Application of Decision Trees in Mining High-Value Credit Card Customers

Jian Wang Bo Yuan Wenhuang Liu

Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, P.R. China
E-mail: gregret124@gmail.com, {yuanb, liuwh}@sz.tsinghua.edu.cn

Abstract

Along with the rapid growth of credit card market in China, each bank has already accumulated a large number of customers. Since it is well known that the majority of the profit usually comes from a small portion of the customers, how to identify high-value customers is an important issue to be addressed in the banking industry. The purpose of this paper is to show how a popular data mining model can be used to help banks predict highly profitable customers based on just a few customer attributes.

Keywords: credit card; customer value; decision tree model; lift curve

1. Introduction

Along with the booming economy in China, all major domestic commercial banks have introduced various kinds of personal financial products and services such as loans and investment funds during the past few years. Among them, credit card products have received special attention and each bank has devoted a significant amount of efforts to maintaining and increasing its market share.

Similar to other personal financial products, customer resources are of crucial importance to the growth of credit card business, especially high quality customers. According to Pareto’s law, each customer’s contribution to a company varies greatly and, as a rule of thumb, around 80% of its total profit is produced by only 20% of its customers. Therefore, how to discover the relatively small portion of high quality customers from the entire customer group is an important question worth of investigation [7, 8].

Data Mining, or Knowledge Discovery in Database, provides powerful tools to extract potentially useful information from massive, inclusive of noise, incomplete and random data [1]. As far as credit card business is concerned, data mining techniques have been mostly applied in areas such as credit evaluation [3, 4] and fraud detection. In this paper, the well known decision tree model is employed as a classifier for the identification of high-value credit card customers.

The main objective is to extract representative features of high-quality credit card customers, making it possible for banks to predict which customers are likely to become profitable ones in the future. By doing so, banks can devote valuable customer service resources more accurately and efficiently to carefully selected customers for the purpose of customer retention and value promotion.

2. Decision Tree Model

Among many popular classification models such as neural networks and support vector machines, the decision tree model was used in this paper due to its ability to derive well-defined rules from the data,

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

which is important in practice for bank staff to comprehend and accept the data mining results [5, 6].

As one of the most well known decision tree algorithms, ID3 [2] was used to create the decision trees in the experiments. The core concept in the ID3 algorithm is information gain based on entropy, which is a statistics measuring a system's degree of chaos. For example, large entropies indicate high degree of disorder in a system. Therefore, the process of building a decision tree is to iteratively partition the original data set into subsets in which data are more uniform in terms of class label.

Given a set S of P samples and m class labels: C_1, C_2, … , C_m with each class having P_i samples ($1 \le i \le m$). The entropy of S relative to this classification is:

$$E(S) \equiv \sum_{i=1}^{m} -\frac{P_i}{P}\log_2\frac{P_i}{P} \qquad (1)$$

In (1), P_i/P is the proportion of the i^{th} class in S. The effectiveness of an attribute A in terms of classification is measured by information gain, which is the expected reduction in entropy should attribute A be used to classify the samples.

The information gain of attribute A in the above scenario is:

$$G(S,A) = E(S) - \sum_{i\in V(A)}\frac{|S_i|}{|S|}E(S_i) \qquad (2)$$

In (2), $V(A)$ is the set of all possible values that attribute A may take and S_i is a subset of S where attribute A takes its i^{th} value. After the information gain of all candidate attributes is calculated, the attribute with the highest information gain is selected as the most favorable attribute (root node) and branches are created according to its possible values.

This process is repeated along each branch until the subset of samples belonging to the leaf node all share the same class label or there is not unused attribute left along that branch.

3. Experimental Results

3.1 Description of Data Set

In this paper, we used a collection of credit card customer data from a leading commercial bank in China. The data set originally contained records of more than 40,000 customers. After eliminating samples with invalid or missing attribute values, 17,952 samples were used in the experiments. Each sample came with a list of basic customer attributes as well as a record of consumption point, which is a major indicator of each customer's contribution to the bank. A simple analysis showed that 3,592 samples (around 20% of 17,952) contributed 80% of total consumption points. Consequently, they were labeled as high-value customers and the objective was to build a predictive model of high-value customers based on the attribute values of these samples.

Three customer attributes: *annual income*, *financial asset* and *education background* were selected after consulting with some senior bank managers to build the classification model. In the data set, a*nnual income* and *financial asset* were continuous variables while *education background* was a discrete variable.

For the sake of clarity, the two continuous attributes were transformed into discrete ones. The *annual income* (A) attribute was divided into four intervals: <70,000, [70,000, 500,000], [500,000, 1,000,000] and >1,000,000, which were represented by 1, 2, 3 and 4 respectively. The *financial asset* (F) attribute was divided into three intervals: <100, 000, [100,000, 300,000] and >300, 000, which were represented by 1, 2 and 3 respectively. In the meantime, the e*ducation background* (E) attribute was also simplified into three levels: junior (primary school), senior (secondary school) and higher (university or above), which were represented by 1, 2 and 3 respectively.

3.2 Building the Decision Tree

The original set of 17,952 samples was divided into two subsets *S*1 (training set, 10,000 samples) and *S*2 (test set, 7,952 samples) with approximately identical class ratio (high-value customers vs. low-value customers = 1:4).

According to (1), the entropy of *S*1 is:

$$E(S1) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} \approx 0.72$$

The information gain of each of the three attributes with regard to *S*1 can be calculated according to (2):

$$G(S1, A) \approx 0.057$$

$$G(S1, F) \approx 0.0081$$

$$G(S1, E) \approx 0.050$$

It is clear that *annual income* had the highest information gain among the three attributes. Consequently, it was chosen to be the root node of the decision tree and *S*1 was divided into four subsets according to the four possible values of the *annual income* attribute.

3.3 The Power of the Decision Tree

Once the complete tree was built, a quick check of the leaf nodes showed that most of the subsets of samples associated with the leaf nodes contained samples of different class labels. As a result, it is difficult to make a deterministic decision on the class label given an unknown sample. It should be noted that this result is very likely to happen as in practice there are many factors that can play a role in the value of a customer, which by no means can be all included in the data set.

Since the purpose of using classification models in this paper is to identify as many as possible potentially high-value customers, an informative performance measure is the *lift* of the model generated instead of its prediction error as in common classification problems (the data set was also clearly unbalanced).

Table 1: The Q Values of 36 Leaf Nodes

A	F	E	Q
1	1	1	0.0276
1	1	2	0.0262
1	1	3	0.0446
1	2	1	0.1111
1	2	2	0.0811
1	2	3	0.1289
1	3	1	0.1795
1	3	2	0.1232
1	3	3	0.3023
2	1	1	0.6943
2	1	2	0.6595
2	1	3	0.6768
2	2	1	0.0464
2	2	2	0.1224
2	2	3	0.2088
2	3	1	0.2486
2	3	2	0.3060
2	3	3	0.4102
3	1	1	0.5389
3	1	2	0.5568
3	1	3	0.3824
3	2	1	0.2308
3	2	2	0.0
3	2	3	0.0
3	3	1	0.4266
3	3	2	0.4340
3	3	3	0.2563
4	1	1	0.9794
4	1	2	0.7742
4	1	3	0.4712
4	2	1	0.20
4	2	2	1.0
4	2	3	0.50
4	3	1	0.4681
4	3	2	0.3478
4	3	3	0.2772

The *lift* of a data mining model in the scenario of this paper is measured by the proportion of high-value customers identified by the model given a certain proportion of the whole customers selected by the model. For instance, given *P* unknown samples, the model can assign each sample a likelihood of being a high-value customer. The question is how many samples in the top *M* % samples are actually high-value customers? More intuitively, the proportion (*N* %) of high-value customers identified is measured. Obviously, a model is expected to identify as many as possible high-value customers by testing as few as samples. The *lift* value is then defined as *N*/*M*.

In our case, it was known in advance that 20% customers were high-value ones. As a result, an idealized model is able to identify all high-value customers by testing exactly 20% customers (*lift* value =5) while the *lift* value of a model based purely on random selection should be 1.

In order to calculate the likelihood, a variable Q was defined for each leaf node. Let k_1 be the number of high-value samples in the subset associated with the leaf node and k_0 be the number of rest samples in this subset. Q is defined as:

$$Q = \frac{k_1}{k_0} \tag{3}$$

The values of Q of all 36 leaf nodes are shown in Table 1. Note that if k_0 is zero for some leaf nodes, Q can be defined similarly as $k_1 / (k_1 + k_0)$.

For each of the 7,952 test samples in *S*2, a Q value was assigned to it based on its corresponding leaf node. All samples were then sorted according to their Q values. A *lift* chart is shown in Fig.1 where the horizontal axis shows the proportion of sampled selected and the vertical axis shows the proportion of high-value customers discovered. The diagonal line is the *lift* curve of a purely random model, which was used as the baseline.

Fig. 1 shows that, for example, by selecting the top 50% of unknown samples based on their Q values, around 78.7% of high-value customers can be correctly identified, giving a *lift* value of 1.574. By contrast, the random model can only identify 50% high-value customers in the same situation.

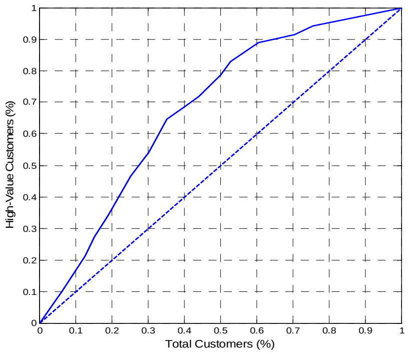

Fig. 1: The *lift* curve of the decision tree model with three attributes.

As mentioned in Section 3.1, three attributes were selected based on the experience of domain experts to build the decision tree model. In order to show whether these three attributes are sufficient for mining high-value customers, additional experiments were conducted with different combinations of attributes. Fig. 2 shows the *lift* curve (dashed line) of the decision tree model built with an extra attribute *customer age*. It is clear that the inclusion of this extra attribute in the model building process did not help improve the *lift* value of the decision tree model, which gives some empirical evidence on the goodness of the original three attributes used in this paper.

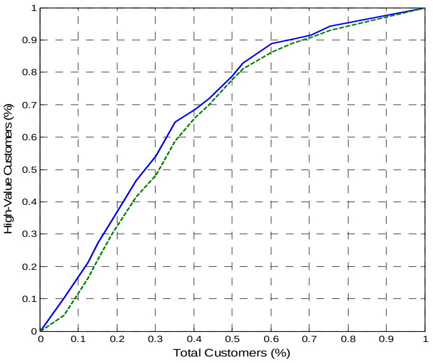

Fig. 2: A comparison of the *lift* curves of two decision tree models with three and four attributes respectively.

4. Conclusions

This paper focused on a challenging issue in the credit card business: how to effectively identify potential high-value customers? For this purpose, a decision tree model was employed to construct a classification (prediction) model, which can be used to estimate the likelihood of a customer being a high-value customer. Experimental results showed that, with only three customer attributes, this model could correctly identify nearly 80% of high-value customers by selecting only half of the candidate customers, which was more than 50% better than a model selecting customers at random.

In addition to the preliminary work presented here, there is still a lot of room for further improvement. One of the important directions is to incorporate more customer attributes into the model in order to achieve higher prediction accuracy. Currently, only three static attributes have been explored while it is believed that much more useful information can be taken advantage of by exploiting the dynamic features of each customer such as transaction records and usage behaviors. Certainly, the achievement of this goal requires both the availability of relevant data as well as more sophisticated techniques for extracting discriminative features from mass transaction data.

Another issue of high practical implication is to discover a set of *interesting* and *nontrivial* rules from the data, which can be readily accepted and used by bank managers.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC Project No.: 70671059).

References

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", *China Machine Press*, pp. 291-309, 2006.

[2] T.M. Mitchell, "Machine Learning", McGraw-Hill, 1997

[3] H. Yeh, M. Yang and L. Lee, "An Empirical Study of Credit Scoring Model for Credit Card", *Proceedings of the Second International Conference on Innovative Computing, Information and Control*, pp. 216-219, 2007.

[4] F. Li, J. Xu, Z. Dou and Y. Huang, "Data Mining-Based Credit Evaluation for Users of Credit Card", *Proceedings of the third International Conference on Machine learning and Cybernetics*, pp. 2586-2591, 2004.

[5] L. Cheng, Z. Xu, "Study on Predictive Model of Air Cargo Customer Defection based on Decision Tree", *Proceedings of International Conference on Wireless Communications, Networking and Mobile computing*, pp. 3693-3696, 2007.

[6] Q. Wang, Y. Wu, J. Xiao, and F. Guang, "The Applied Research Based on Decision Tree of Data Mining In Third-Party Logistics", *IEEE International Conference on Logistics*, pp. 1540-1544, 2007.

[7] F. Meng, L. Shuai and C. Jiang, "An Application of Decision Tree to Analyze the Value of Customer", *Computer and Development*, vol.4, pp.60-63, 2007.

[8] K. Yan, L. Zhang and Q. Sun, "The Application of Decision Making Tree Classification ID3 Algorithm in Customer Value Segmentation in Aviation Market", *Commercial Research*, vol.3, pp. 24-28, 2008.