Improving your Sales with Data Fusion

Pascal van Hattum*and Herbert Hoijtink

Utrecht University, The Netherlands

December 23, 2008

*Address for correspondence: Pascal van Hattum, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. Box 80140, NL-3508 TC Utrecht, The Netherlands, Tel.: +31 30 2537983, Fax.: +31 30 2535797.
E-mail: p.vanhattum@uu.nl

Improving your Sales with Data Fusion

Abstract

This paper shows how an European mail order company uses data fusion in order to improve sales. To select the best data fusion algorithm, two traditional data fusion methods, that are, polytomeous logistic regression and nearest neighbor algorithms, are compared with two model based clustering approaches. Finally, it is shown how internal and external validation criteria are used in order to evaluate the results of the data fusion algorithms.

Key words: Data Fusion, Differentiated Marketing, Nearest Neighbor, Logistic Regression, Model Based Clustering, Internal Validation, External Validation.

1 Introduction

In this paper, the following problem is addressed: an European mail order company, specialized in gardening products, wants to send differentiated catalogues to all the customers in the database. In order to do this, information about customer gardening preferences and interests from an external database is fused to the customer database. Using the fused information, differentiated catalogues can be send to all the customers.

Above described problem can best be illustrated using the schematic representation in Figure 1. In this representation data set **A** is the customer database and contains knowledge and information (represented by J items) from all customers. External data set **B** contains knowledge and information (represented by $J + 1$ items) from a group of customers, that are not in data set **A**. The knowledge and information represented by the first J items is available for each customer in each data set. However, for the group of customers in external data set **B** there is some additional knowledge and information, that is, item $J + 1$. The goal of this paper is to fuse the extra knowledge and information in external data set **B**, that is, item $J + 1$, to data set **A**. As a result of this data fusion, the knowledge and information about item $J + 1$ becomes 'known' for all customers in the database, data set **A**.

The structure of this paper is as follows. Section 2 describes how data fusion can be used in the context of the European mail order company. Also the marketing goals and the data sets used are described. Section 3 shows how data fusion is evaluated using internal and external validation criteria. This paper concludes with a discussion in Section 4.

2 Application

2.1 Improving sales

An European mail order company, specialized in all kind of gardening products, that are, flowers, bulbs, plants, et cetera, wants to increase the number of buying customers[2,3]. In order to do this, they decide to develop a direct marketing strategy. From a external motivational research study, the mail order company knows that each individual has a different attitude towards gardening and gardening products (www.tuinbeleving.nl).

From the motivational study, it is known that there are actually four groups (or clusters) of customers who have more of less the same attitude towards gardening and gardening products. Short descriptions (see www.tuinbeleving.nl for detailed descriptions) of these four motivational clusters are:

- cluster 1: gardens in this clusters are different from other gardens. They are surprising, wild, romantic and stylish; gardens meant for relaxation and unwinding. For customers in this cluster, gardening brings that relaxation;

- cluster 2: gardens in this cluster are more often large patios, easy to maintain and cluttered. Gardens are outdoor spaces to hang out with family and friends. Customers in this cluster think of gardening as strenuous, rather than a relaxing activity;

- cluster 3: in this cluster, the true gardener can be found. Gardens in this cluster are nice, neat, full of atmosphere and fit in with the rest of the neighborhood. Gardening is relaxing, a passion and the main hobby;

- cluster 4: gardens in this cluster are practical and easy to maintain. Customers in this cluster don't feel like gardening and can't find the time for gardening.

These motivational clusters provide a basis for developing a company's vision and/or a company's marketing directions on the strategic, tactical and operational levels, aligning the total marketing mix around the consumers needs in the domain gardening. Table 1 displays the frequencies of the resulting motivational clustering. Furthermore, each individual deals, handles and perceives gardening catalogues in a different way. The mail order company assumes that giving customers (some sort of) a tailor made offer, will eventually increase the number of buying customers.

Using the descriptions of the four motivational clusters, for each of the four clusters, a separate catalogue can be made by a specialized communication agency. The content of the catalogues, that is, the gardening products offered to the customers, is the same for each catalogue. Only the lay out (colors and pictures used on the front page and the back page) and the tone-of-voice of the catalogue's introduction are different for the cluster specific catalogues.

Because the mail order company wants to send differentiated catalogues to their customers, data fusion is used. Using the common items in both the supplier's customer database and the external motivational research study, data fusion methods are used to fuse the motivational cluster to the supplier's customer database.

2.2 Description of the data sets used

The available data sets are the customer database of the company, and the data set with the external motivational research study. Or translated to Figure 1, data set **A** and data set **B**, respectively. The content of data set **A** is data set $\mathbf{X}^A$. This data set $\mathbf{X}^A$ contains $J = 7$ items, that are, house ownership, number of vehicles, education, socio-demographic typology, household stage, prosperity and spending behavior, from all $N = 66,549$ customers.

The content of data set **B** is data set $\mathbf{X}^B$ and data set **Y**. Data set $\mathbf{X}^B$ contains the same $J = 7$ items as data set $\mathbf{X}^A$. The content of data set **Y** is one item, that are the motivational clusters. In total $M = 1,141$ respondents has participated to the motivational study, and for them the motivational clusters are known. Note that these $1,141$ customers are not a fraction of the $66,549$ customers from data set **A**.

As described in Section 1 and illustrated in Figure 1, the goal of the data fusion process is to fuse the information in data set **Y** to data set **A** using the common items in $\mathbf{X}^A$. Or, in the context of this application, using data fusion methods, all the $66,549$ customers are classified to one of the four motivational clusters using the 7 common items. How good or bad this data fusion process is done, is described in the following subsections.

3 Validation

3.1 Internal validation

Van Hattum and Hoijtink[4] compare four data fusion methods, that are polytomeous logistic regression, a nearest neighbor algorithm, a fusion value specific probabilities method and a model based clustering approach. In the nearest neighbor algorithm, the missing motivational clusters in data set **A** are duplicated from data set **B** using cases with similar values on the ex-

planatory items. In polytomeous logistic regression the relationship between motivational cluster and the explanatory items as determined in data set **B**, is used to fuse information to data set **A**. The fusion value specific probabilities model is based on latent cluster analysis, where the role of the latent clusters is taken by the fusion value, in this case the motivational cluster, and the explanatory variables are the items. The model fitted in data set **B** is used to predict the motivational clusters in data set **A**. The assumption in the model based clustering approach is that the data are generated from a mixture of fusion values specific probabilities models. As a result of the model based clustering approach, there will be a fusion value specific probabilities model for each latent cluster found. Translated to the application at hand, the number of latent clusters found is 9; for each of the 9 latent clusters a fusion value specific probabilities model is estimated. The interested reader is referred to Van Hattum and Hoijtink[4] for a full description of the four data fusion approaches.

In order to select the best data fusion method, the statistics TCCR (Total Correct Classification Rate) and model lift are calculated. The TCCR is the percentage of respondents that are classified to the right motivational cluster. Furthermore, a percentage of correct classifications based on a random chance model, can be obtained. Or in other words, the percentage of correct classifications that can be expected when the motivational clusters are randomly assigned to the respondents. This percentage is called TCCR_{chance}. The statistic model lift is calculated as $\frac{TCCR}{TCCR_{chance}} * 100\%$, and can be interpreted as the percentage of more correct classifications than would be obtained by chance. All statistics can be calculated for the overall model and for each motivational cluster separately.

Above described statistics are calculated for each data fusion method, applied to one training data set (about 50% of the cases) and two test data sets (each about 25% if the cases). Models fitted on a data set tend to predict much better for that data set than for a new data set sampled from the same population. Since we want to fuse values from data set B to data set A, we will evaluate the predictive performance of the four fusion methods by fitting them on the training data set, and using the resulting models to make predictions in the two test data sets. As will be illustrated below, this prevents against model overfitting. The interested reader is referred to Verstraeten[5] for a further discussion of model overfitting.

Frequencies, TCCRs and model lifts can be used to determine which data fusion method performs the best. There must be a good comparison between

the actual and the predicted frequencies of the motivational clusters. These frequencies are displayed in Table 2. Looking at this table, it is not clear which data fusion method to choose. When looking at, for example, the data fusion method 'logistic regression', it can be seen that the predicted frequency for motivational cluster 1 can be compared with the actual frequency. However, the difference between the predicted frequency for motivational cluster 2 and the actual frequency for this motivational cluster, is quite large. Similar observations can be made for other data fusion methods.

As can be seen from the TCCRs (displayed in Table 3) and the model lifts (displayed in Table 4) the data fusion method 'nearest neighbor' performs the worst of all. Both the TCCRs and the model lifts are the lowest compared to the other methods. The method 'model based clustering approach (with 9 latent clusters)' has the highest statistics applied to the train data set, but, due to model overfitting, these statistics drop, when applied to the test data sets. The other data fusion methods also suffer from model overfitting, but not as dramatic as the model based clustering approach. From the two tables it can be seen that the statistics for the data fusion methods 'logistic regression' and 'fusion value specific probabilities approach' are among the highest. Both methods are good models, however, the statistics for the latter model are more consistent on the train data set and the two test data sets.

Like in Van Hattum and Hoijtink[4], the fusion value specific probabilities approach turns out to be the best performing data fusion method. Consequently, this method is used to fuse the motivational clusters to the company's customer database. As a result of this data fusion, the motivational clusters become known for all the customers in the database. This is the starting point for differentiated marketing strategies as described in the next subsection.

3.2 External validation

In the case of the European mail order company the initial goal was to increase the number of buying customers. From past experiences the mail order company knows that 3.58% of the customers, who received a catalogue, bought something from this catalogue within four weeks after receival. Using differentiated catalogues, the goal of the mail order company is to increase this number of buying customers.

As a results of the data fusion, the total customer database, with $66,549$ customers, is classified. The columns 'Frequency customers' and 'Percentage

customers' in Table 5 show the resulting motivational cluster frequencies of the fused data set $\widehat{\mathbf{Y}}$. For 6,056 (=9.1%) customers there are no or insufficient common items available in order to classify to one of the four motivational clusters. From the percentages between brackets in Table 1 and 5 it is clear that the mail order company has more customers classified to motivational cluster 1 and 3, compared with the external motivational research study. This is completely consistent with the description of these two motivational clusters; both motivational clusters contain, in general, more true and passionate gardeners.

Using the descriptions of the four motivational clusters, for each cluster a separate catalogue can be made. However, in the first step of the differentiated marketing strategy, the mail order company wants to concentrate on just two of the four motivational clusters, that are motivational clusters 1 and 3. So, for only these two motivational clusters, cluster specific catalogues are made by a specialized communication agency. The content of the catalogues, that is, the gardening products offered to the customers, is the same. Only the lay out (colors and pictures used on the front page and the back page) and the tone-of-voice of the catalogue's introduction are different for the cluster specific catalogues. The focus of the catalogue for motivational cluster 1 is on getting inspired by the catalogue, self creation of gardens, exotic and adventurous gardens. The focus of the catalogue for motivational cluster 3 is on traditional and hobby gardening, and on the amount of information that is giving about gardening products and services.

Furthermore, in order to validate the two cluster specific catalogues, the mail order company decides to compare the results with the standard catalogue. This is done using a randomized experiment among a sample from the customers who are classified to either motivational cluster 1 or 3. The set-up for the randomized experiment is as follows (see also Table 6):

- To 6250 customers, who are classified to motivational cluster 1, standard catalogues are sent;
- To 6250 customers, who are classified to motivational cluster 1, cluster 1 specific catalogues are sent;
- To 6250 customers, who are classified to motivational cluster 3, standard catalogues are sent;
- To 6250 customers, who are classified to motivational cluster 3, cluster 3 specific catalogues are sent.

From the 25,000 customers, who are in the randomized experiment, the percentages of customers, as displayed in Table 7, bought something from the (un)differentiated catalogues. From this table it is clear that the differentiated catalogue approach rendered more buying customers, than with the standard catalogues. Not only compared with the customers who received undifferentiated catalogues, but also with the 3.58% buying customers from past experiences. It can be concluded that the company's goal with the differentiated marketing strategy is attained.

Also from Table 7 it is interesting to see the difference between the two motivational clusters. From the table it is clear that the effect of the differentiated catalogue is larger for motivational cluster 1, than for motivational cluster 3. This difference can be explained by the following three arguments. First of all, from past experiences with the motivational clusters, it is known that customers classified to motivational cluster 1 are, in general, more sensitive to (perceived) tailor made offerings, than customers classified to motivational cluster 3. Secondly, from past experiences with the motivational clusters, it is known that customers, classified to motivational cluster 1, buy, in general, more products from mail order companies. And, finally, the specialized communication agency thinks that the lay out of the standard catalogue and the tone-of-voice of the standard catalogue's introduction, are more likely to attract customers, classified to motivational cluster 3.

However, for both motivational clusters 1 and 3, the percentages of buying customers is higher using a differentiated marketing approach, than with a undifferentiated or standard approach. With the differentiated catalogue, the European mail order company is able to get more buying customers, and, consequently, increase his turnover.

The researcher must keep in mind, that it is impossible to conclude that the increase in number of buying customers can be fully dedicated to the differentiated catalogues. When sending the catalogues it was impossible to control for all kind of side effects that may be associated with customer buying behavior. However, for this application the goal of increasing the number of buying customers, is attained. Furthermore, by using the results of an external motivational research study, instead of conducting there own research study, the mail order company saved dollars on marketing research activities.

4 Discussion

In this paper, the customer database of a mail order company was fused to a motivational research study about gardening. In order to fuse the data sets, different traditional and new data fusion methods were used in order to fuse the data sets.

In the internal validation step, the different data fusion methods were compared and the fusion value specific probabilities approach was found to be the 'best' method. The TCCR for the overall model was around 45-46%, whereas the TCCR with a random chance mode was around 25%. The resulting model lift was around 180%, which means that the fusion value specific probabilities approach provided around 80% more correct matches than would be obtained by chance. The conclusion that the fusion value specific probabilities approach performed the 'best' was not only drawn in this paper, but also in past research[4]. This makes this data fusion method, a method with stable results.

As a result of the internal validation step, the motivational clusters were estimated for all the customers in the database. This was the starting point for differentiated marketing strategies, or in the case of the mail order company, differentiated catalogues were made by a specialized company.

Using a randomized experiment different marketing goals were tested. In the case of the mail order company, more customers bought something from the catalogue, when receiving the right (differentiated) catalogues.

Given the large number of customers involved, the increase in buying behavior gave the company a tremendous amount of extra sales. Furthermore, by using an external domain study, a lot of dollars were saved on marketing research costs. In all cases the data fusion project was profitable, and, consequently, was successful.

References

1. Van der Putten, P., Kok, J.N. and Gupta, A. (2002). Data Fusion Through Statistical Matching. *Paper 185*. MIT Sloan School of Management.

2. Lattin, J.M. and Bucklin, R.E. (1989). Reference Effects of Price and Promotion on Brand Choice Behavior. *Journal of Marketing Research*, 26(3), 299-310.

3. Feinberg, F.F., Krishna, A. and Zhang, J.Z. (2002). Do We Care What Others Get? A Behaviorist Approach to Targeted Promotions. *Journal of Marketing Research*, 39(3), 277291.

4. Van Hattum, P. and Hoijtink, H. (2009). The Proof of the Pudding is in the Eating. Data fusion: An Application in Marketing.*Journal of Database Marketing & Customer Strategy Management*, 16, xx-xx.

5. Verstraeten, G. (2005). Issues in Predictive Modelling of Individual Customer Behavior: Applications in Targeted Marketing and Consumer Credit Scoring. PhD. thesis, Marketing, Gent University, Belgium (2005).

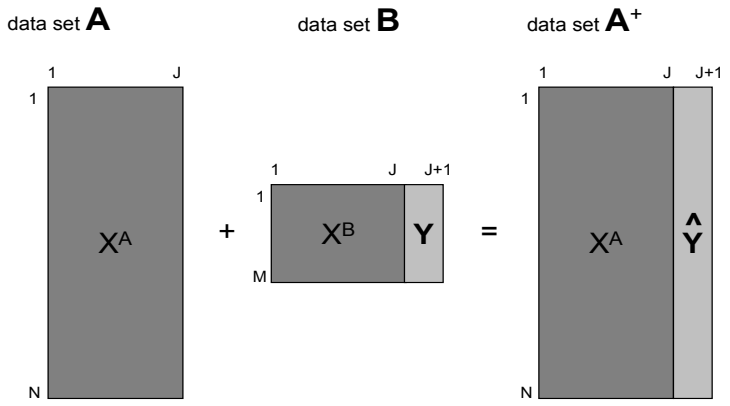

Figure 1: Schematic representation of data fusion in marketing (derived and adjusted from Van der Putten et al.[1])

Table 1: Frequency of respondents in motivational clusters concerning gardening (between brackets are the percentages based on the total number of respondents classified to one of the four motivational clusters)

Cluster	Frequency respondents	Percentage respondents
Cluster 1	246	21.6% (23.0%)
Cluster 2	228	20.0% (21.3%)
Cluster 3	343	30.1% (32.0%)
Cluster 4	254	22.3% (23.7%)
No cluster	70	6.1%
Total	1,141	100.0%

Table 2: Classification percentages after applying data fusion methods for domain gardening

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4
	actual	train	539	23.0%	18.2%	34.3%	24.5%
		test1	257	21.4%	22.2%	30.4%	26.1%
		test2	275	24.4%	26.5%	29.1%	20.0%
nearest neighbor	predicted	train	539				
		test1	257	25.7%	17.1%	36.2%	21.0%
		test2	275	25.5%	18.5%	33.5%	22.5%
logistic regression	predicted	train	539	24.7%	10.9%	38.6%	25.8%
		test1	257	23.3%	11.3%	38.9%	26.5%
		test2	275	25.8%	12.7%	39.3%	22.2%
fusion value specific approach	predicted	train	539	33.2%	13.7%	31.7%	21.3%
		test1	257	31.1%	16.3%	29.6%	23.0%
		test2	275	32.0%	15.3%	34.5%	18.2%
model based clustering approach	predicted	train	539	25.4%	13.2%	39.3%	22.1%
		test1	257	23.3%	16.0%	34.2%	16.5%
		test2	275	21.8%	17.8%	35.3%	25.1%

Table 3: Total correct classification rates after applying data fusion methods for domain gardening

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	total	chance
nearest		train	539						26.4%
neighbor	predicted	test1	257	31.8%	20.5%	38.7%	37.0%	33.5%	25.5%
		test2	275	38.6%	35.3%	37.0%	22.6%	33.8%	25.4%
logistic		train	539	46.6%	50.8%	55.8%	51.1%	51.8%	26.4%
regression	predicted	test1	257	43.3%	48.3%	43.0%	52.9%	46.3%	25.5%
		test2	275	42.3%	57.1%	40.7%	37.7%	42.5%	25.4%
fusion value		train	539	39.1%	45.9%	55.0%	47.0%	46.8%	26.4%
specific	predicted	test1	257	41.2%	45.2%	44.7%	55.9%	46.3%	25.5%
approach		test2	275	44.3%	52.4%	46.3%	36.0%	44.7%	25.4%
model based		train	539	48.9%	47.7%	58.5%	52.9%	54.7%	26.4%
clustering	predicted	test1	257	40.0%	31.7%	42.0%	44.1%	40.5%	25.5%
approach		test2	275	43.3%	42.9%	39.2%	31.9%	38.9%	25.4%

Table 4: Model lifts after applying data fusion methods for domain gardening

method		data set	#records	cluster 1	cluster 2	cluster 3	cluster 4	total
nearest neighbor	predicted	train	539					
		test1	257	149	92	128	142	131
		test2	275	158	133	127	113	133
logistic regression	predicted	train	539	203	280	162	209	196
		test1	257	202	218	142	203	182
		test2	275	173	215	140	189	167
fusion value specific approach	predicted	train	539	170	253	160	192	177
		test1	257	193	204	147	215	182
		test2	275	182	197	159	180	176
model based clustering approach	predicted	train	539	213	318	170	216	208
		test1	257	187	143	139	169	159
		test2	275	178	161	135	159	153

Table 5: Frequency clusters in application gardening (between brackets are the percentages based on the total number of customers classified to one of the four motivational clusters)

Cluster	Frequency customers	Percentage customers
Cluster 1	16,938	25.5% (28.0%)
Cluster 2	8,409	12.6% (13.9%)
Cluster 3	24,076	36.2% (39.8%)
Cluster 4	11,070	16.6% (18.3%)
No cluster	6,056	9.1%
Total	66,549	100.0%

Table 6: Number of catalogues sent

		Catalogue sent	
		Standard	cluster specific
Predicted	cluster 1	6,250	6,250
	cluster 3	6,250	6,250

Table 7: Percentage buying customers

		Catalogue sent	
		Standard	Cluster specific
Predicted	cluster 1	3.70%	4.46%
	cluster 3	3.65%	3.83%