Data Mining: An Overview of Methods and Technologies for Increasing Profits in Direct Marketing

C. Olivia Rud, President, OptiMine Consulting, West Chester, PA

ABSTRACT

Data Mining is a new term for the common practice of searching through data to find patterns, which will predict future outcomes or define measurable relationships. Statistical and machine learning methods are the favored tools of many businesses that utilize direct marketing. This paper will explore some of the typical uses of Data Mining in direct marketing with discussions and examples of measuring response, risk and lifetime customer value. The paper will conclude with highlights of SAS® Enterprise Miner and its ability to transform the Data Mining process.

INTRODUCTION

Increasing competition and slimmer profit margins in the direct marketing industry have fueled the demand for data storage, data access and tools to analyze or 'mine' data. While data warehousing has stepped in to provide storage and access, data mining has expanded to provide a plethora of tools for improving marketing efficiency.

This paper details a series of steps in the data mining process, which takes raw data and produces a net present value (NPV). The first step describes the process used to extract and sample the data. The second step uses elementary data analysis to examine the data integrity and determine methods for data clean up. The third step defines the process to build a predictive model. This includes defining the objective function, variable preparation and the statistical methodology for developing the model. The next step overlays some financial measures to calculate the NPV. Finally, diagnostic tables and graphs demonstrate how the NPV can be used to improve the efficiency of the selection process for a life insurance acquisition campaign. An epilogue will describe the ease with which all of these steps can be performed using the SAS® Enterprise Miner data mining software.

OBJECTIVE FUNCTION

The overall objective is to measure Net Present Value (NPV) of future profits over a 5-year period. If we can predict which prospects will be profitable, we can target our solicitations only to those prospects and reduce our mail expense. NPV consists of four major components:

1) Paid Sale - probability calculated by a model. Individual must respond, be approved by risk and pay their first premium.
2) Risk - indices in matrix of gender by marital status by age group based on actuarial analysis.
3) Product Profitability - present value of product specific 5-year profit measure usually provided by product manager.
4) Marketing Expense - cost of package, mailing & processing (approval, fulfillment).

THE DATA COLLECTION

A previous campaign mail tape is overlaid with response and paid sale results. Since these campaigns are mailed quarterly, a 6-month-old campaign is used to insure mature results.

The present value of the 5-year product profitability is determined to be $553. This includes a built in attrition and cancellation rate.

The marketing expense which includes the mail piece and postage is $.78.

The risk matrix (see Appendix A) represents indices, which adjust the overall profitability based actuarial analysis. It shows that women tend to live longer than men, married people live longer than singles and course, one of the strongest predictors of death is old age.

To predict the performance of future insurance promotions, data is selected from a previous campaign consisting of 966,856 offers. To reduce the amount of data for analysis and maintain the most powerful information, a sample is created using all of the 'Paid Sales' and 1/25th of the remaining records. This includes non-responders and non-paying responders. The following code creates the sample dataset:

```
DATA A B;
SET LIB.DATA;
IF PREMIUM > 0 THEN OUTPUT A;
ELSE OUTPUT B;

DATA LIB.SAMPDATA;
SET A B (WHERE=(RANUNI(5555) < .04));
SAMP_WGT = 25;
RUN;
```

This code is putting into the sample dataset, all customers who paid a premium and a 1/25th random sample of the balance of accounts. It also creates a weight variable called *SAMP_WGT* with a value of 25.

The following table displays the sample characteristics:

	Campaign	Sample	Weight
Non Resp/Non Pd Resp	929,075	37,163	25
Responders/Paid	37,781	37,781	1
Total	966,856	74,944	

The non-responders and non-paid responders are grouped together since our target is paid responders. This gives us a manageable sample size of 74,944.

THE DATA CLEAN-UP

To check data quality, a simple data mining procedure like PROC UNIVARIATE can provide a great deal of information. In addition to other details, it calculates three measures of central tendency: mean, median and mode. It also calculates measures of spread such as the variance and standard deviation and it displays quantile measures and extreme values. It is good practice to do a univariate analysis of all continuous variables under consideration.

The following code will perform a univariate analysis on the variable income:

```
PROC UNIVARIATE DATA=LIB.DATA;
VAR INCOME;
RUN;
```

The output is displayed in Appendix B. The variable *INCOME* is in units of $1000. N represents the sample size of 74,944. The mean value of 291.4656 is suspicious. With further scrutiny, we see that the highest value for INCOME is 2000. It is probably a data entry error and should be deleted.

The two values representing the number of values greater than zero and the number of values not equal to zero are the same at 74,914. This implies that 30 records have missing values for income. We choose to replace the missing value with the mean. First, we must delete the observation with the incorrect value for income and rerun the univariate analysis.

The results from the corrected data produce more reasonable results (see Appendix C). With the outlier deleted, the mean is in a reasonable range at a value of 49. This value is used to replace the missing values for income. Some analysts prefer to use the median to replace missing values. Even further accuracy can be obtained using cluster analysis to calculate cluster means. This technique is beyond the scope of this paper.

Because a missing value can be indicative of other factors, it is advisable to create a binary variable, which equals 1 if the value is missing and 0 otherwise. For example, income is routinely overlaid from an outside source. Missing values often indicate that a name didn't match the outside data source. This can imply that the name is on fewer databases and hence may have received fewer pieces of direct mail. This will often lead to better response rates.

The following code is used to replace missing values:

```
IF INCOME = . THEN INC_MISS = 1;
ELSE INC_MISS = 0;

IF INCOME = '.' THEN INCOME = 49;
```

MODEL DEVELOPMENT

The first component of the NPV, the probability of a paid sale, is based on a binary outcome, which is easily modeled using logistic regression. Logistic regression uses continuous values to predict the odds of an event happening. The log of the odds is a linear function of the predictors. The equation is similar to the one used in linear regression with the exception of the use of a log transformation to the independent variable. The equation is as follows:

$$\log(p/(1-p)) = B_0 + B_1X_1 + B_2X_2 + \ldots\ldots + B_nX_n$$

Variable Preparation - Dependent

To define the dependent variable, create the variable *PAIDSALE* defined as follows:

```
IF PREMIUM > 0 THEN PAIDSALE = 1;
ELSE PAIDSALE = 0;
```

Variable Preparation - Independent: Categorical

Categorical variables need to be coded with numeric values for use in the model. Because logistic regression reads all independent variables as continuous, categorical variables need to be coded into n-1 binary (0/1) variables, where n is the total number of categories.

The following example deals with four geographic regions: north, south, midwest, west. The following code creates three new variables:

```
IF REGION = 'EAST' THEN EAST = 1;
        ELSE EAST = 0;
IF REGION = 'MIDWEST' THEN MIDWEST = 1;
        ELSE MIDWEST = 0;
IF REGION = 'SOUTH' THEN SOUTH = 1;
        ELSE SOUTH = 0;
```

If the value for *REGION* is 'WEST', then the values for the three named variables will all have a value of 0.

Variable Preparation - Independent: Continuous

Since, logistic regression looks for a linear relationship between the independent variables and the log of the odds of the dependent variable, transformations can be used to make the independent variables more linear. Examples of transformations include the square, cube, square root, cube root, and the log.

Some complex methods have been developed to determine the most suitable transformations. However, with the increased computer speed, a simpler method is as follows: create a list of common/favorite transformations; create new variables using every transformation for each continuous variable; perform a logistic regression using all forms of each continuous variable against the dependent variable. This allows the model to select which form or forms fit best. Occasionally, more than one transformation is significant. After each continuous variable has been processed through this method, select the one or two most significant forms for the final model. The following code demonstrates this technique for the variable AGE:

```
PROC LOGISTIC LIB.DATA:
WEIGHT SMP_WGT;
MODEL PAIDSALE = AGE AGE_MISS AGE_SQR
AGE_CUBE AGE_LOG / SELECTION=STEPWISE;
RUN;
```

The logistic model output (see Appendix D) shows two forms of AGE to be significant in combination: AGE_MISS and AGE_CUBE. These forms will be introduced into the final model.

Partition Data

The data are partitioned into two datasets, one for model development and one for validation. This is accomplished by randomly splitting the data in half using the following SAS® code:

```
DATA LIB.MODEL LIB.VALID;
  SET LIB.DATA;
IF RANUNI(0) < .5 THEN OUTPUT LIB.MODEL;
ELSE OUTPUT LIB.VALID;
RUN;
```

If the model performs well on the model data and not as well on the validation data, the model may be over-fitting the data. This happens when the model memorizes the data and fits the models to unique characteristics of that particular data. A good, robust model will score with comparable performance on both the model and validation datasets.

As a result of the variable preparation, a set of 'candidate' variables has been selected for the final model. The next step is to choose the model options. The backward selection process is favored by some modelers because it evaluates all of the variables in relation to the dependent variable while considering interactions among the independent or predictor variables. It begins by measuring the significance of all the variables and then removing one at a time until only the significant variables remain. A reasonable significance level is the default value of .05. If too many variables end up in the final model, the signifiance level can be lowered to .01, .001, or .0001.

The sample weight must be included in the model code to recreate the original population dynamics. If you eliminate the weight, the model will still produce correct ranking-ordering but the actual estimates for the probability of a 'paid-sale' will be incorrect. Since our NPV model uses actual estimates, we will include the weights.

The following code is used to build the final model.

```
PROC LOGISTIC LIB.MODEL:
WEIGHT SMP_WGT;
MODEL PAIDSALE = AGE_MISS AGE_CUBE EAST
MIDWEST SOUTH INCOME INC_MISS LOG_INC
MARRIED SINGLE POPDENS MAIL_ORD//
SELECTION=BACKWARD;
RUN;
```

The resulting model has 7 predictors. (See Appendix E) The parameter estimate is multiplied times the value of the variable to create the final probability. The strength of the predictive power is distributed like a chi-square so we look to that distribution for significance. The higher the chi-square, the lower the probability of the event occurring randomly (pr > chi-square). The strongest predictor is the variable MAIL_ORD. This has a value of 1 if the individual has a record of a previous mail order purchase. This may imply that the person is comfortable making purchases through the mail and is therefore a good mail-order insurance prospect.

The following equation shows how the probability is calculated, once the parameter estimates have been calculated:

$$\text{prob} = \frac{\exp(B_0 + B_1X_1 + B_2X_2 + \ldots\ldots + B_nX_n)}{(1+ \exp(B_0 + B_1X_1 + B_2X_2 + \ldots\ldots + B_nX_n))}$$

This creates the final score, which can be evaluated using a gains table (see Appendix F). Sorting the dataset by the score and dividing it into 10 groups of equal volume creates the gains table.

The validation dataset is also scored and evaluated in a gains table (see Appendix G).

Both of these tables show strong rank ordering. This can be seen by the gradual decrease in predicted and actual probability of 'Paid Sale' from the top decile to the bottom decile. The validation data shows similar results, which indicates a robust model. To get a sense of the 'lift' created by the model, a gains chart is a powerful visual tool (see Appendix H). The Y-axis represents the % of 'Paid Sales' captured by each model. The X-axis represents the % of the total population mailed. Without the model, if you mail 50% of the file, you get 50% of the potential 'Paid Sales'. If you use the model and mail the same percentage, you capture over 97% of the 'Paid Sales'. This means that at 50% of the file, the model provides a 'lift' of 94% {(97-50)/50}.

Financial Assessment

To get the final NPV we use the formula:
NPV = Pr(Paid Sale) * Risk * Product Profitability - Marketing Expense

At this point, we apply the risk matrix and product profitability value we discussed earlier. The financial assessment shows the models ability to select the most profitable customers based on (See Appendix H). Notice how the risk index is lower for the most responsive customers. This is common in direct response and demonstrates 'adverse selection'. In other words, the riskier prospects are often the most responsive.

At some point in the process, a decision is made to mail a percent of the file. In this case, you could consider the fact that in decile 7, the NPV becomes negative and limit your selection to deciles 1 through 6. Another decision criteria could be that you need to be above a certain 'hurdle rate' to cover fixed expenses. In this case, you might look at the cumulative NPV to be above a certain amount such as $30. Decisions are often made considering a combination of criteria.

The final evaluation of your efforts may be measured in a couple of ways. You could determine the goal to mail fewer pieces and capture the same NPV. If we mail the entire file with random selection, we would capture $13,915,946 in NPV. This has a mail cost of $754,155. By mailing 5 deciles using the model, we would capture $14,042,255 in NPV with a mail cost of only $377,074. In other words, with the model we could capture slightly more NPV and *cut our marketing cost in half!*

Or, we can compare similar mail volumes and increase NPV. With random selection at 50% of the file, we would capture $6,957,973 in NPV. Modeled, the NPV would climb to $14,042,255. This is a lift of over 100% ((14042255-6957973)/ 6957973 = 1.018).

Conclusion

Through a series of well designed steps, we have demonstrated the power of Data Mining. It clearly serves to help marketers in understanding their markets. In addition, it provides powerful tools for improving efficiencies, which can have a huge impact on the bottom line.

Epilogue

SAS® has developed a tool called the SAS® Enterprise Miner, which automates much of the process we just completed. Using icons and flow charts, the data is selected, sampled, partitioned, cleaned, transformed, modeled, validated, scored, and displayed in gains tables and gains charts. In addition, it has many other features for scrutinizing, segmenting and modeling data. Plan to attend the presentation and get a quick overview of this powerful tool.

References

Hosmer, DW., Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

SAS Institute Inc. (1989) *SAS/Stat User's Guide, Vol.2, Version 6, Fourth Edition*, Cary NC: SAS Institute Inc.

AUTHOR CONTACT

C. Olivia Rud
OptiMine Consulting
1554 Paoli Pike #286
West Chester, PA 19380

Voice: (610) 918-3801
Fax: (610) 918-3974
Internet: Olivia Rud@aol.com

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

APPENDIX A

	MALE				FEMALE			
	M	S	D	W	M	S	D	W
< 40	1.22	1.15	1.18	1.10	1.36	1.29	1.21	1.17
40-49	1.12	1.01	1.08	1.02	1.25	1.18	1.13	1.09
50-59	0.98	0.92	0.90	0.85	1.13	1.08	1.10	1.01
60+	0.85	0.74	0.80	0.79	1.03	0.98	0.93	0.88

APPENDIX B

Univariate Analysis

Variable=INCOME

Moments

N	*74,944*
Mean	*291.4656*
Std Dev	*43.4356*
Num ^= 0	*74,914*
Num > 0	*74,914*

Quantiles

100% Max	*2000*
75% Q3	*57*
50% Med	*47*
25% Q1	*41*
0% Min	*6*

Extremes

Low	*High*
6	*74*
13	*75*
28	*77*
30	*130*
32	*2000*

APPENDIX C

Univariate Analysis

Variable=INCOME

Moments

N	*74,944*
Mean	*49*
Std Dev	*6.32946*
Num ^= 0	*74,913*
Num > 0	*74,913*

Quantiles

100% Max	*130*
75% Q3	*56*
50% Med	*46.5*
25% Q1	*38.5*
0% Min	*6*

Extremes

Low	*High*
6	*73*
13	*74*
28	*75*
30	*77*
32	*130*

APPENDIX D

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	10.1594	27.1690	0.1398	0.7085	.	.
AGE	1	-23.2172	0.3284	0.0057	0.9358	-4.287240	0.000
AGE_MISS	1	-3.8671	1.7783	4.7290	0.0297	-0.997359	.
AGE_CUBE	1	0.0033	1.3594	5.9005	0.0411	0.851626	
AGE_LOG	1	1.9442	0.2658	0.0633	0.8013	0.936637	
AGE_SQR	1	0.8499	0.7291	1.5507	0.2130	0.672450	

APPENDIX E

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.5744	0.0169	0.1398	0.0001	.	.
AGE_CUBE	1	-0.0166	0.0059	0.0057	0.0049	-0.030639	0.000
MIDWEST	1	0.0263	0.0063	4.7290	0.0001	0.043238	1.027
LOG_INC	1	0.0620	0.0085	5.9005	0.0001	0.081625	1.064
INC_MISS	1	0.0291	0.0105	0.0633	0.0055	0.038147	1.030
MARRIED	1	0.0353	0.0081	1.5507	0.0001	0.046115	1.036
POPDENS	1	-0.2117	0.0057	0.0633	0.0001	-0.263967	0.809
MAIL_ORD	1	0.0634	0.0062	7.5507	0.0001	0.079093	1.065

Association of Predicted Probabilities and Observed Response

Concordant = 57.1%	Somers' D = 0.209
Discordant = 36.2%	Gamma = 0.224
Tied = 6.6%	Tau-a = 0.030
(7977226992 pairs)	c = 0.604

APPENDIX F

Model Data

DECILE	NUMBER OF ACCOUNTS	PREDICTED % OF PAID SALES	ACTUAL % OF PAID SALES	NUMBER OF PAID SALES	CUM ACTUAL % OF PAID SALES
1	48,342	11.47%	11.36%	5,492	11.36%
2	48,342	8.46%	8.63%	4,172	9.99%
3	48,342	4.93%	5.03%	2.429	8.34%
4	48,342	2.14%	1.94%	935	6.74%
5	48,342	0.94%	0.95%	459	5.58%
6	48,342	0.25%	0.28%	133	4.70%
7	48,342	0.11%	0.11%	51	4.04%
8	48,342	0.08%	0.08%	39	3.54%
9	48,342	0.00%	0.00%	2	3.15%
10	48,342	0.00%	0.00%	1	2.84%

APPENDIX G

Validation Data

DECILE	NUMBER OF ACCOUNTS	PREDICTED % OF PAID SALES	ACTUAL % OF PAID SALES	NUMBER OF PAID SALES	CUM ACTUAL % OF PAID SALES
1	48,342	10.35%	10.12%	4,891	10.12%
2	48,342	8.44%	8.16%	3,945	9.14%
3	48,342	5.32%	5.76%	2.783	8.01%
4	48,342	2.16%	2.38%	1,151	6.60%
5	48,342	1.03%	1.07%	519	5.50%
6	48,342	0.48%	0.56%	269	4.67%
7	48,342	0.31%	0.23%	112	4.04%
8	48,342	0.06%	0.05%	25	3.54%
9	48,342	0.01%	0.01%	5	3.15%
10	48,342	0.00%	0.00%	1	2.83%

APPENDIX H

Financial Analysis

DECILE	NUMBER OF ACCOUNTS	PREDICTED % OF PAID SALES	RISK INDEX	PRODUCT PROFITABILITY	AVERAGE NPV	CUM AVERAGE NPV	SUM CUM NPV
1	96,685	10.35%	0.94	$553	$58.27	$58.27	$5,633,985
2	96,686	8.44%	0.99	$553	$46.47	$52.37	$10,126,713
3	96,686	5.32%	0.98	$553	$26.45	$43.73	$12,684,175
4	96,685	2.16%	0.96	$553	$9.49	$35.17	$13,602,084
5	96,686	1.03%	1.01	$553	$4.55	$29.05	$14,042,255
6	96,685	0.48%	1.00	$553	$0.74	$24.33	$14,114,007
7	96,686	0.31%	1.03	$553	($0.18)	$20.83	$14,096,406
8	96,685	0.06%	0.99	$553	($0.34)	$18.18	$14,063,329
9	96,685	0.01%	1.06	$553	($0.76)	$16.08	$13,990,047
10	96,686	0.00%	1.10	$553	($0.77)	$14.39	$13,915,946