CRM in A Segmented Ebanking/Ecommerce Marketplace – New Approaches to Data Analysis

Jonathan Cottrell
Principal, Solid-Data, Oldham, UK
Postal Address: **Jonathan Cottrell, 44 Purdy House, Eldon Street, Oldham, OL8 1NH, UK**
Author's Personal/Organizational Website: **www.solid-data.com**
Email: **jcc@solid-data.com**
Jonathan Cottrell is the founder of Solid-Data, a commercial software company based in Oldham UK. Graduating in Mathematics and Computer Science from the University of Cambridge, Jonathan was introduced to data analysis on large scientific equipment at Micromass (now Waters Corporation), a global leader in mass spectrometry solutions. Subsequently, Jonathan identified opportunities to develop these techniques for enterprise-based commercial data analysis, leading to the formation of Solid-Data.

Abstract

We explore the development of a probabilistic model of segmented markets, and go on to show how such a model adds value in various common e-banking and e-commerce scenarios. We compare this modern treatment with classical approaches, and demonstrate how such a probabilistic model reduces the cost of acquiring new customers, improves ROI on marketing campaigns, and helps identify the correct point to intervene in the customer lifecycle to retain previously loyal customers.

A NOTE ABOUT TERMINOLOGY

Data mining professionals come from a wide variety of backgrounds, including computer science, IT, statistics, pure mathematics, machine learning and engineering. As a result, there are often a plethora of terms in use to describe the same concept. Most often, we see data in spreadsheet format, with columns representing variables in the data (that is, a value for an observable quantity), and rows representing an instance of several variables (Figure 1).

	A	B	C	D	E	F	G
1	GENDER	AGE	HOMEOWNER	BEDS	CHANNEL	PRODUCT	
2	F	75	HO	4	ON	D	
3	M	24	TE	2	PR	C	
4	F	42	TE	3	PR	A	
5	M	56	HO	4	PR	B	
6	M	37	HO	3	TV	A	
7	F	73	HO	6	ON	D	
8	M	46	TE	4	TV	A	
9	F	54	HO	4	TV	B	
10	M	29	HO	4	PR	B	
11	M	39	TE	2	PR	C	
12	M	24	TE	3	ON	B	

Figure 1; A simple dataset containing some retail demographics and purchasing information. See Table 1 for an explanation of the abbreviations.

We will use the term "variable" to mean "column" and "data point" to mean "row". More often than not, a row holds observable data representing a customer's details; where this is not the case we will make it clear. Further, we will use the term "segmentation" where others may write "cluster analysis".

INTRODUCTION

To the professional data analyst, segmentation is a hard problem, which is considered unsolved according to the state of the art. Gian-Carlo Rota wrote in 1997 [1]:

"… Or the subject is important, but nobody understands what is going on; such is the case with quantum field theory, the distribution of primes, pattern recognition, and cluster analysis."

It seems not much has changed. We will go on to show how a minimal set of assumptions leads to a probabilistic segmentation model, which is invaluable for solving real-world e-banking and e-commerce problems.

DATA ANALYSIS AND SEGMENTATION

What do we mean by "*segmentation*"? In a marketing sense, we may have a good intuitive feel that customers are separated according to market demographics, with customers within each segment lying in some sense close together, and customers in different segments far apart (see Figure 2).

Unfortunately, this notion proves slippery for data analysis. If all the data were numerical variables (age, income, years at present address, etc.), then it would indeed be possible to define some notion of *distance* between any two points in the dataset, or more usefully, between any data point and the centre of each cluster.

There are classical analysis algorithms that perform adequately in this case. Such an algorithm partitions the datasets into, in this case, two subsets, one for each segment. The aim is to minimize the total distance between each point in both segments and the centre of the appropriate segment, the partitioning is considered optimal when this condition has been met.
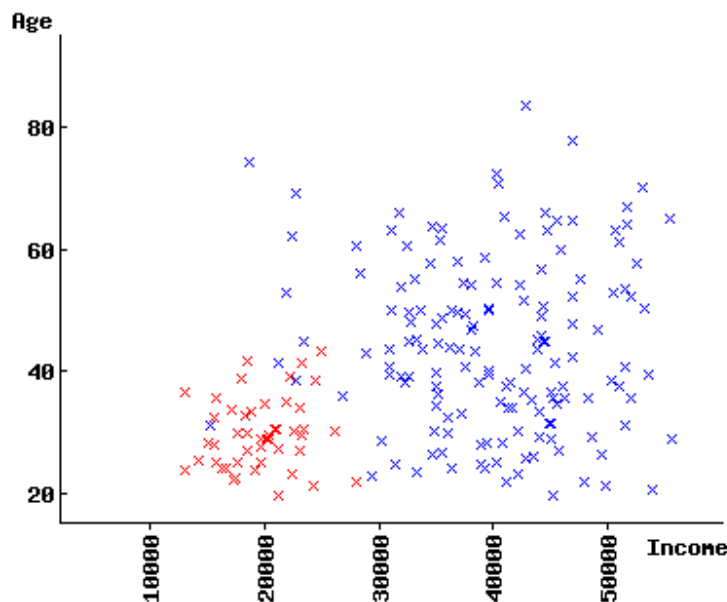

Figure 2; Two hundred observations of Age and Income sampled from a dataset in two segments.

Far more typically however, data contains a mixture of *numerical and categorical variables*, where a *categorical variable* takes on one of a discrete set of values (gender, sales channel, retail outlet, etc.) It is impossible to define distance without at least some concept of order amongst the categories. Sales channel is a good example. A sales lead may have been referred from any of 10 or 20 websites or campaigns, but what is the "distance" between *Google* and the *New York Times*?

Consider the following example, a dataset representing observations of *Sales Channel* against *Promotion* in a retail context (see Figure 3). *Sales Channel* has categories indicating from where each lead was referred, *Promotion* has categories that specify which of a number of special promotions was on offer. Again, there are two segments here and again colored red and blue. The brighter each cell, the more customers came via the corresponding sales channel and promotion. A cell that is more red in color represents more customers from the red segment than the blue, and vice-versa.

We may think we see some structure in this data, but it is illusory. Remember, there is no order among the categories - *Google* is neither "greater than" nor "less than" the *New York Times*. Hence the order of the cells arranged along the axes is completely arbitrary, and we can legitimately change the apparent shape of the dataset by permuting the categories.
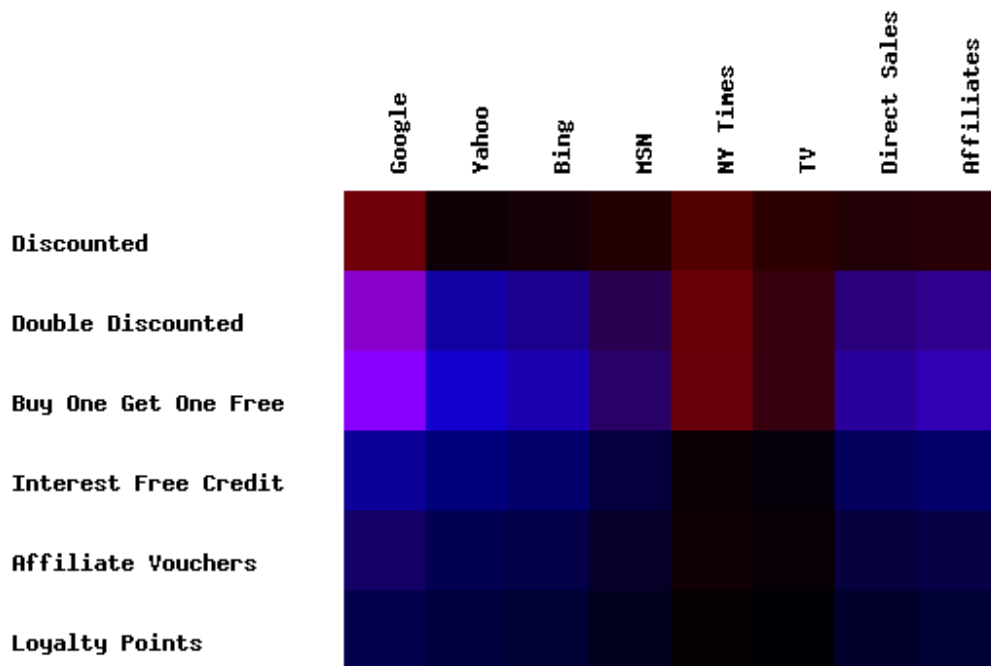

Figure 3; A dataset sampled in two categorical variables, Sales Channel and Promotion, again containing two segments.

Given that the shape of the data is mutable, it is meaningless to talk about cluster

"centers", or "distance to center". Hence admitting even a single categorical variable to the data spells disaster for any segmentation method that relies on a distance metrics.

There have been many approaches to building segmentation models that work round this problem. Some require that the number of segments be known in advance (impossible in practice); others that every variable in the dataset be categorical or continuous (unlikely in practice), or make unwarranted assumptions (such that all variables be mutually independent). The last of these assumptions is particularly insidious, as mutual independence effectively precludes modeling segmentation. Further, it is usually the case that more than one of these assumptions is required to build a model that captures something of the shape of the data in a simple form. It is interesting to note, however, that some useful results have been produced despite these limitations [2, 3].

If we are to build a model that accurately captures the shape of a segmented dataset whilst making minimal assumptions, we must think about how little we can assume about the properties of an arbitrary variable, irrespective of its shape. Any variable has a probability distribution; and the probability calculus, built on well-founded axioms, gives us a way to manipulate and combine these distributions. A segmented joint distribution that represents the shape of the whole dataset is indeed possible to compute from just the fundamentals of probability theory, with much weaker assumptions than independence over all variables.

In the past, such a probabilistic model would have been hard to compute in practice, owing to the somewhat large number of parameters involved in the model. A continuous distribution has 2 parameters, *mean* and *width*, which is quite manageable, but a categorical distribution has a total of 1 parameter (probability) for each category, less 1. Because the total of a properly-formed probability distribution is constrained to 1, so knowing *n-1* of the *n* parameters gives us the *nth*. Additionally, each segment has a *relative importance*; it is clear from the data in Figure 1 that there are about twice as many blue data points as red. This adds another *k-1* parameters for *k* segments, again because the total is constrained to 1.

Looking again at the data in Figure 1, we can compute the number of parameters needed for the distribution of each variable from Table 1:

Table 1
Variable types used in data shown in Figure 1

Column Name	Meaning	Variable Type	Possible Values	Number of Parameters
GENDER	Customer's gender	Categorical	M, F	1
AGE	Customer's age	Continuous	18-95	2
HOMEOWNER	Does the customer own their own home ?	Categorical	HO (homeowner), TE (tenant)	1
BEDS	Number of bedrooms in customer's home	Discrete	1-7	1
CHANNEL	Sales Channel	Categorical	ON (online), PR (print), TV (television), DI (direct)	3
PRODUCT	Product Code	Categorical	A, B, C, D	3

You may notice that we have introduced a new variable type, "*Discrete",* for the BEDS variable. Although it takes numerical values, this variable follows a "counting distribution", which means that it only has one parameter, *mean.*

So adding up the number of parameters in the final column, we get a total of 11, plus one for the *segment relative importance*, thus 12 per segment. For 4 segments, this would be a total of 4 x 12 − 1 = 47. The total is reduced by 1 because of the overall constraint that the *relative importance* must sum to 1.

In this simple example, considering any reasonable number of segments leads to a fairly modest number of parameters, definitely within the grasp of classical optimization techniques. However, real-world applications often contain several categorical variables with 20 or 30 possible values, leading to a couple of hundred parameters per segment, and up to thousand parameters for a multi-segmented model.

Fortunately, however, recent advances in algorithm technology [4, 5] allow computation of any model in a number of steps proportional to the square of the number of parameters, and at this point, these techniques become feasible in practice. Of course, the ever-increasing power of computer hardware helps as well.

Such a model is not just possible, but desirable, for the following three reasons. Firstly, models constructed using these techniques are very robust. Missing data values, that is, where the value of one or more variables is undefined, are easily and correctly dealt with

by the modeling algorithm, and so are the outlier points that indicate faulty data, erroneous procedures, or areas where our assumptions might be invalid. Secondly, very efficient use is made of the data – taking a random sample of a few tens of thousands of data points is enough to well-define the shape of the dataset, which helps in reducing the computation time. Thirdly, the form of the resultant model is very quick to compute on new data after the model has been defined, which means that the model can be easily applied to huge new datasets cheaply and quickly.

The power of the probabilistic model comes from estimating parameters "forwards"; that is, exploring possible values for all the parameters to evaluate the fit to the data. Classical techniques essentially work "backwards"; that is, trying to evaluate a single "best" set of parameters directly from the data. Much can go wrong with the classical approach, and these techniques are inherently less robust and stable.

The probabilistic model would be especially useful for improving ROI on advertising campaigns. It's essential that the number of market segments be well-defined, or the data will be "blurred" over some incorrect shape, and it will be impossible to accurately determine either the value of each segment, or the demographics that represent it. For example, a full analysis using a probabilistic algorithm of the data from which Figure 1 is drawn reveals three distinct market segments. The age demographic for each segment is shown in Figure 4.
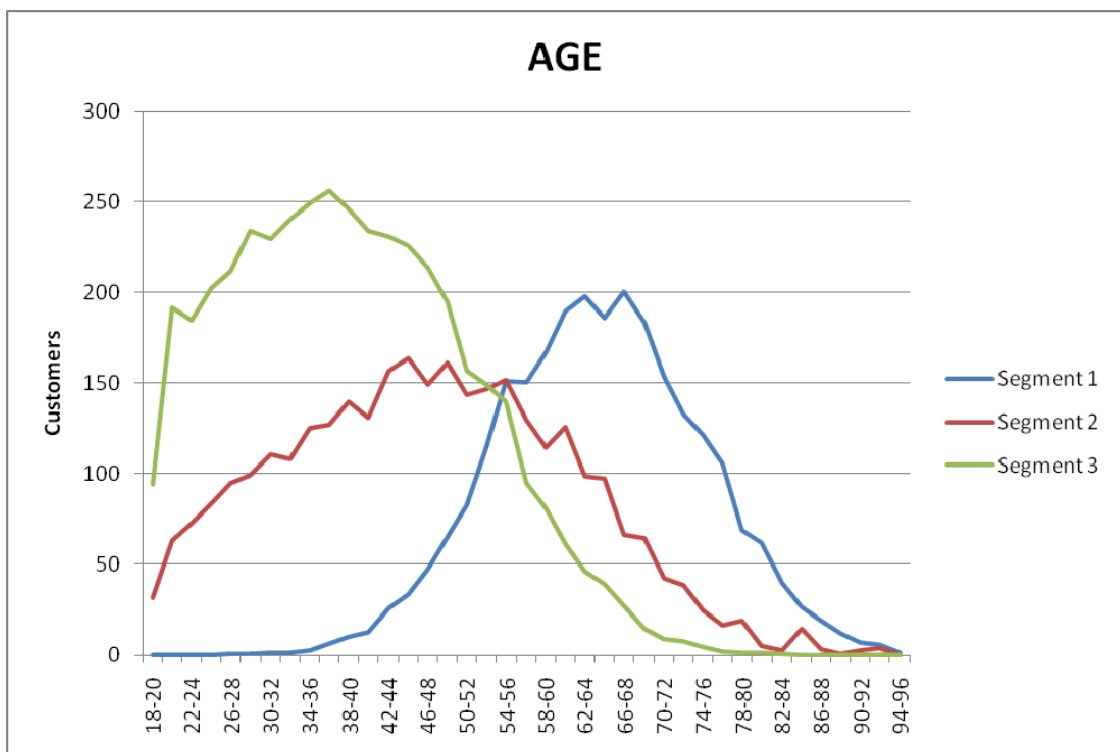

Figure 4; Age demographic of the data in Figure 1, revealing 3 segments

These segments, although quite distinct when plotted in this fashion, are not well-separated. The damage done by starting with the data and working "backwards" would almost certainly be enough to obscure this result.

SEGMENTATION MODELS IN E-BANKING AND E-COMMERCE

Market segmentation is essentially the process of partitioning the customer base into demographic groups, and then deciding which demographic group best represents each individual customer. Precisely what has been recorded in the data might determine the uses that this information can be put to.

Modern banks provide a wide range of financial products, designed to service a variety of market demographics. For the purpose of applying segmentation models such range of financial products can be considered in the context of a bank as a retailer.

Segments have a value, which ultimately can be measured financially. Some segments will be more valuable than others; this may be due to the fact that they represent customers who have been loyal for many years, or they may represent customers who have opted for more profitable products. In an e-banking context, a segment may be considered high value simply because it represents customer accounts with a high savings balance (or conversely, a high loan balance).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	QDATE	QTIME	GENDER	AGE	INCOME	HOMEOWNER	PROP TYPE	BEDS	LOAN	CHANNEL	SOLD		Segment
2	19/07/2008	12:52:55	F	81	47805	HO	BU	5	1929	ON	N		3
3	08/03/2008	13:01:45	M	56	26126	HO	FL	4	5678	PR	N		1
4	02/02/2008	11:34:00	M	45	29578	HO	DH	4	5453	PR	N		1
5	06/09/2008	13:07:12	M	34	30478	TE	SH	4	6060	PR	N		1
6	09/02/2008	13:08:04	F	35	16328	HO	SH	3	6364	PR	N		1
7	01/03/2008	17:51:29	M	29	20668	TE	FL	1	5959	ON	N		2
8	28/06/2008	12:32:50	F	67	32219	HO	TH	4	6783	PR	N		1
9	19/07/2008	11:26:13	M	43	30156	TE	TH	3	6882	PR	N		1
10	17/05/2008	12:52:49	F	43	27451	TE	TH	1	6615	PR	N		1
11	29/11/2008	12:29:35	M	34	24656	TE	SH	1	7986	ON	N		1
12	19/04/2008	12:59:12	M	50	22376	HO	TH	4	5860	PR	N		1
13	22/11/2008	12:48:55	M	65	25862	TE	TH	1	7894	TV	N		1
14	16/08/2008	12:56:25	F	34	19730	TE	MA	3	4621	TV	N		2

Figure 5; An example spreadsheet containing loan quotations, and showing segment assignments (three segments in total).

The observed variables in the data determine how the segments are defined, and which questions we can sensibly hope to answer. The data in figure 5 contain a variable indicating whether a sale was made from each quotation ("SOLD"). Hence a conversion ratio can be calculated for each segment (see Table 2). It is clear that segment 3 is the most valuable, with a conversion ratio approximately double either of the other 2 segments, so marketing activity should be directed towards the corresponding

demographic to maximize ROI.

Table 2
Conversion ratios and demographics for the data shown in Figure 5

Conversion Ratio	Segment Number	Age (mean)	Gender		Income (mean)
11.7%	3	64	M	38.8%	$38,900
			F	61.2%	
6.3%	2	40	M	64.1%	$21,900
			F	35.9%	
6.2%	1	44	M	81.3%	$25,000
			F	18.7%	

It costs advertising dollars to acquire new customers; therefore each customer initially has a negative value. We hope the customers' value to us will rise over time, and ideally reach a peak and stay there (see Figure 6).
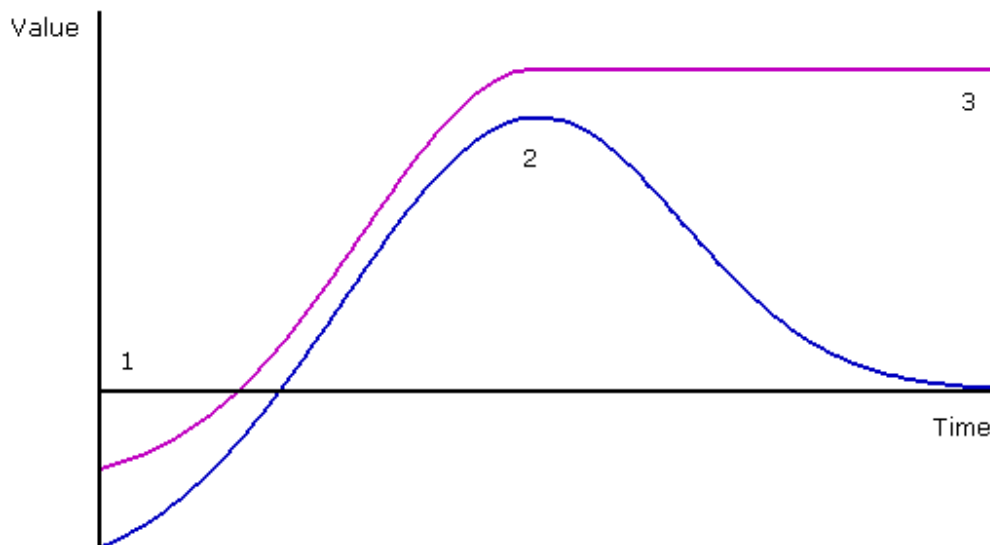
Figure 6; Typical customer lifecycle (blue), and optimized (purple)

As figure 6 shows, customer value is initially negative (labeled point 1), but cost can be reduced by optimizing the marketing campaign to target more valuable segments. Cross-selling opportunities can increase peak customer value (labeled point 2), and intervening with a promotion at the correct point in the customer lifecycle can maintain value over time (labeled point 3). These concepts are now discussed in more detail.

Looking first at how to minimize the cost of acquisition, we would need data from a marketing campaign, one data point per sales lead, that contains a variable indicating whether a sale was made. A segmentation model would then reveal customer demographics, in whatever other variables were collected as part of the campaign, and also yield conversion ratios for each segment. Subsequent campaigns could then be targeted to the more profitable segments. The data in Figure 5, and the results collated in Table 2 illustrate this.

Improving customer value over time can be as simple as identifying cross-selling opportunities. This could be achieved by using a dataset that represents sales of several products, one data point per customer per sale, and a variable indicating the type of product. A segmentation model would yield the associations between each type of product in each segment, that is, for a given demographic, the probability that each customer would be interested in car insurance given that they have already taken out home contents cover. The data in Figure 1 contain the necessary information; each customer record specifies which of four possible products was purchased (labeled A, B, C or D). These data represent actual sales, so the simplest possible metric for segment value is just the relative customer volume represented by each segment.

Table 3
Relative volumes from an analysis of the data in Figure 1

Relative Volume	Segment Number
42.8%	3
31.3%	2
25.9%	1

Looking at the distribution of product sales across the segments allows us to identify cross-selling opportunities (see Figure 7).
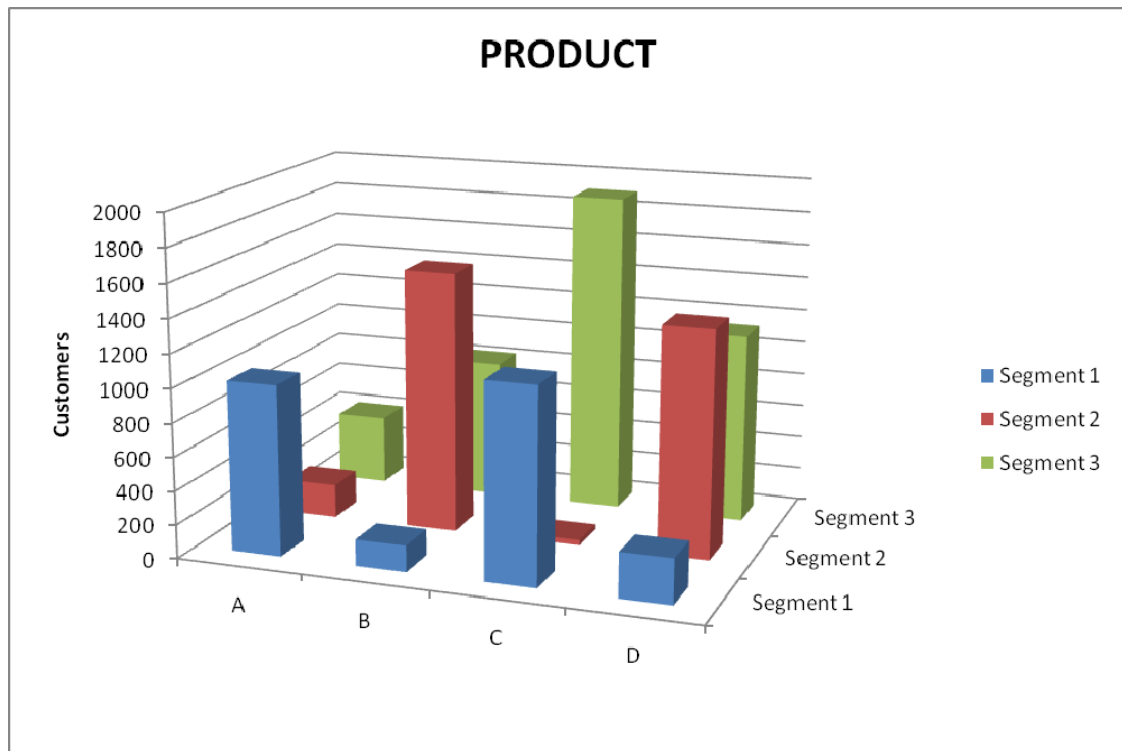
Figure 7; Distribution of product purchases for the data in Figure 1

It is apparent from Figure 7 that there are good opportunities in segment 1 to cross-sell products A and C, and in segment 2 to cross-sell products B and D. In segment 3, there is less clear opportunity, because the distribution of product is more even.

A final example of a pure segmentation technique is focused on optimizing marketing ROI. If marketing data were collected containing a variable indicating sales channel (in a simple example such as the one that follows, this could be "print", "online", "tv" or "direct sale", or more likely in a real-world application, something like referring website), by examining the high-value segments of a segmentation model, the optimal distribution across sales channels for future campaigns would be immediately apparent.

Referring again to the data in Figure 1 and summarized in Table 3, it would be wise to focus our efforts on segment 3 as representing the largest slice of the market. The distribution of sales channel across the segments is shown in Figure 8.
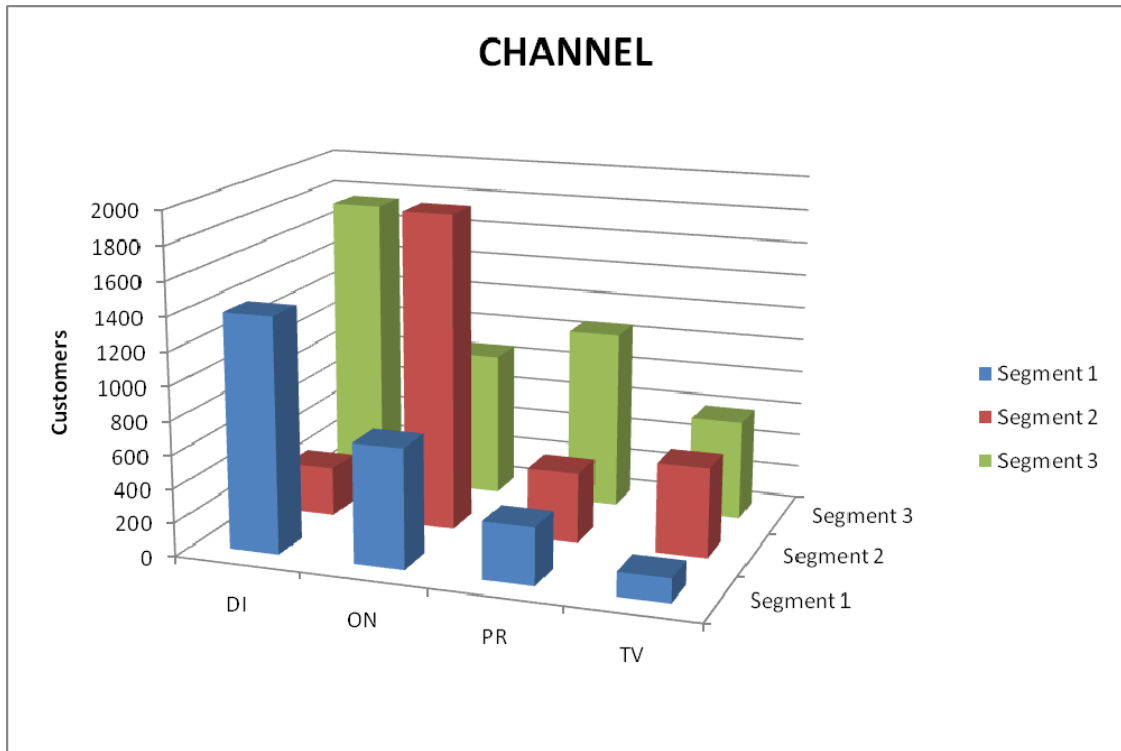
Figure 8; Distribution of sales channel across segments for the data in Figure 1

If we were to focus on segment 3, we would spend approximately almost half (actually more like two-fifths) the marketing budget on a direct marketing campaign, and spread the remainder about evenly across the remaining three channels.

PREDICTIVE MODELS – THE CUSTOMER LIFECYCLE

Predictive models require data that contains a transaction history for each customer – that is, multiple data points, one per transaction, for each customer (see Figure 9).

	A	B	C	D	E	F	G	H
1	GENDER	AGE	HOMEOWNER	BEDS	CHANNEL	VALUE	CUSTOMER	DATE
2	M	26	HO	3	PR	$151.54	X2141	14/12/2009
3	M	22	TE	1	PR	$32.71	X2508	07/06/2009
4	M	51	HO	5	PR	$51.24	X0586	02/07/2009
5	M	23	TE	4	PR	$30.34	X1583	12/04/2009
6	M	41	TE	4	DI	$186.89	X0355	23/03/2009
7	M	66	HO	1	ON	$259.75	X1397	14/08/2009
8	M	41	TE	4	ON	$47.47	X0924	15/09/2009
9	M	61	HO	3	PR	$115.66	X1586	27/07/2009
10	M	77	HO	5	ON	$163.50	X2282	27/05/2009
11	F	49	HO	2	TV	$64.94	X2522	03/08/2009
12	F	75	HO	4	PR	$243.55	X0837	04/08/2009
13	F	46	TE	4	PR	$124.59	X0955	21/07/2009

Figure 9; Sample transactions, one row per transaction, with customer reference and dollar values.

Classically, customer loyalty has been computed according to a Recency-Frequency-Monetary (RFM) model [6]. Customers who have purchased Recently, Frequently, or have spent large sums of Money are deemed to have a higher value. In the e-banking world, we usually look at purchases and standing orders – a large number of transactions on a e-banking website for a given month indicates a high likelihood that a customer will repeat this pattern the following month.

Recorded in the data are transaction date, and transaction value (dollar amount). To create a classical RFM analysis, a number of histogram bins must be manually selected for Recency, Frequency and Value variables. For instance, the Recency values might be separated into bins representing: customers with purchases within the last 7 days; between 8 and 30 days; and longer than 30 days.

Segments are then created from the intersection of the bins. If there were three bins for each variable, then the resulting matrix would have 27 possible combinations. The resulting segments can be ordered from most valuable (highest recency, frequency, and value) to least valuable (lowest recency, frequency, and value). See Table 4.

Table 4
Classical RFM analysis of the data in Figure 9, sorted by segment value, highest value first.

Recency (days)	Frequency (days)	Value ($)	Relative Volume (%)
less than 7	less than 7	greater than 200	0.00%
less than 7	less than 7	100 to 200	0.51%
less than 7	7 to 21	greater than 200	0.28%
7 to 21	less than 7	greater than 200	0.05%
less than 7	7 to 21	100 to 200	3.32%
7 to 21	less than 7	100 to 200	0.70%
7 to 21	7 to 21	greater than 200	0.47%
less than 7	less than 7	less than 100	2.15%
less than 7	greater than 21	greater than 200	0.66%
greater than 21	less than 7	greater than 200	0.05%
7 to 21	7 to 21	100 to 200	6.79%
less than 7	7 to 21	less than 100	2.67%
7 to 21	less than 7	less than 100	2.81%
less than 7	greater than 21	100 to 200	2.15%
7 to 21	greater than 21	greater than 200	1.69%
greater than 21	less than 7	100 to 200	2.20%
greater than 21	7 to 21	greater than 200	1.40%
7 to 21	7 to 21	less than 100	4.21%
7 to 21	greater than 21	100 to 200	4.26%
greater than 21	7 to 21	100 to 200	19.52%
less than 7	greater than 21	less than 100	0.37%
greater than 21	less than 7	less than 100	9.36%
greater than 21	greater than 21	greater than 200	3.79%
7 to 21	greater than 21	less than 100	0.51%
greater than 21	7 to 21	less than 100	15.87%
greater than 21	greater than 21	100 to 200	11.89%
greater than 21	greater than 21	less than 100	2.29%

There are three main problems with the classical approach. Firstly, it's very difficult in practice to choose appropriate bin start points and bin widths, without damaging the shape of the data. Any approach that requires a histogram be computed from continuous variables should be viewed with suspicion, in our opinion. The data are continuous, so use a continuous distribution (see Figure 10).
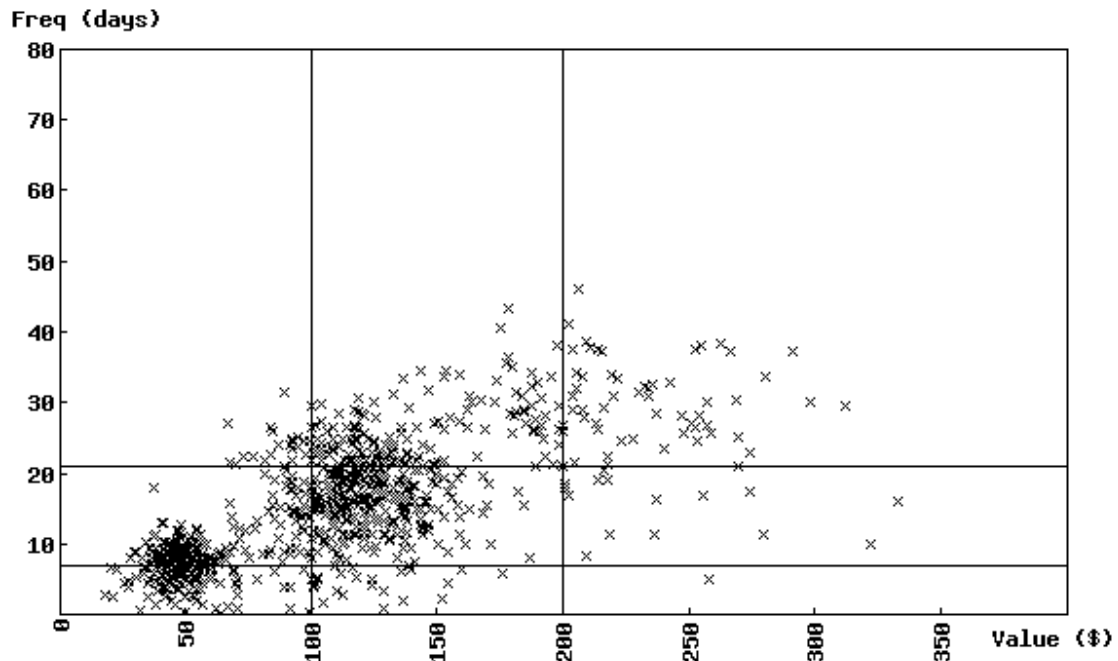
Figure 10; Frequency – Value plot for a classical RFM analysis of the data in Figure 9. Note that here the data points represent customers, not transactions, and the frequencies have been averaged for each customer.

Secondly, the classical approach ignores customer demographics in building the RFM model. There may well be significant information in the data that helps to define purchasing segments. It should be apparent that affluent, 20-something homeowners will have different spending patterns than retired, 60-something tenants, yet the RFM algorithm will treat all the data as one demographic segment.

Thirdly, this approach identifies the best, most reliable customers. These customers probably don't need an incentive to make a repeat purchase; we want to focus our efforts on the customers who are on the point of defection, to win them back as cheaply as possible. Unfortunately, the classical algorithm can't tell us how likely these customers are to remain with us.

A probabilistic algorithm can perform a good deal better. We proceed with a dataset containing not just transaction history, but also customer demographics. Probabilistic RFM proceeds by computing the frequency of the transactions (difference between transaction dates), and adding frequency as a separate dimension to the dataset. Now a segmentation analysis of the type we have been discussing will separate market segments not just according to customer demographics, but also transaction value and frequency; that is, demographic spending pattern. Finally, the date and value of each customer's final transaction to date is tested against this model, to yield a metric whose value is indicative of the customer's likelihood of making another transaction.

Table 5
Segmentation analysis of demographics and spending patterns for the data in Figure 9

Segment No	Relative Importance	Age (mean)	Gender		Frequency (days)	Value ($)
1 (Red)	27.8%	39.8	F	39.4%	7.0 +/- 5.4	44.38 +/- 15.57
			M	60.6%		
2 (Green)	11.5%	63.6	F	60.6%	26.1 +/- 22.1	224.22 +/- 69.10
			M	39.4%		
3 (Blue)	60.7%	44.5	F	19.4%	16.8 +/- 15.4	112.61 +/- 40.26
			M	80.6%		
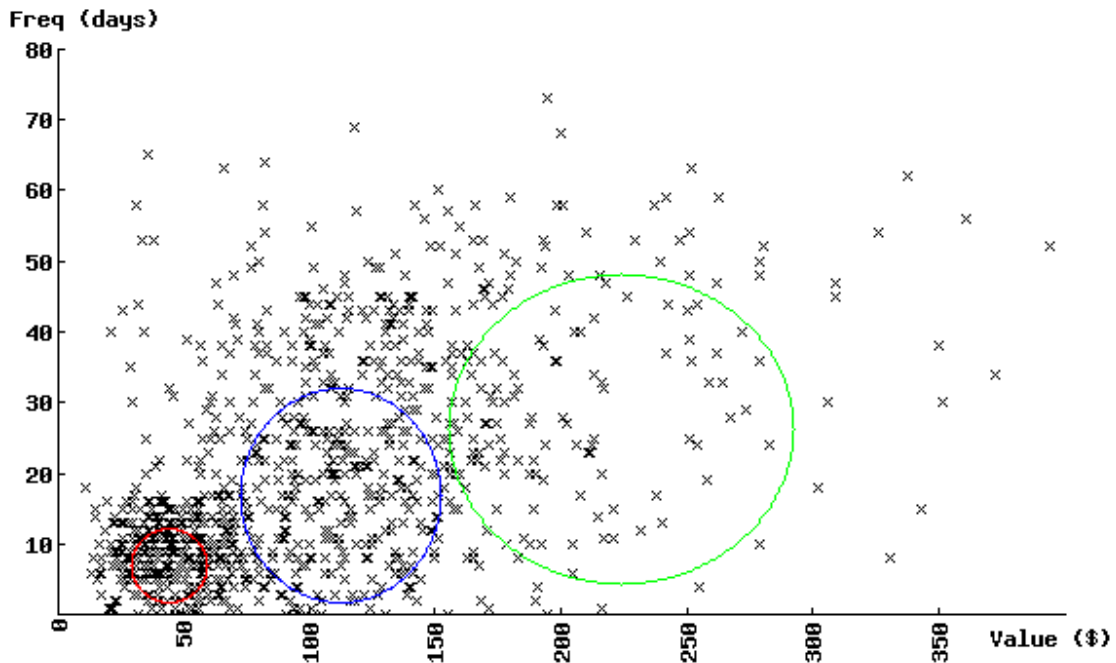
Figure 11; Segmented Frequency – Value model for the transaction data in Figure 9, summarized in Table 5. The ellipses represent 1 standard deviation contours around the center of each segment.

Each customer's demographic can then be tested simultaneously along with the customer's transaction history against the segmentation model. If the segmentation model is probabilistic in nature, rather than getting a single "best" segment, our analysis yields the distribution across segments, that is, probabilities that each customer falls within each segment (see Figure 12).
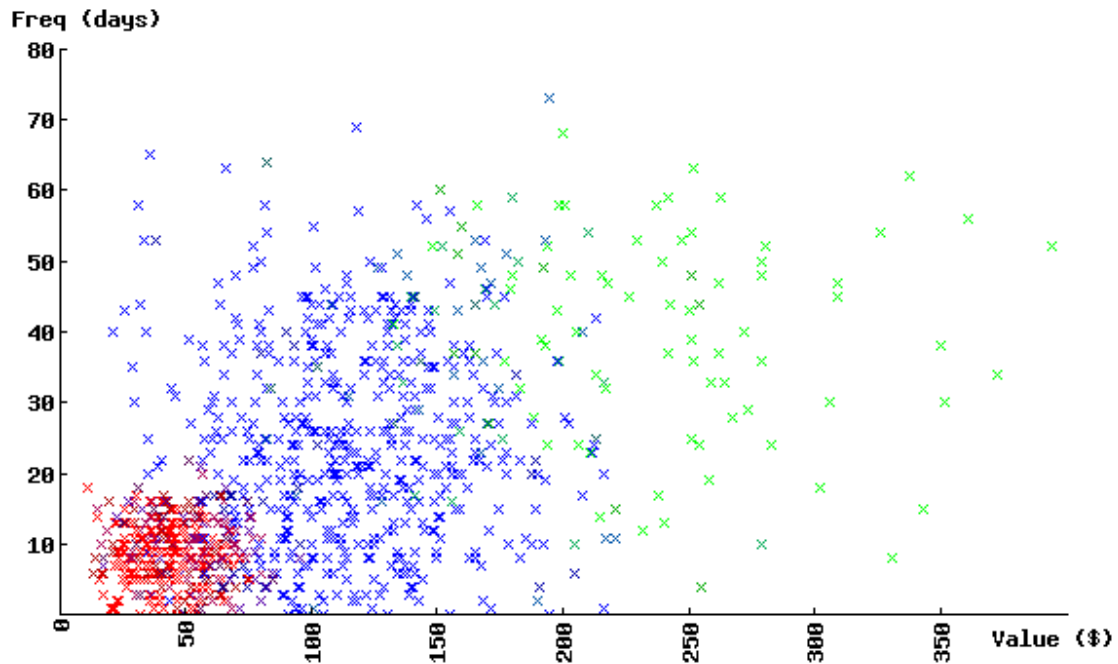
Figure 12; Probabilistic segment assignments for the transaction data in Figure 9. Color blending round the boundaries of the segments indicates assignments that are spread over multiple segments.

Given this distribution, and using the axioms of probability calculus [7], we can simply and quickly test the date of each customer's final transaction against all the segments, with the net effect that we get the best possible estimate of the loyalty metric, as all possibilities have been considered at differing levels of probability.

A final pay-off is that because the segmentation model is probabilistic, and we have computed the loyalty statistic by following the probability calculus, the loyalty metric is itself a properly defined probability, ranging from 0% (perfectly disloyal) to 100% (perfectly loyal). What you do with this metric is then up to you – we recommend a starting point of 50% loyal as a good place to try to influence the customer with fresh incentive, but the optimum actual figure is necessarily different in each case. Referring to Figure 13, we would like to influence the customers who have not yet defected (plotted red), and there is no need to target the customers still loyal (plotted green), so we aim at those customers plotted in a mid red/green shade.
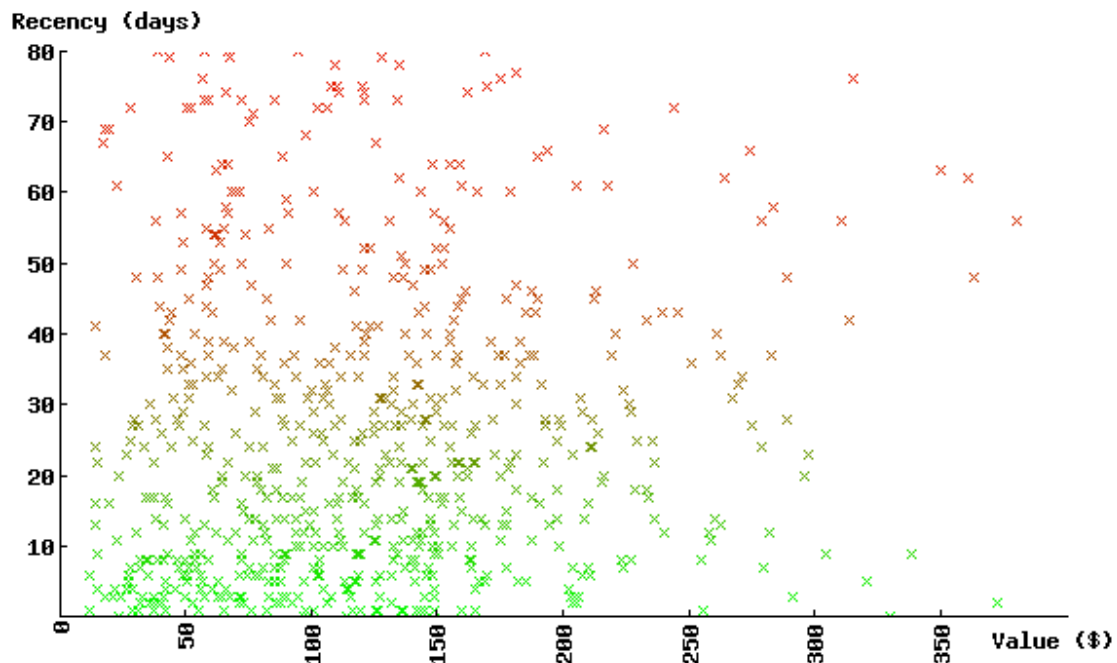
Figure 13; Loyalties for the data shown in Figure 9. Note that in this chart, the data points represent customers, not transactions. A data point in green represents a loyal customer, red represents a disloyal customer.

CONCLUSION

Probabilistic models hold a number of advantages over classical techniques, including efficient use of data and tolerance to undefined data values. Further, a probabilistic model will not merely assign each customer a single demographic segment, but will compute the distribution of each customer over all segments. This avoids data points on the boundary between segments being "forced" into either, but admits the possibility of all possible assignments. A robust segmentation technique is essential to draw actionable conclusions about the effectiveness of a given marketing campaign, to identify cross-selling opportunities, or minimize the cost of acquiring new customers. Such a segmentation technique, given a purchasing history for each customer, can also yield better insights into overall spending patterns than classical techniques are able to provide.

REFERENCES

[1] Gian-Carlo Rota (1997) Indiscrete Thoughts, Chapter XX (p. 216), Birkhäuser. Boston, Massachusetts, USA.

[2] A. L. McCutcheon (1987). "Latent class analysis". Quantitative Applications in the Social Sciences Series No. 64, Sage Publications, Thousand Oaks, California, USA

[3] Caruana, R. and Niculescu-Mizil, A. (2006), "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning,

[4] Bernd A. Berg (2004), "Markov Chain Monte Carlo Simulations and Their Statistical Analysis", World Scientific, Singapore

[5] Siddhartha Chib and Edward Greenberg (1995), "Understanding the Metropolis–Hastings Algorithm". American Statistician, 49(4), 327–335

[6] Wikipedia, "RFM", http://en.wikipedia.org/wiki/RFM

[7] Mathworld, "Probability Axioms", http://mathworld.wolfram.com/ProbabilityAxioms.html