Customer Analytics

RFM

Lecture 2

Tilburg University

Agenda

1. RFM scores
2. RFM segments
3. Making rollout decisions based on RFM
4. Extensions

What is RFM?

Recency

When was the last time (relative to some date) that a customer purchased?

Frequency

How often did the customer purchase?

Monetary value

Total cumulative amount spent since first purchase

or

Average amount spent per order

RFM is a 3-number summary of customer behavior

Why RFM

Targeting all customers in database is too costly

- Some customers will never respond or purchase

Use a model to select good customers who will respond/buy

Goal: predict the response or purchase probability of each customer

- Use these predictions to target only, for example 20%, of customers rather than all

Example 1:
Transaction-level dataset (CDNow)

Each row represents a transaction

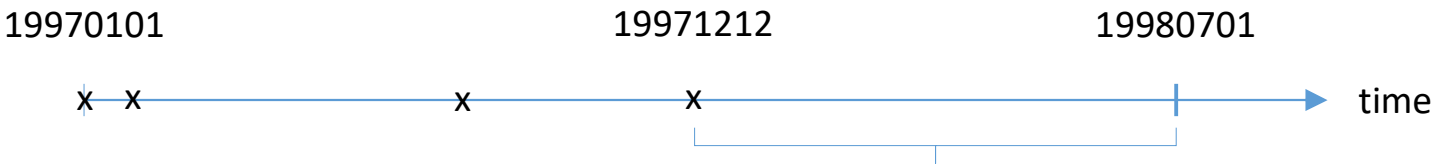
- Dates: 19970101 – 19980630
- Analysis time is 19980701 (July 1, 1998)

Customer ID	Date (yyyymmdd)	Amount
1	19970101	29.33
1	19970118	29.73
1	19970802	14.96
1	19971212	26.48
2	19970101	63.34
2	19970113	11.77
3	19970101	6.79
4	19970101	13.97
5	19970101	23.94
6	19970101	35.99
6	19970111	32.99
6	19970315	77.96
6	19970416	59.3
6	19970424	134.98
6	19970623	91.92
6	19970722	47.08
6	19970726	71.96
6	19971025	78.47
6	19971206	83.47
6	19980118	84.46
6	19980215	123.96
6	19980221	32.98
6	19980226	23.06
6	19980510	72.99
6	19980620	55.47

Example 1: Transaction-level dataset

Customer ID	Date (yyyymmdd)	Amount
1	19970101	29.33
1	19970118	29.73
1	19970802	14.96
1	19971212	26.48

Timeline for customer 1



F = # of transactions = 4

M = total amount / # of transactions = 25.13

or

M = total amount = 100.50

R = time between last transaction and end of database
= 201 days
= 28.7 weeks
= 7.2 months

The units are not that important

Transaction-level dataset

Customer ID	Date (yyyymmdd)	Amount
1	19970101	29.33
1	19970118	29.73
1	19970802	14.96
1	19971212	26.48
2	19970101	63.34
2	19970113	11.77
3	19970101	6.79
4	19970101	13.97
5	19970101	23.94
6	19970101	35.99
6	19970111	32.99
6	19970315	77.96
6	19970416	59.3
6	19970424	134.98
6	19970623	91.92
6	19970722	47.08
6	19970726	71.96
6	19971025	78.47
6	19971206	83.47
6	19980118	84.46
6	19980215	123.96
6	19980221	32.98
6	19980226	23.06
6	19980510	72.99
6	19980620	55.47

Each row represents a transaction

Customer-level dataset

ID	R (days)	F	M (average)
1	201	4	25.13
2	533	2	37.56
3	545	1	6.79
4	545	1	13.97
5	545	1	23.94
6	10	16	69.19

Each row represents a customer

Example 2: customer-level data set (ebeer)

ï..acctnum	gender	R	F	M	firstpur	age_class	single	student	mailing	respmail
10001	1	30	10	35.7	50	3	1	1	1	0
10005	0	16	1	149.0	16	0	0	1	1	0
10010	0	12	1	123.0	12	0	1	0	1	0
10011	0	6	2	147.0	8	0	1	1	1	0
10014	1	6	3	96.0	18	0	0	1	1	1
10020	0	12	2	150.5	14	0	1	1	1	0
10021	1	10	2	101.0	16	0	1	1	1	0
10029	1	8	2	64.0	10	0	1	0	1	0
10031	0	28	12	16.9	62	3	0	1	1	0
10032	0	8	9	32.8	60	2	0	1	1	0

RFM for segment-level prediction

Rather than use the raw customer-level RFM scores, marketers typically **group them into segments** based on **ranking** and **sorting** them.

First we'll do this to one variable, recency.

Then we'll do the full analysis.

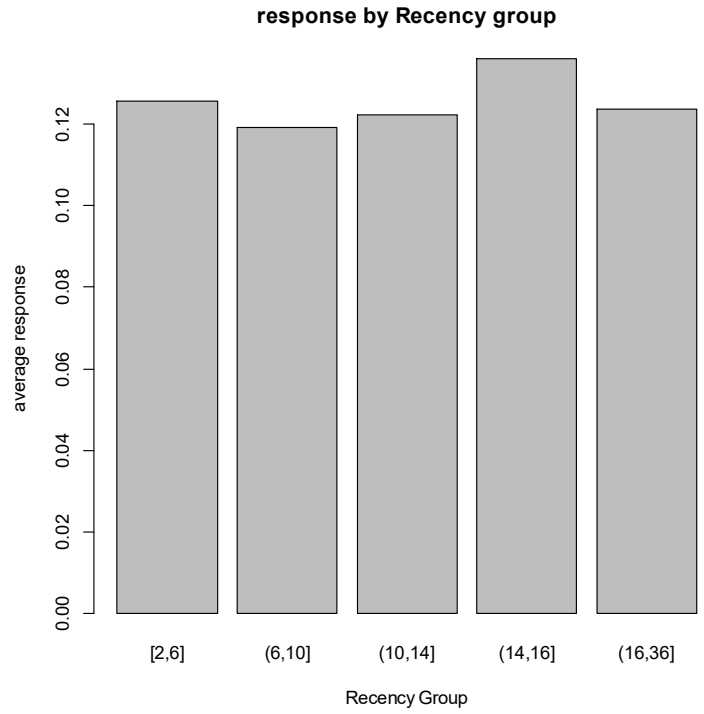
First let's look (separately) at R

We can rank customers individually on R by creating groups:

- Group 1: top 20% most recent customers
- Group 2: 2nd most 20% recent customers
- ..
- Group 5: bottom 20% recent customers

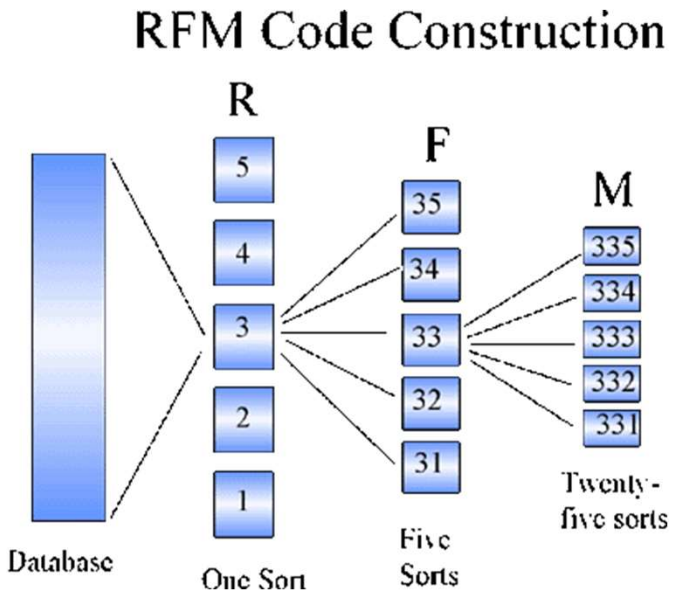
Rgroup	mean	sd	n
[2,6]	3.97	1.616	2229
(6,10]	9.22	0.975	2069
(10,14]	13.00	1.000	2509
(14,16]	16.00	0.000	1231
(16,36]	26.80	5.465	1926

How does the response probability change across segments?



TILBURG UNIVERSITY

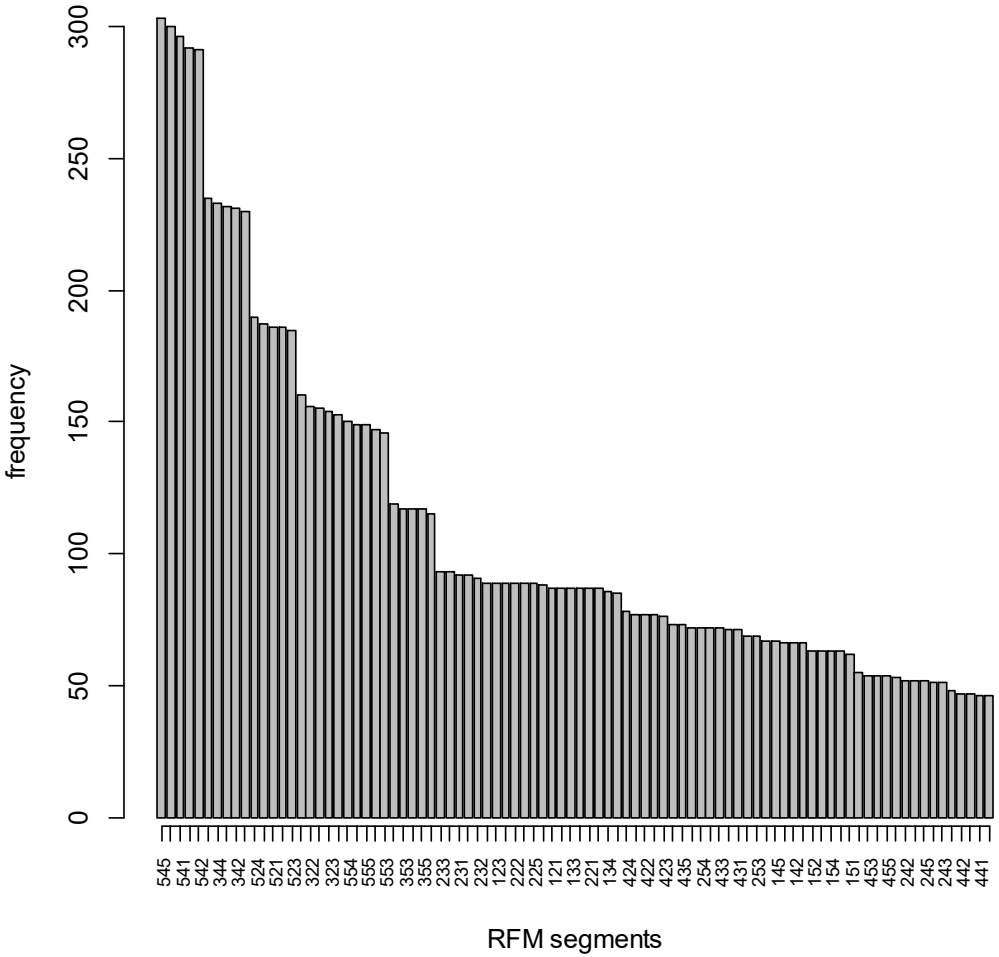
Full RFM segmentation



1. Sort by R, create 5 groups (we just did this!)
2. For each R group (5), sort by F & create 5 groups
3. For each RF group (25), sort by M & create 5 groups

You can also do this with different numbers of groups per var (e.g., 3) or independently rather than nested.

[Quick Profits with RFM Analysis]



125 total possible combinations

In this specific case (ebeer) only 90 segments, because of **bunching**. Some groups may have only 4 possible values of frequency, so it makes no sense to create 5 groups.

Also unequal numbers in groups due to bunching.

<u>But</u> discrete distribution means unequal cells.

For each RFM segment, estimate response rate

n_z = # of segment z tested

s_z = # of responses to test in segment z

All members of segment s have the same (unknown) response probability p_z; this implies that S_z is a binomial random variable

$$P(s \mid n_z, p_z) = \binom{n_z}{s} {p_z}^s (1 - p_z)^{n_z - s}$$

$$\hat{p}_s = \frac{s_z}{n_z} \qquad se(\hat{p}_z) = \sqrt{\frac{\hat{p}_z(1 - \hat{p}_z)}{n_z}}$$

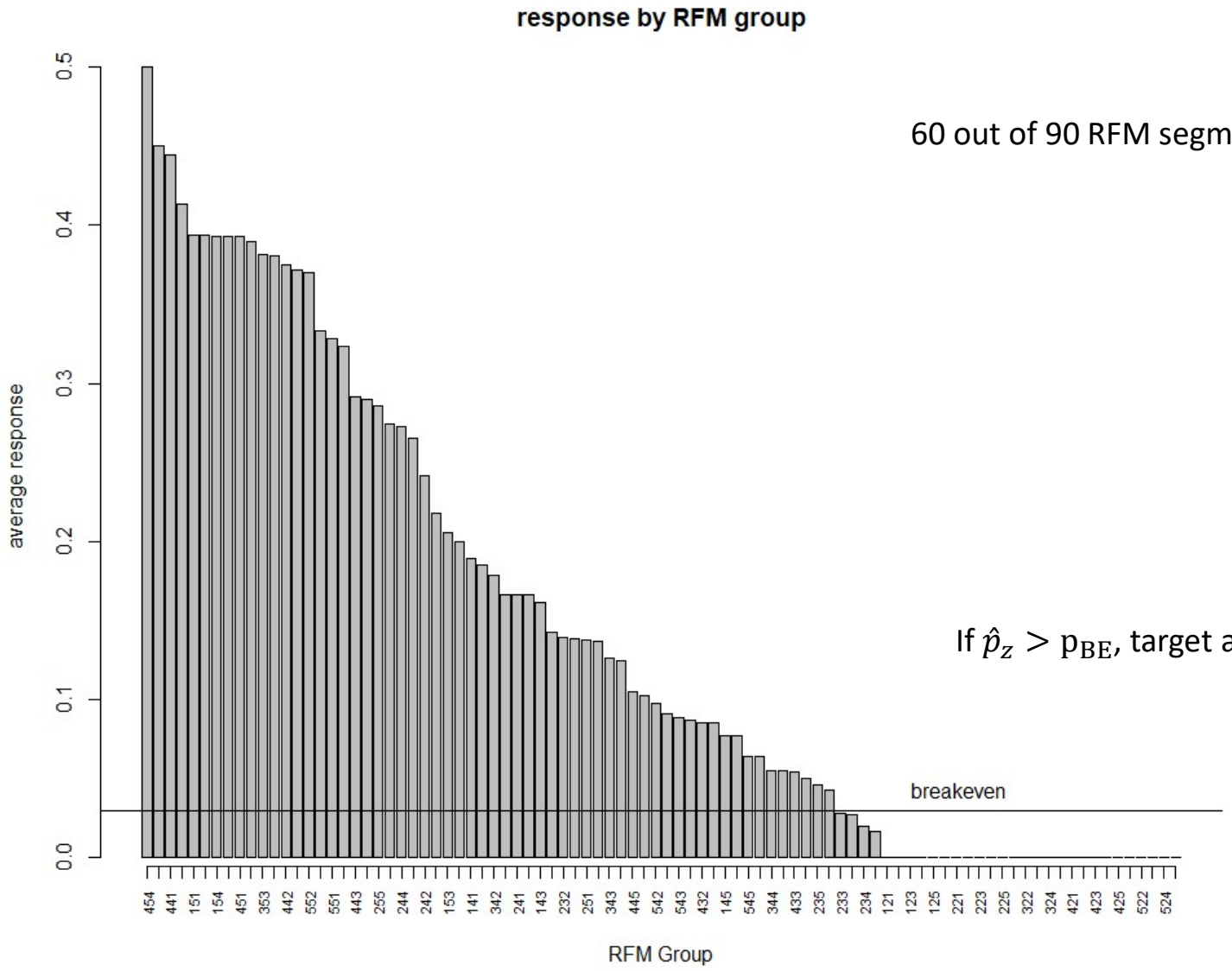
Using the results: next steps

- Apply estimated segment response rates to rollout customers
  - (The point is to predict response for rollout customers, not the test customers who have already responded!)

- Roll out marketing to customers in RFM segments with response rates above the threshold.

- Calculate expected profits of segments you roll out to, costs and return on investment

$$\text{ROI} = \frac{\text{total (expected) profits}}{\text{total costs}}$$



60 out of 90 RFM segments targeted

If $\hat{p}_z > \mathrm{p_{BE}}$, target all customers in segment z

$$\mathrm{p_{BE}} = \frac{c}{m} = \frac{1.50}{50} = 0.03$$

Expected profits and ROI

Expected profit per customer in segment z

$$\max\{(m \cdot \hat{p}_z - c), 0\}$$

Expected profit in rollout sample

$$\text{Total rollout profit} = \sum \max\{(m \cdot \hat{p}_z - c), 0\}$$

Total rollout profit = 26109 in our sample. We target 3222 customers, so total cost = 1.50 * 3222 = 4833.

ROI is then 26109/4833 = 5.4, or 540%.

If targeted all customer, ROI is 23691/7518 = 3.15 or 315%. 10% higher profit for 60% cost

Advantages of RFM

Easy to understand, easy to compute and easy to give managerial recommendations

- Target RFM segments in rollout sample above breakeven response probabilities

- One of the oldest customer metrics still popular today (and for a reason!)

Advantages of RFM

Common types of data used in marketing models

Readily available

- **Transaction data**: past purchases, amounts, dates, discounts, …
- **Demographics**: gender, ethnicity, age, income, family size, occupation, marital status, education, homeowner or renter, length of residence (typically available for prospects)
- **Marketing:** past mailings, content mailings, date, costs
- **(Survey data, e.g. Psychographics)**: attitudes, interests, activities

See BKN chapter 8

Problems with RFM

- No other variables apart from RFM.
  - Not so problematic, can always add control variables.

- RFM test segments, on which we base decisions, may be too small

RFMgroup	n_resp	n_nonresp	n_mail	resp_rate
431	1	35	36	0.0278

- for RFM segment 431:

$$\hat{p}_{431} = \frac{1}{36} = 0.0278 < 0.0300$$

$$se(\hat{p}_{431}) = 0.0274$$

Can repeat analysis with fewer RFM categories than 5

2 x 2 x 2 RFM or 3 x 3 x 3 RFM

Lower variance, higher bias

Extension 1: Bayesian approach

Right now, we assume that these segments response rates are entirely independent of each other.

- Knowing what the response is overall tells us nothing about segment 431.

We need a model that relates the segment response rates to each other

- That way we can borrow information from other segments to help understand our segment

We assume that the segment response rates come from the same prior.

$$p_z \sim \text{beta}(a, b)$$

Empirical Bayes

In lecture 1, we talked about the prior as something that you think about before seeing the data.

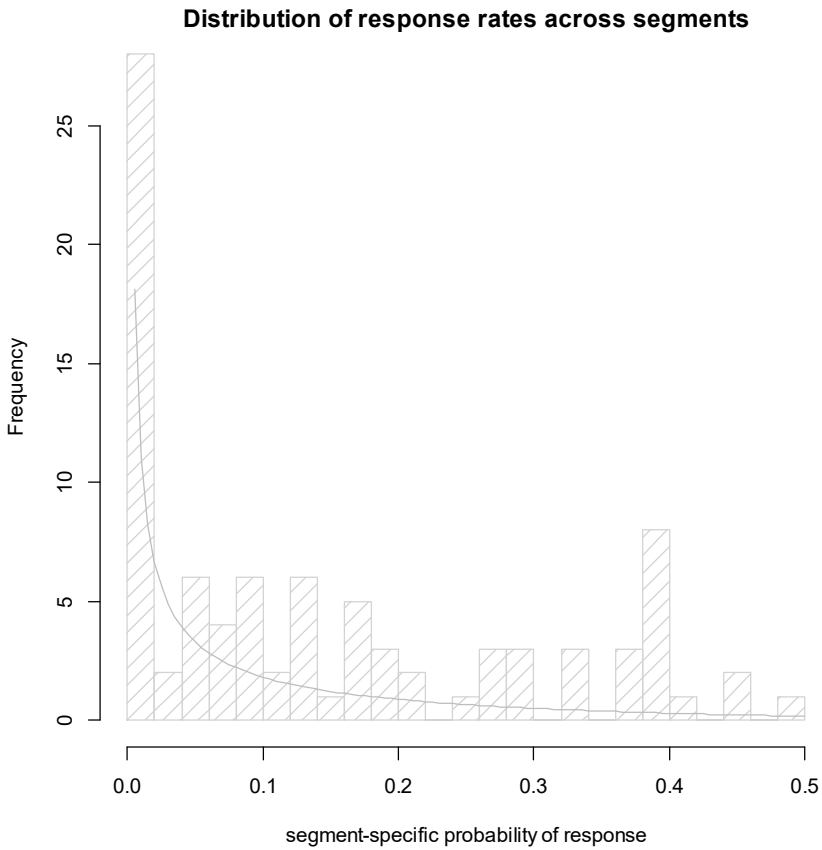
- **Previous experience**: past campaigns' response rates range from about 0-10%, on average slightly below 3%
- **Or truly no idea:** every value is equally likely (**flat** or **diffuse** prior)

happens a lot with Big Data!

But in this case, we have many parallel potentially similar groups (segments). We can use them to see what the prior distribution looks like (in order to pick parameters a and b). We can choose the prior distribution to look most like the actual distribution, across segments.

Using data to estimate the prior distribution (rather than in L1, just assuming some form) is called **Empirical Bayes.**

Distribution of segment specific response rates



$\text{beta}(a = 0.3, b = 3.0)$

Extension 1: Beta-binomial model

All p_z comes from a common Beta distribution with two parameters a & b:

$$p_z \sim \text{beta}(a, b)$$

It follows that the aggregate distribution of responses to a mailing of size n_z is given by:

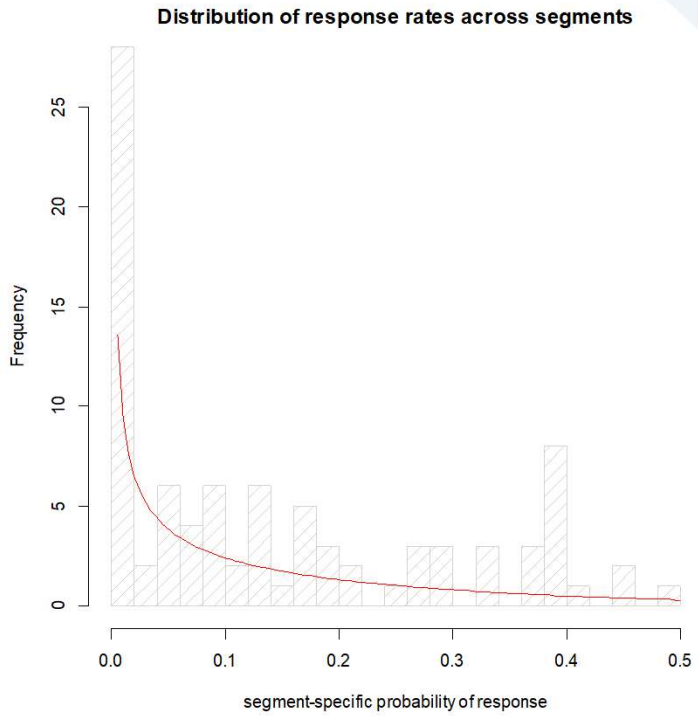
$$P(s_z|n_z, a, b) = \int_0^1 P(s_z|n_z, p_z)\, g(p_z|a, b)\, dp_z$$

$$= \binom{n_z}{s_z} \frac{B(a + s_z, b + n_z - s_z)}{B(a, b)}$$

Estimation

Estimate the model: find a & b that maximize log-likelihood

$$LL(a, b|s_z, n_z) = \sum_z \ln\big(P(s_z|n_z, a, b)\big)$$



$$\text{beta}(\hat{a} = 0.493, \hat{b} = 3.113)$$

Bayes rule

If we observe s_z "successes" out of n_z trials in segment z: what is the posterior for the segment-specific response rate p_z

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}$$

$$\text{f}(p_z) \propto p_z^{a+s_z-1} (1-p_z)^{b+n_z-s_z-1}$$

$$p_z \sim \text{beta}(a+s_z, b+n_z-s_z)$$

Posterior mean segment response rate

For $p \sim \text{beta}(a, b)$,

$$\text{E}[p] = \frac{a}{a+b}.$$

Applying this to our case, $p_z \sim \text{beta}(a + s_z, b + n_z - s_z)$

$$E[p_z] = \frac{a + s_z}{a + b + n_z}$$

12.8 in BKN

Bayes as compromise

We can re-write the last expression to reveal that the posterior mean is a weighted average:

$$E[p_z] = w\left(\frac{a}{a+b}\right) + (1-w)\frac{s_z}{n_z}$$

where $w = \frac{a+b}{a+b+n_z}$.

In general, the direction of shrinking to the mean of the distribution, $\frac{a}{a+b}$. The larger n_z, the less shrinking.

"The more data, the more separate."

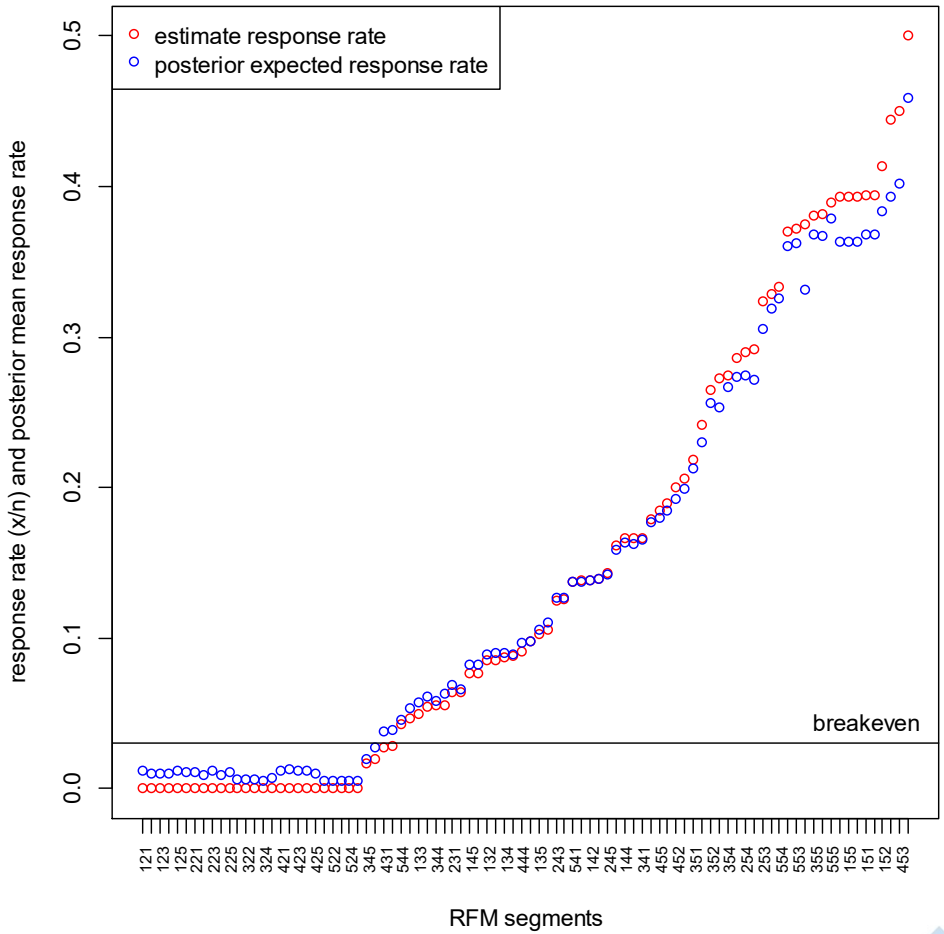
Shrinking estimates

- For segment 431,

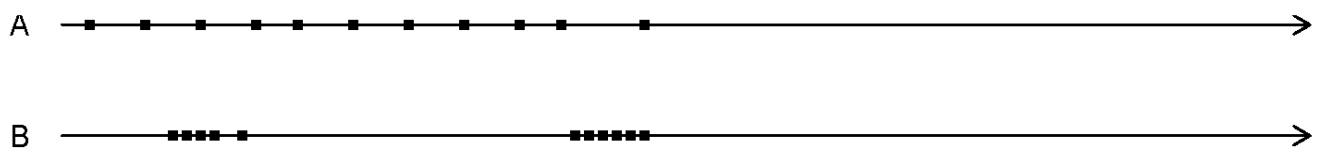
$$\hat{p}_z = 0278$$

$$E[p_z] = .0377$$

This is a segment that would be targeted using the beta-Binomial model, but not using the simple average, $\hat{p}_z$.


estimate response rate
posterior expected response rate
breakeven
response rate (x/n) and posterior mean response rate
0.0
0.1
0.2
0.3
0.4
0.5
121 123 125 221 223 225 322 324 421 423 425 522 524 345 431 544 133 344 231 145 132 134 444 135 243 541 142 245 144 341 455 452 351 352 354 254 253 554 553 355 555 155 151 152 453
RFM segments

Extension 2: Clumpiness



- Both of these customers have the same values of R and F
- From data perspective, the user is called clumpy when more or larger clumps of visits are observed.
- Clumps of visits indicate non-constant visiting rate, specially temporary elevations of propensity --- i.e., periods during which the user is more likely to visit/purchase than his/her average level.

Conclusion

- RFM perhaps oldest example of customer analytics
- Still used, because easy-to-compute, understand, and guide targeting decisions
- Drawbacks: small sample size, no other variables included