Data Mining And Strategic Marketing In The Airline Industry

Lisa Pritscher and Hans Feyen

Atraxis AG, Swissair Group, Data Mining and Analysis, CKCB
CH-8058 Zurich-Airport, Switzerland
epritsch@atraxis.com
hfeyen@atraxis.com

Abstract. In the airline industry, data analysis and data mining are a prerequisite to push customer relationship management (CRM) ahead. Knowledge about data mining methods, marketing strategies and airline business processes has to be combined to successfully implement CRM. This paper is a case study and gives an overview about distinct issues, which have to be taken into account in order to provide a first solution to run CRM processes. We do not focus on each individual task of the project; rather we give a sketch about important steps like data preparation, customer valuation and segmentation and also explain the limitation of the solutions.

1. Introduction

The focus of many industries changes towards a more clearly customer orientation. While competition is increasing, there is a need to better understand customers, and to quickly respond to their individual needs and wants. Thus, the objective is to increase the share of wallet for each individual customer and save costs by focusing on more targeted promotions.

In the airline industry this development is rather at a starting point. Nevertheless, due to frequent flyer programs there is a wealth of data available, which allows to get a better understanding of customer types and customer behavior. Most airlines introduced a frequent flyer program in order to increase and award loyalty of their customers. The key program features are mileage accrual (members can earn miles for air travel, but also for activities like hotel stays, car rental and credit card usage) and mileage redemption (members can spend miles for air travel, hotel stays etc.). Thus, the currency of such a program is miles. Moreover, the program is also used to identify high value customers and provide them with special services and benefits (like lounge access and upgrades) by the means of a top tier program. However, since these data are collected only for administrative usage, features like monetary measures, which are relevant for CRM, are missing. A second issue, which has to be noticed, is that using this data source only part of the current customers are considered, since there are also many customers of an airline, which are non-members of the frequent flyer program.

In a recent project, our objective is to explore the available databases by use of data mining methods and to provide a suggestion of how to prepare an environment for implementing efficient customer relationship management processes. To enable successful CRM, the first task is to identify market segments containing customers with high profit potential (see Thearling [4]). Thus, our main objective is to provide customer segmentation with respect to all important dimensions of customers needs and value. One requirement for this are significant data. The exploration and preparation of the data landscape is a major task, since it was the first analysis on a customer level. One important task is to derive an appropriate measure for customer value, in order to assess the utility and the gain of later marketing campaigns (see Gallant at al. [1]).

Our focus on the project is to develop a rather straightforward solution that can be quickly implemented in an operative system and that can also be scaled-up in future. Thus, we mainly apply rather popular data mining methods like K-means, Kohonen self-organizing networks and classification trees (see Ripley [2] for a detailed description about the methods). Saarenvirta [3] also

describes a case study, which is concerned with similar source of data and gives solution for slightly different problems.

During the project we are concerned with all typical stages, characterizing a data mining process. The paper is organized according to CRISP-DM process cycle, and gives a sketch of the complete project. However, we describe some parts in more details in order to avoid only giving a high level report.

2. Business understanding

For data mining to impact a business, it needs to have relevance to the underlying business process. Data mining is part of much larger series of steps that takes place between a company and its customers. The way in which data mining impacts a business depends on the business process, not on the data mining process. Thus, the key success factor is to start with clarifying the stage of the CRM solution and the current business objectives.

2.1 Business Goals

The marketing department was in an initial phase to build-up a CRM environment, which allows to serve airline customers individually. Thus, the main objective was to enrich the knowledge about individual customers leading to new strategic customer segments. These segmentation results are mainly needed for marketing concerns (e.g. promotions, targeted campaigns). Additionally, the new knowledge can also improve the customer service (e.g. makes the information available for the customer call centers).

There are several business requirements, which have to be taken into account. The marketing experts must be convinced to use the newly provided customer segments. Therefore, the segmentation must be explainable and the added value must be evident. One key issue is to assign a monetary value to each passenger, which can be used to calculate profitability based on segmentation results and also allows to identify core customers. There are some underlying structures, which need to be reflected in strategic customer segments in order to be considered within marketing campaigns.

- Region: The needs of the customers can be satisfied more successfully, if the geographical preferences of customers are well known.
- Market: Based on a sales view, there are distinct markets that have to be served with their own means. We distinguish between home market, third market serving connecting passengers, third market serving stopping passengers and other markets.
- Travel preferences: To be able to give customers the right offer, it is necessary to know which travel conditions they prefer.
- Travel behavior: One important information about customers is to know which type of tickets is purchased regularly. The type of ticket is highly correlated with the purpose of the journey, e.g. leisure or business concerns.

The last requirement concerns the existing data landscape. Since the results of the study should be available as soon as possible, the first solution must be derived within the current data landscape, but there must be the opportunity to expand the solution.

2.2 Data mining goals

The current customer value of a passenger is based on individual mileage, but it is known that mileage is a rather rough measure for customer profitability. Thus, the first data mining goal is to combine distinct data sources and to derive a forecasting model that yield a reliable revenue value for each customer, based on flight activities and the booking history.

The strategic customer segmentation is a classical unsupervised learning problem. In order to provide segments, which can be explained to marketing experts, we focus on data preparation and an exploratory data analysis. This allows us to identify few but substantial attributes as input for the data modeling phase.

We make separate customer segmentations, to ensure that all distinct requirements are taken into account. The result of this approach is an assignment of all members to a segment in each of the views. Since we treat the segmentations separately, we are able to discover more homogenous segments and the information given by each segment is more appropriate. Moreover, using the results from multiple segmentation views, a more detailed characterization of the customers is possible and one has the flexibility to design marketing actions based on each segmentation view. Nevertheless, the multiple view segmentation can still be summarized into few 'grouped' segments.

The following list gives some ideas about how the input of the distinct segmentation views looks like:

- Region: Customers can be separated according to their favorite origins and destinations of their journeys.
- Market: Customers can be separated according to the market, they are using, and the place, they are living.
- Travel preferences: Customers can be segmented according to the ratio of longhaul (intercontinental) flights to shorthaul flights, according to the ratio of high to low fare tickets and according to the preferred combination of both.
- Travel behavior: There is no information of the type of ticket and its price in the database available. Since the level of the price is mainly based on rules concerning time and booking restriction, some evidence is obtained by investigating the temporal pattern of a journey. Moreover, this information often reveals the travel purpose.

3. Data understanding

Given the data mining goals, we have to investigate which data are available and might be useful for achieving the goals. In this case most of the data are collected for administrative reasons and we have to decide how to close the gap between data requirements for marketing reasons and the data situation given.

3.1 Data situation

Our primary source of data is the frequent flyer program database. In focusing on this group of customers, we have a selection bias compared to all airline customers.

For each member, data about demographic figures as well as the current state within the program are collected in a database. A second database contains a list of all single past flight legs, which contains departure and arrival airports, some booking information as well as the member id. Supplier activities like credit card usage are also gathered in an additional database.

There is no information available on passengers' revenue, since the frequent flyer database is used for administrative issues only. In order to complete the flight activities in the frequent flyer database with revenue data, we have to assign the corresponding values to the individual flight segments from a sales information database, where revenue information about individual bookings is available.

3.2 Data preparation

The target population of our project contains not all customers, but the members of the frequent flyer program of an airline. For this population, the data quality is reliable and the data about flight activities are rather complete.

We draw a random sample out of all members of the frequent flyer program database (n =70’000). For these members, we extract all flight activities (about 700’000) as well as all supplier activities (about 300’000) for a 24-month period. We restrict the database to members of certain program countries and members which had at least one flight activity in the period given (about 60 % of all members in the sample). The final sample size is about 40’000 customers. For the segmentation, one quarter of the target population is used as training sub-sample; another quarter is used as validation sub-sample. The remaining data are preserved for further validation studies.

In an explanatory data analysis, we investigate the content of the distinct data sources. It turns out that the most valuable information (with respect to the observed variation between customer groups, and correlation to customer value measures) is given by the historical flight activities. However, data transformations and aggregation, as described in the next section, are needed.

An important issue is to derive a monetary customer value. We extract sales information concerning the flight activities observed within the sample and try to assign a revenue value to each flight activity. A key identification variable, which is necessary for a unique match, is only available in 45 % of all flight activities. In Section 4, we discuss how we complete the revenue information for all flight activities considered.

4. Data transformation and aggregation

Within this section, we will focus on two examples of the data preparation task. Valuable information about the business can often only be revealed, if subsequent observations are combined. Especially, this is important in our case, where we have to derive key measures about behavioral data aggregated on member level. The success of modeling can only be measured, if the customer value in terms of revenue is available for all flight activities, i.e. we have to decide how to impute missing values.

4.1. Requirements

Considering single flight segments gives little indication on the type of customer and the travel purpose. Particularly, one cannot decide whether a passenger flying the segment Munich to Zurich arrives at the final destination or at the first stop of a multisegment journey like Munich - Zurich - New York, nor whether the flight segment is part of a round trip. However, patching together single flight segments, which are likely to belong to the same itinerary, enables one to get information about the routing and the duration of a journey. Using this kind of information yield an improved estimate of the prices, paid for the tickets, as well as gives an impression about the travel purpose of the customer. This more specific information is even more important, since we have to aggregate data.

The main driver for a successful targeted marketing is knowledge about customer value. Financial measures attributed to each customer are better suited to estimate customer value than miles. This becomes clear from the following example. The price range for an economy round trip Zurich-New York is 590-2,897 CHF, but each passenger receives the same mileage reward of 3,919 miles per leg. The price range for economy round trip Zurich-Helsinki ranges from 449-2,531 CHF. Once again, the same mileage reward of 1,105 for each leg is awarded. Passengers, who fly business class, will receive 1658 miles per leg while paying 2812 CHF. In the mileage scheme, a passenger flying to New York is always worthier (in terms of miles) than a passenger flying to Helsinki even if the latter flies business class.

As mentioned above, we have the complete information only in 45 % of all flights. It is essential to develop an estimation model for two reasons. Firstly, we need to assign a revenue value to each flight activity in order to provide a reliable customer valuation for the complete individual history. If we assign a zero value or a mean value, the results will be biased (see Saarenvirta [3] for a detailed discussion about the missing value task). Secondly, a model provides a first revenue estimate until the actual sales information is available, usually some time after operating the flight.

4.2 Solutions

Combing subsequent flight segments allows to reveal similar behavioral patterns of customers and allows to derive further key figures describing this itinerary. We developed an algorithm to identify 'trips': any logical sequence of flight segments belonging to the same itinerary. The member's travel purpose (e.g. leisure or business or mixed) is revealed by the properties of their 'trips'. The fact of returning within the week or stay over the weekend is highly correlated to the ticket price and gives an indication for the purpose. The definition of a 'trip' consists of three major steps:

1. Ordering sequence of segments: For each customer, the flight segments are sorted by date and departure time. If departure time is missing, then the algorithm still provides a sensible ordering of the flight segments.
2. Rules for sequential segments within one country: Often flights within a certain area (e.g. country) are connecting flights. Thus, all segments sequentially flown within one country are gathered, except certain rules are not valid.
3. Trips are only allowed to have following country patterns (A/B/C/D): Most itineraries can be described with rather simple patterns. We only consider following sequences:
   - 1 country - return: A(R)
   - 1 country - no return: A
   - 2 country - return: AB(R) ABA (B with/without R)
   - 2 country - no return: AB
   - 3 countries - return: ABC(R) ABCA ABCB ABCBA (C with/without R)
   - 3 countries – no return: ABC
   - 4 countries – no return: ABCD

   A new trip starts if either the rules are violated or some local and temporal conditions are exceeded. The restrictions are necessary in order to avoid collecting too many segments to one itinerary. For each trip, the following attributes are derived:

RET_WW	return trip & trip within 1 week & no weekend
RET_WE	return trip & trip within 1 week & includes weekend
RET_LS	return trip & trip takes more than 1 week (long stay)

There are distinct recommendations, how to impute missing values of key attributes. One widely recommended approach is to build a regression model based on highly correlated variables (e.g. miles, compartment, region) and to forecast the missing value using the prediction and the distribution of the nonmissing records. Nevertheless, we choose an alternative approach due to the origin of the data. Revenue values on a certain flight segment result from several fares, which are published for this flight segment and are usually constant within a certain period. For this reason, it is more accurate to base a model on observed values for a given flight (and robustify the aggregation level), rather to develop a model which try to explain all flights. Since fares are no random values, it is better to use observed values within a subgroup than to build a model based on main effects.

To introduce a suitable customer valuation, we developed the following procedure:

1. Merge sales information to flight activities: The revenue data is attributed to those flight activities, where a unique match is possible.
2. Applying the 'Trip Builder' algorithm to differentiate between local and connecting Flights: The revenue of a flight largely differs for local and connecting flights. Thus, it is necessary to reconstruct the trip made by a customer to identify whether a flight was local or connecting.
3. Development and application of a model to prepare appropriate look-up tables: The forecasting model essentially calculates mean revenue values for a flight activity by using historical data of similar flight segments. The results of the model are various look-up tables by which revenue values are matched to flight activities with unknown revenue. An entry of such a look-up table is the average revenue of a certain subpopulation, determined by attributes like airport pair, season, booking class, connecting flag and round trip flag. The matching procedure is applied in several steps, while the matching criteria are sequentially relaxed.

5. Model building and evaluation

The main objective of the project was to derive strategic customer segments. We apply a clustering algorithm in order to identify groups which are different from each other according to their product mix as well as to their value, but whose members are very similar to each other. Clustering algorithms are appropriate, if there is no predefined segmentation. We illustrate our model building and model evaluation approach in more details by means of an example, the travel behavior of customers.

5.1 Nature of the task

Airline customers can be characterized by their travel purpose (i.e. business or leisure travel) and by the type of tickets they purchase regularly. As mentioned above, the newly derived attributes, which describe the trips of customers and are aggregated on member level, give a good indication for this view. Further potential attributes are properties, which are given by the booking process. The objective of the segmentation is to identify typical customer groups, which are similar with respect to the frequency of the product usage as well as with respect to the product mix, but which significantly differ between the groups. To solve this data mining problem, clustering methods are appropriate.

One important requirement for the segmentation is that the derived segments can be explained to marketing experts: If they can understand what has been discovered, they will trust it and put it into use. One important part of this problem is to closely interact with business experts, allowing them to interact with the output.

The second concern is the validity of results: the results are only valuable, if they are representative for the complete database and the rules are robust. Moreover, the results are only relevant to the business problem, if they differ between high and low potential customer groups. An alternative way of confirmation is to perform plausibility checks based on additional important variables.

5.2 Approach

We use some of the common algorithm to perform clustering: K-means is an algorithm, which pass through customers, assigning each to the closest existing cluster center. Kohonen feature maps develops segments on a planar grid with an explicit notion of neighboring segments (see Ripley [2] for a detailed description). We apply both methods, in order to be certain that the results are not depending on the method applied. Thus we only concentrate on consistent solutions. Particularly, we are applying several runs of K-means, given between 6 to 10 clusters, and we are also applying the Kohonen network with a grid of 6 x 6 (36) output layers. If the K-means clusters are well separated in the Kohonen feature map and the properties of neighboring clusters are sensible, then we assume to obtain a stable clustering solution.

One critical task of clustering algorithm is to choose the right variables and the right scale, respectively. We only use normalized counts and numeric variables and proportion of occurrences instead of categorical variables in order to ensure that the influence of all variables is similar. Using forward selection for variable screening, leads to clusters, which rather simple and can be explained in business terms. We also closely work together with marketing experts, especially for judging the final clustering results.

The input variables of the final models are number of trips and segments, the number/proportion of return trips and the number/proportion of within week, weekend, and long stay return trips. It turns out that the gain of including the second relevant dimension booking information is rather small compared to the increase in complexity. Th information about the time period between booking and departure is often missing, and the distribution of late business tickets and late sales offers can not be distinguished. There is also few information contained by the flag of promotional offers.

Finally, the segmentation based on travel behavior leads to six customers segments, which are described in Table 1. Segments identified based on travel behavior. Figure 1 illustrates the segments by means of input variables.

Segment	Description	Perc.
LS_low	Few long stay returns	26.5%
LSWE_med	Medium amount of trips, both weekend and long stay returns	20%
No_ret	Few segments, no return flights	5%
WE_low	Few weekend flights	8%
WW_high	High within week returns, no other returns	19%
WWLS_top	High amount of trips, mainly within weeks, but also long stays as well as weekend flights.	21.5%

Table 1. Segments identified based on travel behavior**.**

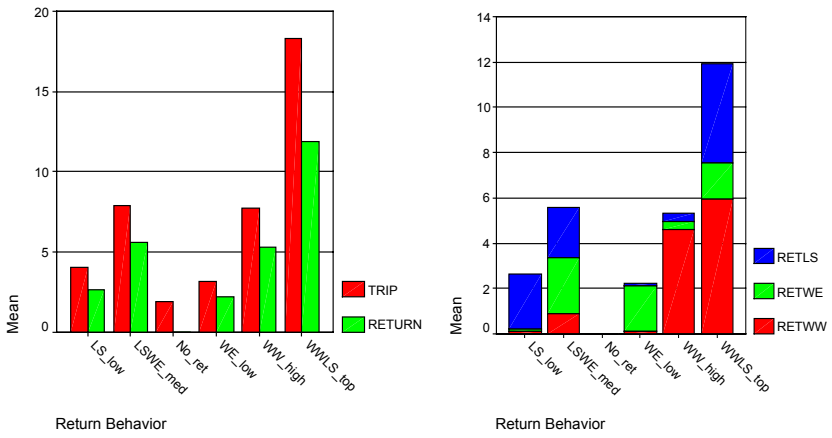

Fig. 1. Mean activities within each segment: Number of trips and return trips, respectively (*left*), and return trips resolved into sub-types (RETLS: long stays, RETWE: journey which contains a weekend, RETWW: within week journey).

The segmentation leads to sensible clusters concerning the business problem. There is a clear separation between the 'No_Return' segment and the other segments, which have comparable rates of return trips. There are 3 classes of frequencies: low (LS_low, WE_low), medium & high (LSWE_med, WW_high) and top (WWLS_top). The behavior is rather homogeneous within the segments. The customers of LS_low are mainly leisure holiday travelers, whereas WE_low are customers, which are exclusively traveling on weekends. WW_high contains the pure business customers. LSWE_med and WWLS_top use all types of return trips, but the number of flight activities can distinguish them. The WWLS_top class is the high value core customer segment and the LSWE_med are the low value core customers.

To further evaluate the results, we apply the segmentation rules to the validation set. Again, the rules separate the date in homogenous subgroups. This means that the identified segments are representative.

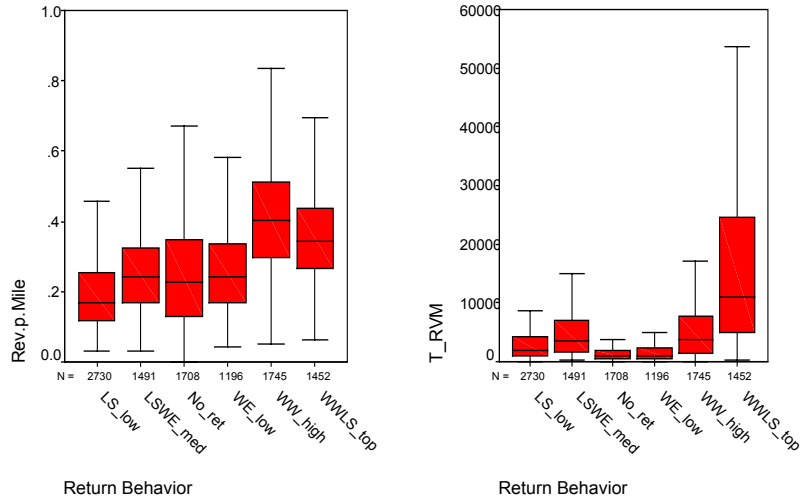

Fig. 2. Boxplots of revenue per mile and total revenue within each segment**.**

There is no actual quantitative definition of a good segmentation, thus assessing the groups by investigating their revenue distribution (customer value) is worthwhile. It confirms the consistency and the relevance of the results. From a revenue point of view, the dominating group is WWLS_top. But the segments LSWE_med and WW_high are also of interest. If we take the criterion revenue per mile then WW_high and WWLS_top are highest. As expected, the segments having within week trips (WWLS_TOP and WW_High) have a different fare level usage. Members falling in the segment 'No_Return' purchase more expensive fares than the remaining segments. These are customers, which are likely to fly a flight segment by chance. This segment is clearly an interesting prospect population, since their fare level is higher than average.

Further confirmation is obtained, if we take a look on the distribution of the type of fares: There is significantly higher portion of high fares (Business class: FCH, FCM, and economy full fare FYH) in the segments WWW_high and WWLS_TOP.

We also compare the distribution of further key attributes (e.g. average mileage) for the training set and the validation set. This also confirmed that the found segments are meaningful and reflect the dimension in a robust way.

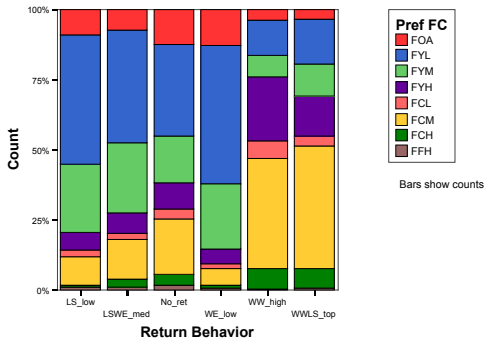

Fig. 3. Description of segmentation view 'Return behavior' with Preferred Fare Level (FFH: first class, FCH: business class high, FCM: business class medium, FCL: business class low, FYH: economy class high, FYM: economy class medium, FYL: economy class low, FOA: other fares).

6. Deployment

The next step is to profile the clusters to obtain SQL code. This allows to assign the segmentation to the complete database. Then it can be flexibly used for special marketing strategies.

6.1 Nature of the task

The segmentation results lead to better knowledge about certain customers. This knowledge is very useful for tactical marketing. However, the rules behind the segments are rather 'black boxes'. Therefore, it is necessary to derive rules, which are as simple as possible and which can be applied on the entire member base, using SQL statements.

Some requirements have to be considered. The attributes used needs to be easy to calculate. Moreover, in order to establish robust result for a certain period in time, proportion are preferred instead of absolute values. The considered time period has to be constant.

6.3 Approach

By means of classification trees (C5.0), we derive a simple rule set to uniquely classify the complete database into six segments. Again, we have to generate the attributes, resulting from the sequence of flight segments. The accuracy of the forecast for each segment is provided by balancing the training set according to equally sized clusters. We regulate the number of subsequent rules, while determining a minimal numbers of records given within each subgroup.

The most important quantities that are used in the rule set are the proportion of return trips to all trips and the percentage of within week return flights, trips over the weekend and long stays. The application of the classification rules on the validation set is convincing. Almost 98% of the members in the sample are predicted correctly. Table 2 shows the distribution of the segments given in the validation set and in the total customer database. Obviously, the approach yields a consistent result.

The segments are meaningful and reflect the dimension in a robust way.

Seg. #	Label	Target Value (study)	Real Value (complete db)
1	WWLS_top	19.40	19.67
2	WW_high	19.25	19.41
3	LSWE_med	20.65	20.82
4	LS_low	27.63	27.20
5	No_ret	4.90	4.71
6	WE_low	8.24	8.25
99	Non		

Table 2. Comparison of segments in the validation set and in the complete database

7. Conclusions and outlook

7.1 Conclusions

In this paper, we show that data mining is very useful to support CRM in the airline industry. In an initial phase of CRM, customer segments based on individual patterns are found, describing groups of customers with distinct needs and value. The segmentation results are very useful for marketing concerns and for improving customer services. For instance, the derived strategic segments can be used to derive some high-level business strategies and to perform tactical marketing actions, respectively.

7.2 Next steps

There are some open issues, which we consider in future projects. Firstly, the customer value can be improved while considering operational costs in assessing a flight segment. There is an accounting system, where all kinds of costs (e.g. operational costs as well as overhead costs) are gathered. The information is available for each single flight leg, but it is rather difficult to give a sensible breakdown to single passengers costs.

Members of the frequent flyer program are only one part of airlines customer. A further task is to find strategies to bundle information about flight activities of all customers. The gained information can also be used to identify future potential customers beyond prospects.

A further focus of future work is to develop a monitoring system, which is able to identify trends within customer segments, to discover outliers and to control the quality of the segmentation model. This is necessary, since customer behavior is strongly influence by exogenous factors (e.g. competitors' offer, new travel policies of companies).

References

1. Gallant, S., Piatesky-Shapiro, G. and Pyle, D. (2000): Successful customer relationship management in financial applications. Tutorial PM-1. KDD-2000, ACM SIGKDD 7th annual conference on Data Mining and Knowledge Discovery
2. Ripley, B.D. (1996): Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge UK.
3. Saarenvirta, G. (1998): Mining customer data. A step-by-step look at powerful clustering and segmentation methodology. DB2 magazine. http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.shtml
4. Thearling, K. (2000). Data mining and customer relationships. http://www3.shore.net/~kht/text/whexcerpt/whexcerpt.htm. Excerpted from *Building Data Mining Applications for CRM* by Alex Berson, Stephen Smith, Kurt Thearling (McGraw Hill, 2000).