Identification of factors that can cause mobile phone customer churn with application of symbolic interval-valued logistic regression and conjoint analysis

Marcin Pełka[1], Aneta Rybicka[2]

Abstract

The Polish mobile market is a fast growing one: according to forecasts, in 2020 the whole telecommunication market will be worth 47.78 bln PLN. At the end of 2017 the number of active sim cards reached 53.3 mln. At the same time, post-paid services were more popular than pre-paid services. As customers can easily change their current phone operator to another, there arises a key question: what factors can cause customer churn on the Polish mobile phone market? To identify the main factors, conjoint analysis and symbolic logistic regression (with centers approach for model estimation) are used. Both techniques allow different groups of factors to be identified. Conjoint analysis focuses on preferences, while symbolic logistic regression allows main factors of customer loyalty to be identified. Both results complement each other and allow a more comprehensive look at mobile customer churn.

Keywords: *customer churn, conjoint analysis, symbolic interval-valued logistic regression, mobile phone market*

JEL Classification: *C01, C81, D12*

1. Introduction

Customer churn, known also as customer turnover, customer attrition or customer deflection, is a major concern for a number of industries (e.g. banks, internet service providers, insurance companies and telephone service companies). Customer churn is particularly acute in the competitive and quite liberalised mobile telecommunication industry (Keaveney 1995).

The costs of gaining new customers are usually five to even six times higher than the costs of retaining an existing customer (Bhattacharya, 1998). Such costs vary when considering different countries and can vary from 300 USD to 600 USD (Athanassopoulos, 2000).

There are many studies concerning customer churn in general and mobile customer churn (see for example Xia and Jin 2008; Richter, Yom-Tov and Slonim 2010; Neslin et. al. 2006; Burez and Van den Poel 2009; Śmiatacz 2012). But usually these studies are focused only on one side of the customer churn problem – customer loyalty – without taking into account elements of the current offer (especially customer preferences).

As Polish customers can change their mobile phone operator to another one quite easily, the following question arises: what factors can cause churn on the Polish mobile phone market? The main aim of the paper is to identify main factors that can cause mobile phone customer churn. To obtain this goal, conjoint analysis and symbolic logistic regression are used (with the

[1] Wroclaw University of Economics, Faculty of Economics, Management and Tourism, Department of Econometrics and Computer Science, 118/120 Komandorska St., 53-345 Wroclaw, e-mail: marcin.pelka@ue.wroc.pl.

[2] Wroclaw University of Economics, Faculty of Economics, Management and Tourism, Department of Department of Econometrics and Computer Science, 118/120 Komandorska St., 53-345 Wroclaw, e-mail: aneta.rybicka@ue.wroc.pl.

centers method for model estimation). Both techniques allow different groups of factors to be identified. Conjoint analysis focuses on preferences, while symbolic logistic regression allows main factors of customer loyalty to be identified. The results are supplementary to each other and allow us to get a more comprehensive look at mobile customer churn.

2. Symbolic logistic regression and ensemble symbolic logistic regression

Symbolic objects can be described by the following variable types (see e.g. Bock and Diday 2000; Diday and Noirhomme-Fraiture 2008; Billard and Diday 2006; Noirhomme-Fraiture and Brito 2011):

1. Quantitative (numerical) variables:
   a) numerical single-valued variables,
   b) numerical multi-valued variables,
   c) interval-valued variables,
   d) histogram variables.
2. Qualitative (categorical) variables:
   a) categorical single-valued variables,
   b) categorical multi-valued variables,
   c) categorical modal variables.

In general, this kind of data allows objects to be described more precisely, but it requires new, special methods and algorithms. More details about symbolic variables and objects can be found in e.g. Bock and Diday (2000), Billard and Diday (2006), Diday and Noirhomme-Fraiture (2008), Noirhomme-Fraiture and Brito (2011).

In logistic regression for symbolic data, as in logistic regression for classical data, we model binomial (binary, dichotomous) variables (e.g. y – employment status: 1 – employed, 0 – unemployed), while explanatory variables are symbolic interval-valued variables. The general multivariate regression model can be written as:

$$Y_t = b_0 X_{0t} + b_1 X_{1t} + \cdots + b_m X_{mt} + e_t = \sum_{j=0}^{m} b_j X_{jt} + e_t \,, \tag{1}$$

where: Y – dependent variable, $X_0, X_1, \ldots, X_m$ – explanatory (dependent) variables, $b_0, b_1, \ldots, b_m$ – model coefficients, e – model error, $t = 1, \ldots, T$ – observation number, $j = 0, 1, \ldots, m$ – variable number.

In the logit model we assume that we are dealing with a latent variable y^* that cannot be directly observed. However we can observe:

$$y_i = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leqslant 0 \end{cases} . \tag{2}$$

The probability that the independent variable y_i will be 0 or 1 is defined as follows:

$$P_i = F(x_i^T b) = \frac{1}{1 + \exp(-x_i^T b)} = \frac{\exp(x_i^T b)}{1 + \exp(-x_i^T b)} . \tag{3}$$

Symbolic interval-valued data is represented by an interval $[\underline{x}_i\ \overline{x}_i]$, where $\underline{x}_i$ – is the lower bound of an interval-valued variable, and $\overline{x}_i$ – is the upper bound of an interval-valued variable.

De Souza et. al. (2011) describe four different approaches of probability estimation (de Souza et. al. 2011, 275–278):

1. The centers method, where each interval is represented by its center (midpoint) $\frac{\underline{x}_i+\overline{x}_i}{2}$. The probability (3) is estimated for centers of explanatory (independent) variables.
2. The lower and upper bounds method, where each interval is represented by its $\overline{x}_i$ lower and upper bound, respectively. The probability (3) can be estimated:
   a) conjointly for lower and upper bounds, so we get 2*m* variables, and the final probability is estimated for them,
   b) separately for lower and upper bounds (so two different models are estimated – one for lower, and one for upper bounds). Final probability is calculated as the mean of these two probabilities.
3. The vertices method, where instead of symbolic interval-valued variables we have a combination of lower and upper bounds that are represented by an **M** matrix. If we have one object and two interval-valued variables $[\underline{x}_{11}\ \overline{x}_{11}]$, $[\underline{x}_{21}\ \overline{x}_{21}]$, then the **M** matrix is:

$$\mathbf{M} = \begin{bmatrix} \underline{x}_{11} & \underline{x}_{21} \\ \underline{x}_{11} & \overline{x}_{21} \\ \overline{x}_{11} & \underline{x}_{21} \\ \overline{x}_{11} & \overline{x}_{21} \end{bmatrix}. \qquad (4)$$

The final probability in the vertices method is calculated as the mean, maximum or minimum probability for these combinations (de Souza et. al. 2011, 277).

Usually the centers, both lower and upper bound approaches obtain better results than the vertices method for symbolic interval-valued logistic regression (de Souza et. al. 2011).

In this paper the centers method will be used for model estimation.

In the ensemble model *D* different models $D = (D_1, ..., D_l)$ are combined to obtain one single (aggregated) model (D*) that obtains better results (in terms of lower error) than any of the models that are the part of the ensemble.

In this paper, the bagging approach will be used. The initial data set is divided into $U_1, ..., U_l$ subsets (subsamples) that are drawn with replacement from the initial data set. Then each subset is used to build a model (obtain regression results). In the case of this paper, 20 subsets will be taken.

Final results (final probabilities in the case of logistic regression) are obtained by averaging all results (see for example Polikar 2007, 60–61).

3. Conjoint analysis

Conjoint analysis is a powerful market research technique that measures how people make decisions based on certain features of a product or service. This method is well-described in the literature (see for example Orme 2006; Sagan 2013; Gustafsson et. al. 2013; Rao 2014; Wiley et. al. 2014), so only the most important elements will be presented in this paper.

The conjoint method originated in mathematical psychology and was also developed beginning in the mid-sixties by researchers in marketing and business. Conjoint analysis is a statistical method for finding out how consumers make trade-offs and choose among competing products or services. It is also used to predict (simulate) consumers choices for future products or services (see for example: Sagan 2013).

The main aim of conjoint analysis is to estimate part-worth utilities for all attribute levels. The part-worth utilities are estimated for each respondent separately, and as an average values for the whole sample. Estimated utilities allow the following to be estimated: total utilities of a profile for all respondents, average total utilities in the sample, average attribute importance and average total utilities in the segments (clusters, groups) of respondents. In conjoint analysis, attributes (also called factors) are used to describe explanatory variables describing goods or services. Attribute levels describe values of attributes and profiles (stimuli, runs, treatments) that are variants of goods or services.

The most important feature of conjoint analysis based on a full profile method is that the number of attributes taken into consideration is usually limited to six. The profiles are described using all attributes and are presented to the respondents to be assessed. Profiles are generated on the basis of the orthogonal factor system and are maximally and mutually varied. All respondents evaluate the same set of profiles. Conjoint analysis, which represents the decomposition approach, can take into account main effects and the effects of an attribute interaction (Green and Srinivasan 1978).

4. Empirical results

For the purposes of symbolic interval-valued logistic regression, 109 mobile phone users (a convenience sample) from Lower Silesia were asked to answer the following questions (opinions):

y – Would you consider changing your current mobile phone operator (0 – no, 1 – yes),

x_1 – “I use the services of my mobile phone operator as they are the best choice for me”,

x_2 – “If I could, I would make the same choice again” (probability of making the same choice again),

x_3 – “I use the services of the same mobile phone operator over time”,

x_4 – “I would consider using services of another mobile phone operator if the prices of my mobile network rose slightly”,

x_5 – “If had a chance, I would try the services of other mobile phone operator”,

x_6 – “I would consider using the services of another mobile phone operator if the current one had some technical issues”,

x_7 – “My current mobile phone operator provides better services than others”,

x_8 – “In my opinion, the services of my current mobile phone operator are not better than competitors”,

x_9 – “I’m sharing positive opinions about my mobile phone operator with other people”,

x_{10} – “I would recommend my current mobile phone operator to other people”,

x_{11} – “I feel that I’m emotionally bound to my current mobile phone operator”,

x_{12} – “I use the services of my mobile phone operator as I want to (I don’t have to)”.

All of the questions had the following intervals that reflect the willingness (probability) of the respondent to agree with a question (statement): [0; 15], [7; 25], [25; 70], [45; 80], [60; 100].

Besides that, two open-ended questions were taken into account:

x_{13} – “I usually spend not less than … (insert an amount), but surely not more than … (insert an amount) monthly on my mobile services”,

x_{14} – “My mobile phone calls usually last from … (insert an integer) to … (insert an integer) minutes”.

The model was estimated using glm function from the stats package of R software. The lrtest function from the lmtest package of R (see Hothron et. al. 2018) software was used to check the model.

The estimated model obtained the results that are shown in the Table 1.

Table 1. Results of estimation

Coefficients	Estimate	Standard error	z-value	p-value	odds ratio
Intecept	0.032936	3.617657	0.009	0.9927	1.0334840
x_1	0.052520	0.029475	1.782	0.0748*	**1.0539238**
x_2	0.046198	0.025659	1.800	0.0718*	**1.0472815**
x_3	−0.053805	0.025669	−2.096	0.0361*	0.9476170
x_4	−0.020768	0.024264	−0.856	0.3920	0.9794458
x_5	−0.021036	0.022383	−0.940	0.3473	0.9791837
x_6	0.045450	0.025999	1.748	0.0804*	1.0464985
x_7	−0.092530	0.047409	−1.952	0.0510*	0.9116219
x_8	0.069309	0.042100	1.646	0.0997*	**1.0717672**
x_9	0.029991	0.030101	0.996	0.3191	1.0304456
x_{10}	−0.016937	0.041599	−0.407	0.6839	0.9832058
x_{11}	−0.006965	0.019349	−0.360	0.7189	0.9930596
x_{12}	0.011584	0.030175	0.384	0.7011	1.0116510
x_{13}	−0.032541	0.024372	−1.335	0.1818	0.9679829
x_{14}	0.027575	0.019164	1.439	0.1502	1.0279585

Variables significant at α=0.1 level are marked with “*”

Highest odd ratios for significant variables are in bold.

The null deviance for the centers model was equal to 90.927 (on 108 degrees of freedom), while residual deviance was equal to 51.091 (on 94 degrees of freedom). The Akaike information criterion (AIC) reached 81.091.

Variables x_1, x_2, x_3, x_6, x_7 and x_8, are the only relevant ones ($\alpha = 0.1 > p - value$). Consequently, they are the only ones to be interpreted in the further analysis.

The most important variable, in terms of the odds ratio, is the x_8 ("In my opinion, the services of my current mobile phone operator are not better than competitors"). Increasing midpoint values for this variable raise the chance that a customer will consider changing his/her current mobile phone provider. Variables x_1 ("I use the services of my mobile phone operator as they are the best choice for me"), x_2 ("If I could I would make the same choice again"), x_6 ("I would consider using the services of another mobile phone operator if my current one had some technical issues"), have a slightly lower, but positive impact on the probability that the customer will consider changing his/her current mobile phone provider. Variables x_3 ("I use the services of the same mobile phone operator over time") and x_7 ("My current mobile phone operator provides better services than others") have a negative impact on the chance of changing mobile phone providers. Increasing midpoint values for these variables lower the probability that the customer will consider changing his/her current mobile phone provider.

The regression model was checked if it is relevant with the application of the lrtest function from the lmtest package of R software. As the whole regression is relevant.

All model pseudo R^2 values (Efron's, McFadden's, Nagelkerke's and square of correlation coefficient between empirical and theoretical values) were equal for the single model 0.8098423 (Nagelkerke's pseudo R^2), and all other others were similar, reaching a value of around 0.70. These values were slightly better for ensemble model (0.8728378 in terms of Nagelkerke's pseudo R^2) and around 0.74 in terms of other pseudo R^2 values.

The same 109 respondents also evaluated 17 profiles with the following attributes and levels:

1) phone operator with the following levels – T-Mobile, Orange, Play, Other
2) offer type – with phone, without phone,
3) contract type – post-paid, pre-paid, duet,
4) monthly fee (rate) – up to 30 PLN, 30–50 PLN, 50–100 PLN, more than 100 PLN,
5) supplementary promotions – all services without limitations, some services are unlimited.

The customers were asked to evaluate profiles using a scale from 1 to 10, where 1 signifies, "I would not choose this" and 10, "I'll absolutely choose this."

In the case of this paper the output variable differs slightly from the typical cases, as it reflects a negative, not positive, action that can be performed by customers.

The conjoint model was estimated using the conjoint package of R software. The average importance of all factors is presented in Figure 1.

When considering all factors, the monthly fee was the most important one (36.07% with the highest utilities for two first levels (fee up to 30 PLN, fee 30 to 50 PLN) then mobile phone operator (28.01% with positive values for Plus and Play and negative for others). The least important was the offer type (9.72% with positive values for offer with phone, while the offer without the phone has a negative value).

Positive values were also recorded for services without limitations (in supplementary promotions factor) and Duet and pre-paid services (in terms of contract type). All part-worth utilities for all attribute levels are shown in the Table 2.

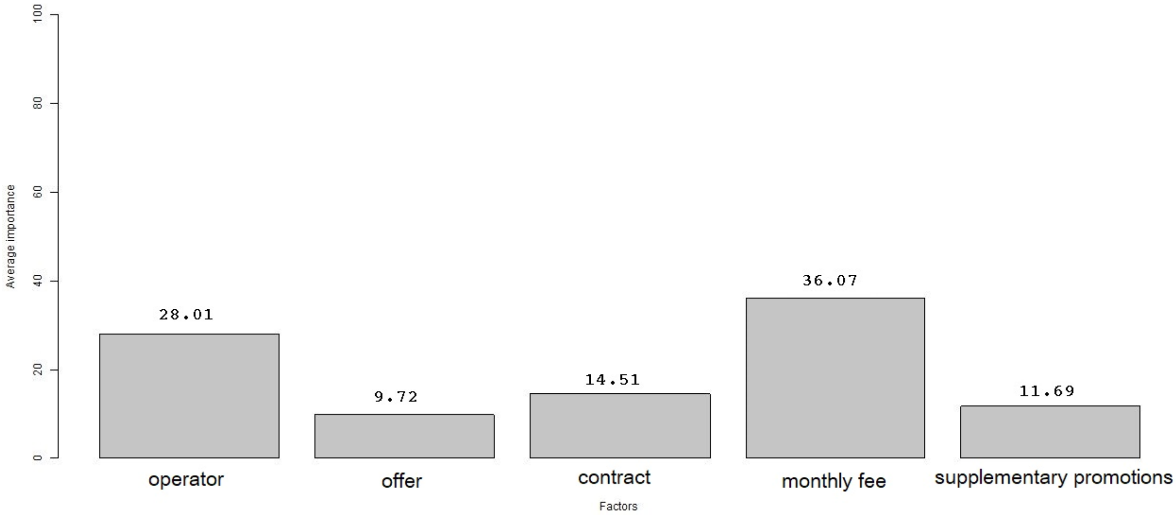

Fig. 1. Average importance of all factors

Table 2. Part-worth utilities

Levels	Utilities	Levels	Utilities
Intercept	3.6456	Duet	0.0825
Plus	0.4944	up to 30 PLN	1.2696
T-Mobile	−0.1389	30–50 PLN	0.5294
Orange	−0.1496	50–100 PLN	0.4480
Play	0.3071	more than 100 PLN	−1.3510
Other	−0.5130	all services without limitations	0.2525
with phone	0.2852		
without phone	−0.2852	some services without limitations	−0.2525
post-paid	−0.1191		
pre-paid	0.0365		

Conclusions

Symbolic logistic regression discovered that customer opinion about services when compared to other competitors (“In my opinion, the services of my current mobile phone operator are not better than competitors”) is the most important factor. The increasing midpoint values for this variable raise the chance that a customer will consider changing his/her current mobile phone provider. Furthermore, customers were more likely to change their provider depending on their loyalty to their present operator, whether there might be technical problems with their operator or whether another operator made a better offer. Continuous use of the same services, and the feeling that current services are better than those of competitors, decrease the probability of mobile phone provider change.

Conjoint analysis found that when taking into consideration customer preferences, the monthly payment is the most important factor (rising values encourage customers to change mobile phone providers and decrease the likelihood that they will choose a more expensive offer). Other important factors include the phone operator, contract type and supplementary promotions and offer type.

References:

Athanassopoulos, A.-D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of business research*, 47(3), 191–207.

Bąk, A., & Bartłomowicz, T., The conjoint package for R software. www.r-project.org.

Billard, L., & Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining.* John Wiley.

Bhattacharya, C.B. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of the academy of marketing science*, 26(1), 31–44.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626–4636.

de Souza, R.M., Queiroz, D.C., & Cysneiros, F.J.A. (2011). Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, 14(3), 273.

Diday, E., & Noirhomme-Fraiture, M. (Eds.). (2008). *Symbolic data analysis and the SODAS software*. John Wiley & Sons.

Diday, E., & Bock, H.H. (2000). Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data.

Green, P.E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research*, 5(2), 103–123.

Gustafsson, A., Herrmann, A., & Huber, F. (Eds.). (2013). *Conjoint measurement: Methods and applications*. Springer Science & Business Media.

Hothron, T., Zeileis, A., Farebrother, R.-W., Cummins, C., Millo, G., & Mitchell, D. (2018), The lmtest package for R software. www.r-project.org.

Keaveney, S.-M. (1995). Customer switching behavior in service industries: An exploratory study. *The Journal of Marketing*, 71–82.

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., & Mason, C.-H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, 43(2), 204–211.

Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, 4(2), 157–170.

Orme, B.K. (2006). *Getting started with conjoint analysis: strategies for product design and pricing research*.

Polikar, R. (2007). Bootstrap-inspired techniques in computation intelligence. *IEEE signal processing magazine*, 24(4), 59–72.

Rao, V.R. (2014). *Applied conjoint analysis*. New York, NY: Springer.

Richter, Y., Yom-Tov, E., & Slonim, N. (2010, April). Predicting customer churn in mobile networks through analysis of social groups. In: *Proceedings of the 2010 SIAM international conference on data mining*, 732–741. Society for Industrial and Applied Mathematics.

Sagan, A. (2013). Market research and preference data. *The SAGE handbook of multilevel modeling. London: SAGE Publications Ltd*, 581–99.

Śmiatacz, K. (2012). *Badanie satysfakcji klientów na przykładzie rynku usług telefonii komórkowej w Polsce*. Wydawnictwo Uczelniane Uniwersytetu Technologiczno-Przyrodniczego w Bydgoszczy.

Wiley, J.B., Raghavarao, D., & Chitturi, P. (2010). *Choice-based conjoint analysis: models and designs*. Chapman and Hall/CRC.

Xia, G.E., & Jin, W.D. (2008). Model of customer churn prediction on support vector machine. *Systems Engineering-Theory & Practice*, 28(1), 71–77.