at the top of the worksheet, adding field names: ID, Date, Spend. (See Figure 1.)

	A	B	C
1	ID	Date	Spend
2	1	19970101	29.33
3	1	19970118	29.73
4	1	19970802	14.96
5	1	19971212	26.48
6	2	19970101	63.34
7	2	19970113	11.77
6917	2356	19970927	31.47
6918	2356	19980103	28.98
6919	2356	19980607	28.98
6920	2357	19970325	25.74

Figure 1: Raw transaction data

[1] See `http://brucehardie.com/datasets/CDNOW_sample.zip`

[2] Note that the following steps are not meant to represent the most elegant approach to the task of creating the RFM summary; Excel experts will be able to identify better approaches.

- Scrolling down the dataset, we note that some customers had more than one transaction on a given day. For example, customer 26 had two separate transactions on 13 January 1997, while customer 46 had two separate transactions on 28 August 1997. Typing

  ```
  =IF(AND(A2=A1,B2=B1),1,0)
  ```

  into cell `D2` and copying it down to cell `D6920`, we note that cells `D2:D6920` sum to 223, indicating that there are a total of 223 such "additional" transactions.

- The transaction-flow models are developed by telling a story about interpurchase times. As we only know the date (and not the time) of each transaction, we need to aggregate the records associated with same-day transactions — we can't have an interpurchase time of 0. We do so in the following manner:

  - Deleting the current contents of column `D`, we type

    ```
    =IF(AND(A2=A1,B2=B1),C2+D1,C2)
    ```

    into cell `D2` and copy it down to cell `D6920`. This creates a running within-transaction-day total spend for each customer.
  - For those situations where a customer has more than one transaction in a day, we wish to identify the all-but-last transactions (and then delete them). We type

    ```
    =IF(AND(A2=A3,B2=B3),1,0)
    ```

    into cell `E2` and copy it down to cell `E6920`. We then copy and "paste special" / "values" cells `D2:E6920` onto themselves. Next we sort the whole block of data by column `E` "ascending" (with a "header row") and delete the rows for which column `E` contains a 1, a total of 223 rows.
  - Finally, resorting the block of data by ID and Date (both "ascending"), deleting columns `C` and `E`, and labeling the new column `C` Spend gives us a raw transaction dataset in which no customer has more than one transaction on any given day.

3. "Frequency" and "Monetary Value"

Now that we have a "clean" raw transaction dataset, we can compute the *frequency* and *monetary value* summaries for each customer.

Most of the previous analyses undertaken using this dataset have split the 78 weeks of data in half, creating a 39-week calibration period (1997-01-01–1997-09-30) and 39-week validation period (1997-10-01–1998-06-30). Furthermore, these analyses have generally ignored each customer's first-ever purchase at CDNOW, which signals the start of the customer's "relationship" with the firm; this means calibration-period "frequency" has usually

been the number of *repeat* transactions, and "monetary value" has been the average dollar value per repeat transaction.

- In order to identify each transaction as being the first-ever purchase for the customer, a calibration-period repeat transaction, or validation-period transaction, we enter

  ```
  =IF(A2<>A1,"first",IF(B2<=19970930,"calib","valid"))
  ```

  into cell `D2`, copy it down to cell `D6920`, labeling the column Period (cell `D1`).

- We can now create the desired frequency summary table using the "pivot tables" feature in Excel. We highlight the cell range `A1:D6697` and select the *PivotTable and PivotChart ...* option under the *Data* menu. We use ID as the *row field*, Period as the *column field*, and ID as the *data item*. (Make sure the "Pivot Table Field" is "Count of ID", not sum or some other summary of the ID field.) We rename the worksheet containing the resulting table `Pivot Table 1`; this reports the number of repeat transactions made by each of the 2357 customers in the calibration and validation periods.

- We now compute the average spend per repeat transaction for the calibration and validation periods in a similar manner. Going back to the `CDNOW_sample` worksheet, we highlight the cell range `A1:D6697`, and select the *PivotTable and PivotChart ...* option under the *Data* menu. We use ID as the *row field*, Period as the *column field*, and Spend as the *data item*. (Make sure the "Pivot Table Field" is "Average of Spend", not sum or some other summary of the Spend field.) We rename the worksheet containing this report `Pivot Table 2`.

4. "Recency"

The next step is to compute *recency*, as well as the length of time over which we have observed each customer's purchasing behavior.

- To identify the date on which each customer made their first-ever purchase at CDNOW, we go back to the `CDNOW_sample` worksheet and type

  ```
  =IF(A2<>A1,B2,0)
  ```

  into cell `E2`, copy it down to cell `E6920`, labeling the column First Purchase (cell `E1`).

- Highlighting the cell range `A1:E6697`, we select the *PivotTable and PivotChart ...* option under the *Data* menu. We use ID as the *row*

field and First Purchase as the *data item*. (Make sure the "Pivot Table Field" is "Sum of First Purchase".) We rename the worksheet containing this report `Pivot Table 3`; this gives us the calendar date of each customer's first-ever purchase at CDNOW

- To identify the date on which each customer made their last calibration-period purchase, we go back to the `CDNOW_sample` worksheet and type

  ```
  =IF(B2<=19970930,IF(OR(A2<>A3,B3>19970930),B2,0),0)
  ```

  into cell `F2`, copy it down to cell `F6920`, labeling the column Last Purchase (cell `F1`).

- Highlighting the cell range `A1:F6697`, we select the *PivotTable and PivotChart ...* option under the *Data* menu. We use ID as the *row field* and Last Purchase as the *data item*. (Make sure the "Pivot Table Field" is "Sum of Last Purchase".) We rename the worksheet containing this report `Pivot Table 4`; this gives us the calendar date of each customer's last calibration-period purchase.

- For modeling purposes, "recency" is not the calendar date of the last observed purchase; rather the time origin for t_x is the start of the observation period. Since we track customers' purchasing from their first-ever purchase at CDNOW, the date of this first purchase is the time origin. Therefore, t_x is the length of time between the first-ever purchase and the last observed purchase (in the calibration period), i.e., Last Purchase − First Purchase.

- In order to perform the difference in dates calculation, we need to convert the data fields into "Excel dates":

  - Going to the `Pivot Table 3` worksheet, we enter

    ```
    =DATE(LEFT(B5,4),MID(B5,5,2),RIGHT(B5,2))
    ```

    into cell `C5` and copy it down to cell `C2361`. Highlighting cells `C5:C2361`, we change the cell *Number* format to *General*. (The resulting numbers represent the number of days since January 1, 1900.)
  - Similarly for the `Pivot Table 4` worksheet, we enter

    ```
    =DATE(LEFT(B5,4),MID(B5,5,2),RIGHT(B5,2))
    ```

    into cell `C5` and copy it down to cell `C2361`. Highlighting cells `C5:C2361`, we change the cell Number format to General.

- The difference between a column C entry in the `Pivot Table 4` worksheet and the corresponding cell in the `Pivot Table 3` worksheet gives us the desired recency number in days. When analyzing this dataset, Fader et al. (2005a,b) have used week as the basic unit of time. We therefore compute *recency* by entering

```
=(C5-'Pivot Table 3'!C5)/7
```

in cell `D5` of the `Pivot Table 4` worksheet and copy it down to cell `D2361`. (We label the column by entering t_x in cell `D4`).

- The final piece of information we need is T, the length of time we observe each customer (i.e., the time between the customer's first-ever purchase at CDNOW and the end of the calibration period, 1997-09-30). We compute this in the `Pivot Table 3` worksheet by entering

```
=(DATE(1997,9,30)-C5)/7
```

in cell `D5` and copying it down to cell `D2361`. (We label the column by entering T in cell `D4`).

5. Bringing It All Together

We can now create a single worksheet that contains all the required "RFM" information.

- We insert a new worksheet (let's call it `Summary`) and label the first five cells of row 1 ID, x, t_x, T, and m_x.
- We enter

```
='Pivot Table 1'!A5
='Pivot Table 1'!B5
='Pivot Table 4'!D5
='Pivot Table 3'!D5
='Pivot Table 2'!B5
```

into cells `A2` – `E2` respectively and copy this block of cells down to row 2358. (See Figure 2.)

	A	B	C	D	E
1	ID	x	t_x	T	m_x
2	1	2	30.43	38.86	22.35
3	2	1	1.71	38.86	11.77
4	3	0	0.00	38.86	0.00
5	4	0	0.00	38.86	0.00
6	5	0	0.00	38.86	0.00
7	6	7	29.43	38.86	73.74
2355	2354	5	24.29	27.00	44.93
2356	2355	0	0.00	27.00	0.00
2357	2356	4	26.57	27.00	33.32
2358	2357	0	0.00	27.00	0.00

Figure 2: RFM summary of calibration-period transactions

This is the required "RFM" summary of calibration-period buying behavior needed to estimate the previously mentioned models of transaction flow and spend per transaction. (Columns A–D contain the same data as given in the `Raw Data` worksheet in the Excel workbook `bgnbd.xls` found at <`http://brucehardie.com/notes/004/`>). We can create a summary of validation-period purchasing behavior by extracting the corresponding number of transactions and average spend per transaction numbers from the `Pivot Table 1` and `Pivot Table 2` worksheets.

References

Fader, Peter S. and Bruce G. S. Hardie, (2001), "Forecasting Repeat Sales at CDNOW: A Case Study," *Interfaces*, **31** (May–June), Part 2 of 2, S94–S107.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005a), ""Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, **24** (Spring), 275–284.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005b), "RFM and CLV: Using Iso-value Curves for Customer Base Analysis," *Journal of Marketing Research*, **42** (November), 415–430.

Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987), "Counting Your Customers: Who They Are and What Will They Do Next?" *Management Science*, **33** (January), 1–24.

Schmittlein, David C. and Robert A. Peterson (1994), "Customer Base Analysis: An Industrial Purchase Process Application," *Marketing Science*, **13** (Winter), 41–67.