Contemporary Engineering Sciences, Vol. 4, 2011, no. 6, 259 - 268

Who are the Likeliest Customers; Direct Mail Optimization with Data Mining

Uroš Bole

Pikarp d.o.o.
Pod gričem 32, SI-5000 Nova Gorica, Slovenia
uros.bole@gmail.com

Gregor Papa

Jožef Stefan Institute
Jamova c. 39, SI-1000 Ljubljana, Slovenia

Abstract

The aim of this article is to present the work and the findings of an effort to improve the response rate of direct mail. The purpose was achieved applying data mining techniques to the available data. The resulting predictive model can, given a set of characteristics of a person, determine probabilistically, whether the person is likely to respond positively to a direct marketing campaign with a purchase of a product or service. The work is based on the data of a Dutch insurance company which was published for the purpose of the CoIL Challenge 2000. The contest had the same aspirations as this article. Its results are discussed briefly in the last chapter.

Keywords: direct mail, data mining

1 Introduction

Direct mailings to a company's potential customers - "junk mail" to many - can be a very effective way to market a product or service. However, as we all know, much of this junk mail is really of no interest to the majority of the people that receive it. Most of it ends up thrown away, not only wasting the money the company spent on it, but also filling up waste sites or needing to be recycled. If the company had a better understanding of who their potential customers were, they would know more accurately who to send it to, so some of this waste and expense could be reduced.

The goal of this article is to propose a solution to help save economic and environmental costs of direct mailing marketing campaigns. The attempt to achieve this purpose was done using data mining techniques. The resulting prediction model will tell a company to which potential customers it should send direct mail in order to have the highest possible probability of it resulting in a sale.

The data to be analyzed was published on the Internet as part of the CoIL Challenge 2000, a contest in which 43 participants had submitted their solutions. They used various data mining techniques besides statistics in order to present their findings. The results of the contest are commented at the end of the article.

2 Data

The dataset used was obtained from the *archive.ics.uci.edu/ml* web site. It served for the EU sponsored CoIL Challenge 2000, a contest which invited data mining experts to contribute their prediction models. The data comes from a Dutch insurance company and contains demographics data and behavioral data about several thousand customers of the company.

The organizers of the contest have wanted to give the participants data which had been pre-processed.

Table 1: A section of the data - the full training dataset has 85 attributes and 5822 instances

Customer subtype	number of houses	household size	average age	customer main type	catholic	protestant	other religion	no religion	married
33	1	3	2	8	0	5	1	3	7
37	1	2	2	8	1	4	1	4	6
37	1	2	2	8	0	4	2	4	3
9	1	3	3	3	2	3	2	4	5
40	1	4	2	10	1	4	1	4	7
23	1	2	1	5	0	5	0	5	0
39	2	3	2	9	2	2	0	5	7
33	1	2	3	8	0	7	0	2	7
33	1	2	4	8	0	1	3	6	6
11	2	3	3	3	3	5	0	2	7
10	1	4	3	3	1	4	1	4	7
9	1	3	3	3	1	3	2	4	7
33	1	2	3	8	1	4	1	4	6
41	1	3	3	10	0	5	0	4	7
23	1	1	2	5	0	6	1	2	1
33	1	2	3	8	0	7	0	2	7
38	1	2	3	9	0	6	0	3	7
22	2	3	3	5	0	5	0	4	7
13	1	4	2	3	2	4	0	3	7
31	1	2	4	7	0	2	0	7	9
33	1	4	3	8	0	6	0	3	9
33	2	3	3	8	0	4	2	3	7
13	1	3	2	3	1	7	0	2	7
34	2	3	2	8	0	7	0	2	7
13	2	4	3	3	0	4	2	4	8
33	1	3	3	8	0	6	1	2	6
37	1	3	3	8	0	5	0	4	7
40	1	3	3	10	0	3	0	6	9
31	1	4	2	7	0	9	0	0	5
33	2	2	3	8	0	7	1	2	5

2.1 Data description

2.1.1 Number of instances

The challenge organizers have provided two sets of data. The first is the training data which contains 5822 customer records (partially presented in Table 1). The second is the data set for predictions. This dataset contains 4000 customer records.

2.1.2 Number of attributes

- numerical: 85
- nominal: none

The number of attributes and the column orders is identical. It should be mentioned that the organizers of the contest had pre-processed the data, so that all had discrete numerical values. For example instead of a real-valued feature giving precise monetary amount that a customer pays for car insurance, the CoIL datasets include only a discrete-valued feature that categorizes this amount into one of seven different discrete levels. Similarly, for the attribute Customer Subtype, in the first column of the data screen-shot below, 41 different nominal values have been transformed into integer values from 1 to 41.

2.1.3 Target variable

The target variable had two numerical values (0 for the customers who do not own a caravan insurance policy and 1 for those who do). The distribution of the target variable is highly unbalanced:

- 6% caravan insurance policy owners (1)
- 94% have no caravan insurance policy (0)

2.1.4 Missing values

Due to the pre-processing of the data by the organizers of the CoIL Challenge, the data had no missing values.

2.2 Data understanding

The dataset to train and validate prediction models consists of two radically different subsegments of attributes. The first (attributes 1-43) contains socio-demographic data while the second consist of insurance product ownership (attributes 44-85). The socio-demographic data is derived from zip codes. All customers living in areas with the same zip code have the same socio-demographic

attributes. Attribute 86 is the target variable - number of mobile home policies. Being able to predict the target variable has important economic implication for the insurance company (or any other company engaged in the activity of direct mailing to a large customer base). The cross-industry, international average for the customer response rate of such a mailing is claimed to be below 1%. This means that for every 10.000 letters that a company sends, less than 100 will result in a sale/purchase. The direct implication is that if, through better prediction, the response rate increases, the direct mailing variable costs are reduced significantly as shown in Figure 1.

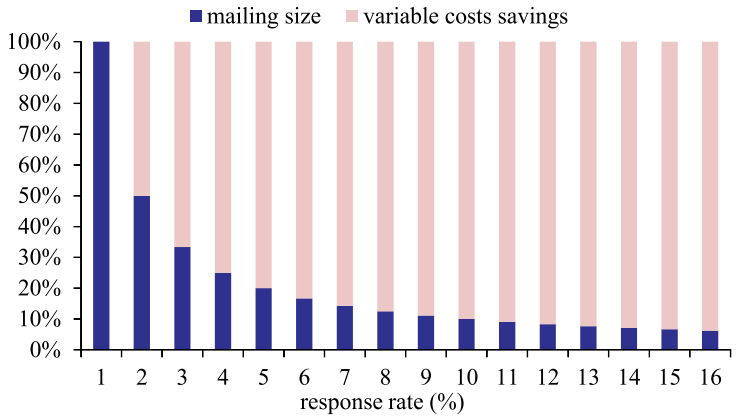

Figure 1: The size of a mailing needed to obtain the same number of responses from customers for different response rates and the associated savings of variable costs as compared with the costs incurred when the response rate to a mailing is 1%.

It should also be noted that, from the business point of view two prediction models are needed. The prediction model which uses some or all behavioral data could only be applied to the company's existing client base, for whom all data is available - the so called cross sale. On the other hand it makes sense to build a separate prediction model using only the socio-demographic data if the company were interested in acquiring customers that have not yet established a business relationship with the company. (This task was not a part of the CoIL Challenge.)

2.3 Data preprocessing

The data was downloaded into a table from a text format. In order to use it with the Weka data mining tool, it was necessary to save it to a *.csv file and also exchange semicolumns with comas. In addition column headings needed to be added since they were provided in a separate file. For the convenience of

the user and to be able to apply the J48 and naïve Bayes algorithms the target variable was changed to nominal values "YES" and "NO" for "1" and "0" respectively. Given that there are only two possible outcomes this change did not result in loss of information, as would be the case if the target variable could take more than two numerical values. All of the abovementioned processing was done manually either in text or spreadsheet editors. In addition the search of the best prediction model involved trying to use different algorithms on full dataset but also omitting some attributes. This elimination of columns was performed in Weka.

3 Machine Learning Methods Used

Different prediction methods (ZeroR, naïve Bayes, k nearest neighbors (kNN) and J48 pruned tree) were considered and tried given the relatively low "cost" of running them in Weka. At the first sight ZeroR and kNN gave "excellent" results with 94% accuracy, however a closer look quickly revealed that the reason for this high accuracy was the fact that the target attribute distribution was so unbalanced (only 6% YES instances). ZeroR always predicted NO, and that was useless for the task at hand. kNN on the other hand would similarly encounter neighbors with a NO class value with 94% probability which effectively meant that it would similarly always predict NO. In the case of J48 pruned tree algorithm the result was similar and a quick look at the confusion matrix revealed that this method also predicted around 6% of YESes and 94% of NOs. The reasons for such prediction rest in the fact that the data is unbalanced while, at the same time, there are a high number of attributes. A model that would predict well on the training dataset would almost certainly be over-fitted thus producing poor results on the test set.

Naïve Bayes algorithm's results seemed sensible. Besides some important classifier's characteristics were considered. The dataset had a sufficient number of training examples for reliable probability estimation. The method is known to be reliable, it achieves good classification accuracy and is robust. In addition it has been successfully applied in many different domains, some similar to the task considered in this article.

3.1 Brief description of the method used

A naïve Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Bayes' theorem relates the conditional and marginal probabilities of events A and B, where B has a non-vanishing probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

The probability model for a classifier is a conditional model:

$$p(C|F_1, \ldots, F_n)$$

over a dependent class variable C with a small number of outcomes or classes, conditional on several attributes F1 through Fn. The problem is that if the number of attributes n is large or when an attribute can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, we write

$$p(C|F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}.$$

By reformulating the above statement and applying the "naïve" conditional independence, we obtain the statement:

$$p(C|F_1, \ldots, F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C)$$

which the algorithm uses to calculate the probability of a target variable given a set of attributes.

3.2 Brief description of the evaluation criteria

To evaluate the model ten-fold cross validation was used on the training set. Finally the performance of the classification model was tested on the supplied test set with 4000 instances. Since the data is highly unbalanced accuracy does not give a meaningful interpretation of the quality of the model. In such cases F-measure is a better indicator of the models accuracy. It considers both the precision and the recall of the test to compute the score.

$$F = 2\frac{precission \cdot recall}{precission + recall}.$$

Recall (a.k.a. true positive rate) is the proportion of correctly predicted instances of class C' among all actual instances of the same class. Precision on the other hand is the proportion of correctly predicted instances of class C' among the number of all predictions of class C'. The F measure can be interpreted as a weighted average of the precision and recall, where an F score reaches its best value at 1 and worst score at 0.

4 Evaluation

Having opted for Naïve Bayes prediction model, the remaining task was to identify the combination of attributes, which would yield the highest F measure. The best result was achieved when part of the demographic data was removed from the dataset. Specifically attributes such as religion, marital status, number of children were removed. Nevertheless, the model considered the demographic data which describes customers' wealth (e.g. do they own or rent a house, how many cars they own, what their income level is, etc). In the Table 2 the F-values for three different scenarios are shown.

Table 2: The considered attributes

	attributes considered	F-value
(1)	all 85 attributes	0.212
(2)	only behavioral attributes (42 attributes)	0.156
(3)	behavioral and 14 demographic attributes	0.229
(4)	the favorite model (3), on the test dataset	0.203

For the purpose of targeting customers who have not had any business with the insurance company, a similar prediction model was trained using only the demographic data. As expected the result was slightly worse. In this case the F-measure was 0.176 on the training dataset and 0.164 on the test dataset.

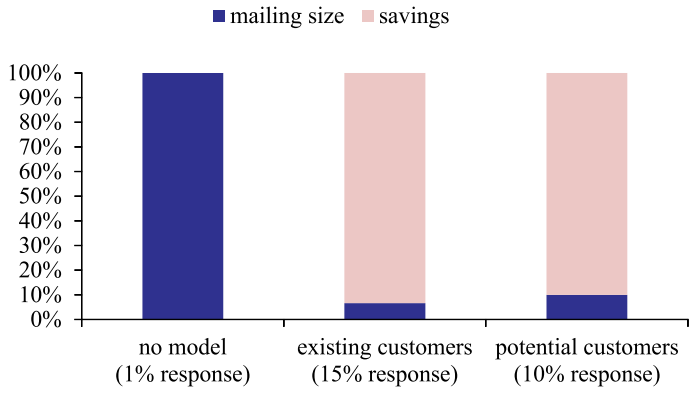

Figure 2: The proportion of a mailing needed to obtain the same number of responses from customers. Current situation as compared with response leveled enabled by the prediction models.

Everything considered the result is encouraging for the client - the insurance company. Using the predictive model proposed in this article they could improve the response rate of their customers to their mailings. Based on the

precision measure, we can expect the response rate to be around 15% for direct mail to company's existing customers and 10% to the potential clients who have not yet purchased insurance from the insurance company. These figures could reduce the variable costs of direct mailing significantly - by around 90% as shown in Figure 2.

Another important economic implication can be deduced from the Figures 1 and 2. Increases in response rate have diminishing returns (savings). Huge savings may be obtained when a company manages to move from 1% response to 2, 3 and 4%. On the other hand moving from 10% response rate to 15% hardly bears any fruit. Yet in terms of data mining effort (and associated cost) it is much easier to go from 1% to 5% than from 10% to 11%. This is an important notion for companies which are not yet taking advantage of data mining. They can have important returns (high savings at low cost) if they start using data mining in order to improve their business operations. On the other hand, companies that have already been using data mining and might be above 5% response rate must do a more careful analysis of costs (data mining effort) and benefits (potential savings obtained) to make sure that the data mining exercise is going to bring them positive returns.

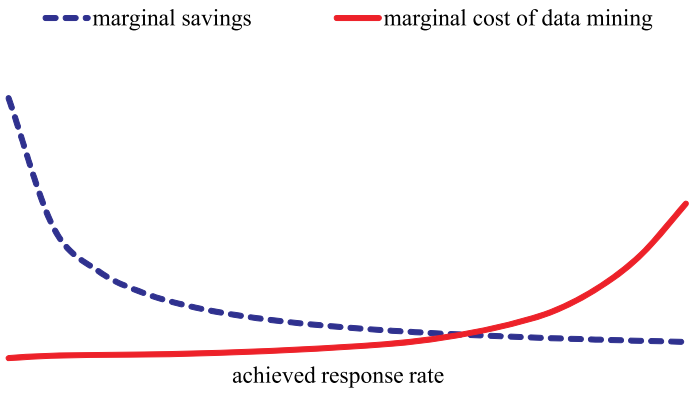

Figure 3: Marginal savings and marginal costs of improving the response rate of direct mail with a data mining effort. The difference between the two is net savings. It is important for a company to determine whether it has already approached the point where the two lines cross. Beyond that point improving predictive model does not make economic sense.

5 The CoIL Challenge 2000 Results

The task was to find the subset of customers with a high probability of having a caravan insurance policy Specifically the organizers asked the participants to

find the set of 800 customers in the test set of 4000 customers that contains the most caravan policy owners. The results of the contest among the 43 entries who submitted the results are presented in the Figure 4.

The winning entry, by Charles Elkan, identified 121 caravan policy holders among its 800 top predictions. The next best methods identified 115 and 112 policy holders. The mean score was 95.4, with a standard deviation of 19.4.

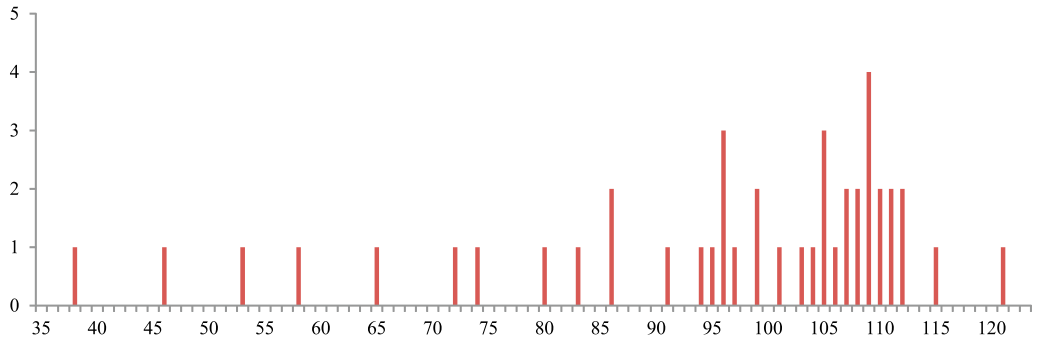

Figure 4: Distribution of scores achieved by 43 entries of the CoIL Challenge 2000.

The data mining algorithm used in the winning entry was standard naïve Bayesian learning. The second best entry also used a naïve Bayesian classifier. Many of the methods described in the 29 reports by participants are more sophisticated. They include combinations of backpropagation neural networks, self-organizing maps (SOMs), evolutionary algorithms, C4.5, CART, and other decision tree induction algorithms, fuzzy clustering and rule discovery, support vector machines (SVMs), logistic regression, boosting and bagging, and more. [1, 2, 3].

The naïve Bayesian predictive model proposed in this report identified 78 caravan policy holders among its 800 top predictions. This score would take place number 37 in the competition, leaving behind 7 contestants.

6 Conclusion

This article gives an overview of the data mining work performed with the aim to build a prediction model, which could help save costs (economic and environmental) to companies that engage in direct marketing activities. The strong performance of naïve Bayesian learning against other methods proposed by the participants of the CoIL Challenge is noteworthy. Perhaps the most important conclusion of this article is that by applying data mining companies could save significantly on direct mail, particularly those who have not yet

applied data mining techniques in their efforts to better target their potential customers.

Acknowledgements

Operation part financed by the European Union, European Social Fund.

References

[1] C. Elkan, Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000, *University of California, San Diego* (2000).

[2] P. van der Putten, M. de Ruiter, M. van Someren, CoIL Challenge 2000 Tasks and Results: Predicting and Explaining Caravan Policy Ownership, Sentient Machine Research (2000).

[3] Reports submitted by 29 participants of the CoIL Challenge 2000, www.liacs.nl/ putten/library/cc2000.

Received: May, 2011