PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1096 — 2005
Ekonometria 15

Adam Kurzydłowski

THE APPLICATION OF THE REGRESSION TREE IN PREFERENCE ANALYSIS OF MILKA CHOCOLATE BUYERS

1. Introduction

Tree-based methods are among the statistical procedures that can be applied to classification and prediction problems. Generally, the classification task is to build a model of some kind that can be applied to unclassified data in order to classify them. Prediction is the same as classification except that the records are classified according to some predicted future behaviour or estimated future value [Aluja-Banet, Nafria 1997, p. 59].

Tree-based methods are an especially good choice when your task is classification of records or prediction of outcomes in a very large database. In the last two decades, they have become popular as alternatives to regression, discriminant analysis and other procedures based on algebraic models. Their advantage is that, in contrast to the above-mentioned methods, they represent rules. Rules can be easily understood, explained or translated into a natural language. Tree-based methods are also used to study the relationships between a dependent measure and a large series of possible predictors that themselves may interact [Berry, Linoff 1997, p. 243].

The growing interest in tree-based methods in marketing research is caused by their other strengths [Berry, Linoff 1997, p. 282; Hamilton et al. 2000; Bao 2001, p. 43-44]:

- ability to handle both continuous and categorical variables;
- ability to cope with incomplete data, including cases of missing attribute values;
- ability to construct a model without knowing the distribution of variables and relations between them;

- ability to provide a clear indication of which fields are most important for prediction or classification.

The tree-structured methods split the cases into two or more subgroups at each node so that the heterogeneity of the subgroups is maximised each time in a certain predefined sense. In other words, it's a rule for predicting the class of an object on the basis of the values of its predictors. The tree is constructed by recursively partitioning a learning sample of data for which the class label and the values of the predictors are known for each case. A node in the tree represents each partition [Aluja-Banet, Nafria 1997, p. 59; Loh, Shih 1997, p. 815]. The result at the end of the tree building process is that we have a series of nodes that are maximally different from one another on the target variable. To illustrate, the fig. 1 shows the procedure of tree-based model constructing [Kurzydłowski 2002, s. 41].

At the first stage, one has to prepare the data *S*. It's formed and named the **root node**.

Then at the **second stage**, a check is done for each acceptable way to split the root node, then nodes $S_1, \ldots, S_K$ into disjoint subgroups based on each predictor. In seeking the right split, each categories of non-binary nominal predictor are merged, whereas for ordinal predictor only neighbourly categories might be merged. However, continuous predictor is categorised for the purpose of the analysis.

At the third stage, the estimation of homogeneity of each node is done with respect to the target variable. If the root node and then nodes $S_1, \ldots, S_K$ are relatively homogenous, the procedure of tree building is stopped.

Otherwise, the procedure is continued and the best significant split is chosen depending on the selected quality measure of partition.

The fifth stage is realisation of partition on selected predictor. The data are divided with regard to the selected predictor. Database is split into as many subsets as there are categories or intervals of the selected predictor.

The sixth stage is repetition of stages 2-5 for each node $S_1, \ldots, S_K$.

In market analysis the tree-based analysis is used for [Kurzydłowski 2002, p. 81]:

- segmentation,
- searching for characteristic of consumer groups with similar attitudes (presence or absence of readiness to buy a product),
- identification of product characteristics which strongly influence the consumers attitudes,
- searching for patterns in consumer behaviour.

As I mentioned before, the objective of tree-based methods is to provide a simple rule for predicting a target variable from a set of predictors. The target variable can be either continuous or categorical, leading to what are called „regression trees" or „classification trees". A classification tree divides the data space, so that in each node of tree, data points distribute homogeneously. However, a regression

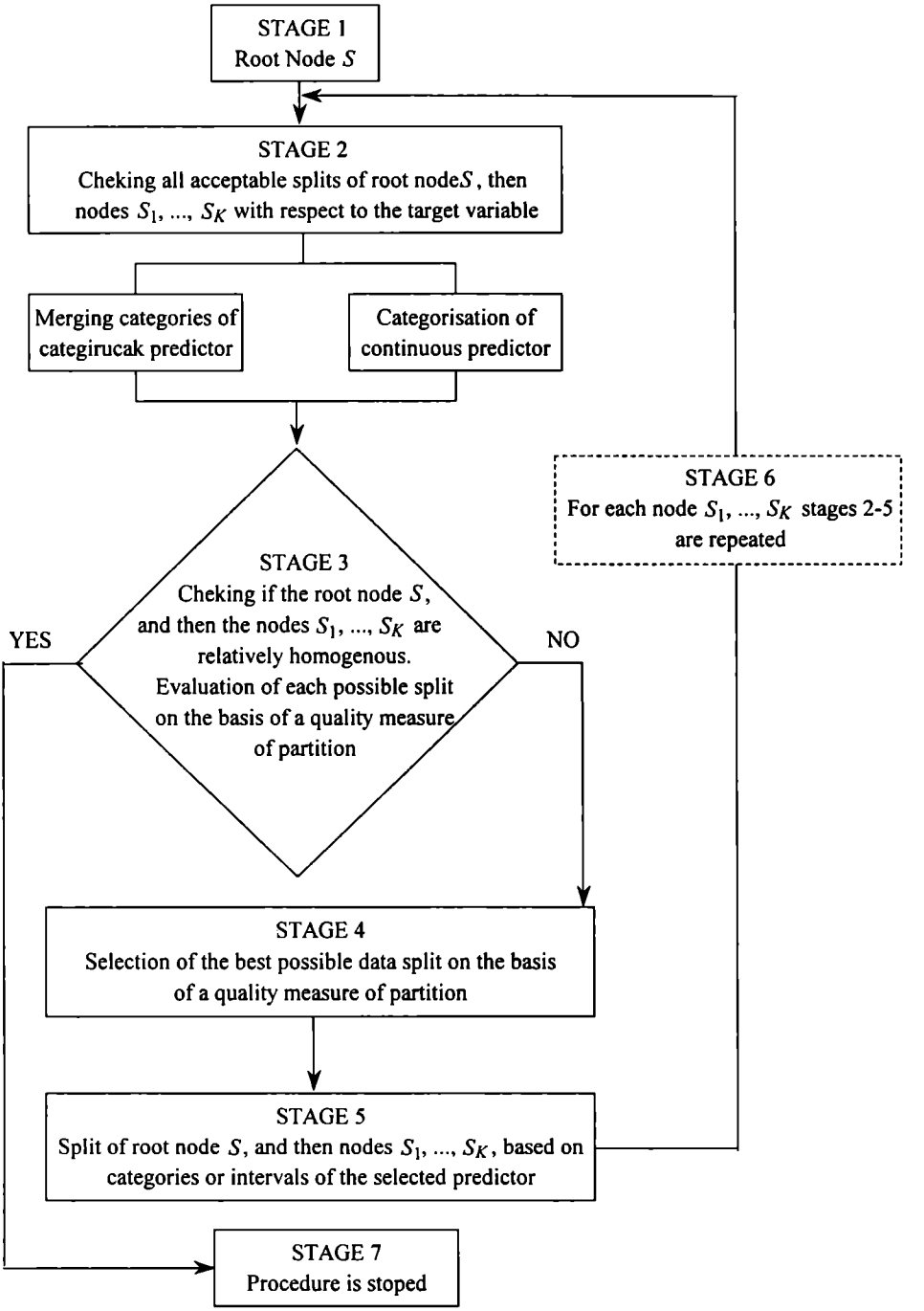

Fig. 1. The procedure of tree-based model constructing

tree divides the data space so that for all the data in each node, values of a particular attribute are close to each other [Chen and McNamee 1991, p. 319; Aluja-Banet and Nafria 1997, p. 59].

2. CHAID Algorithm Profile

CHAID algorithm besides strengths of tree-based methods pointed out earlier has certain advantages as a way of looking for patterns in complicated datasets. Firstly, the level of target variable measurement can be metric or categorical. Secondly, the level of measurement of the predictors can be nominal, ordinal or interval. Thirdly, not all predictors need be measured at the same level (nominal, ordinal, and interval). Finally, missing values in predictors can be treated as a „floating category" so that partial data can be used whenever possible within the tree.

CHAID is an exploratory method used to study the relationships between a target variable and a series of predictors. CHAID selects a set of predictors and their interactions that optimally predict the target variable. CHAID divides the data into two or more mutually exclusive, exhaustive subsets based on the best statistically significant split of predictor.

In order to determine the best split at any node, first of all CHAID algorithm merges any acceptable pair of categories of the predictor variable if there is no statistically significant difference within the pair with respect to the target variable. It happens when the p-value for that pair of categories is greater then 0,05 [Magidson 1993, p. 125].

The process is repeated until no nonsignificant pair is found. The resulting set of categories of the predictor variable is the best split with respect to that predictor variable. The testing for independence between the pair of categories and the dependent variable depends on the measurement level of target variable (Y). If Y is nominal, one can use either Pearson chi-square test or the likelihood-ratio test. If Y is ordinal, one can only use the likelihood-ratio test. If Y is continuous, one can use F test [*AnswerTree* 3.0 2001, pp. 190].

This process is followed for all predictors. The predictor that has the smallest p-value is selected, and the node is split. The process repeats recursively until no further significant splits can be found.

3. Implementation of CHAID Algorithm

In the analysis of average monthly expenditures of students exclusively on Milka chocolate a regression tree is applied. The output of investigation will permit, on one hand, to establish the height of revenue from selling this brand of chocolate in individual segments (their attractiveness), on the other hand it will allow to prepare a relevant advertising campaign.

But before that can be done, one should distinguish homogeneous groups of students (potential segments) based on their monthly average expenditures on Milka chocolate. That's why in this research the target variable was assumed to be a continuous variable whose values indicate the respondents' declared average monthly expenditures.

The chocolate data set is obtained from a survey carried out among the first-to--fourth-year students of the Faculty of Regional Economics and Tourism in Jelenia Góra at University of Economics in Wroclaw in 2001. Data were collected for 508 cases.

The set of predictors consists of variables such as the type of homogenous chocolate, the type of chocolate with additives, the weight of chocolate, student's gender, year of studies, size of average monthly income, the main source of income and permanent place of residence (tab. 1).

First thing you have to do when you want to use CHAID algorithm is to specify three growing criteria: Maximum Tree Depth, Minimum Number of Cases in Parent Node, and Minimum Number of Cases in Child Node. The values of these parameters in my investigation were established as follows: *Maximum Tree Depth* = = 4, *Parent Node* = 90, *Child Node* = 30. The assumed number of four nodes in the constructed tree allows keeping the balance between clear interpretation of the

Table 1. Some of the predictors included in research

Predictor	Categories	Percentage of population
Type of homogenous chocolate	white	19,5
	unsweet	5,3
	milk	75,2
Weight of chocolate	100g	64,8
	180g	9,4
	200g	18,1
	250g	7,1
	300g	0,6
Gender	female	76,0
	male	24,0
Year of studies	the first	26,8
	the second	26,0
	the third	24,6
	the fourth	22,6
The basic source of income	scholarship	8,7
	employment	32,3
	money from parents	49,0
	different sources	10,0
Permanent place of residence	city (over 100 000 inhabitants)	22,8
	town (50 000 - 100 000 inhabitants)	32,3
	town (below 50 000 inhabitants)	29,5
	village	15,4

class structure obtained and the required level of detail. The second parameter sets the minimum number of cases in parent node at 90; below that value further splitting of the subset is impossible. The value was established on the basis of analysing of the surveyed student population. The last parameter guarantees that if splitting a node would result in a child node with number of cases less than 30, the node will not be split.

The collected data were processed under the CHAID algorithm in accordance with the above parameters and values, and the final result is presented in Fig. 2.

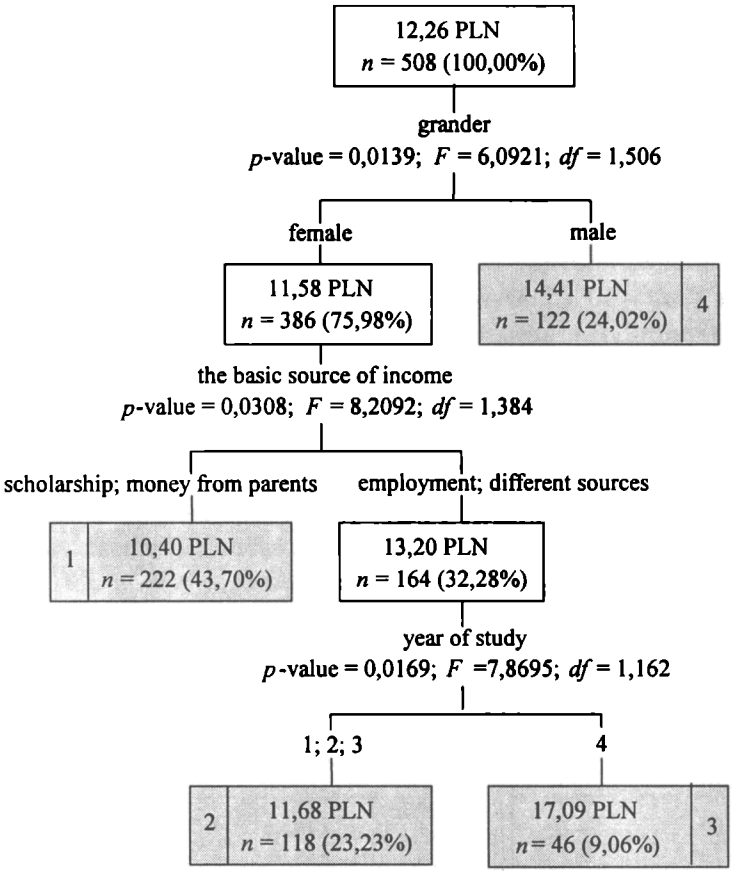

Fig. 2. Results of applying a regression-tree approach

In the overall structure of the tree there are four terminal nodes (segments) of students of the Faculty of Regional Economics and Tourism in Jelenia Góra, which were made by three splits. The split on the first level of chocolate tree is based on gender. By using that variable, most homogeneous subgroups with respect to the target variable were reached.

There is a terminal node of 122 students, whose average monthly expenditures on Milka chocolate amount to 14.41 PLN. The lack of further split is caused by not finding a variable whose *p*-value would be less than predetermined alpha level (0,05). Hence, this group was not split further and the dendrogram terminates along this branch.At the second level, the left branch is divided by the basic source of income. There is a large node including 222 students whose main source of income is scholarship or money from parents. For the subgroup of women whose income comes from own work or different sources, year of study becomes the next signifycant factor. Notice that this process separates two terminal nodes. There are 118 women in the left terminal node. These people are studying in the first, second and third year, and their average monthly expenditures on chocolate amount to 11,68 PLN. The right node (third segment) consists of 46 fourth-year students whose monthly average expenditures on chocolate amount to 17,09 PLN. Both segments seem to be homogenous because each predictor has greater *p*-value than the predetermined alpha level. Synthetic results of application CHAID algorithm are shown in tab. 2.

Table 2. Gains charts based on expected earnings

No	Segment specification	Average monthly expenditures in segment (PLN)	No. of respondents	Expected earnings from segment (PLN)	Percentage of sample
I	Female students, whose basic source of income is scholarship or family support	10,40	222	2308,80	43,7
IV	Male students	14,41	122	1758,02	24,0
II	Female students in first, second or third year of studies whose income comes from employment or other sources	11,68	118	1378,24	23,2
III	Female students in fourth year of studies whose income comes from employment or other sources	17,09	46	786,14	9,1

Analysis with the use of a regression tree defined the characteristics of individual classes. It showed that female students of class III are characterised by the highest average monthly expenditure on Milka chocolate. The lowest level of expenditure, i.e. 10,39 PLN, is typical of female students of class I. Even though the expenditure level is so low there this segment will potentially bring the Milka manufacturer the largest earnings. The reason is the size of this segment (222 students).

From manufacturer's point of view, the expected earnings determine segment attractiveness. With regard to that, the most attractive segment proved to be seg-

ment I (2308,80 PLN). However, if the significant criterion was not only maximum earnings, but also, for instance, increased brand awareness, other – less profitable – segments could also be taken into account provided that they would not bring long--term losses.

Among other predictors, not accounted for in the analysis, the most significant one was the variable of average monthly income of respondents. This is most noticeable in segment I, where 151 respondents (68% of the total) reported monthly income up to 300 PLN, while only three of them reported income over 1000 PLN (mere 1,4%). Other segments are not so much diversified. In segment III, the most numerous group were students with average monthly income over 1000 PLN (over 41%), while the least numerous one were students reporting lowest monthly income level (3 students – 6,5%). In segment IV this proportion was exactly opposite.

Table 3. Indicated diversity in selected segments with regard to average monthly income (%)

	Segment I		Segment II		Segment III		Segment IV	
(0-300]	151	68,0	11	9,3	3	6,5	52	42,6
(300-700]	63	28,4	60	50,8	9	19,6	39	32,0
(700-1000]	5	2,3	33	28,0	15	32,6	22	18,0
(1000-3000]	3	1,4	14	11,9	19	41,3	9	7,4
	222	100	118	100	46	100	122	100

4. Some conclusions

The application of CHAID allows calculating the relative importance of each predictor's impact and establishing how well the predictors, as a group, explain variations in the target variable. Although eight predictors were entered into analysis, besides gender CHAID additionally chose year of study and the basic source of income. These predictors have the largest influence on average monthly expenditures on Milka chocolate. Other predictors were not significantly different with regard to the purpose of the analysis. This means there were no significant relationships between the selected predictor and target variable.

The set of predictors also allowed identification of market attractiveness of individual sectors. Thus, segment III is characterised by the highest level of average monthly respondents' expenditure on Milka chocolate (17,09 PLN). This segment includes female students in fourth year of studies whose income comes from employment or other sources.

Obviously, the obtained results cannot be projected for the total student community of the University of Economics in Wrocław. A parallel analysis should first be carried out at another faculty of the University in order to verify the results. It would also be advisable to repeat the survey to check if respondents' decisions of 2001 were incidental or typical of the population in question.

References

AnswerTree 3.0. (2001), *User's Guide*, SPSS INC, Chicago.

Aluja-Banet T., Nafria E. (1997), *Generalized Impurity Measures and Data Diagnostics in Decision Trees*, [in:] Greenacre M., Blasius J. (ed.), *Visualization of Categorical Data*, Academic Press, San Diego, p. 59-69.

Bao H.T. (2001), *Introduction to Knowledge Discovery and Data Mining*, http://www.netnam.vn/unescocourse/knowlegde/AllChapters.doc.

Berry M.J., Linoff G. (1997), *Data Mining Techniques for Marketing, Sales and Customer Support*, John Wiley & Sons Inc, New York.

Chen M., McNamee L. (1991), *Summary Data Estimation Using Decision Trees*, [in:] Piatetsky-Shapiro G. (ed.), *Knowledge Discovery in Databases*, Massachusetts Institute of Technology. Cambridge, MIT Press, p. 309-323.

Hamilton H., Gurak E., Findlater L., Olive W. (2000), *Course on Knowledge Discovery in Databases*, http://www.cs.uregina.ca/~dbd/cs831/cs831.html.

Kurzydłowski A. (2002), *Drzewa klasyfikacyjne w badaniach marketingowych*, PhD thesis, University of Economics, Wrocław.

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

Magidson J. (1993), *SPSS for Windows CHAID. Release 6.0*, Chicago, SPSS Inc.

ANALIZA PREFERENCJI NABYWCÓW CZEKOLADY MILKA Z WYKORZYSTANIEM DRZEWA O CHARAKTERZE REGRESYJNYM

Streszczenie

W artykule przedstawiono wykorzystanie drzewa klasyfikacyjnego o charakterze regresyjnym (tj. gdy zmienna objaśniana mierzona jest na skali metrycznej) w analizie preferencji. W badaniu wykorzystano algorytm CHAID stosowany w przypadku zbiorów danych, w których przy opisie obiektów, uwzględniane są zarówno zmienne mierzone na skali metrycznej (skala ilorazowa i przedziałowa), jak i zmienne niemetryczne (mierzone na skali porządkowej i nominalnej). Procedura tego algorytmu polega na rekurencyjnym podziale zbioru obiektów na rozłączne podzbiory, do których wyznaczenia jest wykorzystywany określony test. Implementacja miernika oceny jakości podziału uzależniona jest od skali pomiaru zmiennej objaśnianej. W artykule omówiono także zagadnienie identyfikowania atrakcyjności wyodrębnianych klas respondentów (potencjalnych segmentów).

Dr Adam Kurzydlowski jest pracownikiem Katedry Ekonometrii i Informatyki Akademii Ekonomicznej we Wrocławiu.