Advanced Business Analytics - Power of Predictive Modeling 226161-0131

Adrianna Wołowiec

e-mail: *awolow1@sgh.waw.pl*

Foundations for Segmentation and Lifetime Value Models

What is Customer Lifetime Value?

Customer Lifetime Value (CLV) is the discounted sum of future cash flows attributed to the relationship with a customer. It estimates the profit that an organization will derive from a customer in the future. Maximizing CLV is one of the main goals of for-profit organizations.

$$CLV = \sum_{t=1}^{\infty} \frac{E(V_t)}{(1+d)^{t-1}}$$

V_t - the customer’s net contribution in period t

d – the discount rate

Customer Lifetime Value

One of the applications of CLV is in deciding how much an organisation could spend to acquire a customer. The acquisition costs are justified if they are less then the customer’s CLV. The aim is to identify prospective customers with high CLV and avoid aquiring those woth low CLV.

(e.g. media providers, telecommunicationsand cable companies, online retailers)

The CLV provides the best rationale for allocating marketing resources: organisations should invest marketing resources only in activities that increase CLV so that it is more then their costs.

Three main ways to increase the CLV of existing customers:

- Retain them longer
- Increase customers’ revenues
- Decrease the costs of serving them, marketing to them or both

The amount of money that can be spent on such tactics is informed by the change that they will have on a customer’s CLV.

Modelling CLV

Modelling CLV concerns especially businesses in which there is a contractual relationship with a customer – where there is an observable end to the relatioship.

Examples:

- Cell phone customers exit their contracts and stop paying their bill
- Media content subscriptions end when customers cancel their subscription or fail to renew it
- Health-club memebership expires at a certain time

Classes of CLV models

There are two types of CLV models:

1. Gone-for-good models
   - they assume customers who cancel the service will not return
   - in this case the most important issue is retaining customers over time
   - survival analysis models are used to study the time until a customer cancels
   - examples: simple and general retention models
2. Always-a-share models
   - they do not assume that customer inactivity implies the customer will never return
   - a retail customer who does not buy this month might come back next month
   - examples: migration model and data mining approaches to lifetime value

Segmentation models

Heterogeneity of customers

In almost all situations customers have different wants, needs, preferences and so on. Whenever such heterogeneity exists, companies that recognize and accommodate differences can achieve an advantage over competitors in a category. Not all needs of heterogeneous customers will be met with only one offering. A competitor can offer a better-targeted product and attract such customers.

One approach for addressing heterogeneity is segmentation. Customers with similar wants and needs are grouped into segments so that an organization can better meet the different needs. For this purpose companies implement customer segmentation strategies and clustering methods.

Business applications

1. **Market segmentation.**

An entire market is first segmented into homogeneous groups. The organization usually targets one market segment and develops a product or service and brand for this segment.

1. **Customization and personalization to subsegments.**

Within a market segment there will still be heterogeneity. Customization is when a firm allows its customers to configure the marketing mix to meet these heterogeneous needs more closely. The company enables the customization by offering up a menu of options and the customer decides.

Introduction to segmentation models

We have random sample of n people and assume that each person belongs to exactly one of K unobserved groups, which will be labeled $1, \dots, K$. The value K is fixed before estimating a model. Unobserved means that we do not know from which group a person comes. It is job of the model to find these groups and estimate how likely it is that each person comes from each group. These groups are also called clusters, types, segments or subsegments. Let $g_i \epsilon \{1, \dots, K\}$ be the true group membership of person $i = 1, \dots, n$. Instead of observing group membership, we have measured p variables that indicate group membership. Let x_{ij} be the observed value of variable j on person i.

Example 1: Newspaper readership

We have a random sample of $n = 2\,939$ Chicago residents and a crosstab of the time spent reading a newspaper during a week (octile numer) and the number of sections read

Time	**Number of Sections**								**Total**
	0	1	2	3	4	5	6	**7**	
0	370	0	0	0	0	0	0	0	370
1	9	127	119	55	34	55	34	21	454
2	3	72	94	68	48	74	48	29	436
3	1	43	107	48	40	63	52	38	392
4	0	21	58	39	32	47	30	25	252
5	0	15	40	23	25	51	38	32	224
6	0	21	67	54	39	90	59	71	401
7	0	15	47	31	51	103	67	96	410
Total	383	314	532	318	269	483	328	312	2 939

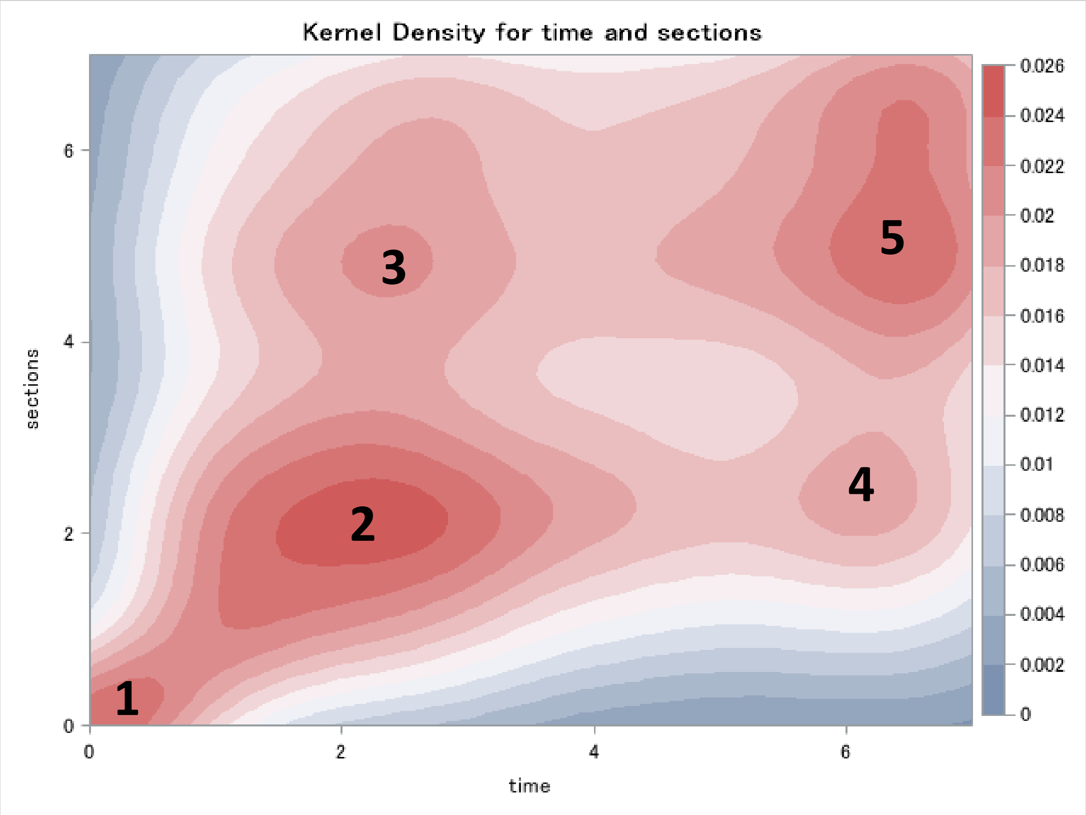
Are there types of newspaper readers?

Solution

With only one or two variables indicating group membership, we do not really need cluster analysis or any other segmentation model to identify segments. We can simply inspect frequency distributions, histograms, crosstabs and other descpritive statistics to identify natural groups.

When there are more than two manifest variables, such visual methods are less possible.

Solution – Kernel density plot



Kernel density plot is like a two-dimensional histogram, showing joint distribution of time and sections by representing areas with higher probabilities with darker shades of red and areas with lower probabilities in blue.

Segments:

1 - nonreaders (spend no time with a newspaper and read no sections)

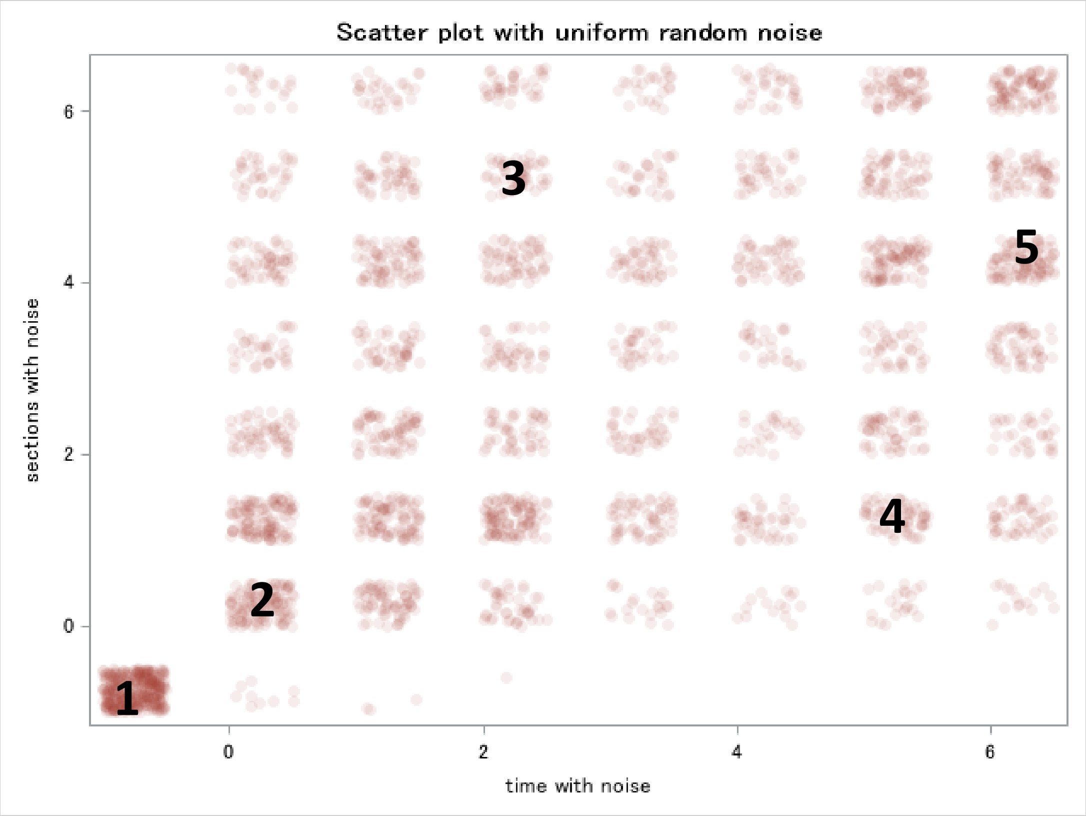
2 - light readers (spend little time with the paper and read only a small numer of sections)

3 - skimmers (skim many sections in the small amount of time)

4 - selective readers (spend sunstantial amount of time with the newspaper, but only read a narrow set of sections)

5 - heavy readers (spend a large amount of time with the newspaper and read most sections)

Solution – Jittered scatter plot



Jittering adds a small amount of uniform random noise to the values with the RANUNI function. Without jittering we would see one point for each combination of time and sections. Darker shades indicates larger counts.

Segments:

1 - nonreaders (spend no time with a newspaper and read no sections)

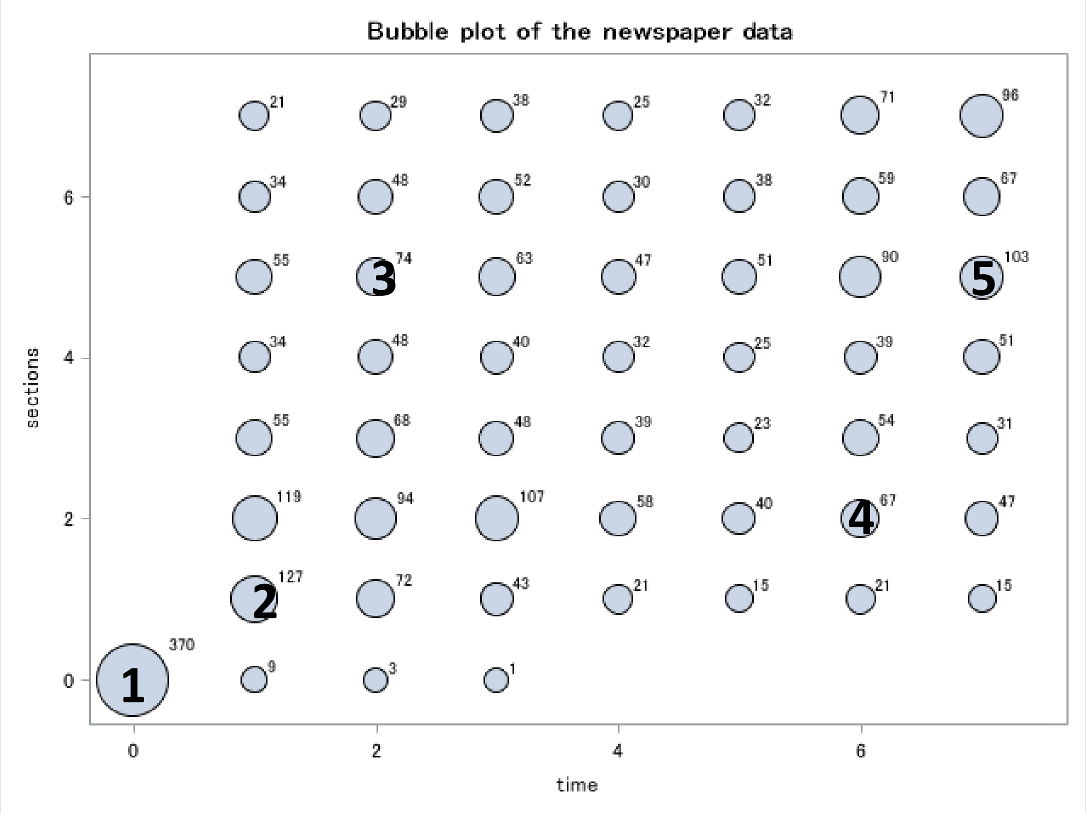
2 - light readers (spend little time with the paper and read only a small numer of sections)

3 - skimmers (skim many sections in the small amount of time)

4 - selective readers (spend sunstantial amount of time with the newspaper, but only read a narrow set of sections)

5 - heavy readers (spend a large amount of time with the newspaper and read most sections)

Solution – bubble plot



The area of the circles indicates the counts.

Segments:

1 - nonreaders (spend no time with a newspaper and read no sections)

2 - light readers (spend little time with the paper and read only a small numer of sections)

3 - skimmers (skim many sections in the small amount of time)

4 - selective readers (spend sunstantial amount of time with the newspaper, but only read a narrow set of sections)

5 - heavy readers (spend a large amount of time with the newspaper and read most sections)

Questions addressed by segmentation models

1. How many segments are there? In practice the analyst will have to decide how many segments are appropriate.
2. For a specified number of segments K and set of variables, what are natural groups? What are their characetristics (eg. class-conditional means – conditional on group membership)
3. How large is each segment? What is its share in the population?
4. To which segment does each respondent belong, considering the information we have on the person?
5. What actions should be taken, based on the segments? (eg. marketing actions, contact points, personalized offers)

Two classes of segmentation models

K-means clustering

- a special case of the more general finite mixture model
- comparatively simple to estimate
- available in commercial statistical software
- an easy transformation of data before estimation

Finite mixture model

- an alternative approach to segmentation
- ensures more advanced estimation and more accurate predictions

K-means model

The K-means model hypothesizes that:

$$x_{ij} = \mu_{jg_i} + e_{ij},$$

where:

- x_{ij} - the observed value of variable *j* for respondent *i*
- μ_{jg}- the true mean of all members of cluster *g* on variable *j*
- e_{ij}- an error term with mean E(e_{ij})=0 and variance σ^2, which is common across all variables and clusters; we assume that the errors have normal distributions

K-means model - implications

Implications of a common error variance (model assumption):

- the distribution of points from some cluster is round or spherical
- all of the clusters have the same shape and dispersion

When the manifest variables do not fall into round clumps of equal size, the K-means model can have problems (eg. clumps have elliptical shapes). The finite mixture model allows for other shapes and varying sizes. Alternatively, transforming the data before estimating the K-means model may also help to make the clusters more spherical (not resolved the equal-size issue).

K-means model vs categorical variables

K-means method generally works best with numerical variables. There are two approaches for incorporating categorical variables in a K-means analysis:

- partitioning – without a model – the observations on important categorical variables and applying K-mean to each partition
- creating dummy variables and applying the variable weighting methods

K-mean algorithm

1. **Initialize.** Select an initial set of cluster centers or seeds $\hat{\mu}_{jk}^{0}$, where k = 1,…,K and the superscript 0 indicates the iteration numer. Initalize iteration counter h = 1.
2. **Assign clusters.** Assign observation *i* to the cluster whose mean is closest. The Euclidean distance between point x_i and cluster center $\hat{\mu}_{k}^{h-1}$ is given by $d_{ik}^{h} = \sqrt{\sum_{j=1}^{p}\left(x_{ij} + \hat{\mu}_{jk}^{h-1}\right)^2}$. Observation i is assigned the cluster that is closest, that is, $g_i^h = argmin_k d_{ik}^h$. The $argmin_k$ indicates that g_i^h equals the value k giving the smallest value of d_{ik}^h.
3. **Compute cluster means.** With cluster assignments fixed, compute custer means. They are also called cluster centroids or centers $\hat{\mu}_{jk}^{h} = average\{x_{ij}\text{: } g_i^h = k\}$. In other words, the new estimate of cluster mean k is the simple average of all observations assigned to it ($g_i^h = k$).
4. **Compute SSE.** This is the sum of squared errors or total within-cluster variation. $SSE = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - \hat{\mu}_{jg_i^h}^{h}\right)^2 = \sum_{i=1}^{n} d_{ig_i^h}^2$. The errors are really distances between each observation and its closest center.
5. **Loop.** Let h = h + 1 and return to step 2 until the convergence criterion is satisfied or the maximum numer of iterations is exceeded

K-mean algorithm

This algorithm attempts to minimize SSE, which is a combinatorial optimization problem and is computationally very difficult. In particular, this algorithm is not guaranteed to converge to the optimal solution, and different starting seeds might produce different solutions. Because of these problems, analysts might want to estimate the K-means solution multiple times with different starting seeds and select the solution with the smallest value of SSE.

Example 2: Newspaper readership

Find the five cluster solution for the newspaper data using PROC FASTCLUS. The number of people in the sample with each unique combination of time and section values is given by the count variable.

As a reminder:

We have a random sample of $n = 2\,939$ Chicago residents and a crosstab of the time spent reading a newspaper during a week (octile numer) and the number of sections read

Statistics' formulas

RMS Std Deviation for cluster *k* = $\boldsymbol{\sqrt{\frac{1}{p(n_k-1)}\sum_{g_i=k}\sum_{j=1}^{p}\left(x_{ij}-\hat{\mu}_{jk}\right)^2}}$,

where $\hat{\mu}_{jk}$ is the final cluster mean of variable *j* in cluster *k*

Total STD of variable *j* = $\boldsymbol{\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_{ij}-\hat{\mu}_{j.}\right)^2}}$,

where $\hat{\mu}_{j.}$ is the overall mean of variable *j*

Within STD for variable *j* = $\boldsymbol{\sqrt{\frac{1}{(n-K)}\sum_{i=1}^{n}\left(x_{ij}-\hat{\mu}_{jg_i}\right)^2}}$,

where $\hat{\mu}_{jg_i}$ is the overall mean of variable *j* in cluster *k* to which observation *i* was assigned

R-square = $\boldsymbol{1-\left(\frac{Within\ STD}{Total\ STD}\right)^2}$

Example 3: Newspaper readership

Earlier we suggested that K-means is sensitive to starting values (initial seeds). Estimate the five-cluster solution to the newspaper example with 100 different random seeds.

Transformation of data

We mentioned earlier that it is necessery before building cultering model to transform data. By newspaper readership example the manifest variables had comparable value range, so there was no need to transform data. Otherwise, we need to preprocess the data before modeling.

Example 4: Theater attitude

A group of theaters want to attract people who do not currently attend theater productions. Therefore the group conducted a marketing research study in which they recruited random sample of ca. 3 000 adults to complete the survey. The respondents were asked 14 questions measuring their attitudes and beliefs about theater using semantic differential scales. As a result of exploratory factor analysis we have five dimensions:

1. **attitude** – the 10 questions averaged (scale from 0 to 7)
2. **planning** – 1 = spur of the moment; 7 = requires planning
3. **parents** – 1 = my parents disliked; 7 = my parents liked plays
4. **goodval** – 1 = too expensive; 7 = good value for the money
5. **getto** – 1 = hard to get to; 7 = easy to get to

What are the types of people with respect to their attitudes toward theater?

Read the data set into SAS, standardize the variables and generate the one-, two- and three-cluster solutions. Interpret the three-cluster solution.

Transformation of variables

Transforming a single variable means applying some function to the variable, such as the logarithm, Z-score formula or both. The transformed values are then used in the cluster analysis. K-means clustering is highly sensitive to variable scaling and it is therefore important to have commensurate units. K-means also produces undesirable results when the distribution of the manifest variables is skewed.

The formula for standardized value (a z-score):

$$z = \frac{X - \mu}{\sigma},$$

where: X is the observed value, μ is the population mean, σ is the standard deviation.

In situations where the variables were measured with commensurate units and the distributions are not highly skewed, there is not a strong need for standarization of any kind, and the analyst might use the raw variables to estimate the cluster model.

PCA – principal components analysis

Another way to visualize a cluster solution is to use principal components analysis (PCA). PCA finds a lower-dimensional representation of multivariate data. In Example 4 (Theater attitude) we have five-dimensional observations. PCA identifies the best two-dimensional (n=2 in PROC PRINCOMP) and projects the five-dimensional points onto the plane. Best has two equivalent meanings: (1) the plane shows as much variation in the original data as possible (that is, the variance of the two-dimensional observations is maximized) and (2) the plane minimizes the amount of information that is lost (that is, the sum of squared orthogonal deviations from the original points to their projections onto the plane is minimized). We can think of PCA as showing the best shadow of the five-dimensional points on a plane.

Summary

A proces for building segmentations:

1. **Select variables**
2. **Transform variables** (if necessery)
3. **Find various cluster solutions and compare them** (eg. using pseudo F statistics)
4. **Choose the best cluster solution and summarize it**