Old Mutual Life Assurance Company (SA)

- Established 1845
- Old Mutual plc
- 3,9m Life assurance clients
- 2,9m Banking clients
- 0,6m Short term insurance clients
- £169 billion funds under management
Developing scoring models to predict and enhance probability of client repurchasing

Improve profitability
DATAFLOW STRUCTURE

OPERATIONAL SYSTEMS

Warehouse

CURRENT AND DERIVED DATA

HISTORY

BASE

SAS

NT SERVER MART

Campaign Segmentor

ENTERPRISE MINER

FUTRIX

OLD MUTUAL
PREPARING THE STUDY BASE

Client Base

Data mining Universe

Descriptive dimensions

Changed dimensions
DEFINING TARGET VARIABLES

PURCHASES = 
- New Policy
- Voluntary Addition
+ Lump sum injection
- Discontinuances
SELECT UNIVERSE

Does actual purchasing represent purchasing potential?
POTENTIAL OF CLIENT \( \times \) EFFICIENCY OF CONTACT = ACTUAL SALES
SELECTING UNIVERSE

- Identify clients most likely to have been contacted
  - Select from active salesmen
<table>
<thead>
<tr>
<th>SEGMENT</th>
<th>Client with change</th>
<th>Client Count</th>
<th>%Clients with a change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top</td>
<td>27,850.00</td>
<td>259,527.00</td>
<td>10.73%</td>
</tr>
<tr>
<td>Middle</td>
<td>14,786.00</td>
<td>304,367.00</td>
<td>4.86%</td>
</tr>
<tr>
<td>Lower</td>
<td>4,085.00</td>
<td>115,797.00</td>
<td>3.53%</td>
</tr>
<tr>
<td>Orphan</td>
<td>26,153.00</td>
<td>1,172,058.00</td>
<td>2.40%</td>
</tr>
<tr>
<td>Summary</td>
<td>74,874.00</td>
<td>1,851,749.00</td>
<td>4.04%</td>
</tr>
<tr>
<td>Latest Servicing Intermediary Code</td>
<td>Client with change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>067937</td>
<td>140.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>065582</td>
<td>103.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>032114</td>
<td>96.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>042262</td>
<td>94.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>016561</td>
<td>83.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>067914</td>
<td>83.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>068553</td>
<td>81.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>013158</td>
<td>80.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>043977</td>
<td>79.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>067931</td>
<td>76.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>018137</td>
<td>71.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>031346</td>
<td>70.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>040903</td>
<td>69.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>060196</td>
<td>63.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>030315</td>
<td>62.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>060908</td>
<td>61.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>042759</td>
<td>59.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>034738</td>
<td>58.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>012855</td>
<td>57.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SELECTING UNIVERSE

- Identify clients most likely to have been contacted
  - Select from active salesmen

- Counteract low repurchase rate
  - Balance purchasers and non-purchasers
PRODUCT REPURCHASE RATES

Create target variable
  • Specific product

Select Universe
  • Specific product sales activity

Select Dimensions
  • Create new variables
CREATE APPROPRIATE DIMENSIONS

Time Variables derived from dates

- Space between events
- Seasonal
- Rhythmic
- Specific Event
% Clients who bought related to birthday month
MISSING VALUES

• Discard

• Average or deducted value?

• Dummy value

• Understand why
<table>
<thead>
<tr>
<th>Product Type</th>
<th>Occupation</th>
<th>Blank</th>
<th>A1 (Doctor)</th>
<th>Other</th>
<th>%Occupation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endowment Assurance</td>
<td></td>
<td>36,508</td>
<td>1,393</td>
<td>12,809</td>
<td>28.01%</td>
<td>50,710</td>
</tr>
<tr>
<td>Endowment Pure</td>
<td></td>
<td>50,211</td>
<td>3,322</td>
<td>29,409</td>
<td>39.46%</td>
<td>82,942</td>
</tr>
<tr>
<td>Endowment Pure RA</td>
<td></td>
<td>28,499</td>
<td>2,406</td>
<td>18,999</td>
<td>42.89%</td>
<td>49,904</td>
</tr>
<tr>
<td>Endowment RA</td>
<td></td>
<td>5,139</td>
<td>370</td>
<td>3,107</td>
<td>40.36%</td>
<td>8,616</td>
</tr>
<tr>
<td>Term</td>
<td></td>
<td>15,569</td>
<td>1,475</td>
<td>11,491</td>
<td>45.44%</td>
<td>28,535</td>
</tr>
<tr>
<td>Whole Life</td>
<td></td>
<td>19,228</td>
<td>2,279</td>
<td>17,002</td>
<td>50.07%</td>
<td>38,509</td>
</tr>
</tbody>
</table>
## MISSING DEMOGRAPHIC CLUSTERS

### Salesmen
- 75% Clustered: 1.324
- 25% Not Clustered:
  - Clients with no address: 0.350
  - P.O. Box address: 0.850
  - Rural areas: 1.790

### Mail
- 75% Clustered: 1.030
- 25% Not Clustered:
  - Clients with no address: 0.000
  - P.O. Box address: 1.150
  - Rural areas: 0.840
GEOGRAPHIC INDICATORS

• Province
  • Weak indicator

• Match Between Agent’s and Client’s Province
  • Variable Impact
<table>
<thead>
<tr>
<th>Telephone indicator</th>
<th>Home Telephone Number</th>
<th>Day Telephone Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>184 B</td>
<td>22843</td>
<td>21001</td>
</tr>
<tr>
<td>185 B</td>
<td>00000000000</td>
<td>0123155627</td>
</tr>
<tr>
<td>186 B</td>
<td>4260</td>
<td>413</td>
</tr>
<tr>
<td>187 B</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>188 B</td>
<td>0153031818</td>
<td>-</td>
</tr>
<tr>
<td>189 B</td>
<td>0137971278</td>
<td>0153451394</td>
</tr>
<tr>
<td>190 B</td>
<td>0152938143</td>
<td>71044</td>
</tr>
<tr>
<td>191 B</td>
<td>7730056</td>
<td>0828968601</td>
</tr>
<tr>
<td>192 B</td>
<td>7730995</td>
<td>7730995</td>
</tr>
<tr>
<td>193 B</td>
<td>0158513036</td>
<td>0158510040</td>
</tr>
<tr>
<td>194 B</td>
<td>7971806</td>
<td>00000</td>
</tr>
<tr>
<td>195 B</td>
<td>0158510350</td>
<td>-</td>
</tr>
<tr>
<td>196 B</td>
<td>0137730387</td>
<td>1122</td>
</tr>
<tr>
<td>197 B</td>
<td>3031701</td>
<td>3030578</td>
</tr>
<tr>
<td>198 B</td>
<td>3031606</td>
<td>1523631606</td>
</tr>
<tr>
<td>199 B</td>
<td>00000000000</td>
<td>015821132</td>
</tr>
<tr>
<td>200 B</td>
<td>24704</td>
<td>33251</td>
</tr>
<tr>
<td>201 B</td>
<td>3032348</td>
<td>510052</td>
</tr>
<tr>
<td>202 B</td>
<td>3031003</td>
<td>2952923</td>
</tr>
</tbody>
</table>

Telephone indicator = B

Total Records in Extract: 9,034

Use Variable Labels: Yes

Record Numbering: Yes
## TELEPHONE NUMBERS

<table>
<thead>
<tr>
<th>Home Telephone</th>
<th>Work Telephone</th>
<th>Model Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Code</td>
<td>No Code</td>
<td>1.054</td>
</tr>
<tr>
<td>No Code</td>
<td>Complete Number</td>
<td>0.885</td>
</tr>
<tr>
<td>No Code</td>
<td>No Number</td>
<td>0.952</td>
</tr>
<tr>
<td>Complete Number</td>
<td>No Code</td>
<td>1.595</td>
</tr>
<tr>
<td>Complete Number</td>
<td>Complete Number</td>
<td>1.518</td>
</tr>
<tr>
<td>Complete Number</td>
<td>No Number</td>
<td>1.307</td>
</tr>
<tr>
<td>No Number</td>
<td>No Code</td>
<td>0.764</td>
</tr>
<tr>
<td>No Number</td>
<td>Complete Number</td>
<td>0.810</td>
</tr>
<tr>
<td>No Number</td>
<td>No Number</td>
<td>0.575</td>
</tr>
</tbody>
</table>
SELECTING STATISTICAL APPROACH

• Neural Networks
  • Good discrimination
  • No reasons for selection

• Regression Analysis
  • Scores Each Factor
  • Interaction between variables

• Decision Trees
  • Demonstrate Splits
  • Unmanageable charts
CHANGE OF ADDRESS

Change of Address
- 1 60.6%
- 0 39.4%

No change of Address
- 1 45.8%
- 0 54.2%
Change of Address

Province Match
1  61.8%
0  38.2%

No Province Match
1  45.7%
0  54.3%
Change of Address

Province Match
- 1: 61.8%
- 0: 38.2%

No Province Match
- 1: 45.7%
- 0: 54.3%

Serviced Client
- 1: 70.2%
- 0: 29.8%

Not Serviced Client
- 1: 50.5%
- 0: 49.5%

Not Serviced Client
- 1: 55.0%
- 0: 45.0%

Serviced Client
- 1: 38.2%
- 0: 61.8%
CHANGE OF ADDRESS PROCESS

• Old
  • Serviced Clients - Inform agent
  • Unassigned Clients - Give to new manager

• New
  • Review all servicing when client moves
PROPENSITY OF CLIENT \( \times \) EFFICIENCY OF SALESMAN = SALES

TIME TO CONTACT PRODUCT \( \rightarrow \) KNOWLEDGE \( \rightarrow \) ENTHUSIASM = IMPROVED SALES
VALIDATION

IN USAGE

SELF FULFILLING PROPHECY ?

CONTROL GROUPS FROM UNIVERSE

ALREADY BIASED ?

BACKSCORING PAST CAMPAIGNS

USEFUL DIRECTIONAL VALIDATIONS
## BACKSCORING

<table>
<thead>
<tr>
<th>Predict</th>
<th>0-25%</th>
<th>26-50%</th>
<th>51-75%</th>
<th>76-100%</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch</td>
<td>%Clients with Change</td>
<td>%Clients with Change</td>
<td>%Clients with Change</td>
<td>%Clients with Change</td>
<td>%Clients with Change</td>
</tr>
<tr>
<td>Umtata</td>
<td>13.81%</td>
<td>13.10%</td>
<td>14.08%</td>
<td>50.00%</td>
<td>13.21%</td>
</tr>
<tr>
<td>East London</td>
<td>4.90%</td>
<td>11.08%</td>
<td>10.19%</td>
<td>30.00%</td>
<td>10.42%</td>
</tr>
<tr>
<td>Queenstown</td>
<td>4.37%</td>
<td>10.19%</td>
<td>10.20%</td>
<td>36.36%</td>
<td>9.58%</td>
</tr>
<tr>
<td>Port Elizabeth</td>
<td>3.57%</td>
<td>6.61%</td>
<td>8.66%</td>
<td>37.50%</td>
<td>6.63%</td>
</tr>
<tr>
<td>George</td>
<td>3.36%</td>
<td>6.24%</td>
<td>8.17%</td>
<td>25.00%</td>
<td>6.17%</td>
</tr>
<tr>
<td>Summary</td>
<td>5.25%</td>
<td>9.26%</td>
<td>9.47%</td>
<td>33.90%</td>
<td>8.91%</td>
</tr>
</tbody>
</table>
## DIRECT MAIL CAMPAIGN

<table>
<thead>
<tr>
<th>Conversion Rate</th>
<th>Backscored Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>4.2%</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>4.1%</td>
</tr>
</tbody>
</table>
KEY STEPS

• Enterprise Miner and SEMMA provide toolkit

• Knowledge and experience make them work

• Results must be used to add value
EFFECT OF DISCRIMINATING VARIABLE

- Improved Average
- Acceptable Limit
- Reduced balance
- New inclusions
- New exclusions
CHANGING PROCEDURES

Gain through follow up
EFFECT OF ADDITIONAL TELEPHONE NUMBERS
EFFECT OF SALES SUPPORT

Potential

- Selected Prospects
  - Basic Potential
  - Effect of better support material
  - Energised salesmen
Data Mining

Improved Models

Improved Actions

Improved Success

DATAMINING FOR PROFIT
Improving data mining results through data selection and preparation

Nigel Wigram and Hein Lochner

OLD MUTUAL – CAPE TOWN

PREAMBLE
This is a joint presentation by a marketer and a statistician to show how datamining and analysis can give really useful information without needing to be absolutely accurate.

We represent a large South African Life Assurance group and describe the process by which we are developing scoring models to predict and enhance the probability that a client will repurchase in a given time-frame. Although we have been working on the models for several years, it is an ongoing process both to refine the models and to take into account the operational changes introduced as a result of understanding the implications of the models.

This is an overview of the approach rather than a technical discussion of statistical analysis.

The approach to the datamining project was to look at a sample from our client base; to look at the descriptive dimensions as they existed at the start of the measurement period; and to see the extent to which the dimensions act as predictors for sales that occur within the measurement period.

We also look at changes in these dimensions because we hypothesised that these would be even stronger predictors than the actual dimensions themselves.

TOOLSET
Datamining is dependant on having the right toolset and access to relevant data. In this case the data is being drawn from the company’s data warehouse using base SAS® and passed through to SAS Enterprise Miner™ operating on a NT Server. Our data warehouse is still in the process of development and does not yet include all the data which could be relevant, and much of it is not in the perfect condition that theoreticians desire. It is critical to start getting results from the intelligent use of available data rather than waiting for Utopia which may never arrive.

THE OBJECTIVE AND THE TARGET VARIABLE
The starting point of any model must be to agree on its objective – in this case to predict and enhance the probability that a client will purchase from us again in a given time-frame so that we can improve our operational efficiencies. Datamining scores are not ends in themselves.

This leads to the definition of the target variable to be predicted. All too often it isn’t that obvious, and it may not be a variable which is directly available from the data held in the database or datawarehouse. In this case the definition of purchases is
not limited to starting a new contract but also includes making a voluntary injection into an ongoing program.

It should also exclude the cases where the initial sale was subsequently cancelled or the product discontinued. If the target had been all “purchases" the model might have accentuated those regular “triers" or coupon clippers who utilize the 21 days cancellation period as a mechanism to get free cover! The “Triers" represent a huge cost to us, not profit.

**DEFINING THE UNIVERSE – PEOPLE WHO HAD THE OPPORTUNITY TO BUY**

Next we must look at people whose behaviour we want to model the universe for the study. Again it is tempting to say that everyone in the database should be considered. We must however try to ensure that people had the appropriate opportunities to repurchase.

Ultimately we want to predict the potential across our whole client base, but in the life assurance industry there is often a big difference between purchasing potential and actual purchasing. Clients often do not take the initiative in making purchases and are dependent on the extent to which they have actually been contacted, and the efficiency of that contact.

The factor for the efficiency of the contact will differ depending on the nature of the campaign which you are utilizing. If you are looking at a direct mail campaign, it refers to the actual execution, whilst if you are giving leads to salesmen it will refer to the skill and enthusiasm with which the salesman follows up.

That will make the mining operation very difficult when you have thousands of salesmen each displaying different selling patterns, and not all your clients are actually assigned to an agent.

**FOCUSBING THE MODEL**

In our initial studies we reduced the impact of the salesmen by creating classes of clients in terms of the sales agent who is on record as being responsible for the client’s sales and service.

Some clients are not assigned to any agent. This fact in itself suggests that they will not have been canvassed recently, and as a result will show a very low repurchase rate. Although a key objective of the model will be to estimate these people’s potential the probable lack of sales activity that they have experienced will make them inappropriate for the initial modelling of what might have happened had they been approached.

Similarly clients who have been introduced by or assigned to salesmen who do not make a practice of repeat selling are probably not demonstrating their true potential and are not the ideal candidates to study.
Finally we have those clients who are serviced by salesmen who have a good record of repeat sales. They are likely to service their clients regularly and thus make repeat sales whenever the client’s propensity is at its highest.

The Futrix® analysis system was used to create an array of the salesmen by repeat purchase rate and draw the sample from the client bases of the top quartile of agents. This creates a universe of people who are likely to have been considered for repeat sales and who are also showing the highest repurchase rates.

Even these select clients only buy relatively infrequently. Datamining works most effectively when the probability of purchasing is more or less equal to the probability of not purchasing. Working with a sample where 90% do not repurchase the model will identify non-repeat purchasers much more accurately than purchasers.

We therefor use the technique of over-sampling to take all the cases where a repeat purchase was identified and to combine them with a random sample of those who have not bought to generate a combined pool containing equal numbers of buyers and non buyers.

PRODUCT SPECIFIC ANALYSES
When studying the purchasing of a specific product type as opposed to the general repurchase rate we use the same approach but define the target variable more specifically and draw the modelling samples from the client bases of those intermediaries who actually sell the products. Even our most active sales intermediaries do not necessarily sell the full product range.

SELECTING THE DIMENSIONS AND CREATING VARIABLES
The textbooks suggest that you should use all possible dimensions in your initial model but there is a great deal of data preparation to be done. This may involve creating new variables as interpretations of available data.

The most obvious ones initially are time variables and the treatment of missing variables. This is where the interaction between the analyst and an experienced practitioner can be important.

The warehouse carries the dates at which various events take place. For mining purposes we can hypothesise that time can have a variety of effects.

Firstly we can predict that there will be some time gap between purchases, but that certain product purchases will be seasonal, for example to co-incide with the end of the tax year; whilst others will be rhythmic based on the previous policy anniversary or the client’s birthday. Some purchases will also be linked to a specific event such as a child starting school which may appear seasonal to the start of the school year.

It is thus necessary to set up time models both by product class and by client, and we must model both absolute dates and relational dates.
MISSING VALUES
Many analysis tools discard cases which have missing values or try to insert some form of average or deduced value. This is extremely dangerous and it is often preferable to insert a dummy variable. The fact that the data is missing may be a critical factor in the weighting. You also need to know why the data is missing.

An example of a potentially misleading correlation was the finding that people with identified occupations tended to buy products with a high life assurance content whereas those with no occupation made investments. The reality is that the client’s occupation is important to life assurance underwriters and is therefore a compulsory field for the purchase of those products, but is often not required or asked for when purchasing investments.

When we look at the importance of socio-demographic clusters we found that we could only cluster about 75% of our prospects. There are three reasons for the missing values:

a) Clients who had no viable address
b) Clients who had a Post Office box address which could not be directly linked to a residential area, and
c) Clients who lived in rural areas not covered by our clustering system.

We put in 3 different dummy variables and found that they had different weightings in the models.

Interestingly these are also variables which had different weightings when modelling for direct mail purposes rather than for salesmen’s activities. Perhaps logically salesmen are less likely to follow up on a case which they know the physical address and it is easier to look up the current telephone number. This is not a problem when using direct mail where postal deliveries to a Post Office box is faster and more reliable than a street delivery.

Socio-demographic indicators were not the only geographic indicators which we used. The file included a province code, and it was hypothesised that the economic conditions would lead to different areas showing different repeat purchase rates. This turned out to be a comparatively weak indicator.

We then developed a “geographic match indicator” to show whether the client lived in the same region as the intermediary. This had little effect for some intermediaries but lower repurchase rates for others. We will come back to that later.

Client contactability obviously will play a considerable role. If you can’t contact the client you can’t sell to him but the presence or absence of a telephone number on our records didn’t initially achieve as significant a rating as we expected.
When we investigated by reaching through to some of the records we found that the telephone marker field was not being interpreted correctly. The marker was set whenever there was anything in the field whether or not it was a valid number. A row of noughts, hieroglyphics or even the words “not available” appeared in the field and were scored as though there was a telephone number. When we broke this down we created three categories: no reasonable number; a number but no dialing code; a number and dialing code. The model then showed that the better the number the better the success rates.

This served a dual purpose since it allows us to start to put a value on the elements of a telephone number and thus justify the expense of enhancing information and changing administrative procedures.

The true value of an accurate telephone number will be higher than that shown by the telephone model. Some of the numbers currently on our database will not be valid. Where our call centres have shown that they have been able to contact the client at the given number subsequent success rates rise dramatically even where the initial contact does not provide a positive result!

SELECTING THE STATISTICAL APPROACH
Alongside setting up the dimensions you must decide which statistical approach to use. Through the SAS Enterprise Miner™ suite we can use Neural Networks; Regression Analyses and/or Decision Trees. Each has its strengths and weaknesses, but we will demonstrate that by using them in combination we can improve both our predictions and the usefulness of the result.

NEURAL NETWORKS
Using a neural network model we get a series of scores which discriminate relatively well in regard to purchasing (or sales) habits. This is extremely useful if you have a very large base and your objective is to select the best prospects for attention. It is also useful in direct mail where you want to establish a mailing list of those most likely to respond and to exclude those unlikely to buy.

Unfortunately the “black box” nature of neural networks mean that you may not know why the client is selected. This has two drawbacks.

a) The sales channel does not necessarily know which appeals will lead to the highest probability of a sale, and
b) You don’t know which factors to change in order to improve probabilities.

REGRESSION ANALYSIS
Regression Analysis looks at each dimension in the model and scores its discriminating effect but this doesn’t fully take into account the interaction between dimensions. Some dimensions can be positive in one set of circumstances but negative in others. Where you believe that this is likely to happen it is often appropriate to create a new combined dimension.
DECISION TREES
Decision trees theoretically overcome this problem by demonstrating all the splits which occur in your data. This works well to identify the relationships between the major drivers but it rapidly becomes unmanageable when you are looking at a large number of dimensions. The technique is however very useful when you want to explore the relationship between a few selected dimensions, particularly those elements that you would instinctively expect to have an effect but which do not initially come through in the models.

CHANGE OF ADDRESS
One such example is the role of a change of address. This was particularly interesting to us since our existing administrative practices were proven to be counterproductive.

When we first looked at change of address it played a relatively small role in the probability of repurchase, which was somewhat surprising. When we looked at it in conjunction with the geographic match indicator we found that cases where the client was moving within or into the agents' home area gave a significantly higher propensity whilst cases where the client moved out of the area gave us a somewhat lower propensity.

CHANGING THE SAMPLING UNIVERSE
This caused us to review the role of the servicing intermediary. The initial sample which we had used was defined only to include clients who were being serviced by selected intermediaries who had a record of regularly contacting those clients whom they could.

To explore this further we drew another sample containing clients of weaker agent and unassigned clients.

Here we found the surprising result that our unassigned clients showed opposite trends to the best practice clients. Unassigned clients who moved area were much more likely to repeat purchase, whilst those staying in the same area showed little probability.

In terms of our existing administrative rules where a serviced client changes his address we immediately inform the agent to update his records and to take appropriate action. The agent would follow up on his local clients but would be much less likely to follow up on a client who had moved hundreds or thousands of miles away.

If however the client has unassigned and moved into a new area the local manager would issue the “lead” to one of his agents to go and welcome the client and this resulted in a good sales success rate.
When the information dealt with a change within an area there was a suspicion that the changes were often “amendments” rather than genuine changes of address and they were thus not treated as priority leads and this became a self fulfilling prophecy.

Simply by changing administrative rules and looking at reassigning serviced clients when the client moves from area to area, and upgrading the priority status of local changes of address could immediately lift success rates.

In fact this administrative change is an over-simplification since we introduced other checks to protect those intermediaries who do service clients in different areas, or who have a very strong personal tie with the client (a family member perhaps).

**EQUIPPING THE INTERMEDIARY**

This brings us back to the original hypothesis that the sales results will depend on the repurchase propensity of the client and the efficiency of the follow-up mechanism. Our model so far has focused on determining client propensities, but we can also use the output from datamining to improve the efficiency of follow up.

Datamining helps us to identify the best time to approach a prospect and the product he is most likely to buy, and this can help us to ensure that the salesman is properly equipped and hits the right hot button when he first approaches the client.

A very good intermediary does a complete review of his client’s circumstances and gets his sales successes by uncovering real needs. A new intermediary won’t necessarily get the client’s initial co-operation to do the full analysis unless he can get the client’s attention through initial relevant comments. Similarly a mass marketing approach is only able to press one button at a time.

Now we can help to identify the buttons.

The enthusiasm with which a salesman follows up on a lead is largely a factor of his morale and the perceptions of value of the lead. If a salesman receives a lead which he perceives to have a high success factor he is much more likely to follow it up enthusiastically and thus get a sale. Marking high probability leads as such will thus become a self fulfilling prophecy but this should not be used to validate the scoring model.

**VALIDATING THE MODELS**

Traditional datamining checks its consistency against control groups taken from the original universe but since we have already introduced some biases into that universe this does not demonstrate the total applicability of the model.

In order to validate our models we have been able to take target files from past campaigns. Both those where leads have been generated for agents to follow up and also those which have been used for direct mail marketing. This means that we
are dealing with entirely different datasets and certainly not limiting ourselves to the best practice exponents.

In each case we have been able to demonstrate that our models discriminate in the correct directions, but that the quantum is lower than that predicted by the models based on high repeat sellers. Perhaps this backscoring provides a double validation since the success gaps point to poor execution by certain intermediaries.

When we introduce the factors for groups of intermediaries as we discussed earlier we get much closer particularly in predicting non-repeat purchasers.

A simple analysis of success factors by branch level helps us to identify places where the manager is also not pulling his weight.

**EXTRA BENEFITS OF BACKSCORING**

An interesting example of a benefit from backscoring occurred when we were looking at the result from a direct mail campaign.

The campaign had been executed in two languages, and the standard analysis showed that it appeared to have performed more or less equally well in both languages.

When we applied our scoring models we found that most segments performed as predicted, but that there was a relatively difference between the two language groups.

Deeper analysis showed that we had something of a problem with the Afrikaans execution as it had performed less well in each segment. This had been hidden in our standard indicator because of the different socio-economic make up of the language groups.

**SUMMARY**

The key steps which we have taken in our datamining approach fit directly into SAS’s® Semma methodology but we feel that it is the understanding of the particular industry and company practices which are critical. These flow through to understanding the true import of the data. The findings of datamining only become really valuable when you can put them to use.

The neural network score allows us to see which clients are better prospects than others without changing the overall average. If we only have the resources to follow up on a proportion, concentrating on the better cases will give higher successes.

The iterative use of datamining has allowed us to predict probably repeat purchasers and “non-repeaters”. Applying this to direct mail list selection significantly improves the cost effectiveness of direct mail campaigns. Initially we obtain cost savings by reducing the volume of non-responsive prospects, but in the longer term by
identifying viable pocket of prospects from otherwise non-viable segments we can increase our total sales volumes within the same expense parameters.

No one datamining technique on its own meets all our needs but by using a combination of techniques we have been able to refine our scores and also identify some of the driving factors. This information is more important than the absolute accuracy of the model.

The use of decision trees in regard to change of address led us to change our administrative procedures, routing the opportunities to people in a better position to use them and identifying other high probability opportunities which we had previous discarded.

Regression analysis shows us which items are the most important to help focus our data gathering and clean up activities. By establishing the value of current telephone numbers we have been able to justify additional effort to collect and maintain these items correctly. This enhances the utility of the database and creates more prime prospects.

By improving our knowledge of the specific opportunities we have been able to provide better support material for the guidance of intermediaries and thus further enhance the probability of a sale.

Through publicizing the use of the scoring models we have been able to energize our intermediaries to give priority to the better opportunities thus creating self-fulfilling prophecies.

All these operational changes also impact on our models and we must now refine them further and thus set off more rounds of virtuous circles.

Datamining is for profit; not for academic interest.