TARGETING OF ONLINE ADVERTISING USING LOGISTIC REGRESSION

Erik Šoltés[1], Janka Táborecká-Petrovičová[2], Romana Šipoldová[3]

[1] *University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Slovakia, ORCID: 0000-0001-8570-6536, erik.soltes@euba.sk;*

[2] *Matej Bel University in Žilina, Faculty of Economics, Department of Corporate Economics and Management, Slovakia, ORCID: 0000-0003-4351-782X, janka.taborecka@umb.sk;*

[3] *University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Slovakia, ORCID: 0000-0001-5682-6635, romana.sipoldova@euba.sk.*

Abstract: *Recently, the internet became the dominant medium in marketing and comparing the development of expenditures into advertising indicates the dominance of online advertising will be inevitably stronger. Internet advertising compared to traditional media advertising has plenty of advantages hence online marketing exhibits a huge expansion in recent era. To fully utilize the potential of online marketing, it is necessary to effectively target activities of relevant internet users with the real presumption they will purchase promoted products or services. The paper is focused on demographic targeting by the mean of logistic regression models. Explanatory variables in presented application are arising from affinities of internet webpages visited by particular users and areas of their interests that are identified from their online behaviour. Our paper provides binomial logistic mode whose role is to predict the gender of internet user and multinomial logistic model constructed for the estimation of age category the user may be assigned to. The only variables exploited in the model by the mean of stepwise regression are variables with significant influence. The impact of particular factors is quantified via odds ratios that are used for the identification of areas of interests typical for women, men and for considered age categories. The paper demonstrates how it is possible to utilise estimated logistic models for the estimation of probabilities that the internet user is from a target group – in our case, women aged 25–44 years old. Prediction quality of models is assessed by the set of classification measures arising from confusion matrix that is generally acceptable in machine learning. Presented analyses are conducted in statistical software SAS Enterprise Guide on data provided from the real advertising campaign. More than 160,000 statistical units enabled the confirm results gained on training dataset of a relatively huge validation dataset.*

Keywords: *Online marketing, targeting, logistic regression, classification metrics.*

JEL Classification: *M31, C38.*

APA Style Citation: Šoltés, E., Táborecká-Petrovičová, J., & Šipoldová, R. (2020). Targeting of Online Advertising Using Logistic Regression. *E&M Economics and Management, 23*(4), 197–214. https://doi.org/10.15240/tul/001/2020-4-013

Introduction

In the era of big data and digital technologies, marketing and especially digital marketing are rapidly developing. This development is significantly supported by utilisation of various quantitative methods allowing for a lot of useful information from different marketing fields. Digital marketing brings access to mass market for a reasonable price and unlike advertisement in traditional media (TV commercials, print), it enables personalized marketing. Digital marketing applies digital channels, devices and platforms to develop or implement a marketing strategy. One subset of digital marketing is online marketing usually defined as internet marketing.

Expansion of internet marketing is considerable and evident by the increase of expenditures into internet advertisement. In the year 2018, 68% of Slovaks and 75% of Czechs had daily access to internet, whereas in 2012 it was 60% and 44%, respectively. Hence, in 2018 in comparison with the year 2012, there was a relative growth of daily access to the internet of 13% in Slovakia and 70% in the Czech Republic (Eurostat, 2019). According to IAB Slovakia (2018), investments into online advertising in Slovakia were 59.04 m eur in 2018 that is 120% more than in 2012. These expenditures in Czech Republic were even higher with 155% (SPIR, 2019). The situation worldwide indicates similar trends: expenditures into online advertising were on the level 227 bn USD in 2018 and exceeded investments into traditional TV advertisement by 21%. While in 2000 online ad investments represented only 7% of the investments in TV, in 2016 they were both approximately on the same level (180 bn USD). Although expenditures in TV are growing, constantly, this growth is much slower than in case on internet ads. Hence, it is expected the differences between them will continue markedly in favour of the internet (Recode, 2018).

The boom of internet advertising is related to the advantages of online marketing representing by measurability, lower costs, better targeting, global scope and continual availability (Shouters Voice, 2018). Online marketing enables the measurement of achieved results and the success of online advertising can be tracked in real time. Internet marketing communication is much cheaper than TV or radio ads and at the same time, it meets the requirements of effectiveness. Marketers can very easily find and approach selected target market on the internet via online advertisement, even on the level of individuals. With the help of SEO *(Search Engine Optimization)* advertising can appeal to the consumers that search on the web for associated topics with the content of advertising. Moreover, online marketing promotes products and services worldwide and constantly during the whole day.

Online marketing has its peculiarities and also has some shortcomings: banner blindness, ad-blocks utilisation, a huge influence of negative feedback and intensive competition (Shouters Voice, 2018). Banner blindness (Pagendarm & Schaumburg, 2001) is caused by the situation where many customers have learned to ignore online ads due to their overload. Moreover, some internet users apply ad-blocks that prevent the visual display of advertising on web pages. Various experiences from the past demonstrate that behind the failure of online campaign could be the sole post, comment or any negative statement of influencer (e.g. youtuber) on social media, related to the brand, product or service. Parallel pop-up ads of competition that uses similar marketing strategy can also be one of the reasons behind unsuccessful online campaign.

Currently, clients of advertising agencies insist on decreasing costs and target campaign more effectively. To identify the most relevant audience for the advertising campaign in so-called Programmatic advertising, the data are used in a real time. According to Chaffey and Smith (2017, p. 396), programmatic advertising uses data to automate the buying and selling of media inventory (ad space). This helps marketers to target audiences that are more relevant, tailor and personalize ads and remarket or serve an ad to someone who was a previous visitor of the web site. Since the agencies cannot spend finances for impressions of ads to users that are not interesting for the client, they try to identify the relevant user or target audience more precisely. Here, the Affinity Index is monitored to show the weight of a specific target audience compared to the total population in case of a specific programme or medium.

There exist various approaches to better target online advertising, such as contextual targeting, keyword targeting, data targeting (audience targeting), geo-targeting or retargeting (Match2One, 2019). Another form of targeting of online advertisement is demographic targeting for which information are usually gained from online surveys, registration form of users (where they fill-in the personal data on a voluntary basis, frequently motivated by rewards).

Our research examines and uses also behavioural targeting. According to Chen and Stallaert (2014), advertising based on behavioural targeting is becoming a sizable industry, hence also economic consequences of such targeting on main actors (advertisers and online publishers) are analysed in their paper. De Bock and Van den Poel (2010) are focused on prediction of web users' demographic profiles using the classification method Random forest that is similarly as other classification methods based on statistical mechanism. Importance of targeting in marketing is emphasized by Dave and Varma (2014), that devote their attention to computational advertising. Computational advertising is a scientific sub-discipline that aims to develop procedures for statistics and informatics like statistical modelling, data-mining, machine learning, optimization, largescale search and text analysis for the purpose to effectively find context on web that should fit the internet ads displayed for relevant audience. Many online ad campaigns include multiple advertising creatives, that is why e.g. Braun and Moe (2013) examined effects of a given ad impression in the context of an individual's impression history. They showed how advertising impression histories could affect the response to subsequent ad exposures and they illustrated how impression histories could affect advertising targeting decisions.

Our literature review of the previously realized studies demonstrates the fact that targeting in marketing should be based on sophisticated methods and procedures at the intersection of informatics and statistics, demonstrating huge progress, recently.

The main aim of the paper is to investigate and demonstrate possibilities how logistic regression can be used to predict that internet user is from desired target audience characterized by certain demographic features (gender and age) and which should be addressed by online advertising. For this estimation we use also behavioural attributes of respondents identified from their online behaviour. Research results should serve as a base for demographic targeting based on behavioural attributes of internet users. To accomplish given aim we addressed these research questions in our paper:

RQ1: Is logistic regression suitable tool for effective targeting of online advertising in case of market segmentation based on binomial and multinomial demographic variables?

RQ2: Are areas of interests of potential customers revealed from their internet behavior, providing relevant information about their assignment into demographic segments?

RQ3: Which areas of interest on the internet are provably male and female and which areas are significantly associated with younger, middle or older generation?

1. Methodology of the Logistic Regression

1.1 Motivation for Using Logistic Regression in Online Advertising Targeting

In marketing, more advanced quantitative methods are being applied and at the same time examined in scientific literature. A growing importance of statistical methods in the research of direct marketing has been confirmed by Bose and Chen (2009). The authors assess strengths and weaknesses of statistical models based on regression (including logistic regression models) and models based on data mining. Since our attention is focused on application of logistic regression, we reviewed previously realized studies in the marketing field. Referring to these scientific literature resources, we may state that logistic regression is a statistical tool suitable for online marketing. Dalessandro et al. (2015) and Miralles-Pechuán et al. (2018) deal with optimisation of online ad campaigns via utilisation of logistic regression. Yoo et al. (2004) examined the influence of animation in online ads. Bogaert et al. (2017) focused their research on development of system supporting decision-making process that would help sport brands to realize targeted online advertisement. They compared various methodologies and statistical approaches resulting in the knowledge that logistic regression achieved the best results. Lissitsa and Kol (2016) used logistic regression for comparison of trends between generation X and Y during the online shopping of products and services. Thanks to relatively simple implementation, prompt prediction and good results, regression models in general and logistical regression, specifically are frequently used also for predicting the click-through rate (Olivier et al., 2014; Shan et al., 2016). Implementation of logistic regression in online marketing can also have interdisciplinary scope with the overlapping of other social or medical sciences. E.g. Kostelic and Pavlovic (2018) provided psychometric assessment of the personality estimates and traits, as well as econometric examination of correlation to consumer first-choice communication preferences using linear logit model with binomial dependent variable. McClure et al. (2016) investigated (via logistic regression) relationship between perception of internet alcohol marketing by youth and their problems with the potential excessive consumption of alcohol in the future.

1.2 Binomial and Multinomial Logistic Regression

The logistic regression model is a special case of the general linear model (see Littell et al., 2010) and serves to model the categorical dependent variable depending on the explanatory variables of the continuous or categorical type. In the case of binary logistic regression, the logarithmic transformation of the odds of probability p for the desired event to occur $(y_i = 1)$; the event that is being examined) to the probability $1 - p$ of occurrence of the undesired event $(y_i = 0)$, is used. The natural logarithm of the odds is called logit and, unlike probability p, it acquires any real values and can be modelled by a linear regression model (Stankovičová & Vojtková, 2007):

$$\operatorname{logit}(p_i) = \ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \\ + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \qquad (1)$$

where p_i is the probability, so that $y_i = 1$ $(i = 1, 2, ..., n)$, then $\beta_0, \beta_1, ..., \beta_k$ are the parameters of the logit model and $x_{i1}, x_{i2}, ..., x_{ik}$ are the values of the explanatory variables $X_1, X_2, ..., X_k$, which are observed for the i-th statistical unit. To obtain maximum likelihood estimators of parameters of the logistic regression model the Newton-Raphson algorithm is generally used (see Allison, 2012).

After estimating the logistic model, it is important to verify its statistical significance and also verify whether the influence of the individual explanatory variables on probability p is significant. The significance of a logistic regression model is revealed by a zero hypothesis test $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \beta_2, ..., \beta_k) = \mathbf{0}^{\mathrm{T}}$ against an alternative hypothesis – at least one regression coefficient should be not zero, while three different chi-square statistics are mostly used (Likelihood ratio, Score statistics, Wald statistics). Allison (2012) discusses the differences between these statistical methods and at the same time notes that in large samples, there is no reason to prefer any of these statistics and they will generally be quite close in value.

In order to validate the significance of the explanatory variable influence, a Wald test is used. It tests the zero hypothesis showing that the respective explanatory variable does not affect the probability of occurrence of the explored event. To verify the hypothesis, Wald statistic:

$$Wald = \hat{\boldsymbol{\beta}}^{\mathrm{T}} . \mathbf{S}_{\mathbf{b}}^{-1} . \hat{\boldsymbol{\beta}} \qquad (2)$$

is used, where $\hat{\boldsymbol{\beta}}$ is the vector of regression coefficients estimates that stand at dummy variables for the respective factor (categorical explanatory variable) and $\mathbf{S}_{\mathbf{b}}$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. Wald statistic has asymptotically χ^2 distribution with degrees of freedom equal to the number of parameters estimated for a given effect. A special case of the above test is the Wald test, which verifies the statistical significance of one regression coefficient. In this case, Wald statistics is asymptotically distributed as χ^2 with 1 degree of freedom. The test statistic has a formula:

$$Wald = \left(\frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}\right)^2 \qquad (3)$$

where $s_{\hat{\beta}_i}$ is an estimated standard error of the i-th estimated coefficient.

Binary logistic regression is used, if the explanatory variable is binomial. If the dependent variable has more than 2 categories (generally these are *s* categories), we can use a multinomial logit model that is created by logit functions:

$$\ln\left[\frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \\ + \cdots + \beta_{1k}x_{ik} \\ \ln\left[\frac{P(Y=(s-1)|\mathbf{x})}{P(Y=0|\mathbf{x})}\right] = \beta_{(s-1)0} + \beta_{(s-1)1}x_{i1} + \\ + \beta_{(s-1)2}x_{i2} + \cdots + \beta_{(s-1)k}x_{ik} \qquad (4)$$

The effect of the explanatory variable X_j on the dependent variable Y is quantified in logistic regression by the odds ratio (OR) estimated by the formula:

$$OR_j = e^{\hat{\beta}_j} \qquad (5)$$

The odds ratio in binary logistic regression expresses how the odds will change: $Y = 1$ compared to the odds that $Y = 0$, in unit growth of the explanatory variable in conditions ceteris paribus. If the explanatory variable is a dummy variable, the odds ratio compares the odds of occurrence of an event at two different levels of the predictor. In the case of multinomial logistic regression, the odds ratio interpretation is analogous to that of binomial logistic regression, we only have to consider which logit formula from formulas (4) we should take into account, and, therefore, which pair of categories of the multinomial explanatory variable we should compare (most often it is l vs. 0 where $l = 1, 2, ..., (s - 1)$).

In binary logistic regression, the estimation of the probability that for the considered vector of explanatory variables values $\mathbf{x}_i$ binary target variable exhibits value 1, is calculated according to the formula:

$$p_i = \hat{P}(y_i = 1) = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}}} = \frac{1}{1 + e^{-\hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}}} \qquad (6)$$

In multinomial logistic regression the probability that a dependent variable will take the values $0, 1, 2, ... (s - 1)$ for the variable explanatory vector $\mathbf{x}_i$ is estimated by formulas:

$$\hat{P}(y_i = 1) = \frac{e^{\hat{\boldsymbol{\beta}}_1^T \cdot \mathbf{x}_i}}{1 + \sum_{l=1}^{s-1} e^{\hat{\boldsymbol{\beta}}_l^T \cdot \mathbf{x}_i}},\ ...,\ \hat{P}(y_i = s - 1) = \frac{e^{\hat{\boldsymbol{\beta}}_{s-1}^T \cdot \mathbf{x}_i}}{1 + \sum_{l=1}^{s-1} e^{\hat{\boldsymbol{\beta}}_l^T \cdot \mathbf{x}_i}},\ \hat{P}(y_i = 0) = \frac{1}{1 + \sum_{l=1}^{s-1} e^{\hat{\boldsymbol{\beta}}_l^T \cdot \mathbf{x}_i}} \qquad (7)$$

where $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}} = (\hat{\beta}_{l0}, \hat{\beta}_{l1}, ..., \hat{\beta}_{lk})$ while $l = 1, 2, ..., (s - 1)$.

The quality of the logistic model can be evaluated by different measures. Among the criteria that measure a relative quality of statistical models belong AIC – Akaike Information Criterion and SC – Schwarz-Criterion, which are based on the logarithmic transformation of the likelihood function, i.e. $-2\,ln\,L$ (see Littell et al., 2010; Kim & Timm, 2006).

1.3 Confusion Matrix and Classification Metrics

A key factor in evaluating the performance of classifier model such as logistic regression model are classification metrics, from which a lot of them is based on confusion matrix (Tab. 1). The diagonal elements represent correct predictions. If the sample is positive and it is classified as positive, i.e. correctly classified positive sample, it is counted as a True positive (TP). If the sample is negative and it is classified as negative it is considered as True negative (TN). If the real value 1 (positive) of the target variable is classified as 0 (negative) according to the model, it is considered as a False negative (FN) or Type II error. Finally, if the real value 0 (negative) of the target variable is classified as 1 (positive) according to the model, it is counted as False positive (FP), false alarm or Type I error.

The confusion matrix is used to calculate many common classification metrics, such as *Accuracy (ACC):*

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (8)$$

Supplement of ACC into value 1 is called *Error rate* (ERR) or *Misclassification rate*. Two kinds of *Accuracy* can be considered *Sensitivity* (TPR) and *Specificity* (TNR):

$$TPR = \frac{TP}{TP + FN} \qquad TNR = \frac{TN}{TN + FP} \qquad (9)$$

Sensitivity or *True positive rate* (TPR) is the proportion of the positive samples that were correctly classified. *Specificity* or *True*

Tab. 1: Confusion matrix

		True/Actual class	
		1 (Positive)	**0 (Negative)**
Predicted class	**1 (Positive)**	True positive TP	False positive FP
	0 (Negative)	False negative FN	True negative TN

Source: own

negative rate (TNR) represents the proportion of the negative samples that were correctly classified. The ratio of calculated values (value 1), that have been correctly classified towards the overall number of observations that were predicted as Positive (1) explains *Positive prediction value* (PPV) or *Precision:*

$$PPV = \frac{TP}{TP+FP} \tag{10}$$

Supplement of the indicator *Precision* into value 1 is called a *False discovery rate*. Similarly, we could define the *Negative predictive value* (NPV) and its supplement into value 1, so-called *False omission rate*. Among frequently used classification matrices that arise from confusion matrix are involved also *Matthew's Correlation Coefficient* (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \tag{11}$$

and F-measure or F1-score:

$$F - measure = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{12}$$

Matthew's Correlation Coefficient explains correlation between observed and predicted classification and it is in relation with the test statistics chi-square for the verification of the contingency (Baldi et al., 2000, p. 419). MCC receives values from the interval $\langle -1; 1 \rangle$ and value 1 exhibits perfect prediction, value −1 indicates complete unfit between prediction and real values. If this coefficient exhibits value 0, it is a prediction that is not better than random estimation. In case of values F-measure are desirable values close to 1, hence in comparison with MCC by this measure the least desirable values are the ones close to 0. F-measure is unweighted harmonic average of *Sensitivity* and *Precision*. Weighted harmonic average of these measures is $F_\beta - measure$ (Tharwat, 2018).

Whereas in the case of binary classification the process of calculation of the mentioned measures is trivial, in case of multiclassification it is more complicated. To calculate classification measures for multinomial targeted (predicted) variable there can be used micro-averaging and macro-averaging (see ML Wiki, 2015).

Frequently used measure for comparison of classification models and specifically logistic regression models is also ROC (Receiver Operating Characteristic) curve, that demonstrates relationship between the probability of detecting true signal (sensitivity) and false signal (1− specificity) (see Hosmer & Lemeshow, 2000). To compare achievements of predicted classification serves AUC (Area Under ROC Curve), that is with the measures *Sensitivity* (TPR) and *Specificity* (TNR) and frequencies defined in the confusion matrix (Tab. 1) in this relationship (Powers, 2011, p. 41; Idrees et al., 2017, p. 43).

$$AUC = \frac{1}{2}(TPR + TNR) = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{13}$$

In case of multinomial classification of target variable is as a measure of model accuracy more frequently used the volume under ROC surface (VUS), which is the extension of the area under curve (AUC) for binary models. Kapasný and Řezáč (2013) provide approach to estimate VUS based on the confusion matrix. This procedure is used in part 4.

2. Database and Preparing of Statistical Variables

The main aim of the advertising agency when designing and realizing online advertising campaign is to approach the most relevant audience to ensure the campaign effectiveness and minimalisation of advertising costs for the agency. In accordance with this, the aim of our analyses was to construct models of logistic regression that would serve for prediction that visitor of internet webpages belongs to a desired target market group (as an illustration women, aged 25–44 years old). To prevent fragmentation of observations during classification of input dataset according to the target variable we decided to construct two models of logistic regression. The first one will model the probability that the user is a woman (resp. man) and the second one will model probability that user ranks in one of the three age categories (13–24; 25–44; 45–65). If we would directly model target group, 6 (2 × 3 = 6) categories of dependent variable would be created, whereas in this approach target variable will take on 2 values and 3 values, respectively.

Our analyses use data from DSP (Demand-Side Platform) – a real marketing campaign realized within one month. For our disposal, we had 160,544 observations that were divided on training and validation set of data in a ratio 70:30. A list of variables involved in the file of data sources are demonstrated in Tab. 2.

Tab. 2: List of variables in data file

Variable	Explanation
USER_DAY	Day during a week; working days = 1, weekend = 0
USER_HOUR	Hour during the day when user was present on the internet
OPERATING_SYSTEM	ID version of operation system
BROWSER	ID browser
SITE_DOMAIN	Domain
INTEREST	Area of interest of internet user*
GENDER	User gender; female = 1, male = 0
AGE	Age category of user; 13–24 years, 25–44 years, 45–65 years
TARGET	If user belongs to the target group (CS = women, 25–44) or not

Source: own

Note: *According to IAB's (Interactive Advertising Bureau) contextual taxonomy including: Arts & Entertainment, Automotive, Business, Careers, Education, Family & Parenting, Food & Drink, Health & Fitness, Hobbies & Interests, Home & Garden, Law, Gov't & Politics, News, Personal Finance, Pets, Real Estate, Science, Shopping, Society, Sports, Style & Fashion, Technology & Computing a Travel.

Within the models of logistic regression, we will have these numerical explanatory variables: USER_HOUR, *Affinity_female* (model with targeted variable Gender) and *Affinity_Age_25–44* (model with targeted variable Age), explaining affinity of visited internet webpage with the respect to the target market group (woman or age category 25–44-years). Besides, we will count with categorical variables:

- dummy variable USER_DAY;
- dummy variables characterizing user's area of interest, while these particular areas are listed in the explanation under Tab. 2;
- binomial and multinomial categorical variables based on affinity of internet webpage (with the description below).

With the respect to the fact that variable SITE_DOMAIN consisted of 2,128 unique webpages visited by individual users it was necessary to modify this variable and we used it to create new ones: CAT_GENDER and CAT_AGE. CAT_GENDER variable divides individual webpages according to the gender as: male (value 0), female (value 1) and neutral (value 2). The criterion for selection and assignation of internet webpages was affinity of the webpage for male and female gender. If variable *Affinity_female* gained values higher than 1.10, webpage was classified as female one; if variable *Affinity_male* gained values higher than 1.10, webpage was classified as male one. Remaining webpages ranked 0.90–1.10 were classified as neutral. Following affinity for age categories, we applied the same approach and created CAT_AGE variable with values 0, 1, 2, 3. These values classify webpages into: the ones visited especially by individuals aged 13–24 years (value 0), users aged 25–44 years (value 1) and 45–65 years (value 2). For neutral web pages, CAT_AGE variable achieved the value 3. Original variables OPERATING_SYSTEM and BROWSER have been used for creation of numerical variables *Affinity_OS_Female* and *Affinity_B_Female,* expressing affinity of women towards the type of operation system and type of browser. Since the target group was defined as (besides) women also persons aged 25–44 years, other explanatory variables involved in analysis were variables *Affinity_OS_25–44* and *Affinity_B_25–44,* developed by the authors of this paper. These variables characterize affinity of target market group of persons aged 25–44 years towards type of operation system and type of browser.

It is obvious from database description that paper presents results of case study based on quantitative data adopted from marketing campaign. Despite the fact that the file includes internet users from just one campaign, the sample is relatively huge and scarcely to reach in conditions of smaller central European countries. With the respect to sample size and cultural similarity of internet users' behaviour in post-socialistic countries in Central Europe we can generalize our research results to certain extent on market in this geographical space. Although some authors (Kolman et al., 2003) argue that Central European countries should not be treated as a homogeneous group, there exist studies presenting also similarities (together with differences) in terms of their national and cultural identity (Skinner et al., 2008).

3. Results

In this section of the article, we will use binomial (model Gender) and multinomial (model Age) logistic regression for the assessment of the statistical significance of the influence of the considered explanatory variables on the probability that visitor of the webpage is a 25–44 years old woman. The impacts of each relevant factor will be quantified in subsections 3.2 and 3.3 by odds ratios, while the effect of other significant factors being fixed. Subchapter 3.4 will be devoted to the assessment of models' quality by classification measures and to the illustration of prediction.

3.1 Choice of Explanatory Variables into Model Gender and Model Age and Assessment of the Quality of These Models

With the respect to the target group (25–44 years old women) we will create binomial logistic regression with the dependent variable Gender (male = 0 and female = 1), while the model category will be category (female). In the next step, we will create a model of multinomial logistic regression with categorical dependent variable Age that has 3 values (13–24 years; 25–44 years and 45–65 years), while as the reference category we chose targeted category 25–44 years old internet users. From the set of considered explanatory variables we chose regressors that have relevant impact on the probability that the visitor of internet webpage will be from the target group with the method of stepwise regression (Wooldridge, 2013) and settled significance level as 0.05 to assign an explanatory variable into the model. We will present only the results of models – Gender and Age that involve explanatory variables selected by this method. Interests Education and Real estate were not assigned into Gender model what can be interpreted that women and men are equally interested in these areas. Model Age on the other hand does not involve Careers and Pets interests; therefore, it seems that these two areas are comparably attractive with the same intensity for all age categories.

Based on the tests of statistical significance of regression coefficient vector (first part in Tab. 3) we found that both models are statistically significant as a whole. Values AIC, SC a −2 Log L show that both models are

Tab. 3: Assessment of the quality of the model Gender and the model Age

Testing global null hypothesis: BETA = 0						
	Model Gender			Model Age		
Test	Chi-squared	DF	Pr > ChiSq	Chi-squared	DF	Pr > ChiSq
Likelihood ratio	17,351.418	27	<0.0001	30,935.814	56	<0.0001
Score	16,238.072	27	<0.0001	32,108.770	56	<0.0001
Wald	14,248.767	27	<0.0001	25,241.391	56	<0.0001

Model fit statistics				
Criterion	Model Gender		Model Age	
	Intercept only	Intercept and covariates	Intercept only	Intercept and covariates
AIC	154,474	137,177	222,148	191,324
SC	154,484	137,447	222,167	191,882
−2 Log L	154,473	137,121	222,144	191,208

Source: own in SAS EG based on dataset provided by advertising agency substantially better than model involving only intercept.

Since we used stepwise regression method, influence of all variables involved into model Gender and also model Age is statistically significant on the level of significance 0.05, what was confirmed also by Wald test with test statistic (2). Values of Wald statistics (due to the limited scope of the article will not be explained into the details) after considering a degree of freedom shown some interesting facts. The probability that the visitor of internet webpage is a woman is mostly influenced by the fact if the individual webpage is female or non-female oriented; then by affinity of webpage on operation system and browser with the respect to female gender; together with the fact of the webpage being oriented towards Sports, Automotive or Family & Parenting. Other areas of interest have also significant influence, but the previously mentioned ones are so typical for one particular gender that its influence on observed probability was essentially significant. Wald test for the analysis of effects in Age model exhibits that probability that internet user is from 25–44 years old group is mostly influenced by the fact that if this webpage is usually visited by individuals from this target age group, then by affinity of the webpage on operation system with the respect to target age group. When it comes to areas of interest, observed probability is mostly determined by Hobbies & Interest and Arts & Entertainment.

3.2 Quantification of the Impact of Relevant Variables on Target Variable in Binomial Logistic model Gender

In Tab. 4 are exhibited estimates of parameters for binomial logistic regression model, that have been estimated through the method of maximum likelihood and that are significant what is obvious from calculated p-values for Wald test (3). The tool for quantification of the impact of individual regressors will not be estimated regression coefficients but odds ratios calculated from them (5).

With the respect to relatively high number of explanatory variables we will interpret only selected odds ratios. These odds ratios will be interpreted under condition of ceteris paribus; it means that all other explanatory variables considered in the regression model are constant. For example, odds that user affected by advertisement is a woman in comparison with the odds that this advertisement approached man, is 1.11-higher during the working days than during the weekend.

Areas of interest listed in Tab. 4 are ranked from the ones women are more interested in, to the ones typical for men. From the estimation of odds ratios quantified for the areas of interests, we revealed that typical female interests are especially Family & Parenting because corresponding odds (the probability that user targeted by ads is woman against the probability it is a man) is 1.56 higher if the user indicates interest in the area of Family & Parenting in comparison with users without this interest. Among female-oriented webpages are considered also Science, Arts & Entertainment, Style & Fashion and Health & Fitness, where we estimated odds ratios 1.360; 1.312; 1.236 and 1.206. Women dominate also on the webpages from the areas of Food & Drink, Home & Garden, Society, Shopping and even News. However, in case of the user interested in these areas, the odds of a female vs. male is less than 20% higher than in case of user not interested in these areas.

Other areas are more male-focused, though among the typically male areas of interest can be ranked especially Sports and Automotive. In these areas, we can calculate odds ratios for men (male vs. female) as a reversed values of odds ratios that are estimated in Tab. 4. In Sports and Automotive areas is then odds male vs. female 90% higher (more precisely it is 95.3% and 88.7%, respectively) than in case of users that do not show interest in these areas on the internet. Users demonstrating interest in Travel have odds male vs. female approximately 25% higher than internet users not demonstrating interest in this area. We should again emphasize that this statement is valid under the condition of ceteris paribus. In case of individuals exhibiting interest in Law, Gov't & Politics, Hobbies & Interest, Technology & Computing, we quantify odds male vs. female from 1.10 to 1.15-higher than in case of individuals without this type of activities on the internet.

In case of Cat_Gender variable, odds ratio for 1 vs. 2 (see Tab. 4) informs about the fact that if user was exposed to the ad on the webpage classified as female-focused, then the odds female vs. male is 83.1% higher than

Tab. 4: Estimates of parameters and odds ratios for binomial logistic model Gender

Analysis of maximum likelihood estimates						
Effect		Parameter		Odds ratio		
		Estimate	Pr > ChiSq	Estimate	95% confidence limits	
Intercept		−3.745	<0.0001	–	–	–
USER_DAY		0.104	<0.0001	1.110	1.078	1.142
USER_HOUR		0.004	<0.0001	1.004	1.002	1.007
FAMILY_&_PARENTING		0.442	<0.0001	1.555	1.500	1.612
SCIENCE		0.308	<0.0001	1.360	1.319	1.403
ARTS_&_ENTERTAIN.		0.272	<0.0001	1.312	1.266	1.359
STYLE_&_FASHION		0.212	<0.0001	1.236	1.196	1.278
HEALTH_&_FITNESS		0.188	<0.0001	1.206	1.155	1.260
FOOD_&_DRINK		0.173	<0.0001	1.189	1.125	1.256
HOME_&_GARDEN		0.148	<0.0001	1.159	1.115	1.204
SOCIETY		0.142	<0.0001	1.152	1.114	1.193
SHOPPING		0.055	0.0016	1.057	1.021	1.093
NEWS		0.051	0.0022	1.052	1.018	1.087
PETS		−0.026	0.0093	0.974	0.955	0.994
CAREERS		−0.069	0.0019	0.934	0.894	0.975
BUSINESS		−0.070	0.0016	0.933	0.893	0.974
PERSONAL_FINANCE		−0.073	0.0021	0.930	0.888	0.974
TECHNOL._&_COMPUT.		−0.107	<0.0001	0.898	0.861	0.938
HOBBIES_&_INTEREST		−0.124	<0.0001	0.884	0.853	0.916
LAW_GOV'T_&_POLITICS		−0.139	<0.0001	0.870	0.838	0.903
TRAVEL		−0.214	<0.0001	0.808	0.777	0.839
AUTOMOTIVE		−0.635	<0.0001	0.530	0.509	0.552
SPORTS		−0.669	<0.0001	0.512	0.497	0.528
Affinity_Female		0.259	<0.0001	1.295	1.242	1.350
Affinity_OS_Female		1.526	<0.0001	4.598	4.067	5.199
Affinity_B_Female		1.531	<0.0001	4.621	3.869	5.519
CAT_ GENDER	0 vs. 2	−0.573	<0.0001	0.564	0.543	0.586
	1 vs. 2	0.605	<0.0001	1.831	1.772	1.893

Source: own in SAS EG based on dataset provided by advertising agency

in case this ad is present on neutral webpage. From the odds ratios for Cat_Gender variable (listed in Tab. 4) we can simply calculate also odds ratio 1 vs. 0 with the value 3.246. Based on this value we find out that considered odds is in case of female-focused webpages even 3.246-higher than in case of male-oriented ones.

3.3 Quantification of the Impact of Relevant Variables on Targeted Variable in Multinomial Logistic Model Age

Since Age variable has 3 values, the result of multinomial logistic regression are two partial models (model 13–24, vs. 25–44 and model 45–65 vs. 25–44) whose parameters are estimated in Tab. 5. On the basis of results for model 13–24 vs. 25–44, we found out that the fact if online advertising will affect users 13–24 years or 25–44 years does not depend on the fact if the user is interested in webpages oriented towards Home & Garden, Sports alebo Personal Finance. Affinity_B_Age_25–44 factor has also non-significant influence. This conclusion results from p-values for the test of statistical significance related regression coefficient that provides values 0.1553, 0.1326, 0.0823 and 0.4584 for the aforementioned variables. Since these p-values are higher than the significance level 0.05, we do not reject null hypothesis about statistical non-significance of the considered parameter. In case of model

Tab. 5: Estimates of parameters and odds ratios for multinomial logistic model Age

Analysis of maximum likelihood estimates										
Parameter		13–24 vs. 25–44				45–65 vs. 25–44				Interest
		Estimate		Odds ratio		Estimate		Odds ratio		
		Beta	P-value	13–24 vs. 25–44	25–44 vs. 13–24	Beta	P-value	45–65 vs. 25–44	25–44 vs. 45–65	
Intercept		−0.526	<0.0001	–	–	2.246	<0.0001	–	–	–
USER_DAY		−0.260	<0.0001	0.771	1.297	−0.118	<0.0001	0.889	1.125	–
USER_HOUR		0.021	<0.0001	1.022	0.978	−0.001	0.2948	0.999	1.001	–
HOBBIES_&_INTEREST		1.021	<0.0001	2.779	0.360	−0.484	<0.0001	0.617	1.621	A
ARTS_&_ENTERTAIN.		0.630	<0.0001	1.875	0.533	−0.325	<0.0001	0.722	1.385	A
TECHNOL._&_COMPUT.		0.537	<0.0001	1.710	0.585	−0.116	<0.0001	0.887	1.127	A
HEALTH_&_FITNESS		0.353	<0.0001	1.430	0.699	−0.290	<0.0001	0.749	1.335	A
EDUCATION		0.145	<0.0001	1.157	0.864	−0.139	<0.0001	0.872	1.147	A
BUSINESS		0.060	0.0364	1.060	0.943	−0.375	<0.0001	0.693	1.443	A
HOME_&_GARDEN		0.038	0.1553	1.039	0.962	−0.149	<0.0001	0.863	1.159	A
FOOD_&_DRINK		0.456	<0.0001	1.576	0.635	0.129	0.0005	1.137	0.880	B
SPORTS		−0.034	0.1326	0.967	1.034	−0.185	<0.0001	0.830	1.205	C
TRAVEL		−0.065	0.0258	0.937	1.067	−0.122	<0.0001	0.886	1.129	C
AUTOMOTIVE		−0.115	<0.0001	0.892	1.121	−0.340	<0.0001	0.712	1.404	C
STYLE_&_FASHION		−0.169	<0.0001	0.847	1.181	−0.188	<0.0001	0.830	1.205	C
REAL_ESTATE		−0.304	<0.0001	0.739	1.353	−0.298	<0.0001	0.744	1.344	D
FAMILY_&_PARENTING		−0.384	<0.0001	0.680	1.471	−0.096	<0.0001	0.908	1.101	D
SCIENCE		0.294	<0.0001	1.341	0.746	0.305	<0.0001	1.355	0.738	E
SHOPPING		0.179	<0.0001	1.194	0.838	0.200	<0.0001	1.218	0.821	E
SOCIETY		0.085	0.0009	1.090	0.917	0.154	<0.0001	1.167	0.857	E
PERSONAL_FINANCE		−0.057	0.0823	0.944	1.059	0.113	<0.0001	1.122	0.891	F
LAW_GOV'T_&_POLITICS		−0.449	<0.0001	0.639	1.565	0.375	<0.0001	1.455	0.687	F
NEWS		−0.595	<0.0001	0.551	1.815	0.013	0.4678	1.011	0.989	F
Affinity_Age_25–44		−0.327	<0.0001	0.721	1.387	−0.134	<0.0001	0.874	1.144	–
Affinity_OS_Age_25–44		−1.327	<0.0001	0.266	3.759	−1.961	<0.0001	0.141	7.092	–
Affinity_B_Age_25–44		0.072	0.4584	1.073	0.932	−0.486	<0.0001	0.614	1.629	–
CAT_GENDER	13–24	1.128	<0.0001	3.087	0.324	−0.486	<0.0001	0.615	1.626	–
	25–44	−0.030	0.6410	0.969	1.032	−0.573	<0.0001	0.564	1.773	–
	45–65	0.188	0.0034	1.206	0.829	0.220	<0.0001	1.244	0.804	–

Source: own in SAS EG based on dataset provided by advertising agency

45–65 vs. 25–44 we can speak about non-significant influence of the hour during day and the fact the user is interested in webpages from area News. All other factors included in multinomial model of logistic regression, have significant influence at the significance level 0.05 in both models. Influence of these factors was quantified through estimations of odds ratios that are calculated in Tab. 5.

Since we chose age category 25–44 as a reference category, estimated odds ratio compare age category 13–24, respectively 45–65, against users aged 25–44. To achieve results that will more clearly point out the factors that significantly cause increase or decrease of probability that internet user is aged as 25–44 years, we calculated reciprocal of original odds ratios.

Thanks to this procedure we gained probability that internet user will be in the age of 25–44 years in comparison with the probability that he will be 13–24 years, 45–65 years, respectively. Odds ratios originally gained in SAS Enterprise Guide (13–24 vs. 25–44 and 45–65 vs. 25–44) and at the same time, odds ratios (25–44 vs. 13–24 and 25–44 vs. 45–65) derived from them are presented in Tab. 5.

If the internet users of the webpage was affected by ad during the working days then odds 25–44 vs. 13–24 is 1.297-higher than in the case they were affected during the weekend. Considering these facts that it is more effective to target advertisement during working days on 25–44-years old than on 13–24-years old and odds ratio 1.125 (for odds 25–44 vs. 45–65) demonstrates it is more efficient in case of users aged 25–44 years also in comparison to 45–65-years old.

Based on odds ratios we identified 6 types of interests (Tab. 6) from the viewpoint of users' age. Particular types are determined by the ranking of 3 age categories of internet users according to the fact of how intensively these age categories are interested in particular areas. Letter A represents areas where especially young people (13–24 years) are mostly interested; followed by middle-aged ones (25–44 years) and finally older users (45–65 years). Odds ratios (Age I. vs. Age III.) stated in last column of Tab. 6 indicate which age category exhibits the most (Age I.) and the least (Age III.) interest in considered areas. For example, in the F area type, individuals 45–65 years old are dominating the internet and the least interest show 13–24 years old individuals. Ratios of odds that user is from a major age category for certain type (in the case of the F type, it is age category 45–65 years) and odds that user is from a minor age category (in case of F type it is age category 13–24 years) are for particular interest areas (F type: Personal Finance, Law, Gov't & Politics and News) stated in last column. These odds ratios are a product of odds ratios from previous two columns (Age I. vs. Age II. and Age II vs. Age III.).

As we already stated, the greatest discrepancies among observed age categories are in area of interest Hobbies & Interest that was confirmed also by odds ratios that are markedly the greatest just for this area of interest. Odds 13–24 vs. 45–65 is in case of the user interested in Hobbies & Interest even 4.5-higher than in case of user not interested in this area on the internet. Areas that are significantly dominated by young people are also Arts & Entertainment and Technology & Computing and Health & Fitness where odds ratios are 2.597, 1.927 and 1.909, respectively. Young people compared to the other age groups show the greatest interest also in Education, Business and Home & Garden (where on the other hand 45–65 years old are interested at least, similarly as in case of previous areas). Young people most frequently visit also webpages related to Food & Drink that belong under B type, where are areas showing the least interest by individuals of middle age (25–44 years). Since our target market is 25–44 years old individuals we will look at areas they show the least interest but especially on areas where they exhibit the greater activities than other age groups. Besides previously mentioned Food & Drink area, 25–44 years old compared to the other age groups show the smallest interest in Science, Shopping and Society (type E). However, unlike Food & Drink in this area dominate individuals from the oldest age category. The probability that it is the older user (45–65 years) against the probability that he is in middle age (25–44 years) is in case the user exhibit interest in 3 of the aforementioned areas (type E) 1.355, 1.218 and 1.167 higher than if he shows no interest in considered areas on the internet. Areas where the middle age users are most active are areas of interest C and D. The odds 25–44 vs. 45–65 is 1.404-higher in case of user interested in Automotive than in case he is not interested. Odds ratios 25–44 vs. 45–65 in Sports, Style & Fashion and Travel areas are 1.205, 1.205 a 1.129. Middle age individuals have in comparison with younger and older ones

Tab. 6: Types of interests according to the preferences of areas of interests by particular age categories of internet users

Type	Area	Odds ratio		
		13–24 vs. 25–44	**25–44 vs. 45–65**	**13–24 vs. 45–65**
A	Hobbies & Interest	2.779	1.621	4.505
	Arts & Entertainment	1.875	1.385	2.597
	Technology & Computing	1.710	1.127	1.927
	Health & Fitness	1.430	1.335	1.909
	Education	1.157	1.147	1.327
	Business	1.060	1.443	1.530
	Home & Garden	1.039 – n.s.	1.159	1.204
B		**13–24 vs. 45–65**	**45–65 vs. 25–44**	**13–24 vs. 25–44**
	Food & Drink	1.386	1.137	1.576
C		**25–44 vs. 13–24**	**13–24 vs. 45–65**	**25–44 vs. 45–65**
	Automotive	1.121	1.252	1.404
	Sports	1.034 – n.s.	1.165	1.205
	Style & Fashion	1.181	1.020	1.205
	Travel	1.067	1.058	1.129
D		**25–44 vs. 45–65**	**45–65 vs. 13–24**	**25–44 vs. 13–24**
	Family & Parenting	1.101	1.336	1.471
	Real Estate	1.344	1.007 – n.s.	1.353
E		**45–65 vs. 13–24**	**13–24 vs. 25–44**	**45–65 vs. 25–44**
	Science	1.010 – n.s.	1.341	1.355
	Shopping	1.020 – n.s.	1.194	1.218
	Society	1.071	1.090	1.167
F		**45–65 vs. 25–44**	**25–44 vs. 13–24**	**45–65 vs. 13–24**
	Law, Gov't & Politics	1.455	1.565	2.277
	News	1.011 – n.s.	1.815	1.835
	Personal Finance	1.122	1.059	1.188

Source: own in SAS EG based on dataset provided by advertising agency

Note: n.s. – non significant.

higher odds also in case of interests in Family & Parenting and Real Estate (where the smallest interest is shown by young people (13–24 years)). Odds ratio 25–44 vs. 13–24 are in case of users of internet showing interest in these two areas more than 1/3 higher (more precisely 47.1% and 35.3%, respectively) than in case of users where we did not record any interest in them. Areas of interest A and B type are searched on the internet mostly by young people (13–24 years); C and D areas are interesting especially for middle age individuals (25–44 years) and for older generation (45–65 years) are typical visits of internet webpages focused on E and F type. In contrast, among all age categories D and F areas are the least interesting for 13–24-years old; B and E for 25–44 and A and C type for 45–65-years old.

3.4 Classification Metrics for Models Gender and Age and Prediction of Target Variables

The main purpose of the estimated logistic models is to predict probability that the person

Tab. 7: Confusion matrices for model Gender and model Age (for training and validation sample)

Model Gender – training			
INTO	**FROM**		
Frequency	**Female**	**Male**	**Total**
Female	28,608	15,962	**44,570**
Male	21,498	46,313	**67,811**
Total	**50,106**	**62,275**	**112,381**

Model Gender – validation			
INTO	**FROM**		
Frequency	**Female**	**Male**	**Total**
Female	14,622	4,258	**18,880**
Male	6,861	22,422	**29,283**
Total	**21,483**	**26,680**	**48,163**

Model Age – training				
INTO	**FROM**			
Freq.	**13–24**	**25–44**	**45–65**	**Total**
13–24	9,040	2,457	807	**12,304**
25–44	8,274	50,421	14,560	**73,255**
45–65	1,126	8,045	17,651	**26,822**
Total	**18,440**	**60,923**	**33,018**	**112,381**

Model Age – validation				
INTO	**FROM**			
Freq.	**13–24**	**25–44**	**45–65**	**Total**
13–24	3,988	1,233	363	**5,584**
25–44	3,457	21,179	6,344	**30,980**
45–65	537	3,601	7,461	**11,599**
Total	**7,982**	**26,013**	**14,168**	**48,163**

Source: own in SAS EG based on dataset provided by advertising agency

present on the webpage will be from target market group. Firstly, we calculate classification metrics. They are generated from a confusion of matrices (Tab. 7), from which according to Tab. 1 in case of binomial classification and according to e.g. Tharwat (2018) by the mean of micro-averaging (ML Wiki, 2015) in case of multinomial classification we determine number of pairs TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). By substituting them into formulas from (8) till (13) we gain values of classification metrics for Gender model and Age model, and for both cases particularly for training and validation sample (Tab. 8).

Classification metrics show relatively good and comparable quality of both models Gender and Age. Results achieved on the validation sample underline this statement. Classification metrics attain the lowest values in case of Age model estimated in the validation sample, but the difference is negligible hence both models may be used for prediction.

In Tab. 9, there are exhibited values of explanatory variables for which we will estimate that on 2 particular webpages with stated affinities will be user approached by the advertisement of female gender, aged 25–44 years. Moreover we assume that A user is interested in Arts & Entertainment, Business, Careers, Family & Parenting, Food & Drink, Health & Fitness, Home & Garden, Personal Finance, Pets, Science, Society and Style & Fashion, and B user manifests interest in Automotive, Hobbies & Interest, Law, Gov't

Tab. 8: Classification metrics for Gender model and Age model

Model		*Accuracy (ACC)*	*Sensitivity (TPR)*	*Specificity (TNR)*	*Precision (PPV)*	*Matthew's MCC*	*F measure*	*AUC/VUS*
GENDER	**TR**	0.756	0.671	0.824	0.754	0.503	0.710	0.747/–
	VAL	0.769	0.681	0.840	0.774	0.531	0.725	0.761/–
AGE	**TR**	0.766	0.686	0.814	0.686	0.500	0.686	0.750/0.501
	VAL	0.759	0.677	0.808	0.677	0.485	0.677	0.743/0.495

Source: own based on dataset provided by advertising agency

Note: TR – training dataset, VAL – validation dataset.

Tab. 9: **Values of explanatory variables besides the areas of interest used for prediction of targeted variables Gender and Age**

User	USER_DAY	USER_HOUR	Affinity						Cat_Gender		Cat_Age		
			Female	OS_Female	B_Female	25–44	OS_25–44	B_25–44	0 (Male)	1 (Female)	13–24	25–44	45–65
A	1	1	1.680	1.937	1.439	1.479	1.847	1.178	0	1	0	1	0
B	0	14	0.725	0.654	0.803	0.682	0.794	0.747	1	0	0	0	0

Source: own

& Politics, News, Shopping, Sports and Technology & Computing.

We should remark that webpage on which A user is present has higher affinity towards target market than the one where B user is present. Moreover, A user unlike B user has various interests that are typical for target market based on these facts we expect that the probability is significantly higher the user is from the target market will be the A individual rather than the B individual.

Following the binomial logistic model Gender we conduct estimation of probability that user A and user B has female gender, according to the formula (6), while estimation of parameters are calculated in Tab. 4 and values of explanatory variables have been defined in Tab. 9, alternatively by the interests of individuals A and B. Since the age category 25–44-years has be selected as the reference on in multinomial logistic model, then estimation of the probability that user is from target age group, we will use the last formula (7) (more precisely the formula for $\hat{P}(y_i = 0)$), while the estimation of parameters are listed in Tab. 5.

$\hat{P}(A\text{: female} \wedge 25\text{–}44) = \hat{P}(A\text{: female}) \times \hat{P}(A\text{: } 25\text{–}44) = 0.985 \times 0.909 = 0.895;$

$\hat{P}(B\text{: female} \wedge 25\text{–}44) = \hat{P}(B\text{: female}) \times \hat{P}(B\text{: } 25\text{–}44) = 0.985 \times 0.909 = 0.895.$

Person A is with the 98.5% probability of female gender and with the 90.9% probability from age category 25–44. Since we can assume that observed phenomenons (gender, age) are mutually independent, then A user of internet is from the target group with the probability of 89.5%. In case of B user these probabilities are significantly lower and we estimated that there is only 1% probability that this person is from target group.

Conclusions

The aim of the paper is to exhibit possibilities of logistic regression utilization in prediction that internet user is from the target group characterized by particular demographic features. The more accurate will be the identification of internet users belonging into target market of certain marketing campaign, the more effective will be online advertising. To illustrate this reality, target group of 25–44 years female individuals have been chosen. This paper provides a case study based on quantitative data about 160,544 internet users adopted from Polish marketing campaign. Considering the sample size we believe that our research findings can be valid to certain extend for geographical space of post-socialistic countries from Central Europe. They can serve as relevant insight for marketers and companies operating and developing their international marketing strategies with the focus on CEE region (Schuh & Holzmüller, 2003) and preferring perspective of standardization that is popular among Western companies entering CEE (Schuh, 2000). The results of our analyses represent two models: binomial logistic model that estimates probability of the fact that internet user is a woman and multinomial logistic model that predicts that internet user belongs to the target market of the age 25–44. Both models involve relevant factors thanks to which we revealed via analysis of regression models their significant influence on the probability that internet user is a 25–44 years old woman. These partial logistic models serve as a demographic targeting, using behavioural characteristics of users and affinity of webpages. We revealed with the help of models that are areas of interests typical for female and male gender and which ones are typical/not typical for age categories 13–24, 25–44 and 45–65. The influence of individual factors considered in models was quantified through odds ratios. Our research confirmed that areas of interest of potential customers identified on the basis of their online behavior offer relevant information for their assigning into demographic groups. Along with, male and female interest areas have been identified together with areas of interest associated with younger, middle and older generation. These issues are covered in reasearch questions *RQ2* and *RQ3*. Moreover, we quantified to which extend interest area is typical for particular demographic segments defined on the basis of age and gender of internet users.

However, the main intention was prediction; hence it was necessary to verify prediction quality of models. For these purpose commonly respected classification measures have been used, that confirmed a relatively high fit of the real classification and the one estimated by models on training as well as validation sample of data. Results from our analyses indicates that logistic regression is suitable tool for effective targeting of online advertising in case of market segmentation according to both binomial and multinomial demographic attributes *(RQ1).* Procedure introduced in this paper may be applied also to other demographic segments analysed in the paper.

Models based on procedures presented in this paper may significantly help advertising agencies to approach desired target segment, resulting in increased achievements of their online advertising campaign. However, agency has to decide which users will be exposed to online advertising. This decision should be based on probability that internet user is from demographic segment this advertising is focused on. Hence, paper provides also the illustration of prediction by means of binomial and multinomial model. Following the assumption of mutually independent phenomena like a gender and age category of an individual internet user we calculated probability this user belongs to the target group as a product of partial probabilities estimated from mentioned two logistic models.

To the best of our knowledge, we think that thanks to the application of sophisticated statistical methods on a huge data sample acquired from a real advertising campaign our paper provides relevant findings that will be beneficial for online marketing. Presented procedures and achieved results of own authors' analyses may help marketers to ensure better targeting of internet advertising and therefore contribute to its increased effectiveness. Our research also proved that methodologically logistic regression is a statistical tool suitable for online marketing what is in accordance with previously realized studies (e.g. Bogaert et al., 2017; Olivier et al., 2014; Shan et al., 2016). Authors are also aware of the limits of their research. The dataset includes only internet users from one marketing campaign; hence the results can be generalized only with the respect to this fact. Moreover, we have to realize that online market is rapidly changing so presented results have timely limited validity. However, the benefits of the paper are not only empirical results, but the paper also points to the methodology of application of logistic regression in targeting of online advertising. We may conclude that our research demonstrates useful implications for both theory and practice and fill the gap in the specific field of demographic targeting in online marketing.

Acknowledgement: *This paper was elaborated with the support of project VEGA 1/0488/20 Market, marketing, legislative and ethical aspects of biometric technologies utilisation in commercial sector and project KEGA 007EU-4/2020 Interactive and interdisciplinary teaching of Services and Tourism Innovation subjects using information technology.*

References

Allison, P. D. (2012). *Logistic Regression using SAS. Theory and Application* (2nd ed.). Cary, NC: SAS Institute.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics, 16*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Bogaert, M., Ballings, M., Hosten, M., & Van den Poel, D. (2017). Identifying Soccer Players on Facebook Through Predictive Analytics. *Decision Analysis, 14*(4), 274–297. https://doi.org/10.1287/deca.2017.0354

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16. https://doi.org/10.1016/j.ejor.2008.04.006

Braun, M., & Moe, W. W. (2013). Online Display Advertising: Modeling the Effects of Multiple Creatives and Individual Impression Histories. *Marketing Science, 32*(5), 753–767. https://doi.org/10.1287/mksc.2013.0802

Chaffey, D., & Smith, P. R. (2017). *Digital Marketing Excellence: Planning, Optimizing and Integrating Online Marketing*. London: Taylor & Francis.

Dalessandro, B., Hook, R., Perlich, C., & Provost, F. (2015). Evaluating and Optimizing Online Advertising: Forget the Click, but There Are Good Proxies. *Big data, 3*(2), 90–102. https://doi.org/10.1089/big.2015.0006

De Bock, K., & Van den Poel, D. (2010). Predicting Website Audience Demographics forWeb Advertising Targeting Using Multi-Website Clickstream Data. *Fundamenta Informaticae, 98*(1), 49–70. https://doi.org/10.3233/FI-2010-216

Dave, K., & Varma, V. (2014). Computational Advertising: Techniques for Targeting Relevant Ads. *Foundations and Trends® in Information Retrieval, 8*(4–5), 263–418. https://doi.org/10.1561/1500000045

Chen, J., & Stallaert, J. (2014). An Economic Analysis of Online Advertising Using Behavioral Targeting. *MIS Quarterly, 38*(2), 429–449. http://dx.doi.org/10.2139/ssrn.1787608

Eurostat. (2019). *Internet use and activities*. Retrieved April 10, 2019, from http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York, NY: Wiley-Interscience Publication.

IAB Slovakia. (2018). Výdavky do internetovej reklamy [Internet advertising spending]. Retrieved April 12, 2019, from https://www.iabslovakia.sk/vydavky-do-reklamy/objemy-internetovej-reklamy-sk-2017/

Idrees, F., Rajarajan, M., Conti, M., Chen, T. M., & Rahulamathavan, Y. (2017). PIndroid: A novel Android malware detection system using ensemble learning methods. *Computers & Security, 68,* 36–46. https://doi.org/10.1016/j.cose.2017.03.011

Kapasný, J., & Řezáč, M. (2013). Three-way ROC analysis using SAS Software. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 61*(7), 2269–2275. https://doi.org/10.11118/actaun201361072269

Kim, K., & Timm, N. (2006). *Univariate and Multivariate General Linear Models: Theory and Applications with SAS*. Boca Raton, FL: Chapman and Hall/CRC.

Kolman, L. K., Noorderhaven, N. G., Hofstede, G., & Dienes, E. (2003). Cross-cultural differences in Central Europe. *Journal of Managerial Psychology, 18*(1), 76–88. https://doi.org/10.1108/02683940310459600

Kostelic, K., & Pavlovic, D. K. (2018). Econometric assessment of consumers' personality biases and communication preferences correlation. *E&M Economics and Management, 21*(3), 141–154. https://doi.org/10.15240/tul/001/2018-3-009

Lissitsa, S., & Kol, O. (2016). Generation X vs. Generation Y – A decade of online shopping. *Journal of Retailing and Consumer Services, 31,* 304–312. https://doi.org/10.1016/j.jretconser.2016.04.015

Littell, R. C., Stroup, W. W., & Freund, R. J. (2010). *SAS for Linear Models* (4th revised ed.). Cary, NC: SAS Institute.

Match2One. (2019). *What is Programmatic Advertising? The Ultimate Guide (2019)*. Retrieved April 15, 2019, from https://www.match2one.com/blog/what-is-programmatic-advertising/

McClure, A. C., Tanski, S. E., Li, Z., Jackson, K., Morgenstern, M., Li, Z., & Sargent, J. D. (2016). Internet Alcohol Marketing and Underage Alcohol Use. *Pediatrics, 137*(2), e20152149. https://doi.org/10.1542/peds.2015-2149

Miralles-Pechuán, L., Ponce, H., & Martínez-Villaseñor, L. (2018). A novel methodology for optimizing display advertising campaigns using genetic algorithms. *Electronic Commerce Research and Applications, 27,* 39–51. https://doi.org/10.1016/j.elerap.2017.11.004

ML Wiki. (2015). *Precision and Recall*. Retrieved April 18, 2019, from http://mlwiki.org/index.php/Precision_and_Recall#Averaging

Olivier, C., Eren, M., & Rosales, R. (2014). Simple and Scalable Response Prediction for Display Advertising. *ACM Transactions on Intelligent Systems and Technology, 5*(4), 1–34. https://doi.org/10.1145/2532128

Pagendarm, M., & Schaumburg, H. (2001). Why are Users Banner-Blind? The Impact of Navigation Style on the Perception of Web Banners. *Journal of Digital Information, 2*(1).

Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technology, 2*(1), 37–63. https://doi.org/10.9735/2229-3981

Recode. (2018). *Advertisers will spend $40 billion more on internet ads than on TV ads this year.* Retrieved April 12, 2019, from https://www.recode.net/2018/3/26/17163852/online-internet-advertisers-outspend-tv-ads-advertisers-social-video-mobile-40-billion-2018

Shan, L., Lin, L., Sun, C., & Wang, X. (2016). Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization. *Electronic Commerce Research and Applications, 16,* 30–42. https://doi.org/10.1016/j.elerap.2016.01.004

Shouters Voice. (2018). *What Are The Advantages And Disadvantages Of Digital Marketing In 2018.* Retrieved April 10, 2019, from http://www.shoutersvoice.com/advantages-and-disadvantages-of-digital-marketing/

Schuh, A. (2000). Global standardization as a success formula for marketing in Central Eastern Europe? *Journal of World Business, 35*(2), 133–148. https://doi.org/10.1016/S1090-9516(00)00029-8

Schuh, A., & Holzmüller, H. (2003). Marketing Strategies of Western Consumer Goods Firms in Central and Eastern Europe. In H. J. Stüting, W. Dorow, F. Claassen, & S. Blazejewski (Eds.), *Change Management in Transition Economies*. London: Palgrave Macmillan.

Skinner, H., Kubacki, K., Moss, G., & Chelly, D. (2008). International marketing in an enlarged European Union: Some insights into cultural heterogeneity in Central Europe. *Journal of East European Management Studies, 13*(3), 193–215. Retrieved June 9, 2020, from www.jstor.org/stable/23281167

SPIR. (2019). *28,6 miliard korun investovali zadavatelé do internetové reklamy v roce 2018. Více než polovina obchodů proběhla programaticky [The clients invested 28.6 billion crowns in Internet advertising in 2018. More than half of the transactions took place programmatically].* Retrieved April 10, 2019, from http://www.spir.cz/28-6-miliard-korun-investovali-zadavatele-do-internetove-reklamy-v-roce-2018-vice-nez-polovina

Stankovičová, I., & Vojtková, M. (2007). *Viacrozmerné štatistické metódy s aplikáciami [Multidimensional statistical methods with applications].* Bratislava: Iura Edition.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics* (In press, corrected proof). https://doi.org/10.1016/j.aci.2018.08.003

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: Thomson South-Western.

Yoo, C. Y., Kim, K., & Stout, P. A. (2004). Assessing the Effects of Animation in Online Banner Advertising: Hierarchy of Effects Model. *Journal of Interactive Advertising, 4*(2), 49–60. https://doi.org/10.1080/15252019.2004.10722087