Implementation study: Using decision tree induction to discover profitable locations to sell pet insurance for a startup company

Received (in revised form): 23rd July, 2007

Rosskyn D'Souza
received his bachelor's degree in Computer Science (2003) and master's degree in computer applications (2006) from the University of Mumbai. Currently, he is an MS candidate in Computer and Information Science (Engineering) at the University of Pennsylvania. He has worked as a Teaching Assistant for 'Project Management' at the University of Pennsylvania (Fall 2006). His research interest focuses on Artificial intelligence, Machine learning, Data mining and analysis.

Michal Krasnodebski
received his bachelors degree in Systems Engineering from the University of Pennsylvania's Engineering School and his bachelors of Science in Economics in 2006 from the Wharton School at the University of Pennsylvania. His concentrations at Wharton were in Operations and Information Management and Finance. He joined The Boston Consulting Group in New York in 2007. His research interests include Machine learning, Data mining and Entrepreneurial finance.

Alan Abrahams
is an assistant professor at Virginia Polytechnic Institute and State University. He has previously taught in the Department of Operations and Information Management at The Wharton School, University of Pennsylvania, and in the Department of Informatics at the University of Pretoria. He holds a PhD in Computer Science from the University of Cambridge, and a Bachelor of Business Science with honours in information systems from the University of Cape Town. His primary research area is entrepreneurial decision support, including identification and valuation of new technology applications, societal welfare initiatives (HIV/AIDS clinical decision support), automated contract management, and simulations for modelling knowledge growth and transfer.

Keywords *direct marketing, data mining, customer prospecting, database marketing*

Abstract We demonstrate the use of decision tree induction, employing both C4.5 and Profit Optimal (SBP) algorithms, to discover profitable locations for a young startup firm to sell their product, pet insurance. We use publicly available data including US Census data and veterinary surgery location data as our data sources and use the potential profits generated by each of the algorithms as key performance metrics. We show how our findings link to general business behaviour and performance, by describing the implications of our findings for marketing strategy at the pet insurance company.
Journal of Database Marketing & Customer Strategy Management (2007) **14,** 281–288.
doi:10.1057/palgrave.dbm.3250059

Alan Abrahams
Virginia Polytechnic Institute
and State University
1007 Pamplin Hall
Virginia Tech
Blacksburg, VA 24061-0235
USA
Tel: +1 540 231 5887
e-mail: abra@vt.edu

INTRODUCTION AND RELATED WORK

Businesses have been searching for insight from the massive amounts of data that they generate for some time. Initially, this search led to the development of standard online analytical processing (OLAP) tools. Standard OLAP tools, while excellent at performing their reporting function, are not capable of generating the kinds of insights that businesses today require. This need has led to substantial research into data mining; a limited, though expanding, amount of this work has been done with a focus on utilising large business databases for marketing-related efforts.[1]

Data mining tasks can be separated into categories, depending on the type of knowledge generated; Shaw[1] argues that these tasks are Dependency Analysis, Class Identification, Concept Description, Deviation Detection and Data Visualisation. We will focus on discussing research that falls into the Class Identification category as this is the type of knowledge that we were seeking. Within Class Identification the focus of research has been the classification of clusters in business databases into predefined categories.[1] Once discovered/created, the descriptions of the classifications are then generally used to understand and classify new data in order to allow managers to optimally respond. In marketing, this generally means deciding whether a customer or set of customers (grouped by geographic, demographic or other means) is worth spending valuable resources on reaching,[2] with some methods going so far as to evaluate the potential costs and benefits of these customers.[3,4]

Researchers have been prolific in developing different methods for extracting knowledge from databases.[5–11] Numerous studies have illustrated how data mining techniques can be specifically used to identify prospective customers.[1,2,5,12-14] Decision tree algorithms, such as C4.5,[15] SBP[16,17] and others have also been used by many researchers to extract knowledge from databases that can be used by managers to make decisions.[18] In the case study in this paper, we make use of decision tree techniques for assisting a new pet insurance company to characterise its target market.

PetCoverCo (the name of our client has been anonymised) is the new and exclusive US franchise of a well-established global pet insurance provider, which is looking to enter the United States market. In this study, we help PetCoverCo target prospective customers via data analysis based on vets present in zip codes and other census/demographic data. Using decision trees on our data set, we will develop a marketing strategy for PetCoverCo by suggesting which areas and demographics are best to roll-out to. It always helps a marketing team to know their target customer types. Our analysis will provide rules that identify the demographic characteristics of customers that are highly probable to purchase pet insurance. This should help PetCoverCo's marketing team to minimise advertisement and other associated costs, and to maximise the cost effectiveness and impact of their campaigns. The presence of vets in zip codes is an important criterion, as without vets in, or near, the neighbourhood one would not be able to visit a vet and hence would have no need for pet insurance. Since many zip codes in the states have a single vet, we assumed that a zip code would only be profitable if it had two or more vets and our models are based on this assumption.

METHODOLOGY

Decision trees are excellent tools that help to choose between several courses of action. They describe a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications. This is a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options. They also help you form a balanced picture of the risks and rewards associated with each possible course of action. A decision tree can be learned by splitting the source data set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when either splitting is non-feasible or a singular classification can be applied to each element of the derived subset.

The first algorithm we used in this case study is SBP.[16,17] SBP is a profit-based algorithm and classifies data with the aim of earning more profit rather than to be more accurate. The second algorithm we use is

C4.5.[15] This algorithm is based on the ID3 algorithm.[15] It contains several improvements like choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs and handling continuous attributes.

In the first step, Information Gathering, we gathered all the relevant information that was of use in the data mining process. In our case we had Census information and vet location data. Next, we went through the process of Data Preparation to prepare the data in the required format so that it could be fed to the data mining software to produce the decision rules. Thirdly, we ran a data mining software package on a training set. We used Decision Tree Builder Pro (DPTR) to run both the SBP and C4.5 decision tree algorithms.[15–17] We could also have employed further algorithms to create more models.

In the next section, we show the results generated by application of the algorithms. Following that, we conduct a Results Analysis by creating Lift charts, Gain charts and chi-squared analyses. Finally, we go through the process of Knowledge Exploitation and show how the knowledge obtained through data mining can be used to inform PetCoverCo's behaviour and improve its performance.

RESULTS

Model 1

We ran the DTPR software on a training set of 2,000 records. The software used the profit-based SBP algorithm and was limited to three splits. The following rules were output:

Rule 1.1:
IF the 'total household size' >2,650, then predict YES.

Rule 1.2:
IF the 'total household size' ⩽2,650 and the percentage of people in the 'household income range of 150–200k' is >0.01 and the percentage of 'children aged 5 or less' is ⩽0.06, then predict YES.

Rule 1.3:
IF the 'total household size' ⩽2,650 and the percentage of people in the 'household income range of 150–200k' is ⩽0.01, then predict NO.

Rule 1.4:
IF the 'total household size' ⩽2,650 and the percentage of people in the 'household income range of 150–200k' is >0.01 and the percentage of 'children aged 5 or less' is >0.06, then predict NO.

We can make the following general inferences from the induction rules:

1 The household size (number of households in the zip code) is a prime factor that determines if PetCoverCo should consider marketing to and entering the zip code.
2 The percentage of households with an income between 150 and 200k helps to determine which locations are profitable to advertise to, given the household size.

To confirm the validity of the rules generated, we tested them on three unseen test data sets, each containing 2,000 records. The profits generated by applying the rules were \$22,029,580, \$22,066,047 and \$21,841,730 for each of the three sets of test data; the mean was \$21,979,119 and the standard deviation was \$120,371. If we assume a roughly normal distribution, 95 per cent of profits are within two standard deviations of the mean, so this standard deviation is low for profits of \$22m, as it means profits are unlikely to be more than \$240,742 below the estimate (provided the assumptions we used when creating the model are correct). This shows that the payoffs of this model are robust and that the model is good.

Model 2

We ran the DTPR software on a training set of 2,000 records again; this time the software used the C4.5 algorithm, limited to six splits. The following rules were output:

Rule 2.1:
IF the predicted household size for 2008 > 4,139, and IF the percentage of households with income between 20 and 25 k is >0.11, then predict NO.

Rule 2.2:
IF the predicted household size for 2008 > 4,139, and IF the percentage of households with income between 20 and 25 k is ⩽0.11 and IF the percentage of Caucasian in 2000 is >0.116, and IF the percentage of males aged 5 or less is >0.054, predict 'NO'.

Rule 2.3:
IF the predicted household size for 2008 > 4,139, and IF the percentage of households with income between 20 and 25 k is ⩽0.11 and IF the percentage of Caucasian in 2000 is >0.116, and IF the percentage of males aged 5 or less is ⩽0.054 and IF the percentage of households with income between 45 and 50 k is >0.076, predict 'NO'.

Rule 2.4:
IF the predicted household size for 2008 > 4,139, and IF the percentage of households with income between 20 and 25 k is ⩽0.11 and IF the percentage of Caucasian in 2000 is >0.116, and IF the percentage of males aged 5 or less is ⩽0.054 and IF the percentage of households with income between 45 and 50 k is ⩽0.076, and IF the percentage of people aged between 15 AND 20 >0.028, predict 'YES' else predict 'NO'.

Rule 2.5:
IF the predicted household size for 2008 >4,139, and IF the percentage of households with income between 20 and 25 k is ⩽0.11 and IF the percentage of Caucasian in 2000 is ⩽0.116, predict 'NO'.

We can make the following general observations from the induction rules presented above:

1 The household size predicted for 2008 is a prime factor that determines if PetCoverCo should consider marketing and entering a zip code.
2 The percentage of people aged 15–20, percentage of males in aged 5 or less helps us to determine which locations are profitable to advertise given the household size.
3 Rule 2.4 is an important rule as it predicts 'YES' with 55 per cent accuracy and most of the Yeses fall in this category. For the training set, over 95 per cent of the Yeses fell in this category.

Given the previously introduced test data sets, each of 2,000 unseen records, the profits generated by applying these rules are $21,843,333, $20,957,076 and $21,322,323 for each of the three sets of test data; the mean was $21,374,244 and the standard deviation was $445,404. If we assume a roughly normal distribution, 95 per cent of profits are within two standard deviations of the mean; hence, standard deviation is average for profits of $21.374m, as it means profits are unlikely to be more than $890,808 below the estimate (provided other assumptions used when creating the model are correct).

ANALYSIS

Lift chart

A lift chart is used to display the amount of 'lift' the model provides. '*Lift*' is a measure of the improvement in response that we get by selecting top prospects, as scored by the model, instead of just selecting randomly. 'Lift' measures the effectiveness of a predictive model.

The maximum lift is 2.61 and is achieved at the first decile, followed by 2.59 and 2.53 at the second and third decile (Figure 1). This tells us that the model is good: taking the top 30 per cent (three deciles) of

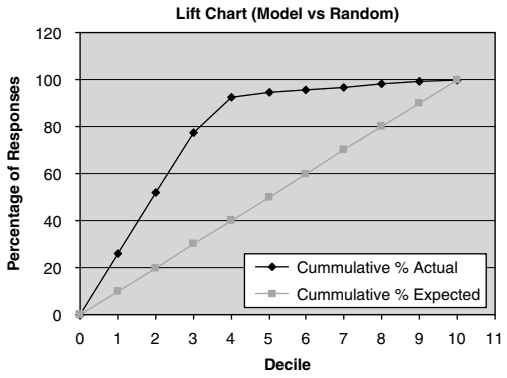

Figure 1: Lift chart for SBP

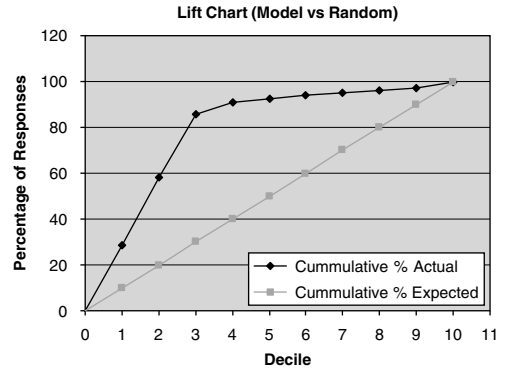

Figure 2: Lift chart for C4.5

prospects, we find 2.5 times as many buyers using the model than we would through random selection of the same number of prospects. The distance between the two curves (actual model vs expected at random) is large, showing the model is far better than a random selection.

For C4.5, choosing the top 30 per cent of prospects gives us 85.5 per cent of the respondents (Figure 2). The maximum lift is 2.95 and is achieved at the second decile. This tells us that the model is good as, for the top 20 per cent of prospects, it produces almost three times as many buyers as a random selection of the same number of prospects.

Gain chart

Gain Charts give us an idea of the profitability of a model, in terms of likely costs and revenues from using it. Furthermore, Gain Charts allow us to easily see what number or percentage of prospects we ought to mail in order to maximise profits. We can then set the selection threshold appropriately in order to choose that number of prospects. Gain charts are particularly helpful when the marketing budget is limited, and we cannot afford to canvas all possible prospects.

To evaluate a model, it is best to look at the profit we could earn by mailing the highest scored prospects. Gains Charts, which show the cumulative profit (or, alternately, the percentage Return on Investment), are useful for this. The charts are created from the consolidated test data.

For the SBP results the gain chart tells us that maximum marginal profit at a given decile is achieved in decile 3 ($19,111,056 for that decile), though decile 2 ($18,404,659 for that decile) and decile 1 ($18,302,439 for that decile) are not far behind (Figure 3).

We observe a monotonic increasing curve. If PetCoverCo is looking to maximise return on investment and is not bothered about market share, or if its marketing budget is limited, it should target the zip codes that fall in the first three deciles. If, on the other hand, PetCoverCo is looking to maximise market share and has unlimited marketing budget, the model is futile and they should go ahead and market their product to all zip codes.

The C4.5 model, on the other hand, provides 78.22 per cent ($61,421,237) of the possible profit by choosing the top 30 per cent of prospects. The gain chart tells us that maximum marginal profit is achieved in decile 2 ($20,872,857 for that decile), though decile 3 ($20,334,306) and decile 1 ($20,214,073) are not far behind (Figure 4).

Again, we observe a monotonic increasing curve. If PetCoverCo is looking to maximise their profits with minimal expense, it should target the zip codes that fall in the first three deciles. If, on the other

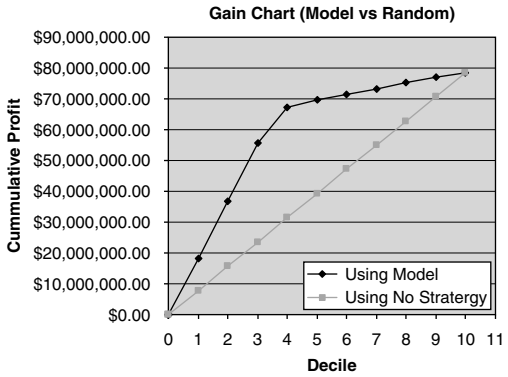

Figure 3: Gain chart for SBP model, compared to random selection

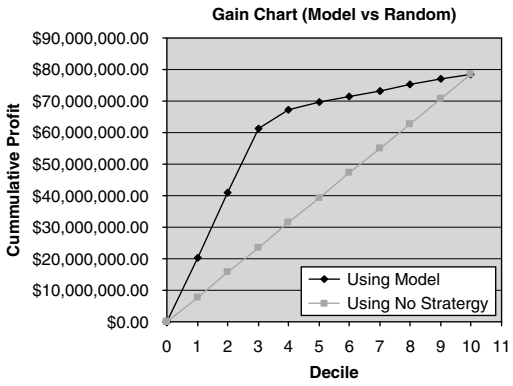

Figure 4: Gain chart for C4.5 model, compared to random selection

hand, PetCoverCo is looking for maximum market share and has unlimited marketing dollars, the model is futile and they should go ahead and market their product to all zip codes. With C4.5, choosing the top 30 per cent of prospects results in over $60,000,000 in profits, which is indicative of a good model.

CHI-SQUARED ANALYSIS

Chi-squared statistics show whether sub-populations are different in a statistically significant way. It is possible for sub-populations to vary just because a random sample of the population was taken, and random samples will of course vary randomly. Chi-squared statistics show the likelihood that the variance in the sub-population is more than just random. The chi-squared statistic can therefore be used to evaluate whether a predictive data mining model performs better than a random selection of individuals. Our analysis indicates that both of the models we created are statistically significant at 82.5 per cent.

KNOWLEDGE EXPLOITATION: LINK TO BUSINESS BEHAVIOUR AND PERFORMANCE

The intention of this data mining study was to determine the characteristics of individuals who are the most promising candidates for pet insurance, so that marketing campaigns could be targeted at individuals of this type. In this section we describe how the knowledge we have discovered can be exploited, and how it informs business behaviour and improves business performance.

Geographically, we found that, of 52 US territories (including the 50 states, the District of Columbia and Puerto Rico), 11 states accounted for more than half of the veterinary surgeries. These were: California, Texas, Florida, New York, Ohio, Michigan, Illinois, Pennsylvania, North Carolina, Georgia and Virginia. The top four alone accounted for more than one quarter of veterinary surgeries, and the top 25 states accounted for more than 80 per cent of veterinary surgeries. This finding proved useful, as PetCoverCo is required to obtain licensing to sell insurance in each state that it enters, and we were able to identify which states were likely to be the most lucrative markets for pet insurance.

Demographically, our findings provided PetCoverCo with useful and actionable rules for the selection or rejection of cost-effective media channels. PetCoverCo has a vast array of media channels available to it, including print media such as various newspapers and topical magazines, audio-visual media such as local or satellite television or radio channels, or online media such as different websites. In order to maximise bang-per-buck of their advertising dollars, target media need to be chosen

wisely. Obvious choices for target media include magazines and television shows targeted specifically at dog and cat lovers. More specifically, however, all chosen media should conform to the demographic criteria identified by our models. Media with alternate demographics should be rejected. For instance, one of our findings above (see *Rule 1.1*) was that PetCoverCo should only advertise in channels that target zip codes with more than 2,650 households. PetCoverCo was therefore able to define one of its prime target markets as urban dwellers. This has important implications for business behaviour: for instance, in order to maximise advertising performance, an urban home and garden magazine would be more appropriate than a rural farming publication. Similarly, if advertising in an airline travel magazine, PetCoverCo would need to choose magazines on board flights originating from and bound to large metropolitan destinations, rather than those on board 'pond-skipper' services to or from small towns. Figures from the Audit Bureau of Circulation provide a useful resource for PetCoverCo to check whether a particular publication meets or fails their target market criteria.

CONCLUSION

Based on the lift charts, choosing the top 30 per cent of prospects given by C4.5 gives us 86 per cent of the respondents. For SBP, however, choosing the top 30 per cent of prospects gives us 77 per cent of the respondents. We observe that C4.5 provides a better lift at each of the first three deciles (when compared to SBP) implying better precision, thus this model should be considered if one is looking for precision and higher recall (for top 30 per cent of prospects).

According to the gain charts, the C4.5 model provides 78 per cent ($61,421,237) of the possible profit by choosing the top 30 per cent of prospects. SBP on the other hand provides 71 per cent ($55,818,155) of the possible profit by choosing the top 30 per cent of prospects and requires choosing 40 per cent of prospects to achieve profits over $60,000,000. Clearly C4.5 provides better profits than SBP within the first three deciles. If PetCoverCo is looking to maximise their profits with minimal expense, then the findings from the C4.5 are the way to go.

The SBP Model and the C4.5 Model could be combined to improve prediction. There are several techniques available for combining models, such as Genetic Algorithms, Boosting, Stacking and Collaborative Learning. Genetic Algorithms work well for combining knowledge/rules. We could combine the rules from the models generated above to see if we can improve the results even further. Using rules from both models above, we could implement crossovers to come out with better rules. In general, a GA comprises the following steps:

1 Start with an initial population of rules.
2 Use a fitness function (function used to decide how good a rule is) to assign a fitness to each rule.
3 Use crossovers to combine rules that have higher fitness values.
4 Use mutation to introduce some form of randomness.
5 Keep trying the above steps, which should increase the population of rules till we obtain a very good set of rules.

It is clear from the analysis of results that, in the case of PetCoverCo, they should use the rules generated by the C4.5 and SBP algorithms to guide the launch of their new pet insurance product. The rules generated by the algorithms could, however, be further improved through the use of Genetic Algorithms, though we leave this as the topic of a further implementation study.

References

1 Shaw, M. J., Subramaniam, C., Tan, G. W. and Welge, M. E. (2001) 'Knowledge management and data mining for marketing', *Decision Support Systems*, Vol. 31, pp. 127–137.

2 Chou, P. B., Grossman, E., Gunopulos, D. and Kamesam, P. (2000) 'Identifying prospective customers', Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Boston, pp. 447–456.
3 Chen, Y. L., Chen, J. M. and Tung, C. W. (2006) 'A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales', *Decision Support Systems*, Vol. 42, pp. 1503–1520.
4 Piatetsky-Shapiro, G. and Masand, B. (1999) 'Estimating campaign benefits and modeling lift', Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, pp. 185–193.
5 Bhattacharyya, S. (2000) 'Evolutionary algorithms in data mining: Multi-objective performance modeling for direct marketing', Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Boston, pp. 465–473.
6 Chen, G., Liu, H., Yu, L., Wei, Q. and Zhang, X. (2006) 'A new approach to classification based on association rule mining', *Decision Support Systems*, Vol. 42, pp. 674–689.
7 Jukic, N. and Nestorov, S. (2006) 'Comprehensive data warehouse exploration with qualified association-rule mining', *Decision Support Systems*, Vol. 42, pp. 859–878.
8 Lawrence, R. D., Hong, S. J. and Cherrier, J. (2003) 'Passenger-based predictive modeling of airline no-show rates', Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining, Washington DC, pp. 397–406.
9 Rosset, S., Murad, U., Neumann, E., Idan, Y. and Pinkas, G. (1999) 'Discovery of fraud rules for telecommunications — Challenges and solutions', Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, San Diego, pp. 409–413.
10 Weiss, S. M., Buckley, S. J., Kapoor, S. and Damgaard, S. (2003) 'Knowledge-based data mining', Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining, Washington DC, pp. 456–461.
11 Wong, M. L. (2001) 'A flexible knowledge discovery system using genetic programming and logic grammars', *Decision Support Systems*, Vol. 31, pp. 405–428.
12 Apte, C., Bibelnieks, E., Natarajan, R., Pednault, E., Tipu, F., Campbell, D. and Nelson, B. (2001) 'Segmentation-based modeling for advanced targeted marketingProceedings of the 7th International Conference on Knowledge Discovery and Data Mining, San Francisco, pp. 408–413.
13 Kim, Y. S. and Street, W. N. (2004) 'An intelligent system for customer targeting: A data mining approach', *Decision Support Systems*, Vol. 37, pp. 215–228.
14 Verhoef, P. C., Spring, P. N., Hoekstra, J. C. and Leeflang, P. S. H. (2002) 'The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands', *Decision Support Systems*, Vol. 34, pp. 471–481.
15 Quinlan, R. J. (1993) 'C4.5 Programs for machine learning', Morgan Kaufmann Publishers, San Mateo.
16 Abrahams, A. S. and Becker, A. (2007) 'Partitioning for profit: An empirical study of methods for handling unequal costs of error in predictive data mining', *Group Decision and Negotiations Journal*, Special Issue on Formal Modeling in Electronic Commerce-Part I, Vol.16, No. 2, pp. 191–209.
17 Abrahams, A. S., Becker, A., Fleder, D. and MacMillan, I. C. (2005) 'Handling generalized cost functions in the partitioning optimization problem through sequential binary programming', Fifth IEEE International Conference on Data Mining (ICDM'05), Houston.
18 Yang, Q., Yin, J., Ling, C. and Pan, R. (2007) 'Extracting actionable knowledge from decision trees', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, pp. 43–56.