Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 176 (2020) 3418–3427

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Application of logistic regression models to assess household financial decisions regarding debt

Agnieszka Strzelecka [a]*, Agnieszka Kurdyś-Kujawska[a], Danuta Zawadzka[a]

[a]*Koszalin University of Technology, Faculty of Economic Science, Department of Finance, Kwiatkowskiego 6e, 75-343 Koszalin, Poland*

Abstract

The paper presents methodological assumptions regarding the logistic regression model and an example of using this research method to evaluate financial decisions taken by households. The aim of the study was to identify and evaluate socio-economic factors determining the debt of households in Central Pomerania using a logistic regression model. The source of data was the results of a survey conducted among 1,000 households of Central Pomerania (Poland). The obtained results prove that the following factors related to the socio-economic characteristics of households: economic education of the head of the household, developmental phase of the household, socio-economic type of the household had a statistically significant positive impact on the likelihood of Central Pomeranian households using external sources of financing: household income and household income. These factors increase the likelihood of households using external sources of financing. In turn, a statistically significant negative impact on the analyzed phenomenon had the household income diversification and the age of the household head.

Keywords: household debt; logistic regression; susceptibility to debt; factors affecting debt; Central Pomerania, Poland

* Corresponding author. Tel.: +48-94-3439-216
*E-mail address:*agnieszka.strzelecka@tu.koszalin.pl

10.1016/j.procs.2020.09.055

1. Introduction

The household, being the basic economic unit, each day undertakes a number of actions and decisions striving to meet the common and individual needs of its members. Among the financial decisions of households there are those that relate to the use of external sources of financing - credit decisions [6]. Household debt is one of the research aspects undertaken in the area of household financial decisions. In recent years, an increase in interest in this issue has been observed, especially in view of the observed effects of the global financial crisis on households [12][22]. The popularity of this research issue is indicated by numerous scientific papers that focus primarily on such issues as: (1) the level of household debt [2] [5] [11] [13] [14] [21] [25] [28]; (2) differentiation of debt among households depending on their socio-demographic characteristics [9] [13] [14] [21]; (3) forms, motives and goals of households' incurring obligations [13] [33]. One of the research areas are also factors determining household debt. The results of previous research in this area prove that the following factors relating to the socio-economic characteristics of the household, such as income, age of the head of the household, sex of the head of the household, education of the head of the household may have an impact on the households s indebtedness , household development phase, size and composition of the household, location of the household and socioeconomic type of the household [1] [7] [8] [11] [18] [19] [22] [31] [32]. This work is part of the research on household debt determinants.

The aim of the research is to identify and assess the socio-economic factors determining the debt of households in Central Pomerania using a logistic regression model.

2. Method

Logistic regression (also known as logit regression or logit model) is a widely used multidimensional method for modeling dichotomous results (see e.g. [4] [17] [24] [34]). It is suitable for models covering decision-making issues, which is why it is often used in statistical analyzes appearing in the economics and finance literature. The regression model serves two purposes: (1) it can predict the result variable for new values of predictive variables, and (2) it can help answer questions about the studied phenomenon, because the coefficient of each predictive variable clearly describes the relative contribution of this variable to the result variable, automatically controlling the influence of other predictive variables [3]. The logistic regression model allows to examine the influence of many independent variables $X_{1,\dots,}X_k$the dependent variable Y. The variable Y takes only two values and is dichotomous. These values are coded as 1 and 0. A value of the variable Y equal to 1 means that a desired event occurs. Otherwise, when an adverse event occurs, this variable assumes a value of 0 [16]. The regression analysis process allows you to determine which factors matter most, which you can ignore, and how they affect each other.

Logit regression uses a logistic function for the description, whose output values are in the range $(0;1)$ and which creates a curve in the formation of the letter S. Three stages of changing the value of the function can be distinguished: initially, up to a certain threshold, they practically do not change the probability, after reaching the threshold value the probability increases to one and stays at this level [29]. The analytical form of the logistic function used in logistic regression is defined by the equation [30]:

$$f(z)=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}},\quad z\in R \tag{1}$$

The logistic regression model for the dichotomous variable Y determines the conditional probability of adopting the distinguished value by this variable and is expressed by the following relationship [26]:

$$P(Y=1/X_1,\dots,X_k)=\frac{e^{\alpha_0+\alpha_1X_1+\dots+\alpha_kX_k}}{1+e^{\alpha_0+\alpha_1X_1+\dots+\alpha_kX_k}}$$

where: $\alpha_0,\alpha_1,\dots,\alpha_k$ are the model parameters, $X_{1,\dots,}X_k$ independent variables that can be both qualitative and quantitative.

Logistic regression model coefficients are sought for by the method of maximum likelihood and by the generalized method of least squares. In practice, the most credibility method is most commonly used. The calculation algorithm of the maximum likelihood method is based on multiple estimation of all regression coefficients so as to maximize the probability of obtaining such results that were achieved in the studied sample [16]. There are two separate formulas for estimating the maximum likelihood ratio [20]. This is the unconditional method and the conditional method.

Due to the non-linearity of the model relative to independent variables and parameters, the logistic model is transformed into a linear model. To this end, the concept of Odds Ratio is introduced, which is the ratio of the probability of occurrence of a specific event to the probability that this case will not occur, i.e.

$$\frac{P(Y=1/X_1,...,X_k)}{1-P(Y=1/X_1,...,X_k)} = \frac{e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k}}{1+e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k}} : \frac{1}{1+e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k}} = e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k} \tag{3}$$

The odds ratio expresses, therefore, how many times the probability that an event will occur increases or decreases if a unit change of the independent variable occurs (with determined values of other independent variables). This ratio is calculated as follows:

$$S(A) = \frac{p(A)}{p(noA)} = \frac{p(A)}{1-p(A)}. \tag{4}$$

The odds ratio refers to the situation when the occurrence of a given phenomenon is studied in two independent groups. It is expressed by the ratio of the chance of this phenomenon occurring in group A or $O(A)$, to the chance of this phenomenon occurring in group B or $O(B)$. The formula for the odds ratio is:

$$OR = \frac{O(A)}{O(B)} = \frac{p(A)}{1-p(A)} / \frac{p(B)}{1-p(B)} \tag{5}$$

where: $p(A)$- probability of occurrence of the event (value of 1 dependent variable) in category A (value of 1 independent variable); $p(B)$ - probability of occurrence of the event (value of 1 dependent variable) in category B of observation (value of 0 of independent variable).

It is assumed that if:

$OR > 1$, then X stimulates the event to occur. OR shows how much the probability of 1 in the dependent variable increases when the predictor value increases by one unit;

$OR < 1$, then X destimulates the occurrence of the event. OR shows how much the probability of 1 in a dependent variable decreases when the predictor value increases by one unit;

$OR = 1$, then X does not affect the occurrence of the event.

In the odds ratio analysis, the confidence interval (CI) is determined for each predictor. This is a numerical range which, with a 1-α probability determined by the researcher and defined as the confidence level, will contain an unknown value of the estimated population parameter.

After estimating the parameters of the logistic regression model, the theoretical values of the Y variable can be determined according to the standard forecast principle:

$$\hat{y}_i = \begin{cases} 1, & \text{gdy } 0{,}5 < \hat{p}_i \le 1 \\ 0, & \text{gdy } 0 < \hat{p}_i \le 0{,}5 \end{cases}$$

where: $\hat{p}_i$ theoretical probabilities obtained from a logistic regression model estimated on the basis of a random sample.

In a situation where the sample is unbalanced, i.e. in which the number of ones significantly differs from the number of zeros, a modification of the standard rule can be used to forecast theoretical values and the forecasts according to the principle of the optimal limit value α can be calculated:

$$\hat{y}_i = \begin{cases} 1, & \text{gdy } \alpha < \hat{p}_i \leq 1 \\ 0, & \text{gdy } 0 < \hat{p}_i \leq \alpha \end{cases}$$

The α limit is set as the share of ones in the sample. You can then assess the correctness of the estimated model by counting the correct and incorrectly classified cases (Table 1).

Table 1. Case classification matrix

Classification of objects based on the logit model	Actual belonging of objects		Sum
	$y_i = 1$	$y_i = 0$	
$\hat{y}_i = 1$	n_{11}	n_{12}	$n_{1\bullet}$
$\hat{y}_i = 0$	n_{21}	n_{22}	$n_{2\bullet}$
Suma	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Source: own study based on [10].

To assess the degree of adjustment of the logistic regression model to empirical data, the R^2 counting measure is used, which takes values from the range $\langle 0,1 \rangle$ defined as follows [26]:

$$R^2_{count} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \ . \tag{6}$$

The closer one value of this measure, the better the logistic model will fit to the empirical data of the studied phenomenon. R^2_{count} indicates the percentage of correctly classified cases. The model works well in forecasting the studied phenomenon when. This means that the classification based on the model is better than random. The quality of the built logistic regression model can also be assessed by other measures, e.g.

$$Pseudo\ \ R^2_{Pseudo} = 1 - \frac{LnL_{FM}}{LnL_0} \tag{7}$$

where: L_{FM} - maximum of the likelihood function of the model with all variables, L_0 - maximum of the likelihood function of the model containing only the free word. It takes values from the range $\langle 0,1 \rangle$. Values close to 1 indicate a very good fit of the model, while values close to 0 indicate no match. Because the R^2_{Pseudo} factor does not take the value 1 and is sensitive to the number of variables in the model, its correct value is determined: $R^2_{Nagelkerke}$ and $R^2_{Cox\text{-}Snell}$.

$$R^2_{Nagelkerke} = \frac{1 - e^{-(2/n)(LnL_{FM} - LnL_0)}}{1 - e^{(2/n)LnL_0}}, \ R^2_{Cox\text{-}Snell} = 1 - e^{-2(LnL_{FM} - LnL_0)/n} \tag{8}$$

The Hosmer-Lemeshow test [15], can be used to assess the quality of the built logistic regression model, which compares the observed number of occurrences in a given subgroup of objects with a distinct O_g feature and the expected counts E_g of a distinguished value for different data subgroups. If O_g and E_g are close enough, then it can be assumed that a well-fitted model was built. Usually the observations are divided into G subgroups for calculations, using e.g. deciles. The hypotheses in the test take the following form:

$H_0 : O_g = E_g$ for all categories,

$H_1 : O_g \neq E_g$ at least for one category.

The value of the test statistics is determined as follows:

$$H = \sum_{g=1}^{G} \frac{\left(O_g - E_g\right)^2}{E_g\left(1 - \frac{E_g}{N_g}\right)} \tag{9}$$

where: N_g number of observations in group g, $g \in \{1,...,G\}$.

This statistic has an asymptotically distribution χ^2 with G-2 degrees of freedom.

In addition, the ROC curve is used to assess the quality of the classification obtained by logistic regression. The ROC curve is a visual measure of the classifier performance. For logistic regression, the standard classification table is a 2×2 matrix that includes decision classes predicted by the model and actually observed [16]. Using the percentage of positive data points that are correctly considered positive and the percentage of negative data points that are incorrectly considered positive, graphics are generated that show the trade-off between the speed at which something can be correctly predicted and the speed of incorrect prediction. An important element of table analysis is determining the percentages of three parameters: accuracy (ACC), sensitivity and specificity. The first determines what percentage of observations are cases correctly classified positively or negatively by the model. The accuracy of the ability to classify models can be measured as the area under the curve (AUC). Sensitivity describes the ability to detect units having a distinguished state [23]:

$$sensitivity = \frac{n_{11}}{n_{11} + n_{21}} \tag{10}$$

However, specificity determines the percentage of cases correctly, but negatively classified by the model among all cases negatively observed [23]:

$$specificity = \frac{n_{22}}{n_{12} + n_{22}} \tag{11}$$

The ROC curve is created by connecting points in a Cartesian coordinate system with coordinates (1-*specificity, sensitivity*). The resulting ROC curve, and in particular the area under it (AUC), illustrates the classification quality of the model. The area under the ROC curve takes values from the range $[0,1]$. The classification quality of the model is good when the area under the ROC curve is much larger than the area under the straight line $x = y$, therefore greater than 0.5. You can verify the hypothesis:

$$H_0 : AUC = 0{,}5;$$
$$H_1 : AUC \neq 0{,}5.$$

Test statistics are used to verify these hypotheses:

$$U = \frac{AUC - 0{,}5}{SE_{0,5}} \tag{12}$$

where: $SE_{0,5} = \sqrt{\frac{n_{1\bullet} + n_{2\bullet} + 1}{12 \cdot n_{1\bullet} \cdot n_{2\bullet}}}$, $n_{1\bullet}$ - size of units having the distinguished characteristic, $n_{2\bullet}$ - size of units not having the distinguished characteristic.

If the curve coincides with the $x = y$ diagonal, it means that the model is not involved in the analysis, because the classifications made by the model are as good as random decisions. The more the curve is closer to the left quadrant, the better results are achieved in the model. A high level of sensitivity (close to 1) means that the model correctly identifies cases, and a low value of 1-S proves that only a few negative cases are classified as positive.

If, as the value of the diagnostic variable increases, the chances of occurrence of the studied phenomenon increase or decrease, the so-called optimal cut-off point, i.e. the predicted probability for which the dependent variable best divides the community into groups: in which the studied phenomenon occurs and does not occur. One way of determining the optimal cut-off point is the value of the diagnostic variable at which the expression: Sensitivity $-m \cdot$ (Specificity $-$ 1) reaches the minimum, where m - the slope of the tangent to the ROC curve [35].

3. Empirical results

3.1. Data Sources

The source of empirical data was the results of a survey conducted among 1,000 households of Central Pomerania in Poland. The study was conducted in the second quarter of 2019, using the direct survey technique. In the course of the research, the number of 746 correctly completed questionnaires was obtained (return rate of 74.6%). The scope of the study covered 2018.

The dominant group of households were entities living in rural areas (46.2%), followed by households from cities with more than 50 thousand inhabitants (27.2%) and smaller towns (up to 50 thousand inhabitants - 26.5%). Considering the developmental phase of the household, it was found that the most numerous groups were marriages / partnerships with dependent children (47.1%). Half of the group were three-person households. Among the entities included in the analysis, those in which the head of the household was predominantly male (62.2%), the average age of the household head was 45 years. Almost 65% of the population were individuals in which the head of the household had at most secondary education. 27.9% of respondents declared higher education

For 61.1% of respondents, the basic source of income was remuneration from employment. Next, the respondents indicated: income from non-agricultural business activity (14.5%), retirement and disability pensions (13.5%) and income from agricultural activity (9.5%). It was also found that 17.6% of the units included in the study had an average monthly net income per capita in the household not exceeding PLN 1,000. In the case of over 1/3 of respondents (36.6%), the amount of income category considered was higher than PLN 2,000 / person. 50.9% of the analyzed households were characterized by a steady increase in income in 2004-2018, while 61.4% of entities were characterized by a steady increase in expenditure in this period. On average, 1/3 of expenditure (34.5%) was allocated to the purchase of food and non-alcoholic beverages. Over half of the population (50.9%) diversified their sources of income.

Among the surveyed household group of Central Pomerania, 34.3% of entities were indebted, allocating an average of 17.6% of income to repay liabilities. The debt structure of the entities surveyed was dominated by mortgage loans, which constituted on average 40.7% of all liabilities. On average, loans from institutions other than banks accounted for 5.1% of all liabilities, while loans from family / friends - 3%.

3.2. Identification of factors affecting household debt - logistic regression model

Logistic regression model was used to empirically verify the factors affecting the debt inclination of the surveyed households of Central Pomerania. Household debt (zero-one variable) was assumed as the dependent variable. In the case when the surveyed household was in debt in 2018, the variable takes the value 1 (256 cases), in the opposite case - the value 0 (490 cases). The explanatory variables were selected on the basis of literature studies. For the assessment

of the probability examined, 9 factors related to the socio-economic characteristics of Central Pomeranian households were adopted and the following 30 independent variables were assigned to them: **(1) age: x_1** - age of the household head [years]; **(2) gender: x_2** - household head gender [female = 1, male = 0]; **(3) education**: because the first of the variables considered referring to the level of household education has several variants, a reference group was distinguished (for variables x_3-x_6) - households with at most basic education; **x_3** - basic vocational [yes = 1, no = 0]; **x_4** - medium [yes = 1, no = 0]; **x_5**- college [yes = 1, no = 0]; **x_6** - higher [yes = 1, no = 0]; **x_7** - economic education of the household head [yes = 1, no = 0]; (4) number and composition of the household: **x_8** - number of members of the household [persons]; **x_9** - share of dependent children [%]; **(5) economic activity of household members: x_{10}** - share of household members engaged in gainful employment in the total number of people in the household [%]; **(6) household development phase:** because the variable in question has several variants, a reference group has been identified - the farm of a lonely young person; **x_{11}** - young marriage / partnership without a child [yes = 1, no = 0]; **x_{12}** - a lonely person with a dependent child / children [yes = 1, no = 0]; **x_{13}** - marriage / partnership with dependent children [yes = 1, no = 0]; **x_{14}** - marriage / partnership in middle or older age without dependent children [yes = 1, no = 0]; **x_{15}** - the management of a lonely elderly person [yes = 1, no = 0]; **x_{16}** - other than the development phase (including multi-generational households) [yes = 1, no = 0]; **(7) socioeconomic type of household:** because the variable in question has several variants, a reference group has been identified - the household of employees; **x_{17}** - farmers [yes = 1, no = 0]; **x_{18}** - self-employed [yes = 1, no = 0]; **x_{19}** - pensioners [yes = 1, no = 0]; **x_{20}** - other (including those living on social benefits) [yes = 1, no = 0]; **(8) household location:** because the variable in question has several variants, a reference group has been identified - a village household; **x_{21}** - city up to 50,000 residents [yes = 1, no = 0]; **x_{22}** - a city of over 50,000 residents [yes = 1, no = 0]; **(9) household income:** since the first of the variables adopted for household income has several variants, a reference group has been separated (for variables x_{23}-x_{25}) - average monthly net income per person in a household up to PLN 1,000; **x_{23}** - from 1001 to 1500 PLN / person [yes = 1, no = 0]; **x_{24}** - from 1501 to 2000 PLN / person [yes = 1, no = 0]; **x_{25}** - above PLN 2,000 / person [yes = 1, no = 0]; **x_{26}** - a steady increase in income since 2004 [yes = 1, no = 0]; **x_{27}** - diversification of sources of income in a household - more than one type of source of income [yes = 1, no = 0]; **x_{28}** - a steady increase in expenses since 2004 [yes = 1, no = 0]; **x_{29}** - share of expenditure on food and non-alcoholic beverages in household expenses [%]; **x_{30}** - have you ever encountered difficulties in accessing financial services or products [yes = 1, no = 0].

Using the backward elimination method, based on the Akaike information criterion (AIC), further predictors were eliminated from the output model and changes in the value of criteria used to assess the quality of the model were assessed. Finally, 13 independent variables, whose impact on the probability was not statistically significant, were eliminated. At each stage, an improvement in the value of the adopted fit measure (reduction in the AIC value) was observed. 17 predictors remained in the final model (Table 2).

Table 2. Results of estimation of model parametres – final model

Variable	Variable parameter	Standard error	z Wald test	Significance level	Odds ratio
x_1	-0.047	0.011	-4.448	<0,001*	0.954
x_7	0.605	0.214	2.823	0.005*	1.832
Household development phase					
x_{11}	0.795	0.359	2.213	0.027**	2.214
x_{12}	1.543	0.570	2.705	0.007*	4.677
x_{13}	1.941	0.360	5.386	<0,001*	6.966
x_{14}	2.026	0.497	4.075	<0,001*	7.587
x_{15}	1.208	0.827	1.461	0.144	3.346
x_{16}	2.050	0.478	4.292	<0,001*	7.770
Socio-economic type					
x_{17}	0.603	0.289	2.084	0.037**	1.827
x_{18}	0.640	0.241	2.655	0.008*	1.897
x_{19}	-0.297	0.408	-0.729	0.466	0.743
x_{20}	-1.070	1.098	-0.975	0.330	0.343
Household Income					
x_{23}	0.002	0.277	0.006	0.995	1.002
x_{24}	0.516	0.287	1.797	0.072***	1.675

x_{25}	0.593	0.273	2.172	0.030**	1.809
x_{27}	-0.295	0.179	-1.647	0.100***	0.745
x_{30}	0.806	0.244	3.304	0.001*	2.239
Intercept	-0.735	0.475	-1.549	0.121	0.480
AIC = 889,56					
Cox-Snell R2 = 0,132					
Nagelkerke R2 = 0,183					
count R2 = 0,705					
AUC = 0,716					
LR = 105,97 (df = 17; p = <0,001)					

* statistically significant variable at the significance level of 1%
** statistically significant variable at the 5% significance level
*** statistically significant variable at the significance level of 10%
Source: own study.

The estimated model of Central Pomeranian households' susceptibility to debt is in the form of:

$$Prob(Y=1) = \Lambda(-0.047x_1 + 0.605x_7 + 0.795x_{11} + 1.543x_{12} + 1.941x_{13} + 2.026x_{14} + 1.208x_{15} + 2.050x_{16} + 0.603x_{17} + 0.640x_{18} - 0.297x_{19} - 1.070x_{20} + 0.002x_{23} + 0.516x_{24} + 0.593x_{25} - 0.295x_{27} + 0.806x_{30} - 0.735)$$

where: $\Lambda(x) = \frac{e^x}{1+e^x}$ logistic distribution function

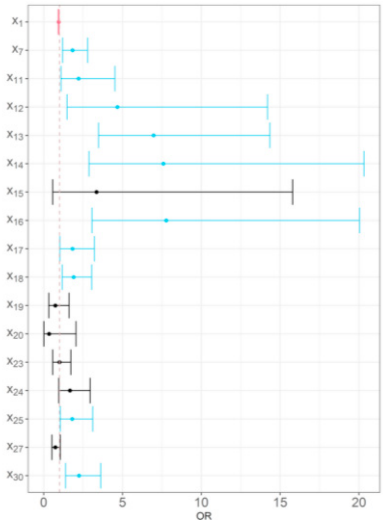

Fig. 1. Odds ratio plot
Source: own study.

Table 3. Matrix of case classification model of Central Pomeranian households' susceptibility to indebtedness

Classification of objects based on the logit model	Real belonging of objects		Sum
	$y_i = 1$	$y_i = 0$	
$\hat{y}_i = 1$	154	136	290
$\hat{y}_i = 0$	102	354	456
Sum	256	490	*746*

Source: own study.

70.5% of cases were correctly classified based on the model (*count*R^2 = 0.705). The quality assessment of the constructed model was carried out based on the Cox-Snell R^2 coefficient (0.132), R^2 Nagelkerk (0.183) and using the ROC curve (Fig. 2).

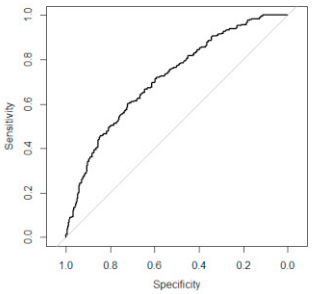

Fig. 2. ROC curve for the model of the tendency to indebtedness of households in Central Pomerania
Source: own study.

The area under the ROC curve (AUC) is 0.716. Since a field greater than 0.5 was obtained, this indicates good quality of the constructed model. The value of LR-statistics is 105.97 (p = <0.001), the critical value of this statistics for 17 degrees of freedom is 33.41.

The results of the study prove that the following factors related to the socioeconomic characteristics of households had a statistically significant positive impact on the studied probability: economic education of the household head (x_7), household development phase (x_{12}, x_{13}, x_{14}, x_{16}), social type -economic household (x_{17}, x_{18}) and household income (x_{24}, x_{25}, x_{27}, x_{30}). In turn, a statistically significant negative impact on the household debt inclination of Central Pomerania had a diversification of income by a household (x_{27}) age of the household head (x_1).

The chance of using external sources of financing by a household is by 83.2% higher in a situation where the head of the household has economic education, in relation to households in which the relationship (*ceteris paribus*) has not been reported. It was also found that the probability among the entities from Central Pomerania had an impact on the development phase in which the household is located (a lonely person at a young age was set as the reference group). With the development of the household and the transition to the next phase (*ceteris paribus*), the likelihood of using external sources of financing increases. This is due to, inter alia, the emergence of new household needs (e.g. purchase of a house / flat, increase in expenses due to the appearance of children) and the search for sources of financing expenses related to satisfying these needs. The study also proved that the impact on debt propensity of the surveyed entities from Central Pomerania has a socio-economic type of household. The chance to use external sources of financing is 89.7% higher in the case of households for which the main source of income is farm income (ceteris paribus) and by 82.7 % higher if the basic source of household income comes from business activity (*ceteris paribus*) than in households for which the basic source of income is income from paid employment. This may result, among others from the use of loans for investment purposes related to the business activity, including agricultural activity. It was also noted that the chance to use external sources of financing is 124% higher in households that encountered difficulties in accessing financial products or services (*ceteris paribus*) at least once, in relation to entities that did not experience this phenomenon. It was also observed that in the case of entities that diversify their sources of income, the chance for the occurrence of the analyzed probability is 25.5% lower than in the case of households in which the indicated characteristic (*ceteris paribus*) was not recorded. The results also prove that with age household heads, the tendency to indebtedness decreases, which is consistent with the assumptions of the life cycle hypothesis [Modigliani 2005].

4. Conclusion

The paper presents methodological assumptions regarding the logistic regression model and an example of using this research method to evaluate financial decisions taken by households. The results of the study presented in the empirical part of the study prove that the following factors related to the socio-economic characteristics of households had a statistically significant positive impact on the likelihood of Central Pomeranian households having: head household economic education (x_7), developmental phase household (x_{12}, x_{13}, x_{14}, x_{16}), socioeconomic type of household (x_{17}, x_{18}) and household income (x_{24}, x_{25}, x_{27}, x_{30}). These factors increase the likelihood of households using external sources of financing. In turn, a statistically significant negative impact on the analyzed phenomenon had a

household income diversification (x_{27}) and household head age (x_1). The results obtained for the households of Central Pomerania confirm the results of research so far presented in the literature on the factors determining the likelihood of using external sources of financing (primarily loans) by households.

The use of the logistic regression model enabled the identification and assessment of factors determining the tendency to indebtedness of Central Pomeranian households.

References

[1] Altundere, Merve Büşra. (2014) "The Relationship Between Sociability and Household Debt." *Adam Academy Journal of Social Sciences/Adam Akademi Sosyal Bilimler Dergisi* **4(2)**:27-58.
[2] André, Christophe. (2016). "Household debt in OECD countries: Stylised facts and policy issues." *OECD Economics Department Working Papers* **1277**, OECD Publishing, Paris.
[3] Bagley, Steven C., White Halbert, Golomb Beatrice A. (2001) "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain." *Journal of Clinical Epidemiology,* **54**: 979–985.
[4] Bennouna, Ghita, and Mohamed Tkiouat. (2019) "Scoring in microfinance: credit risk management tool –Case of Morocco-." *Procedia Computer Science* **148**: 522-531.
[5] Bobolik, Piotr. (2020) "Developments in the household debt-to-GDP ratio across the OECD countries since global financial crisis." *Acta Sci. Pol. Oeconomia* **19(1)**: 5-12.
[6] Bodie, Zvi, and Robert C. Merton. (1998), *Finance*, Prentice Hall, Upper Saddle River, New Jersey, p. 4.
[7] Breuer, Wolfgang; Thorsten Hens; Astrid Juliane Salzmann, and Mei Wang. (2015) "On the determinants of household debt maturity choice." *Applied Economics* **47(5)**: 449-465.
[8] Chien, Yi-Wen, and Sharon A. DeVaney. (2001) "The effects of credit attitude and socioeconomic factors on credit card and installment debt." *The Journal of Consumer Affairs* **35(1)**: 162-179.
[9] Collins, J. Michael, Erik Hambre, and Carly Urban. (2020) "Exploring the rise of mortgage borrowing among older Americans." *Regional Science and Urban Economic*s **83**: 1-23.
[10] Dobosz, Marek. (2004) *Wspomagana komputerowo statystyczna analiza wyników badań.* Akademicka Oficyna Wydawnicza EXIT, Warszawa.
[11] Ebrahimi, Zahra. (2020) "The impact of rising household debt among older Americans." *EBRI Issue Brief* **502**: 1-22.
[12] Elvery, Joel A., and Mark E. Schweitzer. (2020) "Partially disaggregated household-level debt service ratios: construction, validation, and relationship to bankruptcy rates." *Contemporary Economic Policy* **38(1)**:166-187.
[13] Garriga, Carlos, Bryan Noeth, and Don Schlagenhauf. (2017) "Household Debt and the Great Recession." *Review* **99(2)**: 183-205.
[14] Haughwout, Andrew F., Donghoon Lee, Joelle Scally, Lauren Thomas, and Wilbert van der Klaauw. (2019) "Trends in Household Debt and Credit." *FRB of New York Staff Report* **882**.
[15] Hosmer, David W., Lemenshow, Stanley, May, Susanne. (2008) *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley-Blackwell.
[16] Hosmer, David. W., Lemeshow, Stanley. (2000) *Applied Logistic Regression,* New York: John Wiley & Sons.
[17] Jain, Hemlata, Ajay Khunteta, and Sumit Srivastava. (2020) "Churn prediction in telecommunication using logistic regression and logit boost." *Procedia Computer Science* **167**: 101-112.
[18] Khan, Hafizah Hammad Ahmad, Hussin Abdullah, and Shamzaeffa Samsudin,. (2016) "Modelling the Determinants of Malaysian Household Debt." *International Journal of Economics and Financial Issues Econjournals* **6(4):** 1468-1473.
[19] Kim, Kyoung Tae, Melissa J. Wilmarth, and Robin Henager. (2017) "Povert levels and debt indicators among low-income households before and after the Great Recession." *Journal of Financial Counseling & Planning* **28(2)**: 196-212.
[20] Kleinbaum, David, G., Klein, Mitchel. (2002) *Logistic regression – a self-learning text*. New York: Springer.
[21] Korzeniowska, Anna. (2019) "Sources of financing of household debt in Poland." *Folia Oeconomica Stetnensia* **19(2)**: 56-67.
[22] Košťálová, Zuzana. (2019) "Exploring driving forces of household debt in Slovakia." *Economic Review* **48(4)**: 399-414.
[23] Kumar, Rajeev, Indrayan, Abhaya. (2011) "Receiver Operating Characteristic (ROC) Curve for Medical Researchers." *Indian Pediatrics*, **48**.
[24] Kuswanto, Heri, Ayu Asfihani, Yogi Sarumaha, and Hayato Ohwada. (2015) "Logistic regression ensemble for predicting customer defection with very large sample size." *Procedia Computer Science* **72**: 86-93.
[25] Lusardi, Annamaria, Olivia S. Mitchell, and Noemi Oggero. (2018) "The Changing Face of Debt and Financial Fragility at Older Ages." *AEA Papers and Proceedings* **108**: 407-11.
[26] Maddala, Gangadharrao S. (2001) *Introduction to Econometrics*. Third Edition. John Wiley & Sons.
[27] Modigliani, Franco, and Richard Brumberg. (2005) "Utility Analysis and the Consumption Function: An Interpretation of the Cross-Section Data", in Franco Modigliani (eds) *The Collected Papers of Franco Modigliani*, Vol. 6, The MIT Press, Cambridge-Massachusetts-London.
[28] Scott III, Robert H., and Steven Pressman. (2015) "Inadequate Household Deleveraging: Income, Debt, and Social Provisioning." *Journal of Economic Issues (Taylor & Francis Ltd)* **49(2)**:483-492.
[29] Sompolska-Rzechuła, Agnieszka, Świtłyk Michał. (2016) "Factors affecting probability of income increase in agricultural holdings specialised in milk production." *Problems of Agricultural Economics*, **4(349)**: 107-121.
[30] Stanisz, Andrzej. (2007), *Przystępny kurs z zastosowaniem Statistica PL na przykładach z medycyny.* Statsoft, Kraków, Tom 3.
[31] Turinetti, Erin, and Hong Zhuang. (2011) "Exploring Determinants Of U.S. Household Debt." *Journal of Applied Business Research (JABR)* **27(6)**: 85-92.
[32] Wałęga, Grzegorz. (2012) "Socio-economic determinants of household debt in Poland (Społeczno-ekonomiczne determinanty zadłużenia gospodarstw domowych w Polsce)." Research Papers of Wrocław University of Economics/ Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu **245**: 600-610.
[33] Walks, Alan. (2018) "Driving the poor into debt? Automobile loans, transport disadvantage, and automobile dependence." *Transport Policy* 66:137-149.
[34] Zahi, Sara, and Boujemâa Achchab. (2020) "Modeling car loan prepayment using supervised machine learning." *Procedia Computer Science* **170**: 1128-1133.
[35] Zweig, Mark. H, Campbell, Gregory. (1993) "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine." *Clinical Chemistry*, **39**.