Scoring of Bank Customers for a Life Insurance Campaign

by

Brian Schwartz

and

Jørgen Lauridsen

Discussion Papers on Business and Economics
No. 5/2007

FURTHER INFORMATION
Department of Business and Economics
Faculty of Social Sciences
University of Southern Denmark
Campusvej 55
DK-5230 Odense M
Denmark

Tel.: +45 6550 3271
Fax: +45 6550 3237
E-mail: lho@sam.sdu.dk
http://www.sdu.dk/osbec

ISBN 978-87-91657-15-3

Scoring of Bank Customers for a Life Insurance Campaign

Part 1: Response Model

Brian Schwartz[1] and Jørgen Lauridsen[2]

[1]: Chief Analyst, Card Marketing International Ltd, Wellington, New Zealand,

brian.schwartz@cmi.co.nz.

[2]: Corresponding Author: Professor, Institute of Public Health, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark, jtl@sam.sdu.dk.

Abstract

A response model can provide a boost to the efficiency of a marketing campaign by increasing responses and/or reducing mail expenses. The purpose of this paper is to predict who – within a large banks credit card customer database - will be responsive to an offer for life insurance. We develop a response model based on logistic regression. We also discuss how the response model will act as part of a more comprehensive model combining the response model with income elements and a churn model for the purpose of maximizing income from the life insurance campaign.

Introduction

The use of targeting models has become very common in the marketing industry. Many applications like those for response or approval are quite straightforward. But as companies attempt to model more complex issues, such as attrition and lifetime value, clearly and specifically defining the goal is of critical importance. The first and most important step in any targeting-model project is to establish a clear goal and develop a process to achieve that goal.

The goal for the bank is to maximize the long term income generated from a life insurance campaign targeted to the banks existing credit card customers. How this is achieved will be discussed in the following section. However, this paper has been written in order to document the first of two steps towards this goal. It will focus on the technicalities involved in building a **response model** capable of predicting who – within the banks existing cards base - is likely to accept an offer of life insurance by means of direct marketing and subsequent telemarketing. In the second of the two steps an income model will be built and couple to a churn model. Such work, however, is not within the scope of this paper but instead reserved for future work – part 2.

The following sections will cover the creation of the response model, the choice of model applied, choice of variables included, data cleaning and data transformations, logistic regression techniques, model validation and finally how the model is expected to tie in with an income estimating model.

Maximizing Bank Income – Using a Two Step Approach

A response model can provide a boost to the efficiency of a marketing campaign by increasing responses and/or reducing mail expenses. The goal is to predict who will be responsive to an offer for a product or service. It can be based on past behavior of a similar population or some logical substitute.

However, another aspect will need to be considered: Income. Naturally, profit is even more important to the bank than the income stream but given the inability to obtain claims data at this stage the expense elements have been excluded entirely. Predicting future income of a yet-to-be life insurance customer is a matter of predicting two equally important aspects: 1) Expected choice of premium level / sum insured and 2) probability of future lapse. In other words, we want to know the likely level of

income alongside the percentage with which we are likely to loose it. With rates of customer defection reaching ever higher levels in many industries including banking, predicting turnover has become significantly more important to business in recent years.

Combining the income element with the response element enables us to answer the question the bank should be looking for in each approachable customer: Who should be contacted in order to maximize long term income? Again, the word "profit" is deliberately excluded as this part of the project is focusing solely on the income side.

In order to reach the goal of being able to answer the question of income maximization a two step approach will be applied: First, a response model is developed for the purpose of being able to estimate who is likely to accept the offer. For the next 6-12 months time this model will be used in the mail and telemarketing campaigns while the bank is collecting valuable churn and income data. At this stage, a churn model will be built along with a model predicting the income level for customers who agrees to take up the product. These two elements together will form a function which will then be paired with the already existing response model. When combined, it will be possible to apply one score to each customer – a score made up from both expected income and probability to respond. In a later section it will be discussed in more detail how such a model will be constructed. However, for now, focus is on the response model. The following sections and the processes described therein borrow heavily from the process outlined in Ruud (2001).

Constructing the Modeling Data Set

When designing a campaign for a mail/phone offer with the goal of using the results to develop a model, it is important to have complete representation from the pool of eligible customers. It is not cost efficient to mail the entire list so sampling is an effective alternative which is what has been done in this case. From a researchers perspective it is critical, though, that the sample size is large enough to support both model development and validation. In a business environment, however, a compromise will have to be made. How big the sample should be is a question common among target modelers. Unfortunately, there is no exact answer. Sample size depends on many factors. What is the expected return rate on the target group? This could be performance based such as responders. The goal is to have enough records in the target group to support all levels of the explanatory variables. One way to

think about this is to consider that the significance is measured on the cross-section of every level of every variable.

To use every level of these variables to predict response, each cross-section must have a minimum number of observations or values. And this is true among the responders and nonresponders. There is no exact minimum number, but according to Ruud (2001) a good rule of thumb is at least 25 observations. The more observations there are in the cell, the more likely it is that the value will have predictive power. A way to overcome shortcomings in sample size is to group the levels within each explanatory variable. This is a process widely used in the following regressions due to the limited data available at the time of writing.

The sample data contains call results for the customers who was mailed and subsequently called. From this variable it is possible to determine who was contacted and furthermore, who accepted the offer. At the time of writing a total of 7,500 effective contacts were available for model development. 296 of these had resulted in a sale. This is not a large amount and means that we do not have the luxury of assigning a part of these for validation. However, by the time the model is finished a range of new sales would have been made and hence it is expected that model validation can be applied to these.

Types of Data Collected from Within the Bank

Data usually falls into three basic types: demographic, behavioral, and psychographic or attitudinal.

Demographic data includes characteristics such as gender, age, marital status, income, home ownership, dwelling type, education level, ethnicity, and presence of children. Demographic data has a number of strengths. It is very stable, which makes it appealing for use in predictive modeling. Characteristics like marital status, home ownership, education level, and dwelling type aren't subject to change as frequently as behavioral data such as bank balances. Behavioral data is typically the most predictive type of data. For a bank this type of data may include elements like types and dates of purchase, payment dates and amounts, customer service activities, insurance claims or bankruptcy behavior, and more.

A database was constructed where each of the 7,500 customers with a recorded ‘effective contact’ had a wide range of demographic and behavioral data merged onto them. Each record was merged with

customer information from the time of mailing. This helps ensuring that all (especially behavioral) information on each customer is as close to the time of contact as possible.

Process and Methodology Applied for Modeling Response

Statistical Methodology: Logistic Regression

Today, there are numerous tools for developing predictive and descriptive models. Some use statistical methods such as linear regression and logistic regression. Others use nonstatistical or blended methods like neural networks, genetic algorithms, classification trees, and regression trees. Much has been written debating the best methodology. However, it is important to understand that the steps surrounding the model processing are more critical to the overall success of the project than the technique used to build the model. That is why logistic regression was chosen for this project. It is the most widely available technique and is rarely, significantly, outperformed by far more complex models.

Logistic regression is part of a category of statistical models called generalized linear models. This broad class of models includes ordinary regression and ANOVA, as well as multivariate statistics such as ANCOVA and loglinear regression. An excellent treatment of generalized linear models is presented in Agresti (1996).

Logistic regression allows one to predict a discrete outcome, such as acceptance or rejection of insurance offer, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure. Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables – making it unsuitable for this project. Thus, in instances where the independent variables are a categorical, or a mix of continuous and categorical, logistic regression is preferred.

The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success θ, or the value 0 with probability of failure $1-\theta$. This type of variable is called a Bernoulli (or binary) variable.

As mentioned previously, the independent or predictor variables in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group.The relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of θ:

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_i x_i)}}$$

Where α = the constant of the equation and, β = the coefficient of the predictor variables.

An alternative form of the logistic regression equation is:

$$\mathbf{logit}[\theta(\mathbf{x})] = \mathbf{log}\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Several different options are available during model creation. Variables can be entered into the model in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted, called stepwise regression. Three of these methods are used in the process of building the life insurance response model. Each of these will be described in the following section.

Modeling Response in Four Steps

The modeling process will be carried out according to the following four steps:

1) First Round of Variable Reduction

2) Select and Transform the Remaining Variables while applying Second round of Variable Reduction

3) Process the Model while applying third and final Round of Variable Reduction

4) Validate the Model

Each of the above steps is briefly described below.

Step 1: First Round of Variable Reduction

The model dataset contains 233 variables. A thorough analysis of each variable is not an efficient use of time. Instead, and even before looking at response rates – one should remove variables that are either unlikely to have influence on response rates or contains values that are not useful for modeling. An example is a variable which may have the same value for all records in the model dataset. Other examples would be: customer number, surname etc. The result of this exercise reduced the number of variables from 233 to approximately 100 variables.

Step 2: Select and Transform the Remaining Variables while applying Second round of Variable Reduction

The remaining 100 variables were analyzed one by one and in respect to the observed response rates. A wide range of these variables showed no relation to response rate and will hence be discarded. For the remaining group, some 45 variables, simple statistics, frequency tables and rank analyses have been included in this documentation (see following chapter). The variables are also analyzed for outliers and missing values. An outlier is a single or low-frequency occurrence of the value of a variable that is far from the mean as well as the majority of the other values for that variable. In such cases a decision will be made as to what value they should be assigned if not either omitted or grouped separately.

Some variables are continuous, and some are categorical. Logistic regression sees all predictive variables as continuous. So for the non-continuous variables, indicator variables will be used to trick the model into thinking they are continuous. For continuous variables, increased model performance can be obtained if the predictors can be made more linear (because the predictors are linear in the log of the odds - as defined previously). It's not a statistical requirement as with standard regression but merely a benefit to the final model. One way to determine the best transformation or segmentation of a

continuous variable is to create several variations and use a forward logistic regression to select the best fit. The first step is to break the continuous variable into segments. In the following chapter this has been done for all continuous variables. Some people put all continuous variables into segments and treat them as categorical variables. This may work well to pick up nonlinear trends. The biggest drawback is that it loses the benefit of the relationship between the points in the curve that can be very robust over the long term. Another approach is to create segments for obviously discrete groups. Then test these segments against transformed continuous values and select the winners. This is the approach used in this model.

For every continuous variable the response rates are compared and new binary variable are formed where suitable. Along with such binary transformations the original variable and its 20 mathematical transformations have been put into a logistic stepwise regression that estimates the probability to accept the life insurance offer. Whatever significant variable(s) are left in the model (if any) will be kept towards the final model. The rest will be discarded.

The binary formations will vary from variable to variable but the 20 mathematical transformations remain the same:

Squared:	variable_sq = variable_est3**2;
Cubed:	variable_cu = variable_est3**3;
Square Root:	variable_sqrt = sqrt(variable_est3);
Cube Root:	variable_curt = variable_est3**.3333;
Log:	variable_log = log(max(.0001,variable_est3));
Exponent:	variable_exp = exp(max(.0001,variable_est3));
Tangent:	variable_tan = tan(variable_est3);
Sine:	variable_sin = sin(variable_est3);
Cosine:	variable_cos = cos(variable_est3);
Inverse:	variable_inv = 1/max(.0001,variable_est3);
Squared Inverse:	variable_sqi = 1/max(.0001,variable_est3**2);
Cubed Inverse:	variable_cui = 1/max(.0001,variable_est3**3);
Square Root Inverse:	variable_sqri = 1/max(.0001,sqrt(variable_est3));
Cube Root Inverse:	variable_curi = 1/max(.0001,variable_est3**.3333);
Log Inverse:	variable_logi = 1/max(.0001,log(max(.0001,variable_est3)));
Exponent Inverse:	variable_expi = 1/max(.0001,exp(max(.0001,variable_est3)));
Tangent Inverse:	variable_tani = 1/max(.0001,tan(variable_est3));
Sine Inverse:	variable_sini = 1/max(.0001,sin(variable_est3));
Cosine Inverse:	variable_cosi = 1/max(.0001,cos(variable_est3));

Some of the transformations are not applicable to variables with negative values. Hence, these transformations are excluded in such cases.

The chosen transformations differ from variable to variable – something which is reflected in the transformations chosen as candidates for the final model. In the following chapter it will be clearly indicated what variable transformations have been selected for further analysis.

Step 3: Process the Model while applying third and final Round of Variable Reduction

Following the variable reduction and creation processes above the variables for evaluation in the final model was reduced accordingly. In the model processing stage three methods within Logistic Regression will be used: Stepwise, Backward, and Score. By using several methods, we take advantage of different variable reduction techniques while creating the best fitting model. The steps are as follows:

Stepwise. The first step will be to run a stepwise regression with an artificially high level of significance. This will further reduce the number of candidate variables by selecting the variables in order of predictive power. A significance level of .30 is used. This method is very similar to *forward* selection. Each time a new variable enters the model, the univariate chi-square of the remaining variables not in the model is recalculated. Also, the multivariate chi-square or incremental predictive power of each predictive variable in the model is recalculated. The main difference is that if any variable - newly entered or already in the model - becomes insignificant after it or another variable enters, it will be removed.

Backward. Next, a backward regression with the same artificially high level of significance will be run. This method fits all the variables into a model and then removes variables with low predictive power. The benefit of this method is that it might keep a variable that has low individual predictive power but in combination with other variables has high predictive power. It is possible to get an entirely different set of variables from this method than with the stepwise method.

This method begins with all the variables in the model. Each variable begins the process with a multivariate chi-square or a measure of predictive power when considered in conjunction with all other variables. It then removes any variable whose predictive power is insignificant, beginning with

the most insignificant variable. After each variable is removed, the multivariate chi-square for all variables still in the model is recalculated with one less variable. This continues until all remaining variables have multivariate significance.

This method has one distinct benefit over stepwise. It allows variables of lower significance to be considered in combination that might never enter the model under the forward and stepwise methods. Therefore, the resulting model may depend on more equal contributions of many variables instead of the dominance of one or two very powerful variables.

Score. This step evaluates models for all possible subsets of variables. This method constructs models using all possible subsets of variables within the list of candidate variables using the highest likelihood score (chi-square) statistic. It does not derive the model coefficients. It simply lists the best variables for each model along with the overall chi-square.

From the output the final list of variables are selected. This variable list will be used in a final, regular, logistic regression from where the coefficients will be calculated. These coefficients will then need to be inserted into the below logistic regression function to form the predicted probability to respond:

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

Where α = the constant of the equation and, β = the coefficient of the predictor variables.

Step 4: Validate the Model

There are several methods for validating models. One of the drawbacks with the model derived in this documentation is the lack of validation data. Because only 296 sales were available for analysis it was seen as essential to include them all in the model building process. Therefore, a classic model validation check cannot be applied in this case. Instead, a resampling method has been applied in order to apply confidence intervals around the model estimates. In addition, a decile analysis of the estimated response rate will be applied to both the model development data and the final pool data upon which the model eventually will be applied.

Selection and Transformation of Variables

The following section provides a brief description of the variables that proved significant during the regression analysis.

DATEOPEN_mis_cos

DATEOPEN_mis_cosi

A cosine and a cosine inverse transformation of the credit card account open date.

age_mis_sq

A squared transformation of the primary cardholder's age.

AGEMATURE_ind

AGEYOUNG_ind

Indicator variables identifying the very young customer segment (AGEYOUNG) or the more mature customer segment (AGEMATURE).

insstats_ind

An indicator variable identifying credit card accounts that either have – or used to have – credit card insurance associated to the account.

INTEREST_mis

Amount of interest paid last month – as listed on the statement balance.

LUXURY_mis_sq

A squared transformation of the number of times the credit card was used to buy luxury goods in the past 2 years. The variable has been generated by searching transaction data for a group of specific merchant codes.

mosaic_low_ind

Mosaic is a geo-demographic classification that describes Australia's neighbourhoods by allocating them into 41 types. "mosaic_low_ind" is an indicator variable identifying 12 of these types as having a relative low probability of accepting the offer.

SECPURP_ind

The secondary purpose flag indicates whether a customer's information can be used for secondary purposes. In its original form the variable can have three values, "granted", "not granted" or "no value". Customers with no value had a significantly lower response rate than for the other two. As a result these were grouped separately into an indicator variable.

SEGCUST_ind

Indicator variable based on a segment variable which again is derived from credit card behaviour. Certain segments have a relatively higher probability to accept the offer and were hence grouped accordingly.

STATE_ind

Indicator variable based on the customer's residential address. Customers from certain states proved slightly better responders than others. As a result, an indicator variable was used to isolate these.

TITLE_IND

Indicator variable derived in order to identify unmarried women by grouping all customers with titles "Ms" or "Miss". One would think that such indicator could be generated by combining two other variables in the dataset, marital status and gender. However, TITLE is more likely to remain updated as it used for communication to the customer whereas the variable containing marital status may be as old as credit card application date.

Z_MARITAL_ind

Indicator variable identifying customers either separated or in a de facto relationship.

BAL_CL_mis_tan

A tangent transformation of average balance to credit limit ratio over the last 12 months.

BHVRSCRE_mis_sq

Squared transformation of the account's current behaviour score. Each month at cycle date, an account is given a score based on its performance over the previous months. Scores indicate the level of risk of an account going 'bad'.

F_CASH_MED_ind

Indicator variable identifying customers with 2 cash advances in the last 12 months.

F_IND_STUD_ind

Indicator variable identifying customers with occupations such as “Student”, “Unemployed” etc.

F_IND_TRADE_ind

Indicator variable identifying customers that are tradesmen by profession. Examples are “Electrician”, “Mechanic”, “Carpenter” etc.

F_FLIGHT_ind

Indicator variable identifying customers that use their credit card to book flights.

OTHERBANK_IND

Indicator variable identifying customers that have other relations to the bank aside from a credit card.

OTHRCOST_mis_sqrt

Square root transformation of bank costs related to the card account during last month.

PHNEMOBL_IND

Indicator variable identifying customers that have listed a cell phone number as a potential way of getting in contact with the customer.

PROFBAND_EXT_IND

Indicator variable identifying customers in extreme profit intervals, either positive or negative.

SEGMKTNG_ind

Indicator variable identifying customers from certain marketing related segments that have shown a relatively low probability to respond. The original segment variable is based on a range of attributes such as footings, profitability etc.

utilise_mis_sqi

Squared Inverse transformation of card utilisation.

Processing and evaluating the model

Once the transformations have been applied the model is ready to be estimated. The method which has been applied was discussed in previous sections and hence not elaborated on here. In addition, this section has been limited to include only the final model – or in fact two (as will be shown).

Final Model Coefficients

As part of the SCORE option in the logistic regression a list of potential variable combinations is produced – in this case the best combination per number of variables included. Figure A below shows the resulting Chi-Squares.

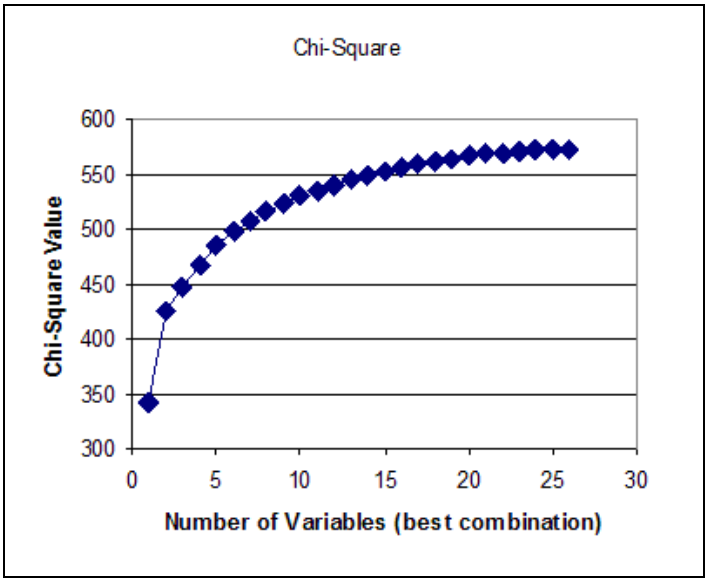

Figure A. Chi-Square for the best Variable Combination given the number of Variable Included

Standard procedure suggests choosing the number of variables where Chi-Square seems to level off. In this case it is somewhat hard to establish exactly when that happens. As a result, two models were chosen, one with 12 variables (model 1) and one with 26 variables (model 2). Going forward, model 1 will be the primary choice and model 2 will be serving as a reference and potential backup.

Coefficients for Model 1

Below are listed the regression coefficients produced by the final logistic regression when the best combination of 12 variables is used (as selected with the SCORE option above):

```
logscore_response =     -5.0953                         -0.6579*DATEOPEN_mis_cos
                        -0.00008*DATEOPEN_mis_cosi      +0.000442*age_mis_sq
                        +1.0339*insstats_ind            +0.00261*INTEREST_mis
                        -0.00252*LUXURY_mis_sq          -0.7196*mosaic_low_ind
                        +2.1059*SECPURP_ind             +0.3965*SEGCUST_ind
                        +0.2891*STATE_ind               +0.5708*TITLE_IND
                        +0.4787*Z_MARITAL_ind
```

Table A below shows the resulting Wald Chi-Squares used to establish which of the variables contribute the most towards the estimated response rate. The higher the value the higher the contribution.

Table A. Wald Chi-Square for Model 1.

Variable:	Wald Chi-Square	Variable:	Wald Chi-Square
SECPURP_ind	210.24	DATEOPEN_mis_cos	10.07
insstats_ind	53.23	LUXURY_mis_sq	8.57
mosaic_low_ind	18.75	SEGCUST_ind	7.83
INTEREST_mis	17.02	Z_MARITAL_ind	7.52
TITLE_ind	14.30	DATEOPEN_mis_cosi	6.78
age_mis_sq	13.74	STATE_ind	5.08

Coefficients for Model 2

Below are listed the regression coefficients produced by the final logistic regression when all 26 variables are included:

```
logscore_response =     -6.6352                      -0.2188*BAL_CL_mis_tan
                        -1.1647*BHVRSCRE_mis_sq      -0.6747*DATEOPEN_mis_cos
                        -0.00008*DATEOPEN_mis_cosi   +0.6895*AGEMATURE_ind
                        +0.7941*AGEYOUNG_ind         +0.000909*age_mis_sq
                        +0.1984*F_CASH_MED_ind       +0.3095*F_IND_STUD_ind
                        +0.2012*F_IND_TRADE_ind      +0.2869*F_FLIGHT_ind
                        +1.0127*insstats_ind         +0.00141*INTEREST_mis
                        -0.00219*LUXURY_mis_sq       -0.7002*mosaic_low_ind
                        -0.3605*OTHERBANK_IND        +0.1462*OTHRCOST_mis_sqrt
                        +0.3486*PHNEMOBL_IND         +0.1967*PROFBAND_EXT_IND
                        +2.1624*SECPURP_ind          +0.2969*SEGCUST_ind
                        +0.2469*SEGMKTNG_ind         +0.3033*STATE_ind
                        +0.6021*TITLE_ind            +0.4401*Z_MARITAL_ind
                        -0.00005*utilise_mis_sqi
```

Table B below shows the resulting Wald Chi-Squares.

Table B. Wald Chi-Square for Model 2.

Variable:	Wald Chi-Square	Variable:	Wald Chi-Square
SECPURP_ind	205.29	F_IND_STUD_ind	4.02
insstats_ind	44.84	OTHERBANK_IND	3.88
mosaic_low_ind	17.49	PHNEMOBL_IND	3.45
TITLE_ind	14.03	F_FLIGHT_ind	2.94
DATEOPEN_mis_cos	10.32	OTHRCOST_mis_sqrt	2.93
age_mis_sq	7.91	INTEREST_mis	2.72
DATEOPEN_mis_cosi	6.94	AGEYOUNG_ind	2.43
LUXURY_mis_sq	6.24	BAL_CL_mis_tan	1.88
Z_MARITAL_ind	6.21	SEGMKTNG_ind	1.72
STATE_ind	5.49	PROFBAND_EXT_IND	1.47
AGEMATURE_ind	5.04	F_IND_TRADE_ind	1.45
utilise_mis_sqi	4.85	BHVRSCRE_mis_sq	1.26
SEGCUST_ind	4.02	F_CASH_MED_ind	1.15

For both models the probability to accept the life insurance offer can now be calculated as:

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i)}}$$

Where α = the constant of the equation and, β = the coefficient of the predictor variables above.

The following section will be analyzing the resulting probabilities and how they compare to the actual, observed response rates from the model data. The section will also deal with the application of the model on the final pool data.

Evaluating Models against Development Data and Pool Data

The following table shows the predicted sales rate to contact sorted by decile – when applied to the model data (7,346 records) as well as the pool data (397,999). It also shows the average, actual response rate as well as the number of life insurance sales in each decile.

Evaluating Model 1

Table C. Actual and estimated Response Rates for Model 1

Model Data:				
Decile	Accounts	Average Predicted Response Rate	*Average Actual Response Rate*	*Number of Actual Sales*
1	734	19.08%	*18.80%*	*138*
2	735	7.68%	*8.16%*	*60*
3	735	4.19%	*5.03%*	*37*
4	734	2.42%	*1.77%*	*13*
5	735	1.64%	*1.90%*	*14*
6	735	1.23%	*1.09%*	*8*
7	734	0.98%	*1.23%*	*9*
8	735	0.76%	*0.41%*	*3*
9	735	0.56%	*0.68%*	*5*
10	734	0.33%	*0.41%*	*3*
Total:	7,346	3.88%	*3.95%*	*290*

Pool Data:		
Decile	Accounts	Average Predicted Response Rate
1	39,799	15.12%
2	39,800	5.83%
3	39,800	3.02%
4	39,800	1.86%
5	39,799	1.38%
6	39,801	1.08%
7	39,800	0.87%
8	39,800	0.68%
9	39,800	0.50%
10	39,800	0.30%
Total:	397,999	3.06%

Figure B illustrates the predicted sales rate to contact for Model 1

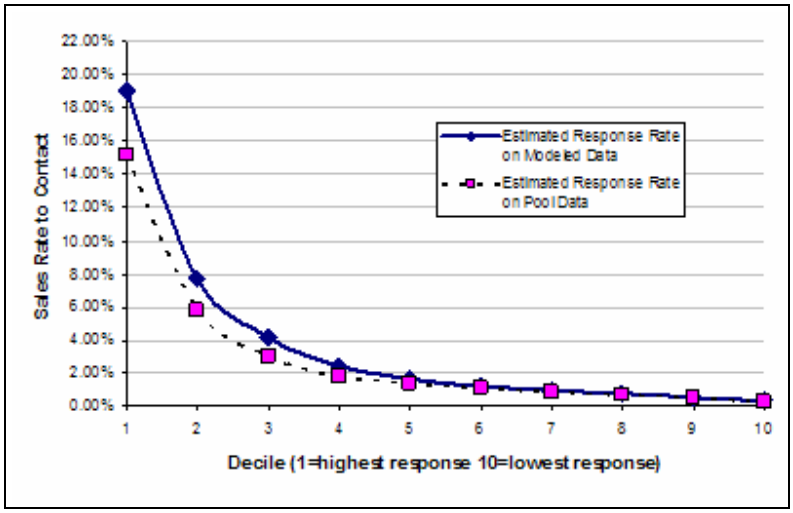

Figure B. Estimated Response Rates for Model 1

The overall downward shift in response rate for the pool data is likely to have been caused by a relative over-representation in the model/development data of records containing SECPURP_ind = 1, AGEYOUNG_ind = 1 and INSSTATS_ind = 1. All three values are strong indicators of a customer with a high likelihood of accepting the offer. There are generally more of these occurrences in the model data than in the pool data and as a result the overall, average, estimated response rate will be slightly overrated on the model data. It does not mean that the model will not be able to identify potential responders. It merely means that - if applying a random selection over the pool data - we should not expect a sales result as high as what occurred in the model data.

Evaluating Model 2

Table D. Actual and estimated Response Rates for Model 2

Model Data:				
Decile	Accounts	Average Predicted Response Rate	Average Actual Response Rate	Number of Actual Sales
1	734	20.12%	21.12%	155
2	735	7.57%	6.53%	48
3	735	4.09%	4.08%	30
4	734	2.46%	2.18%	16
5	735	1.64%	2.45%	18
6	735	1.19%	0.82%	6
7	734	0.89%	0.54%	4
8	735	0.66%	0.54%	4
9	735	0.47%	0.68%	5
10	734	0.26%	0.54%	4
Total:	7,346	3.94%	3.95%	290

Pool Data:		
Decile	Accounts	Average Predicted Response Rate
1	39,799	12.75%
2	39,800	4.43%
3	39,800	2.38%
4	39,800	1.52%
5	39,799	1.09%
6	39,801	0.82%
7	39,800	0.62%
8	39,800	0.47%
9	39,800	0.33%
10	39,800	0.19%
Total:	397,999	2.46%

The comments accompanying model 1 seems even more relevant for model 2. Without looking further into the reasons it is assumed that the higher gap of response rates between modeled and pool data is due to some of the additional variables being distributed somewhat differently in the pool data compared to the model data. The effects of potential overfitting are unlikely to have created this gap simply because an 'overfit' on the model data would merely be carried over to the pool data. In other words the negative impact of overfitting would not have shown until the sales results would come through.

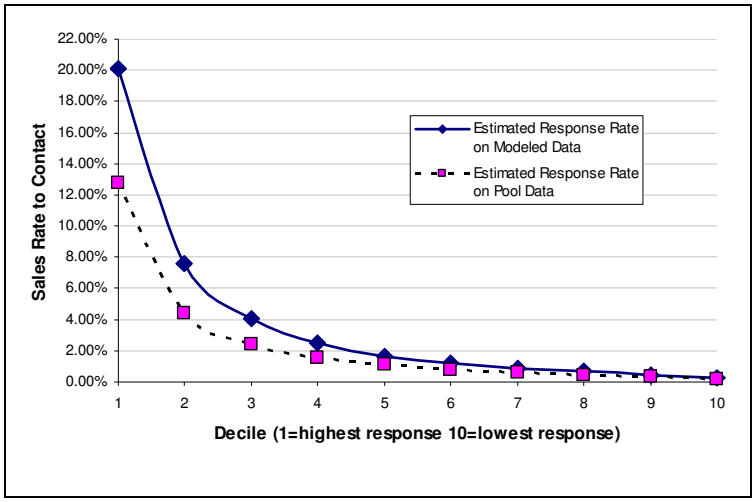

Figure C. Estimated Response Rates for Model 2

Validation of the Models

Validating the model is a critical step in the process. If a model does not validate well it can be due to data problems, poorly fitting variables, or problematic techniques. There are several methods for validating models.

The classic approach is to divide the model data into two, develop the model on the first part and test the model on the second part using for example gains charts or decile analysis. In this case we did not have the luxury of allocating part of the model data for testing. Test data will be available at a later stage but not at the time of writing / implementation of the (first version of the) model. Instead, a resampling method was applied to strees test the model coefficients.

Resampling

Resampling is a common-sense, nonstatistical technique for estimating and validating models. It provides an empirical estimation (based on experience and observation) instead of a parametric estimation (based on a system or distribution). Consider the basic premise that over-fitting is fitting a model so well that it is picking up irregularities in the data that may be unique to that particular data set. Two main types of resampling techniques are used in database marketing: jackknifing and bootstrapping. The following section describes the bootstrapping method in more detail.

Bootstrapping

Bootstrapping is an empirical technique for finding confidence intervals around an estimate. Bootstrapping uses full samples of the data that are pulled from the original full sample *with replacement*. In other words, with a sample random samples are drawn from the original sample. Because the bootstrap sample is pulled with replacement, it is possible for one observation represented several times and another observation to be missed completely. This process is repeated many times.

A variant of the bootstrapping was applied in this model. 1000 regression models were calculated based on random samples drawn from the available model data. The results were used to compare with the model coefficients described in the previous section. Without further elaboration it was concluded that the two models described previously are fairly robust – with model 1 being more robust than model 2. Naturally, a model based on as few observations as 296 sales are prone to be sensitive to the few responders it is based upon. However, as more sales become available it will be easier to build a more stable model.

Combining Response Model with Future Income Model

As indicated previously it is required to build a model that can not only determine which customers are most likely to respond to the life insurance campaign but also determine which customers will be the most profitable in the sense of sum insured, premium paid and probability of lapsing.

The actual income of a customer is not of relevance in this framework. Instead, what is relevant is the ranking of the customers once sum insured, premium paid, probability of lapse and probability of accepting the life insurance offer is incorporated. For this purpose we will define a new function, **RIL** – short for Response, Income and Lapse.

Consequently, the RIL model is chosen to be a function of likelihood to 1) take up product and 2) expected discounted future income calculated as Net Present Value (NPV).

$$RIL = F (Prob[take\text{-}up], E [NPV])$$

where

1) **Probability of take-up** is a function of customer-characteristics historically shown to have an impact on take-up, i.e.:

$$\text{Prob[take-up]} = F(x_1, x_2, x_3, x_4, x_5 \ldots x_n)$$

With x being customer characteristics historically shown to have an impact on take-up. All previous sections in this documentation have been dealing with this task.

2) **Expected Net Present Value** is a function of expected future income as well as expected lapse rate:

$$E[NPV] = F(\text{Prob [lapse]}, \text{expected future income}, \text{discount factor})$$

Please note that the income model does not take into account expenses associated with the customer – including claims and/or probability of claim.

Lapse is lastly a function of customer-characteristics historically shown to have an impact on lapse, i.e.:

$$\text{Prob[lapse]} = F(x_1, x_2, x_3, x_4, x_5 \ldots x_n)$$

Prob[lapse] can be calculated at certain points in time, i.e. probability to have lapsed after 1,2,3 years etc. This will be held against expected income for the same time periods discounted in order to finally form the expected net present value.

By putting both elements together we get the following function defining expected net present value

$$\begin{aligned} E[NPV] = {} & (1 - \text{Prob [lapse after t=1 years]}) * \text{expected income in year 1} * (1-r)^1 \\ & + (1 - \text{Prob [lapse after t=2 years]}) * \text{expected income in year 2} * (1-r)^2 \\ & + (1 - \text{Prob [lapse after t=3 years]}) * \text{expected income in year 3} * (1-r)^3 \\ & + (1 - \text{Prob [lapse after t=4 years]}) * \text{expected income in year 4} * (1-r)^4 \\ & + (1 - \text{Prob [lapse after t=5 years]}) * \text{expected income in year 5} * (1-r)^5 \\ & \ldots\ldots \text{ etc etc .. } \quad \ldots\ldots \text{for } t \rightarrow \text{infinite and where 'r' denotes the discount factor} \end{aligned}$$

For the sake of simplicity it can be assumed that the banks customers' expected income in year 1 is carried through to the future – only appreciated by an inflation factor. To further simplify it can be assumed that the appreciation factor is equivalent to the discount factor used in obtaining the NPV.

Hence, the two elements may equal each other out and leave us with the following expression of income:

E [NPV] = (1 - Prob [lapse after t=1 years]) * expected income in year 1
+ (1 - Prob [lapse after t=2 years]) * expected income in year 1
+ (1 - Prob [lapse after t=3 years]) * expected income in year 1
+ (1 - Prob [lapse after t=4 years]) * expected income in year 1
+ (1 - Prob [lapse after t=5 years]) * expected income in year 1
…… etc etc .. …….for t → infinite

Left is to decide 1) how to estimate attrition and 2) how to estimate the expected income. These questions are outside the scope of this documentation and will be dealt with in part 2 of the project.

We now have both elements of the RIL model. Left is to decide how the two elements should be incorporated in order to form a ranking system.

How to use the RIL Model to rank customers for the Life Insurance Campaign

As probability to take up insurance is part of the model we are facing a trade-off between 1) probability to respond and 2) expected income (NPV). The model needs to be equipped with a measure that can be used for ranking the customers - taking both values into account. It is desirable to have a model that can not only weigh the two elements but also easily change the weights should there be a business need for this.

To accommodate for this we choose to rank the customers by applying weights to the two elements of the RIL model:

1) A weight given to customers ranked by probability to respond and
2) a weight given to customers ranked by expected NPV.

The two weights should equal one. The table below shows a simple example.

The weighted average will be the final RIL ranking model that reflects the weights applied. Hence, the model is not using the actual probability to respond and the actual expected NPV but rather the rank to create a final rank order of all customers. The actual weights will be decided by the bank on a senior

level and supported by simulations showing the estimated income projections under various weight scenarios.

The actual ranking itself cannot be used for financial purposes. It is not a financial value but merely a ranking – a number telling the modeler how the customer ranks compared to the other customers.

Table 1. Emphasize/Weight given solely to probability to take up insurance

Weight given:	100%	0%

Customer:	Rank of Probability to take up Simple Life	Rank of E(NPV)	RIL ( ) using above weighting
Customer A	1	4	1
Customer B	2	2	2
Customer C	3	3	3
Customer D	4	1	4
Customer E	5	5	5

Table A shows that customer A is the most desirable customer to include in the campaign. This is because customer A's probability to take up the product is the highest among the five customers and we have put all importance/weight onto probability to accept the offer. Note how no actual probability to respond or expected NPV is used at this stage.

Table 2. Emphasize/Weight given solely to expected NPV

Weight given:	0%	100%

Customer:	Rank of Probability to take up Simple Life	Rank of E(NPV)	RIL ( ) using above weighting
Customer A	1	4	4
Customer B	2	2	2
Customer C	3	3	3
Customer D	4	1	1
Customer E	5	5	5

Table 2 shows that customer D is the most desirable customer to include in the campaign. This is because customer D's expected income (E[NPV]) is the highest among the five customers and we have put all importance/weight onto this function. Notice how the probability to take up the product is the second lowest among the five customers but we are not giving any importance to response rate in this scenario.

Table 3. Emphasize/Weight given equally to probability to take up insurance and expected NPV

Weight given:	50%	50%	
Customer:	Rank of Probability to take up Simple Life	Rank of E(NPV)	RIL ( ) using above weighting
Customer A	1	4	2.5
Customer B	2	2	2
Customer C	3	3	3
Customer D	4	1	2.5
Customer E	5	5	5

Table 3 shows that customer B is now the most desirable customer to include in the campaign. By weighing the two ranks using a 50% / 50% weight this customer is the overall most desirable customer to include in a life insurance campaign. The probability to respond is quite high and so is the expected NPV - should the customer take up the product.

Conclusion

This document has served as part 1 of the documentation for a project concerned with building a model for selecting a bank's customers for a particular life insurance campaign. Focus has been primarily on building the response model and describing the variables analyzed during this process. An introduction was made to the idea of coupling the response model to a later model which will focus on the income side – once the customer has accepted the life insurance product.

References

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. John Wiley and Sons, Inc.

Hosmer, David and Stanley Lemeshow.1989. *Applied Logistic Regression*. John Wiley and Sons, Inc.

Menard, Scott.1995. *Applied Logistic Regression Analysis*. Sage Publications.Series: Quantitative Applications in the Social Sciences, No. 106.

Ruud, Olivia Parr, "*Data Mining Cookbook. Modeling Data for Marketing, Risk and Customer Relationship Management*", Wiley, 2001

Tabachnick , Barbara and Linda Fidell.1996. *Using Multivariate Statistics*, Third edition. Harper Collins.