QUANTITATIVE METHODS IN ECONOMICS
Vol. XV, No. 2, 2014, pp. 403 – 413

CLASSIFICATION OF POLISH HOUSEHOLDS BASED ON THEIR INCOMES BY MEANS OF DECISION TREES

Krzysztof Karpio, Piotr Łukasiewicz, Grzegorz Koszela, Arkadiusz Orłowski
Faculty of Applied Informatics and Mathematics
Warsaw University of Life Sciences – SGGW
e-mails: krzysztof_karpio@sggw.pl, piotr_lukasiewicz@sggw.pl,
grzegorz_koszela@sggw.pl, arkadiusz_orlowski@sggw.pl

Abstract: Classification trees included in SQL Server 2008R2 Analysis Services package have been used to classify Polish households based on their incomes. The analysis has been performed by means of the three algorithms and their effectiveness has been measured. Using the best algorithm a groups of households with the lowest and the largest incomes have been distinguished. The most important attributes describing households with the lowest and the largest incomes were identified and discussed.

Keywords: income distribution, high incomes, classification trees, entropy, SQL Server

INTRODUCTION

In this paper classification trees included in a SSAS package have been used for classification of Polish households based on their incomes. We search for households with the lowest or the highest incomes. The main aim of this paper is to identify attributes of households with the lowest or the highest incomes. The extreme incomes are defined by the first and the last deciles. We are going to answer the following questions: (1) what is a hierarchy of the attributes based on their influence on the classification results (whether a household has the lowest or the highest income); (2) what attributes and their values best describe households with the lowest or the highest incomes? Studying incomes is preceded by investigation of all algorithms and a choice of the most effective one.

A multidimensional analysis of poverty exists widely in literature about incomes and classification of households. This technique was used to search for households with incomes below social minimum by [Kozera et al. 2013]. The

authors used classification trees based on the CART algorithm with Gini index. The same method was adopted by [Anioła et al. 2012] to analyze factors influencing savings management by Polish households. The authors use also a logistic regression and cluster analysis. Decision trees were also used by [Dziechciarz–Duda et al. 2012]. Factors influencing tendency of households to benefit from social care were identified by means of the CHAID algorithm. The study of factors that determine wages in Poland were performed by [Kompa et al. 2013]. The authors utilized the decision trees based on the QUEST algorithm [Lim et al. 2000].

Noteworhty studies of households performer by [Beckel et al. 2013]. The authors use and compare four types of the classifiers: the *k*-Nearest Neighbor classifier, the Linear Discriminant Analysis classifier, the Mahalanobis classifier, and the Support Vector Machine (SVM). The new classification approach is presented by [D'Ambrosio 2001]. It is based on the index of social distance defined in the paper.

The problem of an identification of households with high incomes is not widely present in literature. However this problem seems to be interesting because of a distribution of high incomes differs from the distribution of remaining incomes. It is clearly visible in the case of personal incomes, where there is an exponential distribution for ca. 95% of incomes, while 5% of the highest incomes follow a power law. This behavior was showed for example for incomes in US and UK [Dragulesku et al. 2001] as well as for EU countries [Jagielski et al. 2013]. It is also known that commonly used economics models of incomes distributions e.g. Dagum and Shing-Maddala [Łukasiewicz & Orłowski 2004] do not describe the highest incomes although they are characterized by a high overall precision and fat tails [Łukasiewicz et al. 2012]. In this paper we deal with those issues and study characteristics of households with high incomes.

Classification trees (decision trees) are one of the methods of multidimensional data analysis, whose beginnings were around 60'ties of the XX century [Morgan & Sonquist 1963]. A very fast development of algorithms used in classification trees took place in eighties and nineties [Breiman et al. 1984, Quinlan 1993, Lim et al. 2000]. Nowadays, classification trees are included in many statistical packages and widely used in biology, sociology, medicine and, economy [Chrzanowska et al. 2009]. They are one of the statistical learning methods. One randomly chooses learning sample from a set of objects characterized by independent variables (attributes). Values of dependent variable (classes) must be known for each selected object. A hierarchy of attributes is determined and rules of splitting objects among subsets of homogeneous class composition are being set out. Based on results of the calculations a tree is constructed and its parameters are evaluated. The hierarchical structure is created, which is often presented graphically as an inverted tree with a root, nodes and leaves (terminal nodes). At this moment one can use the tree to classify other

objects into the classes. For more information see [Koronacki & Ćwik 2008, Gatnar 1998].

Classification trees have been implemented in the biggest decision making system in Business Intelligence (BI) infrastructure based on e.g. Oracle Data Mining, SQL Server Analysis Services (SSAS), SAS Enterprise Miner. In this paper classification trees included in a SSAS package have been used for classification of Polish households. There are implemented three algorithms: *Entropy (E), Bayesian with K2 Prior (BK2P), Bayesian Dirichlet Equivalent with Uniform Prior (BDEUP)* into a SSAS package. The first one is based on a Shannon's Entropy, which acts as a measure of classes' homogeneity in nodes and leaves. Algorithms based on entropy are also present in many other statistical packages. The other two algorithms - *BK2P* and *BDEUP* are newer technics, based on a Bayesian analysis [Cooper & Herskovits 1992, Heckerman 1995]. These algorithms are based on the probability theory, which is used to construct probabilistic networks called Bayesian belief networks. Generally, a Bayesian belief network is a pair (G, P), where G is a directed acyclic graph, and P is a conditional probability distribution of vertices of the graph [Jensen 1996, Olbryś 2007]. It might be interesting that each algorithm built in SSAS has three options: *Binary* (nodes are split into two subsets only, binary tree), *Complete* (nodes are split into maximum number of subsets based on the all possible values of the attribute) and, *Both* (during each split of a node a decision is being made, based on an effectiveness, which of the previous options to use).

DATA

In these studies microdata regarding budgets of households in 2008 have been analyzed. There were 37,107 households in the data set. Households were classified based on their 10 attributes (independent variables) in three groups: 1. variables describing a head of the household (a person with the biggest income); 2. variables describing a household as a whole; 3. variables describing location of a household. All attributes and their possible values are summarized in the Table 1. At this moment we only point out that the socio-economic group is defined as the main source of household's income.

A majority of households are employee's households (about 50%), on the other hand the smallest group consists of households maintained from non-earned sources (about 3.5%).

Table 1. The attributes of the households and their values

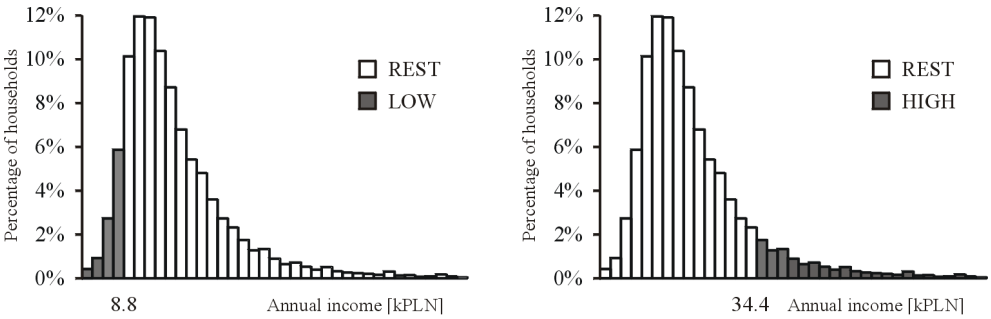
Group	Attribute	Attribute values
1	SEX (Sex of a household's head)	male (1), female (2)
	EDU (Education of a household's head)	tertiary (1), post-secondary (2), upper secondary vocational (3), upper secondary general (4), basic vocational (5), lower secondary (6), primary (7), no formal education (8)
	AGE (Age of a household's head)	16 - 102 (years)
	EGROUP (Economic group of a household)	employed in manual labor position (11), employed in non-manual labor position (12), farmer (2), self-employed (3), retired (41), pensioner (42), maintained from non-earned sources (5)
2	FTYPE (Family type)	marriage without children (1), marriage with 1 to 4 children (2 - 5), mother with children (6), father with children (7), marriage with children and other persons (8), mother with children and other persons (9), father with children and other persons (10), other persons with children (11), singles (12), others (13)
	NPER (Number of persons in a household)	1 – 15
	NCHIL (Number of children)	0 – 9
	NEAR (Number of earners)	1 – 10
3	PRES (Place of residence)	town >= 500 (1), town 200 - 499 (2), town 100 - 199 (3), town 20 - 99 (4), town < 20 (5), village (6) (thousands of residents)
	VOI (Voivodeship)	1 – 16

Source: own study

An annual income of a household is a dependent variable. We study household's income per number of earners, not per number of persons. We consider two cases with two distinct income classes:
(i) variable LOW-REST: LOW (10% of households with the lowest incomes), REST (remaining households);
(ii) variable REST-HIGH: HIGH (10% of households with the highest incomes), REST (remaining households).

Deciles groups (LOW, HIGH) and income limits are shown in the Figure 1.

Figure 1. Income distribution in Poland in 2008 and deciles groups



Source: own study

ANALYSIS OF SSAS ALGORITHMS

The full classification trees have been built based on the 30% random sample for each of the algorithms and options. The remaining part of the data was a validation set. The trees were constructed for each of the dependent variable. The effectiveness of the trees was measured by two parameters obtained for the validation set. The first one is a percentage of properly classified objects. The second one is a percentage properly classified LOW or HIGH objects. Results are in Table 2.

Table 2. The attributes of the households and their values

	Algorythm		
	Entropy	BK2P	BDEUP
Mode	Variable LOW-REST		
Binary	90.6 (9.9)	90.4 (7.0)	90.5 (9.2)
Complete	90.3 (13.8)	90.2 (11.5)	90.2 (11.8)
Both	90.7 (17.3)	90.3 (6.3)	90.5 (8.8)
Mode	Variable REST-HIGH		
Binary	90.9 (17.7)	90.6 (16.7)	90.6 (16.7)
Complete	90.4 (13.0)	90.2 (6.7)	90.0 (0.0)
Both	91.2 (28.9)	90.7 (15.6)	90.7 (15.6)

Source: own study

The first parameter is high (exceeds 90%) and has roughly the same values for all algorithms and options. This behavior is a result of proper classifications of REST objects. Thus the effectiveness of the trees was determined based on the second parameter. The best results were obtained for the Entropy algorithm with Both option for LOW-REST (17.3%) as well as for REST-HIGH (28.9%) variables. The tree for the second variable has a significantly better effectiveness that for the first one. It seems to be related to a shape the income distribution – high incomes, located in the tail of the distribution have much higher dispersion than low incomes. The algorithms based on bayesian networks have relatively low

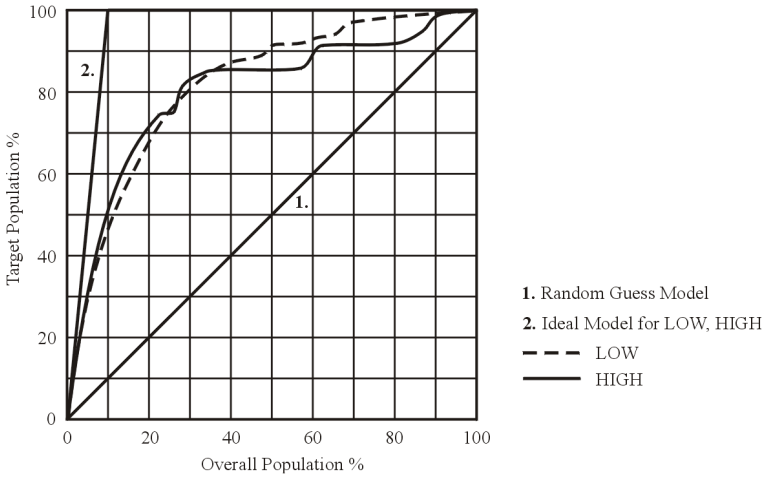
effectiveness. The second parameter has similar values for Binary and Both options, what has not been observed for the Entropy algorithm. All the algorithms with Complete option (bayesian networks in particular) have higher effectiveness for LOW than for HIGH objects.

The results for the Entropy algorithm were compared to the effectiveness of binary entropy based trees implemented in SAS Enterprise Miner package. The first parameter was about 90.6%, the second one was 11.3% for LOW-REST and 19.0% for REST-HIGH variable. The results of SAS are slightly better than for SSAS. However, in SSAS for entropy based trees one can obtain significant improvement of effectiveness by using the Both option.

ANALYSIS OF INCOMES

The following analysis concerns entropy based trees with Both option. Because of a large complexity of obtained trees they will not be presented in graphical form. We will present their global characteristic, attribute rank and characteristic of selected nodes and leaves. The shares of LOW and HIGH objects in a sample as a function of sample size are presented in Figure 2. The random and ideal models are also added to the plot.

Figure 2. Lift chart for LOW-REST and REST-HIGH trees



Source: own study

Using the constructed trees we analyze hypothetic sample data being a part of the data set (horizontal axis). The sample data set will contain percentages of LOW or HIGH objects indicated on the vertical axis. The effectiveness of identification is slightly better for HIGH than for LOW objects for the sample sizes from 0% to 25%. Above 25% the situation is reversed but from 30% increase of sample size causes only a small increase of the effectiveness. Both models become

less effective than random model. The best effectiveness is for small sample sizes. The ideal model reaches 100% for a sample size 10%, what means that all the LOW or HIGH objects are identified. On the other hand random model gives 10%. The trees identify 46% of LOW objects and 50% of HIGH objects in a 10% sample.

During the next stage of analysis the trees were trimmed by setting a node minimal support to 50. The attributes of objects ordered based on their significance are in Table 3. We observe the same three most important attributes for *LOW-REST* and *REST-HIGH* variables. The *EGROUP (Economic group of a household)* is the most important factor in the LOW or HIGH subgroup membership. Based on this attribute the groups of objects are split only on the first levels of the trees (complete split). The remaining two most important attributes are: *EDU (Education of a household's head)* and *NEAR (Number of earners)*. The attributes *EDU* and *NCHIL (Number of children)* have bigger influence on the identification of HIGH than LOW objects. The attributes *PRES (Place of residence)* and *FTYPE (Family type)* are more important in the case of the *LOW-REST* variable. The attribute *AGE (Age of a household's head)* is insignificant for both variables.

After the trimming the tree for *LOW-REST* variable had 58 leaves on 6 levels and the tree for *REST-HIGH* variable had 66 leaves on 8 levels. The majority of the nodes and leaves were of the REST type. A few but the most interested nodes and leaves with a majority of HIGH or LOW objects are described in Tables 4 and 5.

Table 3. The attributes of the households and their values

Variable *LOW-REST*		Variable *REST-HIGH*	
Attribute	Tree level	Attribute	Tree level
EGROUP	1	*EGROUP*	1
NEAR	2, 3, 4, 5	*EDU*	2, 3, 5, 6
PRES	2, 3, 5	*NEAR*	2, 3, 4, 5
EDU	3, 4, 5, 6	*NCHIL*	3, 4, 7
FTYPE	3, 4	*VOI*	3, 5
VOI	4	*SEX*	3, 4, 5, 6, 7
SEX	4, 5	*PRES*	2, 4, 7, 8
NPER	5	*NPER*	4, 6
NCHIL	3	*FTYPE*	6, 7
AGE	–	*AGE*	–

Source: own study

Table 4. Nodes and leaves with majority of HIGH objects

Node / Leaf	Attribute value	% HIGH
N 1	*EGROUP* = 3 & *NEAR* = 1 & *EDU* = 1	70.2
L 1.1	& *LOS* = 1	59.3
L 1.2	& *LOS* ≠ 1	80.8
N 2	*EGROUP* = 12 & *EDU* = 1 & *VOI* = 14 & *PRES* = 1	68.3
L 2.1	& *NEAR* = 1	83.8
N 2.2	& *NEAR* ≠ 1	60.8
L 2.2.1	& *SEX* ≠ 1	50.6
N 2.2.2	& *SEX* = 1	67.7
L 2.2.2.1	& *NCHIL* = 0	57.6
L 2.2.2.2	& *NCHIL* ≠ 0	76.8
L 3	*EGROUP* = 12 & *EDU* = 1 & *VOI* = 14 & *PRES* ≠ 1 & *SEX* = 1	50.1
L 4	*EGROUP* = 12 & *EDU* = 1 & *VOI* = 24 & *NEAR* = 1	50.1
L 5	*EGROUP* = 12 & *EDU* = 1 & *VOI* = 22 & *SEX* = 1	56.2

Source: own study

Table 5. Nodes and leaves with majority of LOW objects

Node / Leaf	Attribute value	% LOW
L 1	*EGROUP* = 5 & *PRES* = 5	54.3
N 2	*EGROUP* = 5 & *PRES* = 6	59.2
N 2.1	& *NEAR* ≠ 2	65.1
L 2.1.1	& *NEAR* = 1	56.4
L 2.1.2	& *NEAR* ≠ 1	73.4
N 3	*EGROUP* = 5 & *PRES* = 4	58.9
L 3.1	& *NCHIL* = 0	52.1
L 3.2	& *NCHIL* ≠ 0	68.1
L 4	*EGROUP* = 42 & *NEAR* ≠ 1 & *PRES* = 6	50.1

Source: own study

In the case of the *REST-HIGH* variable the first group (N 1) was obtained for households with family heads having a tertiary education, self-employed and being the only earner in the household (N 1). More than 70% of households in this group have a high income. We can distinguish two household's models: singles (L 1.1) and households with at least two members, but one earner (L 1.2).

The second group (N 2) are households with family heads employed in non-manual labor position, having a tertiary education and living in Warsaw (*VOI* = 14: mazowieckie voivodeship & *PRES* = 1). About 68% of households in this group have HIGH income. The split of this group into the smaller subgroups is showed in the Table 4. The groups of households characterized by *NEAR* = 1 or *SEX* = 1 (male) are created.

The L 3 group are households with male family heads employed in non-manual labor position, having a tertiary education, living in mazowieckie outside Warsaw. The remaining groups of households (L 4, L 5) are showed in Table 4. In this case we have the same split criteria: household's head employed in a non-manual labor position, has tertiary education and is the only earner or is a male. But those households are located in pomorskie (VOI = 22) and śląskie (VOI = 24) voivodeships. Note at this point that mazowieckie, pomorskie and, śląskie are the voivodeships with the highest mean income in Poland.

In the case of the *LOW-REST* variable there are tree distinguished groups of households with a predominance of LOW objects (L 1, N 2, N 3, see Table 5). The households are maintained from non-earned sources and are located in small towns: 20k–99k (*PRES* = 4), less than 20k (*PRES* = 5) and villages (PRES = 6). The split of N 2 and N 3 nodes into leaves is shown in Table 5. The forth group (L 4) are households of pensioners living in villages who do not maintain their households by themselves.

SUMMARY

In this paper a SQL Server 2008R2 Analysis Services package has been used for the classification of households based on their income. Households with the lowest (LOW) or highest (HIGH) incomes were identified. The effectiveness of the three classification tree algorithms was investigated. For each algorithm the results for three options were compared and discussed. The most effective was algorithm based on a Shannon's entropy with Both option. The effectiveness of identification of HIGH objects (28.9%) turned out to be better than for LOW objects (17.3%). The algorithms based on bayesian networks had lower effectiveness than other methods.

The Economic group of a household (*EGROUP*) was a major classification attribute for both dependent variables. Next most important attributes were: Education of a household's head (*EDU*), Number of earners (*NEAR*), Number of children (*NCHIL*) and Voivodeship (*VOI*) for *REST-HIGH* variable and Number of earners (*NEAR*), Place of residence (PRES), Education of a household's head (*EDU*) and Family type (*FTYP*) for *LOW-REST* variable. Similarly, the high importance of an economic group of households and education was observed by [Kozera et al. 2013]. However the authors also reported a number of persons in the households as an important attribute, what was not a case in our studies. In the other studies [Anioła et al. 2012]: education, family type and a place of residence were proved to be most important attributes.

Despite of relatively low effectiveness of obtained classification trees one distinguished some groups of households with a majority of LOW or HIGH objects. Groups of the biggest share of HIGH incomes are characterized by: incomes coming from a self-employment or non-manual labor position, tertiary

education of a household's head, one earner, mazowieckie, pomorskie or śląskie voivodeships. On the other hand groups of households with the highest share of LOW incomes are characterized by: incomes from non-earned sources or pension, location in small towns or villages.

Noteworthy is high significance of the Number of earners (*NEAR*) attribute. The studied incomes are the incomes per number of earners, so the *NEAR* attribute shall not be so important. Households in the three groups with a majority of HIGH incomes (N 1, L 2.1, L 4) are maintained only by their family heads. It seems to be a characteristic for at least a part of households with high incomes. On the other hand many of the households with low incomes are maintained by more than one person (L 2.1.1, L 4). However a quantitative evaluation of significance of *NEAR* attribute requires further studies.

REFERENCES

Anioła P., Gołaś Z. Zastosowanie wielowymiarowych metod statystycznych w typologii strategii oszczędnościowych gospodarstw domowych w Polsce, NBP, 2012.

Beckel Ch., Sadamori L., Santini S. (2013) Automatic Socio–Economic Classification of Households Using Electricity Consumption Data, e-Energy '13 Proceedings of the fourth International Conference on Future Energy Systems, 75–86.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.

Chrzanowska M., Alfaro E., Witkowska D. (2009) The individual borrowers recognition: Single and ensemble trees, Expert Systems with Applications 36, 6409-6414.

Cooper G.F., Herskovits E. (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data, Machine Learning 9, 309-347.

D'Ambrosio C. (2001) Household Characteristics and the Distribution of Income in Italy: an Application of Social Distance Measures, Review of Income and Wealth 47, No 1, 43–64.

Dragulescu A.A., Yakovenko V.M. (2001) Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States, Physica A, 299, 213–221.

Dziechciarz-Duda M., Król A., Przybysz K. (2012) Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych, Taksonomia Nr 19 (242), 144–152.

Gatnar E. Symboliczne metody klasyfikacji danych, PWN, Warszawa, 1998.

Heckerman D. (1997) Bayesian Networks for Data Mining, Data Mining and Knowledge Discovery 1, 79-119.

Heckerman D., Geiger D., Chickering D.M. (1995) Learning Bayesian networks: the combination of knowledge and statistical data, Machine Learning 20, 197-243.

Jagielski M., Kutner R. (2013) Modelling of income distribution in the European Union with the Fokker–Planck equation, Physica A 392, 2130–2138.

Kompa K., Witkowska D. (2013) Application of Classification Trees to Analyze Income Distribution in Poland, Quantitative Methods in Economics, Vol. XIV, No. 1, 265–275.

Koronacki J., Ćwik J. Statystyczne systemy uczące się, Exit, Warszawa, 2008.

Kozera A., Stanisławska J., Wysocki F. (2013) Klasyfikacja gospodarstw domowych ze względu na stopień zaspokojenia ich potrzeb mierzony kategorią minimum socjalnego, Marketing i Rynek 11, 31-38.

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

Łukasiewicz P., Karpio K., Orłowski A. (2012) The Models of Personal Incomes in USA, Acta Physica Polonica A, Vol. 121, B-82–B-85 (2012).

Łukasiewicz P., Orłowski A. (2004) Probabilistic models of income distributions, Physica A 344, 146.

Morgan J.N., Sonquist J.A. (1963) Problems in the Analysis of Survey Data, and a Proposal, Journal of the American Statistical Association 58 (302), 415-434.

Olbryś J. (2007) Sieć bayesowska jako narzedzie pozyskiwania wiedzy z ekonomicznej bazy danych, Zeszyty Naukowe Politechniki Białostockiej, Informatyka 2, 93-107.

Quinlan J. R. (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos.