Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

Using data mining to estimate customer's buying potential

G. Dragosavac
Senior Consultant, Data Mining and Analytical Intelligence, SAS Institute, South Africa

Abstract

Maintaining the relationship with a customer is a costly exercise, and not all the customers provide the same value to an organization. There are some customers who buy little, the value of their purchase is low and this is unlikely to change, regardless of the type of stimuli. The organization loses money by investing in this type of relationship. Then, there are other customers with whom a stronger relationship and better customer interaction would result in a more profitable relationship and higher value to the organization. These two groups of customers differ in their buying potential. This paper reports on the analysis carried out in an insurance company, with the goal to estimate and realize the purchasing potential of about a million and a half customers who have rarely been contacted in the past.

1 Introduction

Due to the prohibitive cost of the acquisition programs, and the fact that it is cheaper to sell to an existing customer than to a newly acquired customer, this South African insurer has decided to re-acquire a segment of their broad customer base that was rarely contacted in the last 5 years. The total population numbered 1 million clients that had not recently been contacted. The intention was not only to contact the customers who are likely to respond to marketing and sales stimuli, but within that group to contact those whose potential was never fully realized by their previous interactions with the company. Finding *high-potential* clients would be beneficial in itself, but the company wanted to know what to do in order to unlock their hidden potential. Firstly, the underlying question was: what product to offer and to which high-potential client? This

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

would form the basis for the effective cross-selling system. The secondary goal was to describe and profile characteristics of the responsive and active customers. To meet the set deliverables, the project team had at its disposal the company's data, housed in an enterprise data warehouse with SAS *Enterprise Miner* software, and a wealth of skills and experience in analytical, programming, business analysis and project management tasks. The estimated duration of the project was three months, and the key objective was to provide the right products to the right customers – those identified as the ones they wanted to do business with. That would in itself minimize the costs of re-acquisition, and maximize its returns.

2 Methods

2.1 Conceptual overview

Data for this project was extracted from the company's data warehouse and contained mostly demographic information. Client history and product information were excluded because of the sparseness of this information within the population on which this model is to be deployed. Also, the client had the intention of testing this model on the new clients, which wouldn't have historical and product information. There were three distinct groups of clients in regard to the type of product and structure of the premium: *recurring premium* clients, *single premium* clients and *group schemes* clients. For each group a separate *client potential* model was to be built and all results in this paper were related only to the biggest group of *recurring premium* clients. There were in total 182.679 records on which the model was to be built, validated and tested. The only difference between this *(training)* dataset and rest was the population on which the model is to be deployed (*deployment data set*) is that in training dataset we were able to establish the number of times clients have responded to sales and marketing initiatives. This variable we have used to construct *target* (*response*) variable for the first model.

On the basis of this information, models were then assessed and evaluated on the holdout subset of the original data set, and if the results were satisfactory, we would score all *recurring premium* clients within the total population. It was decided to build two models where the output of one would be the input of the other. The purpose of the first model (*"Activity Model"*) was to differentiate between *"active"* and *"passive"* clients, so the input data set consisted of known *"active"* and *"passive"* clients. We defined the *"active"* client as someone whose number of purchases was above a given cut-off, in an observed period of time. Such a client was given a flag "1", others were given a flag "0", and this *"activity flag"* was the target variable in the first model. In the second model task was to predict the total amount of the annual premium for *"active"* clients that were predicted by the first model. Thus, after deployment of the second model on predicted *"active"* clients, we would have two key variables needed to calculate the potential; one that was already present - most recent *"Total Premium Amount"*, and the variable that is based on the output of the second model – *"Predicted Total Premium Amount"*. The *predicted* premium amount would be a

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

reflection of the premiums of the similar clients based on their demographic and technographic characteristics. The next step would be to subtract a known premium from the predicted premium, and the result would be a quantitative approximation of the client potential. After calculating the potential for these *predicted-to-be-active* clients, we needed to profile the clients with the highest potential. This was to be done by dividing the whole range of variable "*potential*" in quartiles and then profiling the top quartile.

Following the calculation of potential would be to develop a *Cross-selling Model* for *high-potential* clients. Here, we would use a different dataset consisting of *known-active-and-high-potential* clients, as an input dataset. Input variables would be similar to the ones used in the calculation of client' potential. Target variables would be multiple (binary) variables, indicating the presence or absence of the particular product, within an insurance portfolio. This model would be deployed on the *high potential* clients, giving the probability of the product' purchase as output. By concentrating on the products with the highest probability of purchase, and excluding ones that are owned already, the insurer would be on the right track of realizing predicted high potential.

2.2 Methodologies deployed

2.2.1 Data Mining Project methodology

The Project methodology used was based on the *SAS Data Mining Projects Methodology* approach. This methodology has been developed by SAS Institute to focus data mining strategies in order to receive a high return on investment through rapid results. The *SAS Data Mining Projects Methodology* consist of six key activities, including:

- *Define the business problem*
- *Evaluate environment*
- *Make data available*
- *Analyse the data by data mining in cycles (SEMMA)*
- *Implementation in production*
- *Review*

2.2.2 The *SEMMA* methodology

The *SEMMA* covers the multi-step, iterative process of the data mining. The acronym *SEMMA* stands for five phases of data mining process, which are:

- **Select/Sample** data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.
- **Explore** data to discover unexpected trends and anomalies to gain understanding and ideas.
- **Modify** data by creating, selecting, and transforming the variables to focus the model selection process.
- **Model** data by, for example, allowing the software to search automatically or interactively for a combination of data that reliably predicts a desired outcome.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

- **Assess** data by evaluating the usefulness and reliability of the findings from the data mining process and thus selecting the most appropriate model.

2.2.2.1 Select/sample The amount of records in this dataset did not require sampling. However, we needed to partition this dataset, and we have used methods of stratified partitioning to split the data set into *training* (60%), *validation* (30%) and *test* (10%) subsets.

2.2.2.2 Explore *Enterprise Miner* offers various advanced visualisation techniques, which enables the user to evaluate issues such as structure of data, multivariate relationships, and the applicability of variables. Using *R-square* variable selection criterion within variable selection following three steps are performed (the last step is not applied if the target variable is non-binary):

- *Squared Correlation* is performed for each variable and ones that have a value less than the cut-off criterion are rejected (the default *Squared correlation* cut-off is set 0.00500)
- Remaining significant variables are evaluated using a *Forward Stepwise Regression*. Variables that have a stepwise R2 improvement less than the cut-off criterion (the default *Stepwise R2 Improvement* cut-off is set to 0.00050) are assigned the rejected role.
- *Logistic Regression* is performed using the predicted values that are output from the forward stepwise regression as the independent input.

2.2.2.3 Modify In this step we have filtered extreme values from the data using interactive filtering functionalities within the *Filter Outlier Node*. The Dataset had a fairly low level of missing values. Variables with the highest percentages of missing values had 5% of values missing. Within a broad range of replacement statistics in the *Replacement Node* we have used mostly *Decision Tree* imputation. Some of the data transformations used to improve the fit of the model were "*Log*", "*Square Root*", and "*Optimal Binning for Relationship to Target*".

2.2.2.4. Model Several modelling techniques were deployed and the ones that gave the best results was the *Ensemble* method in the *Client Activity* model. The *Ensemble Node* creates a new model by averaging the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple models. The new model is then used to score new data. Two models used as input in the ensemble model were *Neural Network* and *Decision Tree* methods. *Neural Network* architecture was *Multi-layer Perceptron* with one hidden layer and 10 neurons within the hidden layer. For this *Neural Network* we have used the *Lavenberg-Marquardt* optimisation technique. For *Decision Tree* we have used *Gini Reduction* as splitting criterion with *Average Profit/loss* as model assessment criterion.

For the second model (total annual premium) the chosen model was *Neural Network* with *Multi-layer Perceptron* architecture with *Quasi-Newton*

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

optimisation technique. The *Cross-selling* model was also *Neural Network* with similar architecture, while for profiling, largely used was the method of *Decision Tree* with *Gini Index* and *Entropy* as splitting criterions

2.2.2.5. Assess In this phase we have compared different modelling architectures and chose ones with the best results on the holdout data set. The common criteria for all modelling and predictive tools are the expected and actual profits for a project that uses the model results. The assessment node within *Enterprise Miner* offers a variety of assessment charts such as: *Lift Charts*, *Profit Charts*, *Return* on *Investment Charts*, *Diagnostic Classification Charts* and others, that we have used in model selection.

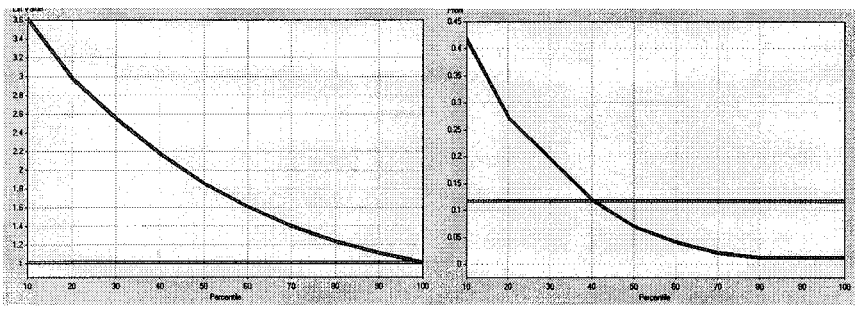


Figure 1: Cumulative lift chart and non-cumulative profit chart.

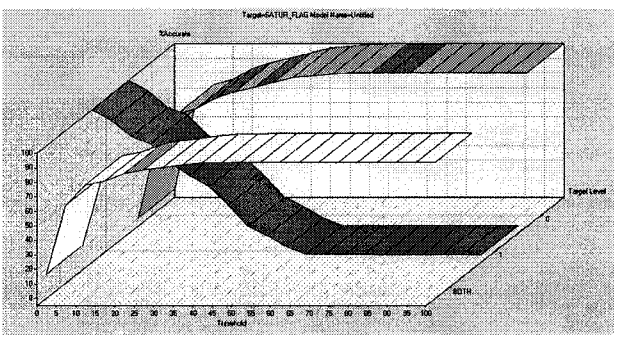


Figure 2: Correct classification chart.

3 Results

3.1 Modelling results

3.1.1. Results of the "Activity" model

This model was surprisingly successful in differentiating between "*Active*" and "*Passive*" customers. *The average Square Error (ASE)* for the holdout dataset was 0.29015 and the *Misclassification Rate* was 0.11522. The model was able to isolate 42% of *active* customers, among the top 10% of customers with the

highest probability of being an *active* client – on the holdout data. This was an improvement on the baseline probabilities of 3.6 times – as the *Cumulative Lift Chart* shows (on the left). The chart on the right is the *Non-Cumulative Profit Chart* which shows profit-generating deciles as well as deciles of negative return.

As the Correct Classification Chart shows, the prediction accuracy for the active customers ("1" = dark curve) is 90% for the threshold level of 5%. The most actives (90%) have a predicted probability of at least 5%. As the threshold level increases for actives, the prediction accuracy begins to tail off (you start to miss some of the *actives*). The reverse condition is true for *non-actives,* ("0" = light grey line). A threshold level of 30-45 could provide optimal prediction accuracy for both *target variable* levels.

In this project we have flagged any client that falls within the top 40% of probability of being an "*active*" client as an input in the following model.

3.1.2. Results of the "Total annual premium" model

In the second model the *Neural Network* achieved the best results, as it can be seen from the assessment table:

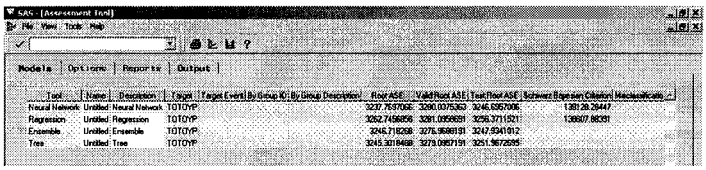


Figure 3: Assessment results for different models in Model Manager window

The value of *Average Squared Error* for the best model, as the test data set was fairly large, which was not surprising bearing in mind that most of the variables in input space were demographic variables. However, we decided to proceed further in calculating the potential and then in the profiling phase, we decided to do further evaluation of the results with help from a team of business analysts.

The two charts below are *Predicted Plot Charts* capturing the relation between predicted values of "*Total Premium Amount*" and "*Assumed Ethnic Group*", "*Marital Status*", "*Gender*" and "*Correspondence language*". Results confirmed the expectancies of the business analysts.

3.1.3 Results of the predictive model done on the calculated potential

As described earlier, potential was calculated by subtracting the known *Total Premium Amount* from the *Predicted Total Premium Amount.* The resulting variable was named *Client Potential,* and from that variable a new binary variable was constructed – *High Potential,* where indicator "1" was given to the highest 25% of the values and where indicator "0" was given to the remaining

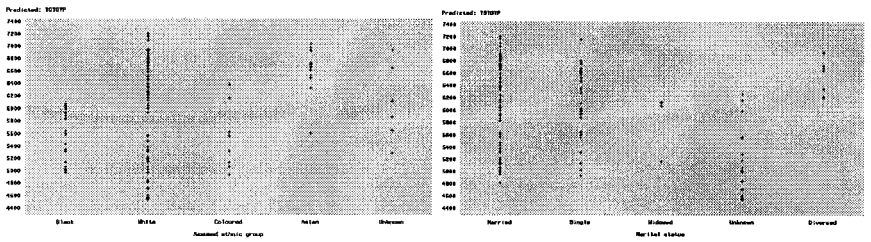


Figure 4: Predicted Plot Charts for "*Assumed Ethnic Group*" and "*Marital Status*" variables plotted against predicted premium amounts.

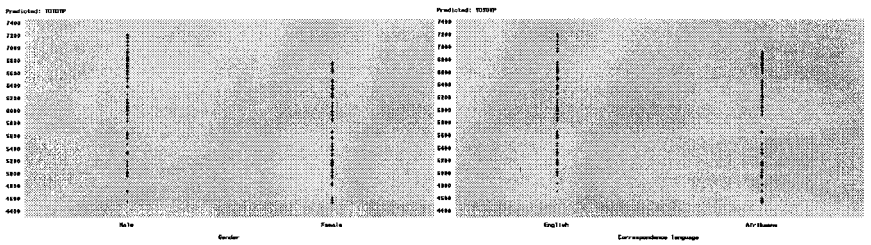


Figure 5: Predicted Plot Charts for "*Gender*" and "*Correspondence language*" plotted against predicted premium amounts.

values. This variable was then used as a target variable in subsequent predictive modelling exercises. The chosen model was *Neural Network*. By knowing *clients' gender, ethnicity, age* and *marital status*, the model was able to find 99.15% of the *high potential clients, (232 of 234 in total).* However, out of the total number (433) of predicted *hp clients, 43% (201)* did not belong to the class of high potential clients *(false positives).* By lowering the optimal cut-off threshold, we would find fewer of these *high potential* clients, but with a higher likelihood of them being an *hp* client. The relatively low cost of implementation of this project did allow us to go with large number of "false positives", as long as we find most of the high potential clients.

3.1.4 Results of the profiling of "high" potential clients

Profiling of *high potential* clients was an important deliverable of this project. If a company could know who are these clients are, and what their characteristics are, they would be able to steer their sales and marketing efforts toward this most lucrative market segment. Business analysts wanted to know to what extent these results confirmed what they thought the *high potential* client should look like. Results have shown no big surprises in terms of basic assumptions; however, they shed more light on their cut-offs and distribution boundaries. Some of the variables investigated were *gender, age, ethnicity, marital status* and *language*. It was no surprise to see that in their market, the value of "*male*" within the variable "*gender*" would be far more dominant for high potential clients, than in the population on average. The contribution of the "*female*" gender variable was

lower than expected. The variable *"language of correspondence"* has shown that the effects on a baseline probability have different languages. Even though the outcome was as expected, there were mixed interpretations on the nature of the interaction between the variables *"language of correspondence"* and the target variable *"high potential"*. The variable *"ethnicity"* provided surprises for some analysts, prompting us to do some testing which in the end proved the original results to be correct. The variable *"age"* brought no surprises, it was intuitively clear that *high potential* clients would be of an older age, and that was confirmed in data.

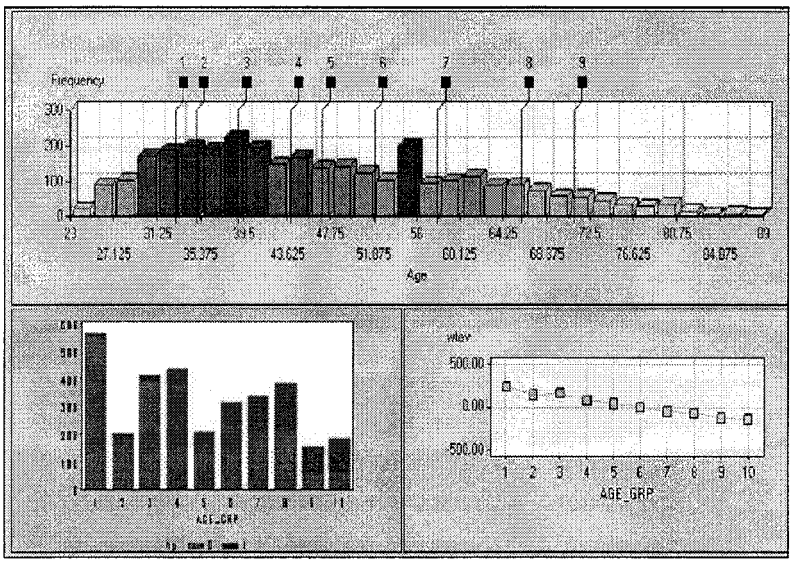


Figure 6: Output from interactive grouping node.

Automatic groupings done by the *Interactive Grouping Node* within *Enterprise Miner* shows the linear increase of *high potential* clients (dark grey) as the age increases.

Decision tree splits show what values of which variables increase or decrease the initial (*baseline*) probabilities of finding high potential clients. From the first split we can see that the probability of finding a *high potential* client increases from the baseline probability of 24.1% to 50.4% on validation data set – if the client is older than 51 years of age.

The final output of profiling analysis was a set of rules that can help the organization to recognize its high value segments. Also, business can use this information to steer its marketing and sales efforts toward better acquisition and retention strategies toward such segments.

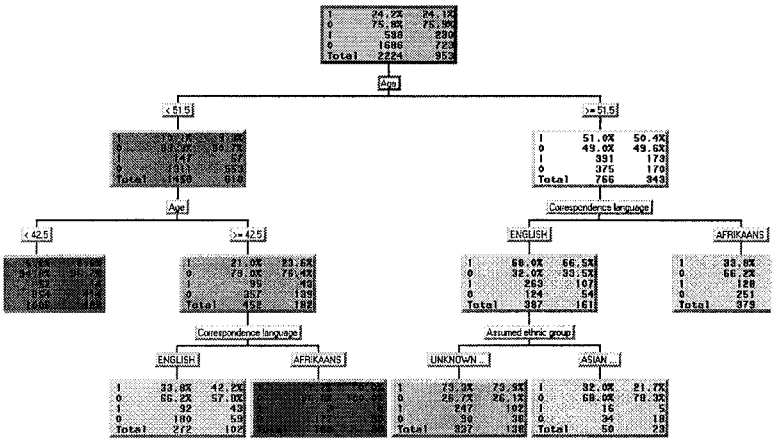
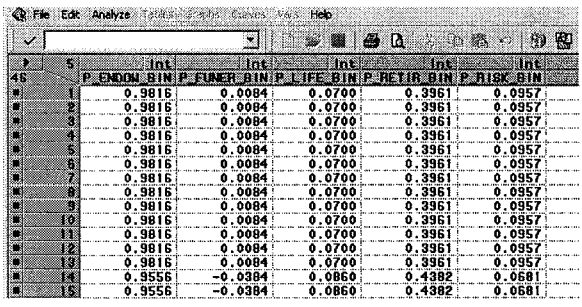


Figure 7: Decision tree splits.

3.1.5 Results of the cross-selling model

The dataset used for the construction of the *Activity*" and *Total Premium Model* was merged with relevant product information, and for each product binary variable was constructed. If a client has an endowment product, the value in the newly constructed binary variable would be "1", otherwise it would be "0". The same was done for all products, and these indicator variables became target variables in the C*ross-selling Model.* Input variables were mostly demographic and purchasing behaviour information. After the C*ross-selling Model* was built and evaluated, it was used to score calculated *high potential* clients on their propensity to purchase particular product.

Following is the section of the output table, containing the probability of purchase of a certain line of insurance products, such as probability of purchase of endowment products (P_ENDOW_BIN), funeral products (P_FUNER_BIN), retirement products (P_RETIR_BIN) and risk products (P_RISK_BIN). The first record in the table shows that the customer is the most likely to purchase an



	S	Int	Int	Int	Int	Int
46		P_ENDOW_BIN	P_FUNER_BIN	P_LIFE_BIN	P_RETIR_BIN	P_RISK_BIN
■	1	0.9816	0.0084	0.0700	0.3961	0.0957
■	2	0.9816	0.0084	0.0700	0.3961	0.0957
■	3	0.9816	0.0084	0.0700	0.3961	0.0957
■	4	0.9816	0.0084	0.0700	0.3961	0.0957
■	5	0.9816	0.0084	0.0700	0.3961	0.0957
■	6	0.9816	0.0084	0.0700	0.3961	0.0957
■	7	0.9816	0.0084	0.0700	0.3961	0.0957
■	8	0.9816	0.0084	0.0700	0.3961	0.0957
■	9	0.9816	0.0084	0.0700	0.3961	0.0957
■	10	0.9816	0.0084	0.0700	0.3961	0.0957
■	11	0.9816	0.0084	0.0700	0.3961	0.0957
■	12	0.9816	0.0084	0.0700	0.3961	0.0957
■	13	0.9816	0.0084	0.0700	0.3961	0.0957
■	14	0.9556	-0.0384	0.0860	0.4382	0.0681
■	15	0.9556	-0.0384	0.0860	0.4382	0.0681

Figure 8: Section of the scoring table.

Transactions on Information and Communications Technologies vol 29, © 2003 WIT Press, www.witpress.com, ISSN 1743-3517

endowment product, and if he has this product already, the company would send him a marketing message concerning retiring products, as this is second most likely product that he would purchase.

3.1.6 Implementation results

So far, the implementation results have shown steady and sometimes dramatic improvement, in response rates of the direct marketing initiatives toward high potential clients. The last two direct marketing campaigns had 3.5% increase in the responses compared to their last "random-mailing" campaign. These results could be improved further if customers are contacted using the right contact channel and additional work to be done is related to the building of a *Channel Allocation Model.* The sales force had increases of 6% in the amount of closed sales, based on the results in last sales quartile. As implementation is on going, it is too early to calculate the total financial impact that this project had on the organization, however preliminary assessments have shown that profits are already being generated.

4 Conclusion

This project has shown what data mining can do when it is supported by proven project and process methodologies. It has shown the value of intuitive and creative thinking and focused on the business issues rather than on technology even when the data used is based on the most basic demographic variables. In conclusion, we think that this application can help an organization to identify their *high-potential client* segments, and to unlock their hidden potential.