ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Jarosław Michalak**

ON THE APPLICATION OF CLASSIFICATION TREES TO ANALYZE CUSTOMER LOYALTY AND SATISFACTION

Abstract. It became more important to recognize customer expectations and to offer product properties, which convince customer to buy the analyzed product.

In customer loyalty and satisfaction analysis methods, classification trees play a very important role.

The aim of the paper is to present an application of tree – structured models to analyze product properties influencing buying decision of the target group.

Key words: classification tree, customer loyalty and satisfaction.

1. INTRODUCTION

Classification tree can be described as tree – like way of representing a collection of hierarchical rules that lead to a class. In other words, we want to predict values of a categorical dependent variable (e.g. group membership) from one or more continuous and categorical predictor variables.

The model building process is based on recursive partitioning the learning set into homogenous subsets considering dependent variable.

Generally, at each node of the tree we do the following steps (Breiman *et al.* 1984):

i) Examine every allowable split on each predictor variable,

ii) Select and execute the "best" of these splits,

iii) Stop splitting on a node when some stopping rule is satisfied.

* Ph.D., Chair of Statistical Methods, University of Łódź.

More formally, let us consider an additive model (Breiman *et al.* 1984; Gatnar 2001)

$$y = \alpha_0 + \sum_{m=1}^{M} \alpha_m g_m(x, \beta) \tag{1}$$

where $g_m(\mathbf{x}, \beta)$ are functions of $\mathbf{x}$ with parameters β.

An approximation of (1) can be written as

$$y = a_0 + \sum_{m=1}^{M} a_m I(x \in R_m) \tag{2}$$

where R_m $(m = 1, ..., M)$ are disjoint regions in the p-dimensional feature space, a_m are real parameters and $I\{A\}$ is an indicator function

$$I\{A\} = \begin{cases} 1, & \text{if the proposition inside the brachets is true} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

For real-valued dimension of the region R_m, characterized by its upper and lower boundary $x_r^{(d)}$ and $x_r^{(g)}$, we have:

$$I\{x \in R_m\} = \prod_{r=1}^{p} (x_{mr}^{(d)} \leqslant x_r \leqslant x_{mr}^{(g)}) \tag{4}$$

For each categorical variable x_r we have

$$I\{x \in R_m\} = \prod_{r=1}^{p} I\{x_r \in B_{mr}\} \tag{5}$$

where B_{mr} is a subset of the set of the variable values (see e.g. Gatnar 2001).

In our analysis, we used two classification trees algorithms: CART described in detail in L. Breiman *et al.* (1984) and QUEST proposed by W.-Y. Loh and Y.-S. Shih (1997).

Briefly, at each step of the CART procedure, the dataset is divided into two purer descendant subsets. The results of the splitting of the data can be depicted as a binary tree. The growing of the tree is stopped when the nodes are very small or pure. That big tree is then pruned using a cost-complexity pruning algorithm to get a decreasing sequence of subtrees. The best tree in the sequence is chosen by cross-validation.

QUEST, in general, is designed to have unbiased variable selection in the splitting process. Variable selection procedure is based on statistical tests: ANOVA F-test for numerical covariates or chi-square test for categorical ones. A modified form of quadratic discriminant analysis is then applied to split the node on the selected variable. The tree is pruned using the CART cost-complexity pruning algorithm and the best tree is chosen due to cross-validation results.

Tree-based models are simple, flexible and powerful tools for classification analysis, dealing with different kinds of variables, including missing values and very easy to interpret.

2. APPLICATION OF TREE-BASED MODELS TO CLASSIFY CUSTOMERS

The main purpose of the research is the evaluation of reactions and decisions of customers. We analyze a group of potential users of the blood pressure monitor, who want to buy this product in the near future.

150 customers, drawn independently from customer databank owned by the producing company, were asked to complete the questionnaire with 15 questions.

The following independent variables were taken into consideration:

I. Features describing product (from 1 – feature of no importance, to 5 – feature of great importance):

1 – design,
2 – price,
3 – warranty,
4 – size of the display,
5 – customer service reaction time,
6 – number of remembered measurements,
7 – number of possible measurements on the one battery,
8 – possibility of working with the feeder,
9 – easy service,
10 – product availability.

II. The product destination (1) – professional use, 2) – own use, 3) – the gift).

III. Demographic variables:

1 – age: (**1** – up to 30, **2** – (30–50), **3** – (50–70), **4** – 70 and more years old,

2 – education (elementary (**e**), vocational secondary (**v**), secondary (**s**), university (**u**)),

3 – sex (man (**m**), woman (**w**)),

4 – dwelling place (**1** – big city, **2** – small town, **3** – country).

The dependent binary variable defines whether or not every person is interested in buying the blood pressure monitor from the Hartmann company: 1 – yes (group 1; 89 persons), 2 – no (group 2; 61 persons).

The potential association of each of the considered features with the response variable was calculated using χ^2 test.

The following results were obtained:

1. Age – $p < 0.001$. In the group of people, who are not interested in buying the blood pressure monitor from the Hartmann company, there are elderly people, more than 70 years old.

2. Education – $p < 0.001$. People from the second group possess worse education.

3. Sex – has no association with decision of buying the blood pressure monitor.

4. Dwelling place – $p < 0.001$. People, who are interested in buying, are mainly from big cities.

5. Destination – $p < 0.001$. Generally, in both groups, there are more people who are going to buy the blood pressure monitor for their own use. However, in the second group, there are 96,7% such people. In the first group (interested in buying) approximately 20% are going to buy the blood pressure monitor for professional use.

6. Design – $p < 0.001$. As far as the first group is concerned, there are more people for whom this feature is important or very important. In the second group – this feature is unimportant in the process of buying.

7. Price – $p < 0{,}001$. The price plays a very important role among people from the second group. In the first group this factor is less important.

8. Warranty – $p < 0.05$. It is more important for people from the first group.

9. The size of projector – is not associated with the willingness of buying.

10. The rapidity of service reaction – $p < 0.001$. It plays more important role for people from the first group. In the second group, customers do not pay attention to this feature.

11. The number of remembered measurements – $p < 0.05$. It is much more important for people in the first group.

12. The number of measurements on one battery – $p < 0.05$. In the second group – it has no importance, but it has great importance in the first group.

13. Working with the feeder – $p < 0.001$, $C = 0.373$. It is very important factor for people from the first group. For people from the second group – rather unimportant.

14. Easy service – the lack of important dependence with willingness to buy the blood pressure monitor.

15. Product availability – $p < 0.05$. It has no importance for people from the first group, but it has great importance for people from the second group.

To find features defining potential customers, the module of Classification Trees was used (from the package of STATISTICA PL). This module gives opportunity for creating classification trees in accordance to two algorithms: CART (Breiman *et al.* 1984) and QUEST (Loh, Shih 1997).

The best results in terms of prediction accuracy (the smallest percentage of incorrect classifications) and ease of interpretation (the smallest number of the terminal nodes) were obtained for the QUEST algorithm with univariate splits.

Trees created by the QUEST algorithm are evaluated as optimal. In contrast to the CART trees, they are unbiased in split variable selection process. T.-S. Lim *et al.* (2000) presents the results of comparison of 33 algorithms (classification trees, classical methods as discriminant analysis, neural networks), due to the prediction accuracy, degree of complexity and the time of computer work necessary to the process of learning for 16 datasets. Among algorithms which create classification trees, QUEST had the highest evaluation.

The obtained tree is presented in Fig. 1. The tree has 8 splits and 9 leaves. The classification error rate was evaluated using cross – validation and resubstitution methods (Tab. 1).

In the cross-validation method, the learning set is randomly divided into V equal-size subsets (usually $V = 10$). A tree is created V times, each time from a different group of $V-1$ subsets. The rule obtained is then used to classify the cases from the subset left out in the tree construction process. The V misclassification rates are then averaged to obtain the CV – error rate.

In the resubstitution method, the learning set is employed to create the classification tree and then is used to test the obtained decision rule.

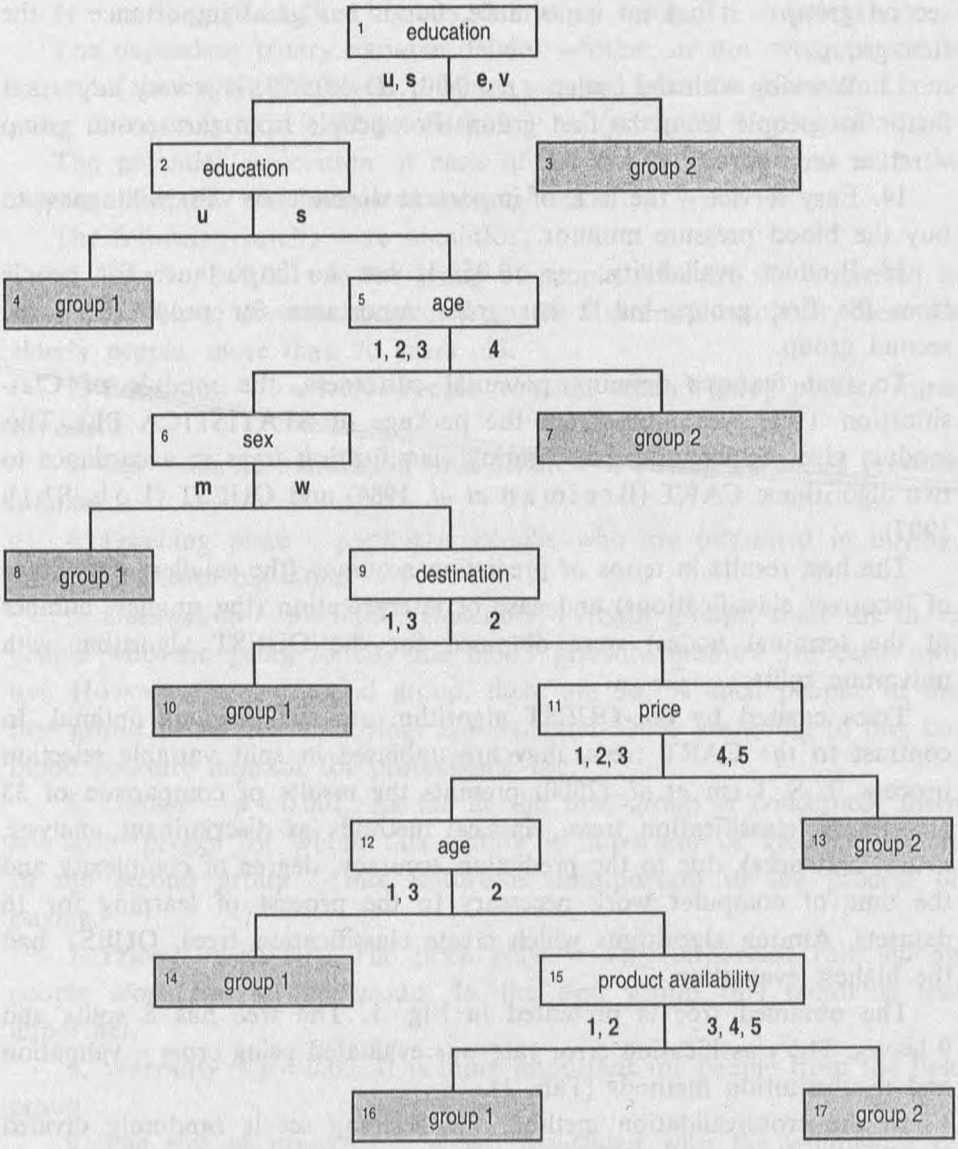

Fig. 1. The QUEST classification tree

Source: own elaboration.

Table 1 presents the results of classification derived from the constructed tree.

Table 1

The results of classification

Predicted group	Actual group	
	group 1	group 2
Group 1	76	7
Group 2	13	54
Total	89	61
Frequency of correct classifications	76/89 (85,39%)	54/61 (88,52%)
The cost of resubstitution	20/150 (13,33%)	
10-CV error rate	22,00%	

Source: own evaluation.

In the analysis of the tree it is easy to notice that the splitting variables are mainly the demographic ones, such as education, age, sex and a few variables which characterize the product: destination, price, product availability.

During the analysis of the classification tree (from the root to the leaves), it is easy to define the classification rules for both groups of potential customers. As far as people who are not interested in buying the blood pressure monitor are concerned, they could be defined as:

1) people with vocational or elementary education;

2) people with secondary education and older than 70 years;

3) women with secondary education and younger than 70 years who buy the blood pressure monitor for their own use or who pay huge attention to the price of product;

4) women with secondary education, aged 30–50 years, who buy the blood pressure monitor for their own use or who do not pay attention to the price of product, but they pay an enormous attention to the product availability.

The fact that people with vocational or elementary education do not buy the product could be caused by not knowing the Hartmann company.

In addition, these people pay attention mainly to price and product availability. With regard to the fact that the blood pressure monitors are not sold in stores and supermarkets, this group of people is not interested in buying this product or they do not know anything about it.

Analogously, we can describe some decision rules for people who are interested in buying the blood pressure monitor producing by the Hartmann company.

The separated features, which characterize both groups of inquired people could be used to decide in cases like: changing the offer, the way of distribution and leading a promotional campaign (for example: reducing the price) etc.

The classification rules, which are obtained, give a good prediction in customers affiliation to the analyzed groups. The percentage of correct classifications for the whole learning set is 86.67%. When it comes to the second group the frequency of correct classifications is slightly better than in the first group – 88.52 and 85.39%.

The *CV*-error rate equals to 22%, so the obtained tree has the ability to accurately predict the group membership for new potential customers, not regarded in the research.

3. CONCLUDING REMARKS

According to L. Breiman *et al.* (1984) there are at least two main objectives of a classification task: 1) to get as accurate prediction as possible on unseen data and 2) to gain insight and understanding into the predictive structure of the data.

The results obtained from the QUEST algorithm are very good in terms of accuracy (considering *CV*-error rate). There are some other advantages of tree-based models over many traditional statistical methods: no requirement of knowledge of the variable distribution, dealing with different types of variables (very important in market research) including missing values and outliers, ease of interpretation of the results, reduction of the cost of the research by selecting only some important variables for splitting nodes.

Recursive partitioning method can be used as a supplement to classical statistical methods to solve numerous decision – making problems in market research. Some other examples of using classification trees in market research are described for instance in E. Gatnar and M. Walesiak (red., 2004).

REFERENCES

Breiman L., Friedman J., Olshen R., Stone C. (1984), *Classification and Regression Trees*, CRC Press, London.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE im. Oskara Langego we Wrocławiu, Wrocław.

Lim T.-S., Loh W.-Y., Shih Y.-S. (2000), *A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-three Old and New Classification Algorithms*, "Machine Learning", **40**, 203–229.

Loh W.-Y., Shih Y.-S. (1997), *Split Selection Methods for Classification Trees*, "Statistica Sinica", 7, 815–840.

Jarosław Michalak

O ZASTOSOWANIU DRZEW KLASYFIKACYJNYCH W ANALIZIE SATYSFAKCJI I LOJALNOŚCI KLIENTÓW

Rozpoznawanie oczekiwań klientów co do jakości oferowanych im produktów odgrywa istotną rolę w planowaniu strategii marketingowej firmy.

W artykule zaproponowano wykorzystanie metody rekurencyjnego podziału w analizie lojalności i satysfakcji klientów firmy Paul Hartmann, zainteresowanych nabyciem ciśnieniomierzy. Celem prowadzonych badań było wskazanie tych cech produktu, które mają największe znaczenie w procesie podjęcia decyzji o jego zakupie oraz opisanie reguł klasyfikacyjnych, dotyczących klientów grupy docelowej.