TRAVELPASS

Data Mining and Data Analysis applied in a customer loyalty programme

María del Rosario Bruera

Travelpass is a customer loyalty programme developed in Argentina and based on a cobranding card integrated by four main shareholders companies: Shell, Banco de Galicia, Telecom and Supermercados Norte, and other twelve companies among different industries (pharmacy, toys, restaurants, optical shop, etc.).

This paper, based on concepts such as customer loyalty, CRM, and shopper understanding, shows the integration of information technology, data mining models and market research tools with the real world business problems, and how these support, drive and measure marketing and CRM actions.

PROGRAM OVERVIEW

With national coverage, as of December 2001 the datawarehouse Travelpass hosts information about:

- 1,159,000 active accounts;
- 1,759,000 issued cards;
- 2,500 million points generated by 3,500 millions pesos in purchases;
- 60 million transactions;
- 2,800 sales points;
- 550 *Canje Express* centers: sales points where rewards could be redeemed immediately online;
- 430,000 redeemed rewards;
- 520,000 accounts able to change points by rewards.

Queries about accumulated points, rewards catalog and redemptions can be made via the website (www.travelpass.com.ar) and call center.

Since its beginning on September 1998, Travelpass has included the Data Analysis department into its organization to support Marketing, CRM and Executive Information Management. The Data Analysis department has grown from simple reports generated using SQL queries and Excel dynamic tables at the beginning to today's to state-of-the-art Data Mining techniques to produce descriptive and predictive models.

Web surveys are also conducted via this website to complete, update and enrich client information.

CUSTOMER SEGMENTATION: A MULTIVARIATE CLUSTERING MODEL

The first Data Mining project was the segmentation of clients in terms of their consumption behavior. The main target was the identification of accounts with the largest loyalty patterns and the subsequent socio-demographic characterization.

From a methodological point of view, this is a "*clustering*" or "*non-supervised learning*" process, as client segments (clusters) are not previously defined but are generated based on the available data.

As the datawarehouse was under construction, data was extracted from the transactional system.

First of all, during the data cleaning process, some activities were performed in order to get the flat data table analyzed:

- transaction aggregation at client (account) level;
- identify and filter invalid transactions (e.g. zero standard and bonus point);
- identify and filter transactions with invalid dates.

Thereafter, some derived indicators (new variables calculated from available data) were generated at account level:

- average days between transactions;
- first and last transaction dates;
- average ticket;
- cumulative standard points;
- Multibrand Index (number of participants where Travelpass clients made purchases);
- months within the programme (difference between present date and first transaction date measured in months).

Based on the above indicators, the three traditional RFM measures (Recency, Frequency, Monetary) and Activation Days were calculated:

- Frequency: average days between transactions;
- Recency: difference between present date and last transaction date;
- Monetary: monthly average points (standard points / months within the programme);
- Activation days: difference between present date and first transaction date.

The final flat data set is shown in table 1.

Table 1
FLAT DATA MINING TABLE

Account	*Avg. Tck*	*Frequency*	*Recency*	*Avg. points*	*Multibrand*
20009	24.50	152.88	34.28	3.92	3
20027	299.70	Missing	601.74	140.00	1
20097	293.35	54.55	89.87	46.35	1
20102	16.69	28.50	515.74	13.16	1
20103	2887.45	40.57	110.18	1066.70	2
20106	87.73	2.46	11.25	515.00	1
20107	10.00	Missing	129.79	5.00	1
20109	19.76	12.04	5.23	24.29	1

Analysis began identifying missing values and outliers that were processed as follows:

1. Missing values in *Frequency* indicator. As *Frequency* cannot be calculated in accounts with only one transaction, they are flagged in two segments:
   - New: Activation less than 120 days
   - Discarded: Activation longer or equal than 120 days
2. Outliers in Recency, Monetary indicators. Variables are discretized with cut-point based on quartiles. Outliers are asigned to extreme categories. Table 2 shows the four ordinal variables and their categories.

Table 2
ORDINAL VARIABLES AND CATEGORY DEFINITIONS

Variable / category	*1*	*2*	*3*	*4*	*5*
Recency	Up to 15 days	16 - 30	31 - 90	91 +	
Frequency	Up to 7 days	8 - 14	15 - 30	31 +	
Monthly avg. points	100 +	50 - 99	20 - 49	Less than20	
Multibrand index	1	2	3	4	5 +

For example, an account (client) qualified as R1 (Recency), F1 (Frequency), Mu5 (Multibrand Index) and P1 (Monthly Average Points) means that

- Recency is less than 15 days
- Frequency is less than 7 days
- Monthly Average Points are 100 or higher
- Multibrand Index is 5 or more

Thus, this account can be characterized as follows:

- Last transaction occurred within two last weeks
- The client makes one purchase per week (on average)
- The monthly purchases generate, at least, 100 points
- The client makes purchases with at least five participants

Figure 1
CATPCA CATEGORY PLOT

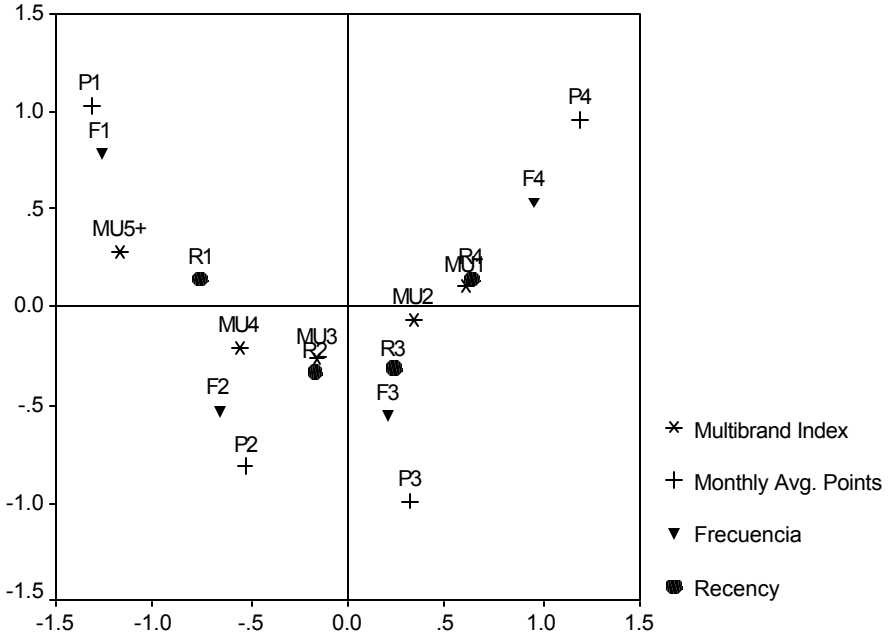
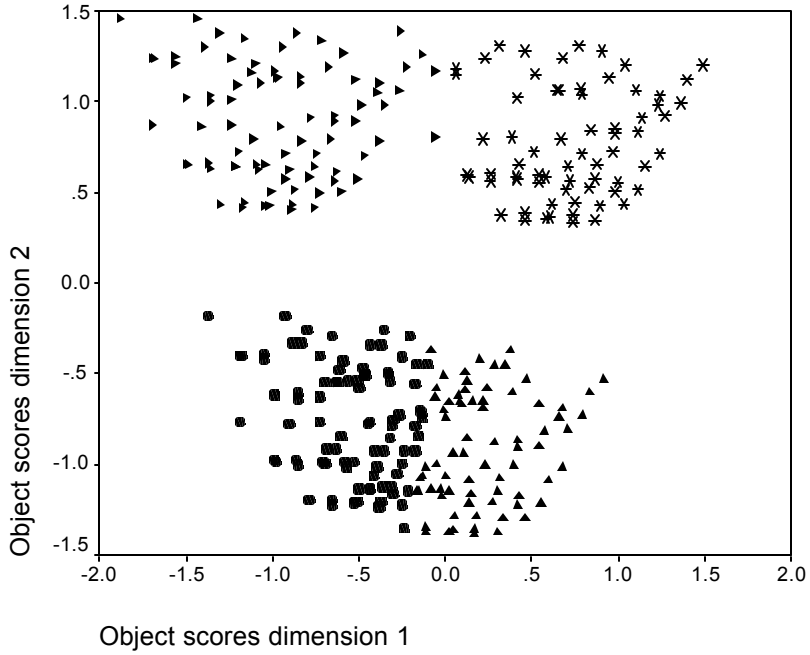


Figure 2
CATPCA CASES PLOT



To generate client segments, K-Means clustering method was applied to the data table. K-Means algorithm requires an Euclidean metric to define the "distance" concept. Thus, original ordinal variables were transformed into factorial scores applying the Categorical Principal Components Analysis (CATPCA) method. Score plots generated by CATPCA for both, variables and cases (accounts), are shown in figures 1 and 2 respectively.

It is interesting to remark that, even though CATPCA procedure was not informed about variables ordinality, variables factorial plot (map) rebuilds this attribute along the first dimension, which explains 61% of the total variability.

The cases plot shows that four clusters can be clearly identified. K-Means clustering method is then applied to the cases (accounts) factorial scores, generating four segments (clusters). Once these four segments were described and characterized with the available variables, they were named as *Top*, *Loyal*, *Medium* and *Light*.

The Pareto analysis shows, in table 3, that 80% of the issued total points are concentrated in two segments (Top and Loyal) which include 35% of the accounts.

Table 3
PARETO ANALYSIS

Segment	*Total points*	*Accounts*	*% Total points*	*% Accounts*
Top	1,573,139,626	225,558	62%	19%
Loyal	463,093,388	189,709	18%	16%
Medium	342,545,921	331,562	13%	28%
Light	110,034,716	256,038	4%	22%
New	4,931,472	19,008	0%	2%
Discarded	48,974,156	165,816	2%	14%
Total	2,552,719,279	1,187,691	100%	100%

Table 4 shows the segment characterization measured in terms of the four original variables based on which they were generated.

For example, Top and Loyal segments show similar patterns in terms of use (Frequency, Recency and Multibrand Index), but a significant difference for the Monthly Average Points. That is, Top clients make greater purchases than Loyal ones.

Table 4
SEGMENTS BEHAVIOUR

Segment	*P25%*	*Median*	*P75%*
Monthly avg. pts.			
Top	126	166	247
Loyal	55	67	82
Medium	27	35	45
Light	7	11	16
New	26	50	95
Frequency			
Top	5	7	13
Loyal	7	10	13
Medium	16	22	32
Light	28	43	73
New	6	10	22
Recency			
Top	12	18	44
Loyal	14	19	30
Medium	22	68	351
Light	26	143	462
New	15	22	31
Multibrand Index			
Top	2	4	5
Loyal	2	3	5
Medium	1	2	3
Light	1	2	2
New	1	1	2

USING CLIENT SEGMENTATION

Client segment is a new "*artificial variable*" which was included as a dimension into the Data Warehouse. Combining original and derived variables with "Client Segment", OLAP techniques were applied to analyze, for example:

- reward redemptions: frequency, rewards preferences, etc.;
- cross-traffic between participants.

Client segmentation was also used to perform sample designs for Customer Satisfaction Surveys.

Thereafter, this segmentation was used in ad-hoc studies for the shareholders Shell, Norte, Banco Galicia and Telecom, assigning special segments according their own clients consumption patterns.

WEBSITE RESEARCH: OPTIMUM SUPERMARKET CONSUMER

Travelpass website capabilities allowed, since March 2001, online client interaction and opened a new communication channel.

In July 2001, clients were invited to register and respond to some questions referring to participant companies. Bonus points were issued as incentive.

Segmentation was used again to target the invited clients, restricting them to those that had a *Multibrand Index* equal to 4 in shareholders companies (that is, they had made purchases in Shell, Norte, Telecom and Galicia simultaneously during year 2001).

These restrictions selected 9,200 clients that were invited – via regular mail – to register in the website. Response rate was 23% and the respondents profile was as follows:

- 90% belonged to Top segment;
- 46% Professionals;
- 59% male;
- Average age 40 years old;
- Average family members 3.4.

One of the survey questions was "Supermarket where you perform your purchases". Even though all respondents made transactions in Norte in 2001, only 67% of them mentioned Norte as the habitual supermarket. The Norte consumption pattern was analysed based on the transactions performed by

respondents during August and September 2001 (survey ended on July 31). (See table 5.)

Table 5
HABITUAL SUPERMARKET

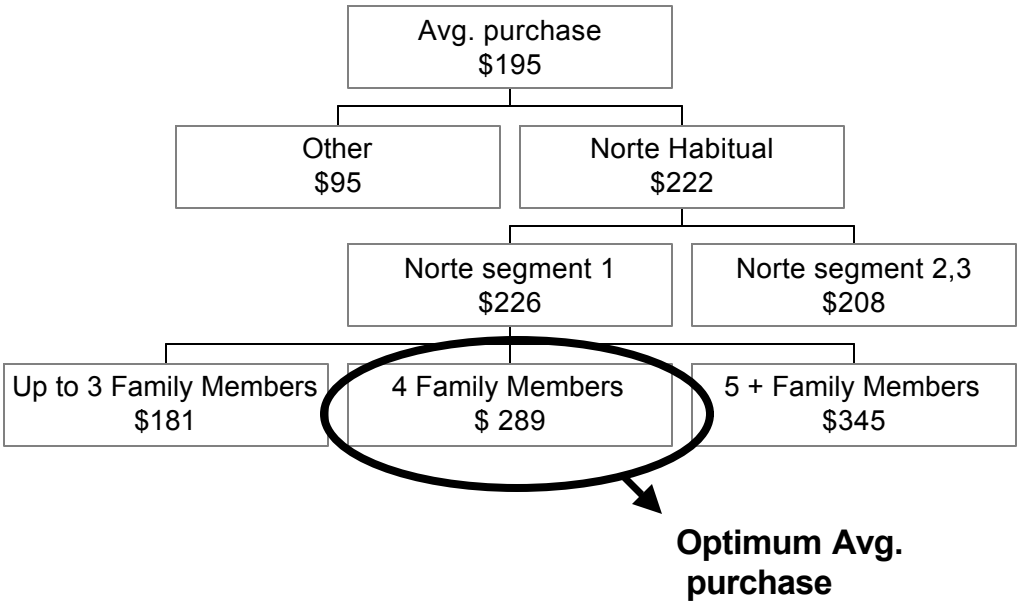
Supermarket	*Accounts*	*%*
Norte	1,424	67%
Coto	210	10%
Carrefour	198	9%
Jumbo	133	6%
Disco	125	6%
Wal Mart	29	1%
Ekono	5	0%
Auchan	4	0%
Total	2,128	100%

Of the respondents:

- 1,520 clients (71%) made purchases in these months;
- 608 clients (29%) did not make purchases in these months;
- Average Monthly Purchase was $195.

As this Average Monthly Purchase shows a heterogeneous distribution a segmentation using a CHAID Tree was applied to identify clients sub-sets with different and homogeneous consumption patterns. In this segmentation the following variables were included: Segment in Norte, Segment in Travelpass, number of family members, habitual supermarket. Main branches of the Chaid Tree are shown in figure 3.

Figure 3
SUPERMARKET OPTIMAL PURCHASE



It is interesting to remark:

- Overall Average Monthly Purchase – $195 – grows to $222 for clients who mentioned Norte as "supermarket where they made their purchases". Habitual supermarket was identified by CHAID method as the first predictor.
- Norte Segment and Family Members are the next predictors to segment Average Monthly Purchase.
- Clients who are "Norte habitual customers" and have four members in their family have an Average Monthly Purchase of $289.

With this maximized Average Monthly Purchase, named "Optimal Monthly Purchase", and considering "Recency in Norte", the following supermarket client segmentation was defined:

- Optimum – Active
  - Average Monthly Purchase: Optimal Monthly Purchase or more
  - Last Purchase: up to 60 days
- Optimum – Lost
  - Average Monthly Purchase: Optimal Monthly Purchase or more
  - Last Purchase: longer than 60 days

- Upgrade – Active
  - Average Monthly Purchase: less than Optimal Monthly Purchase
  - Last Purchase: up to 60 days
- Upgrade – Lost
  - Average Monthly Purchase: less than Optimal Monthly Purchase
  - Last Purchase: longer than 60 days

This segmentation, very easy to calculate, is highly effective because Pareto analysis performed over the totality of the supermarket transactions and clients hosted in the datawarehouse shows that with *11% of clients* (Optimum – Active) *40% of the purchases* are concentrated.

WEBSITE RESEARCH: TELECOM SURVEY RESPONSE RATE ANALYSIS

After the success of this first project, Telecom Group decided to conduct a survey (targeted to their clients) about products and services of cellular and residential telephone and Internet services. Telecom clients were invited to participate via two different channels:

- Regular mailing (all clients) 78,104
- Email (to web site registered clients only) 17,968

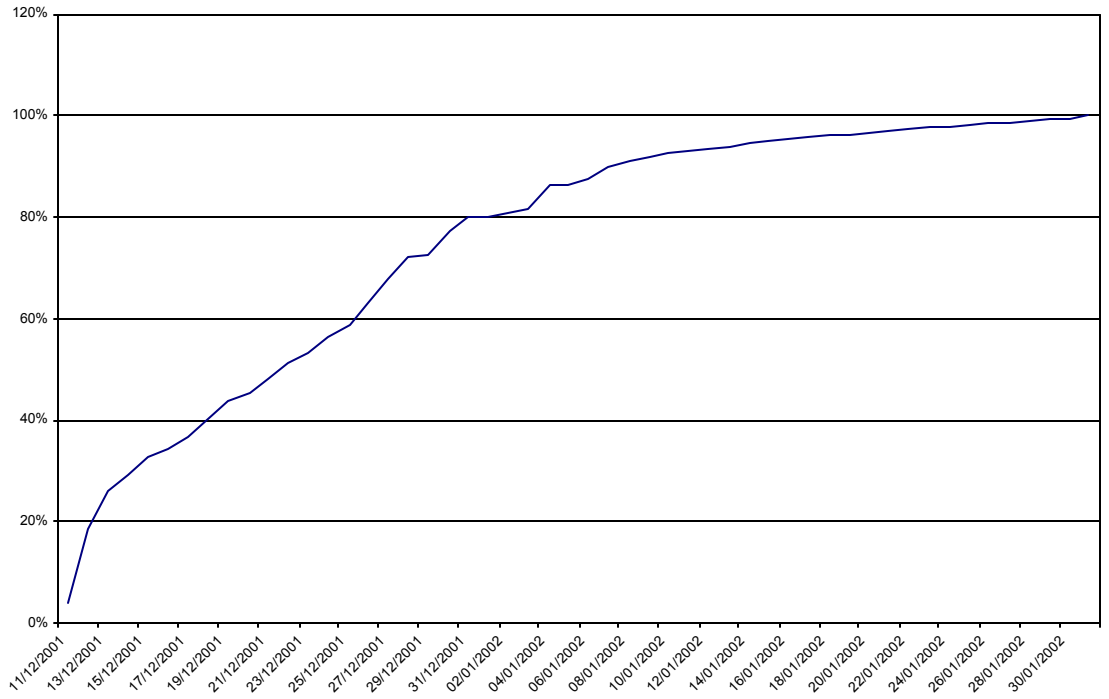
Regular mailing was sent four days after email delivery. The general response rate was 10.62%. A CHAID Tree was applied to explain the response rate in terms of the variables available in the datawarehouse (not in the questionnaire). While the response rate of those who were invited with a regular mailing only was 4.15%, clients invited via regular and email mailing responded with a rate of 32.27%.

The Next predictor was the number of Telecom Group companies (residential - Telecom, cellular - Personal, Internet - Arnet) where clients purchase.

For the clients of the three telecom companies, the response rate grew up to 49.21%. Further predictors were Occupation and Travelpass Segment.

Figure 4 shows the temporal evolution along the data collection time period. Thirty percent of the total responses arrived in the first three days after the email delivery.

Figure 4
CUMULATIVE RESPONSE PERCENT



These conclusions have generated the first business project for 2002: *Campaign to produce a massive client registration in the website.*

CONCLUSIONS

The Travelpasss programme is a great data source to develop research projects based on both Data Mining and Market Research perspectives. Integration of these two disciplines is possible if two main premises are fulfilled:

- data sources are available (datawarehouse, web surveys, etc.); and
- interest exists in developing research activities to support business decisions.

In this context, both disciplines – Data Mining and Market Research – work together to improve the results. Client consumption behaviour – translated in the segments and other derived indicators from the mining of data – did not necessarily match demographic profiles. However these profiles are needed to drive marketing strategies. Thus, integration of Data Mining and Market Research and the feedback between both became a valuable and serious support to orient business strategies.

REFERENCES

Berry, Michael, Linoff, Gordon. (1997). Data Mining Techniques for Marketing, Sales and Customer Support. USA: Wiley

Berry, Michael, Linoff, Gordon. (2000). Mastering Data Mining. USA: Wiley.

Berson, A., Smith, S., Thearling, K. (2000). Building Data Mining Applications for CRM. USA: McGraw Hill

Bigus, Joseph. (1996). Data Mining with Neural Networks. USA: McGraw Hill.

Dillman, Don. (2000). Mail and Internet Surveys: The Tailored Design Method. USA: John Willey & Sons.

Fayyad, U., Shapiro, G. and others (1996). Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, California, USA.

Grabmeier, A., Rudolph, A. (1998). Techniques of Cluster Algorithms in Data Mining version 2.0, IBM Informationssysteme GmBH.

Groth, Robert. (1998). Data Mining: A hands-on approach for business professionals. USA: Prentice Hall

Hair, J.H., Anderson, R.E, Thatam, R., Black, W.C. (1999). Análisis Multivariante, 5ta edición. Iberia, Madrid: Prentice Hall

Pyle, D. (1999). Data Preparation for Data Mining. USA: Morgag Kauffman Publishers Inc.

THE AUTHOR

María del Rosario Bruera, President, Census Data Analysis & Data Mining, Argentina.