CRM FEJLESZTÉSE A GYAKORLATBAN, ÜGYFÉLKAPCSOLATI MENEDZSMENT INNOVÁCIÓ
CRM DEVELOPMENT IN PRACTICE, CUSTOMER RELATIONSHIP MANAGEMENT INNOVATION

[1] Dr. Janos Papp, [2] Dr. Ede Lázár, [3] Ágnes Brix

[1] Associate professor at Szent István University
[2] Associate professor and vice dean at Sapientia Hungarian University of Transylvania
[3] PhD student at Szent István University

Összefogalalás:

A tanulmányban egy olyan ügyfélkapcsolati modellt mutatunk be, amely nem csak innovatív elemeket visz be egy konkrét szolgáltató ügyfélkapcsolat menedzsmentjébe (Customer Relationship Management - CRM), hanem újraszervezi az ügyfélszolgálati folyamatot is. A modell elsődleges célja, hogy növelje az ügyfelek elégedettségét, ezzel összhangban csökkentse a távozó ügyfelek, a lemorzsolódás arányát és elősegítse új ügyfelek szerzését. A modell legfontosabb eredményei az ügyfelenkénti lojalitás-mérés, ami közvetlenül meghatározhatja az ügyfélkapcsolatot, továbbá a magyarázó változók becsült paramétereinek vizsgálata által képet kapunk a különböző kedvezmények, CRM elemek és CRM csatornák hatékonyságáról.

Abstract:

This research shows a customer relationship model, which does not only implement innovative elements into a specific supplier’s customer relationship management (CRM), but also reorganizes the customer service’s procedure. The primary aim of the model is to increase customer satisfaction, and to decrease the number of leaving customers, and furthermore to win new customers. The most important results of the model is the customer loyalty measure, which directly determines the customer relationship. Furthermore the efficiency of the different promotions, CRM elements and CRM channels, is shown by the analysis of the estimated explanatory variable’s parameters.

Keywords: *Costumer relationship, predicting model, CRM*

Introduction

In this paper a customer relationship model is shown, which does not only implement innovative elements into a specific supplier’s, the TIGÁZ LTD, customer relationship management (CRM), but also reorganizes the customer service’s procedure.

The primary aim of the model is to increase customer satisfaction, and with this to decrease the number of leaving customers, and furthermore to help win new customers.

Model description:

- Analytical and predictive: The analytical CRM does not only control and synchronies the customer relationship procedures, but also adds customer related values with the use of mathematical methods. Some of these are also suitable for predictive functions, thus the model can predict the behaviour of the customer and the probability of dropout, based on historical data.

10.18531/Studia.Mundi.2015.02.02.136-148

Dynamic: the model changes according to the newly collected data. This dynamic property means continuity at the data recording level, but at the CRM model level it means that model specification is performed at predefined intervals or ad-hoc. The model is not only dynamic form the method point of view, but in functionality also. The output continuously improves the customer service activity.

Integrated: The model integrates all information coming from the customer relationship channels (form personal, telephone, online, mail sources) and from the customer relationships (inspection, error corrections, etc.). Then loops the results back using differential method.

Customisable according to segments: after the CRM data base, data structure and model specifications problems are solved and done, an analogue CRM model can be created cost efficiently along the customer segments. This type of macro-segmentation results from the model functionality, as there are separate models determining the most attracting promotions for the potential customers, the factors increasing satisfactions and decreasing attrition among the current customers. Furthermore there is possibility to apply techniques, which result a range of more accurate segmentations.

Customer relationship characteristics – research results

The model aiming customer relationship innovation is based on the revealed results from the empirical primer research done by the "GfkHunária". The results describing the detailed process of the customer relationships, where obtained from a wide range mystery shopping, and are grouped separately for potential and current customers according to the model specification requirements. In the following, those customer relationship characters are determined, which help to reach the CRM model's aims and can be appropriately quantified for the model specification aims.

Current customers, the factors influencing satisfaction

An important lesson learned from the research, is that an uniform argument system is necessary, which is used by the customer service managers to convince the unsatisfied customers to stay. The few reasons mentioned during the mystery shopping, can be categorised as ad-hoc.

During the personal administration "the administrators at TIGÁZ tried very rarely to convince the unsatisfied customers to stay in contract with the supplier. In the few cases they did dry to keep a customer, the following reasons were mentioned."

The experience from the mystery shopping over telephone, is that mostly the administrators at TIGÁZ and FŐGÁZ "fought" for their customers. The most often mentioned convincing reasons were:

- The infrastructure is owned by the current supplier, the new supplier would only do the billing, thus it surely won't be any cheaper.
- Neither will the quality be higher nor will the price be lower. It's not worth it.
- The price is the magisterial price controlled by the ministry; it must be applied by everyone.
- The technical administration with the new supplier will be more difficult and circumstantial.
- Premium card, Remote billing."

During the testing of the Back Office, half of the customers who were planning to break contract with Tigáz received a mail, which tried to convince them to stay. One third of these letters were successful according to the customers.

From these observations the following requirement, functions need to be fulfilled by the CRM model:

From the integrated customer relationship database the model makes predictions for every customer about the probability of attrition and determines the factors influencing this probability for the whole database. Such explanatory variables can be data referring to customers (age, gender, address), data referring to the service (size of the bill, length of the service being used, reporting faults due to low heat value, etc.) monitoring frequency. Exogenous data referring to prices, marketing activity of competitors can also be implemented into the model.

The customer service staff needs to be motivated, they need to fell the control. Even the best CRM model cannot work efficiently, if the staff has no interest in the application. From this point of view the CRM model is a control tool, it also represents the technology generating the awareness of control.

Offer the most attractive offer for the customer. Based on the model's analytical properties and integration the most efficient promotions can be determined. Using segmenting tools further distinction can be made between the promotions. The model quantifies and arranges the following promotions in order according to its customer keeping effect: Bill-angel, Remote-bill, telephone/online customer service, Agip-TigázPremio card. This list of course may be changed.

In case of braking contract ask about the reason. During the mystery shopping "none of the administrators (not only at TIGÁZ, but also at the competitors), tried to find out the reason for leaving, breaking up with the supplier." Determining the reason for attrition is a key element for the model to make predictions, it is a question whether is it technically feasible.

Customer acquisition

For customer acquisition a new model is required, which is separately specified, despite that it is similar in many points to the previous model. In this case the output of the model is, whether a potential customer will become a real customer or not. The potential customers are tried to be convinced for changing supplier using the following promotions:

- "complementary, comfort" services
- guaranteed gift
- Agip-TigázPremio card or Tigáz voucher
- gift voucher
- take over the administration with the omitted supplier. As the competitors only provide partial information about administration for the supplier-switch, it is assumed to be a strong reason for those who decided to switch supplier. In case of the telephone customer service this only occurred for 71% of the cases. "The administrators at TIGÁZ could be more helpful in this area"

If the customer explains the reason for changing supplier, it can be an important explanatory variable of the model. This type of information needs to be recorded both at personal and at telephone customer service. "The administrators are practically not interested, why the new arriving customers want to switch supplier."

The overview of the suggested solution's model is shown on the following figure :

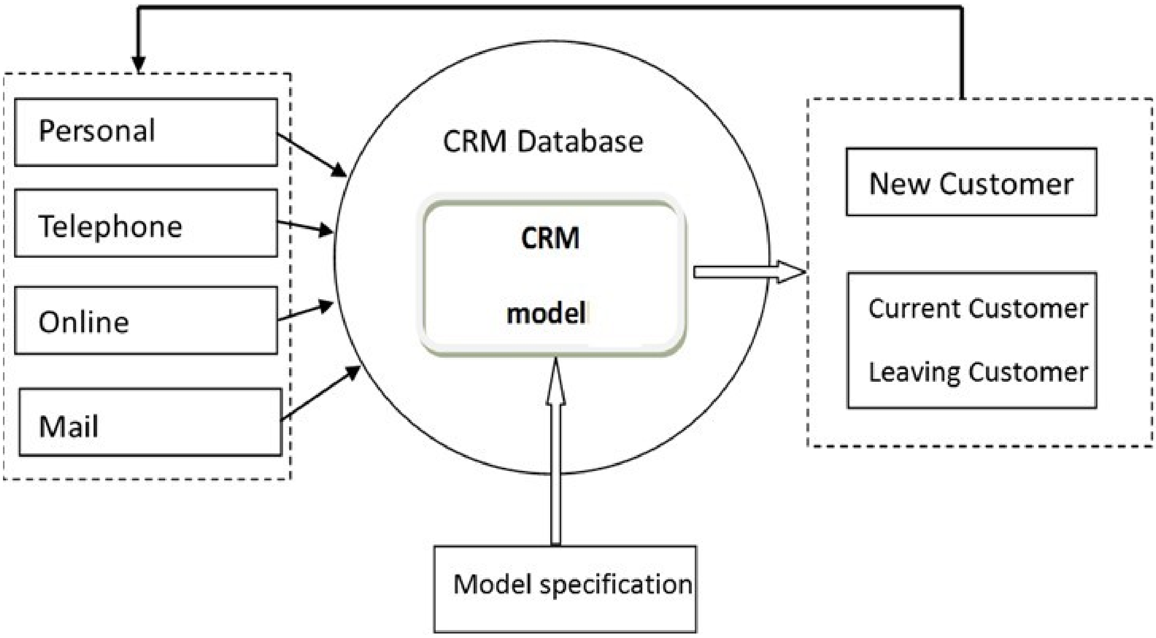


Figure 1.The flow chart diagram of the analytic CRM model

Source: The author

The customer relationship process starts by arranging the information, obtained from the personal, telephone, online and using mail service customer contacts, into the uniform database. This database, which absolutely follows the customer relationship activity and provides the broadest information about the customers, is the base of the CRM model. The model that fulfils the previously described properties and aims is created from the specifications made regularly (monthly, quarter yearly) or ad-hoc (marketing campaign). The CRM model is practically 2 models, because the two aims, customer acquisition and the probability that the current customers stay or leave, use the same method but require different model specifications. The model outputs are looped back to the clients, this way the customer relationship management can be dynamically adjusted according to the requirements.

Analytic CRM model

Customer attrition predicting model

One from the two models, the customer attrition predicting model is detailed, as it requires more complex model specification due to the detailed customer information, than the model estimating the probability of customer acquisition.

The customer attrition predicting model estimates the probability of attrition for the current customers with the help of a predictive mathematical model using data from the previously leaving and staying customers.

The possible solutions of the customer attrition predicting model

Oravecz (2007) suggests the following typology for the predictive models.

I. Conventional methods:
   - Linear probability model
   - probit and logit models
   - discrimination analysis
   - classification tree (recursive partitioning algorithm)
   - linear programming
   - k-th "nearest neighbour" –method

II. Artificial intelligence methods
   - neural networks
   - expert systems
   - genetic algorithms

Another type of grouping can be according to the statistical (econometric) orientedtypology, the parametric and nonparametric models. The difference, from the point of view of measuring the customer service efficiency, is that the nonparametric models predict the outcome usually more precisely as the parametric models. However the parametric models also predict the factors influencing the output. The most commonly used methods for customer attrition predictions are: classification or decision trees, logistical regression model and from the neural networks the parametric logistical regression model.

The classification or decision trees, in other words the Recursive Partitioning Algorithm-RPA, are easy to understand, can be programmed, therefore usually they part of many kinds of business intelligence solutions. The centre node and branches of the classification tree, make a structure which arranges the observation units into homogeneous groups according to the available information. To partition the branches several mechanisms can be used, the models based on the chi-square statistics CHAID (Chi-square Automatic Interaction Detector) are getting more and more widespread.

The neural networks try to reproduce the property of the human brain, according to which the connection strength between the neurons can be varied flexible (Fajszi-Cser, 2004). These changes represent the base of the intelligent learning process. There are several inputs and one target variable, between them are one or several mid layers can be found, which represent activation functions. The algorithm is an iterative process, which changes the parameters of the input variables until the target variables error decreases to possible minimal. To complete this optimization, first a known set of data needs to be selected form the database, where the input variables and corresponding target variables values are known. The input values are multiplied by the weight of the connections, then summed to create the value of the next neuron. The value of the neurons on the output layer gives the estimated value of the target variable. If the target error does not satisfy the requirements, the iterative, learning procedure restarts, so that the weight if the connections are changed by a so called back propagation algorithm.

In this research a customer attrition model based on logistic regression (logit model) was developed. Why is the logistic regression model suggested from the possible mathematical, econometrical models?

These models are widely used for credit rating, bad debit or predicting applications, for a reason. From the method point of view the customer attrition is very similar to the credit rating problem.

“Today the logit model is the most widely used classification process in the credit scoring area. The main reason for this: easily to interpretable, high performance, not only does it classify, but also estimated the probability of default on loans (Bázel II. specification), furthermore the method does not require the explanatory variables to have a normal distribution, this way the category explanatory variables can be easily implemented.” (Oravecz, 2007). Keeping the aims of the analysis in mind, several advantageous features can be listed of the logistic regression based model:

1. Parametric method, thus compared to the neural network the different independent variables effects can be used, with the help of the parameters in the logistic regression equation. This way in the changing market environment, effects of several factors- implemented in the model –can be investigated “ what happens, if...?” using simulations. The predicting capability of the parametric method is slightly weaker than the neutral network’s, because of its internal character. Despite this, the parametric model does not only produce the final out, the probability attrition, but it also gives important information about which factor with what weight is responsible for the attrition.
2. The required statistical preconditions are less strict, than other parametric methods, such as the linear regression model, the linear probability model or the discriminant analysis.
3. Compared to the decision tree or neural network the result is a continuous probability value, from which a dichotomous classification (eq. loyal customer, leaving) can be made. This continuous probability variable can be used in numerous ways, the customers can be classified into more than two groups, the rate oferrors resulting from incorrectly classified customers can also be optimised with respect to their expenses, it can be the base of a customer rating indicator.
4. The fourth advantage is not methodical. Due to the relatively easy specification, easy understanding and fast learning properties, it is relatively cheap.

Logistic regression model

Base equation of the logistic regression model

In the business intelligence one for the most popular model version, from the Categorical and Limited Dependent Variables (CLDV) models, isthe binominal logistic regressive or known as the logit model. Using research methodological definitions, the logistic is such a regression model, in which the dependent variable is a bivalued categorical (dichotomus) variable and the independent variables can be any type: interval, ordinal, nominal. This “technical advantages” has large significance in empirical researches and in variety of applications.

The logistic regression can be considered as the generalization of the lineal regression by extending its limits. One of this important property of the model, is that the target variables value does not exceed the [0-1] interval, produces a probability which is easy to evaluate. The probability of customer attrition can be described with the following logistic regressive equation:

$$\mathrm{P(Y=1)} = \frac{e^{\mathrm{b_0 + b_1x_1 + \cdots + b_kx_k}}}{1 + e^{\mathrm{b_0 + b_1x_1 + \cdots + b_kx_k}}},$$

Where, the independent variables (xi) are the properties of the product and customers, and bi i is the corresponding parameter.

Estimation of the parameters

To estimate the parameters (b_i) of the logistic regression equation, the maximum likelihood (MILE) method is used most commonly. Compared to the ordinary least squares (OLS) method used for the linear regression analysis, where the aim is to minimalise the summed square distances between the observed and estimated value, the parameter estimation of the logistic regression maximises a probability function which is the likelihood function. The likelihood function is the probability which estimates a dependent variables value based on the values of the independent variables. The likelihood function of the discrete dependent variable, it can vary between 0 and 1. The logarithmic version of this function, the log likelihood function can vary between minus infinite and 0.

The maximum likelihood is an iterative algorithm which starts with a random estimation of the logistic equation's parameter, then the direction and magnitude for changing the log likelihood function is determined. After the initial estimation of the function the residues are tested, then the function is estimated again and the process is repeated for about 5 to 8 times until the increment of the function is no longer significant. The initial function of the model is given:

$$LL = -2\{(n_Y=1)\ln[P(Y=1)] + (n_Y=0)\ln[P(Y=0)],$$

where, $n_Y=1$ is the frequency of the event occurrence; P(Y=1) is the probability that the event occurs; Multiplying with -2 is required, because this way the function's distribution will be close to the chi-square distribution and makes it possible to investigate whether the involvement of new explanatory variables significantly increases the value of the likelihood function or not. The testing of the effect is done a chi-square test, which is similar to the linear regression's F-test. The estimation of the parameters is continued till the likelihood functions increment is significant. Amemiya (1985, pp. 110) formally proved, that the log likelihood is globally concave, which means, that according to the Newton-Raphson method whatever value of the initial has it will converge to appropriate maximum and that the ML estimation function is consistent, has asymptotically normal distribution and is asymptotically effective.

All widely spread econometric computer software use the maximum likelihood method for parameter estimation, but there are also two other methods that can be used: weighted non iterative ordinary least squares method and the discriminate function (Hosmer-Lemeshow, 2000.).

Indicators related to the model specification

The predicting property of the attrition-predicting model has high importance, an incorrectly built (specified) model makes false predictions which can result in serious expenses. Therefore in the following the model specification indicators are investigated in detail, in other words the indicators of "model goodness".

The R^2, which shows the explanatory power of the linear regression model cannot be calculated, as the logistic regression model's dependent variable's deviation depends from the variable's distribution also. Greene (2003.) says that the essential difference is that, at the ordinary least squares method the criteria of the b parameter estimation is the maximisation of R^2, whereas during the estimation of the maximum likelihood, not all fitting criteria's maximisation is aimed. Despite this or because of this several indicators were created which are related to the model fitting goodness. Fromm these indicators the ones relevant for the practice, are selected.

These indicators can be classified into two groups: indicators based on the likelihood function's value and the indicators that are based on the models prediction accuracy.

The indicators related to the likelihood function's value

The process of the maximum likelihood's estimation does not have such a clear indicator for the efficiency of parameter estimations, as R^2 is for the ordinary least squares method (Chatterjee-Hadi, 2006.). However several indicators are based on the comparison of the likelihood function's initial and final value.

The quasi or pseudo R^2 formula defined by McFadden:

$$R^2 = 1 - \frac{L_0}{L_1}$$

The Cox-Snell and the Nagelkerke indicator. The Cox-Shell indicator compares the likelihood function's initial a final values, so that the indicator varies between 0 and 1. The problem is, that it never reaches 1, thus the accurate evaluation is not possible

$$R^2 = 1 - \left(\frac{L_0}{L_1}\right)^{2/n}$$

Nagelkerke solved this problem by dividing the Cox-Snell indicator with the maximal value of the sample.

$$R^2 = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - (L_0)^{2/n}}$$

The R^2 developed by McKelvey and Zavoina (1975) differs from the others, it does not calculate based on the Likelihood function, rather calculates the residues similarly to the linear regression R^2. Its advantage is that, it can be used if the dependent variable has more than to values, e.g. for multinomial logit or for probit models.

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i^* - \bar{y}_i^*)^2}{\sum_{i=1}^{N}(\hat{y}_i^* - \bar{y}_i^*)^2 + N\sigma^2}$$

where, $\bar{y}_i^*$is the average of $\hat{y}_i^*$ the estimated values. The value of σ^2 for the logit models are $\sigma^2 = 1/3\pi^2$, and for the probit models $\sigma^2 = 1$. Windmeijer in his analysis (Franses-Paap, 2001.) compared several specification indicators and found the most appropriate indicators are the MCFaddel and the McKelvey-Zavoina, because they do not depend on the sample size yi=1 as much as other indicators.

The informational criterion of Akaike and Bayes. The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) provides information for selection between several models. The two indicators can be calculated the following way:

$$AIC = \frac{1}{N}(-2l(\hat{\beta}) + 2n)$$

$$BIC = \frac{1}{N}(-2l(\hat{\beta}) + n\ln N)$$

Where, n is the number of parameters, N is the sample number and l(β) is the maximum of the β parameters likelihood function. The alternative indicators differing β parameters result different informational criteria values. The indicators on their own have no significance, but when comparing different models, the smaller the AIC and BIC value is, the better the model specification is.

Indicators related to the goodness of the model fitting

The goodness of the model fitting can be defined, by the model's ability to model the dependent variable (Hosmer-Lemeshow, 2000.). The indicators of the fitting's goodness, are all based on the comparison of the dependent variable's real value and the value estimated by the model, therefore they are also called the indicators based on the predictions accuracy.

Classification table. Another indicator of the goodness of the logistic regression model's fitting, is the "classification table", which compares the dependent variable's estimated and actual values. This indicator is very popular in practice, due to easy availability. As it is used very often it is necessary to analyse in detail the pros and cons for its usage.

Based on the model's estimated probabilities, one of the two outputs of the dependent variable is associated to all cases. For this a threshold value (k) needs to be defined, if the probability value is above the threshold product purchase is expected, if the probability value is below the threshold the rejection of the purchase is expected. Let's say that the $\hat{y}_i$ estimated dependent variable's two values are:

$$\hat{y}_i = \begin{cases} 0 & \text{ha } \widehat{\mathrm{P}}(\mathrm{y}=1) \leq \mathrm{k} \\ 1 & \text{ha } \widehat{\mathrm{P}}(\mathrm{y}=1) > \mathrm{k}\,, \end{cases}$$

Where, $\widehat{\mathrm{P}}(\mathrm{y}=1)$ is the estimated probability based on the model. The k threshold is usually 0.5, but the computer software containing the logistic regression analysis allow this value to be changed. The general form of the classification table:

Table 1. Classification Table

Classification Table				
		Prediction		Percent Correct %
		0 - no	1 - yes	
Actual	0 - no	n_{11}	n_{12}	$n_{1.}$
	1- no	n_{21}	n_{22}	$n_{2.}$
Total		$n_{.1}$	$n_{.2}$	n

Source: The author

The rate of the correct predictions can be defined with one value, it is called the classification R_k^2:

$$R_k^2 = \frac{(n_{11} + n_{22})}{n}$$

Where, n is the sample size, n_{11} and $n_{22}$is the frequency of the two types of correct predictions.

Greene (2003.) writes about his reservations about the indicators, it is highlighted, that in case of uneven samples, when the ratio of the ones or zeros are far greater than the category's, the method isn't reliable. The classifications can be modified, by changing the threshold , but while one type of error is reduced the other type increases.

However this problem had been already corrected earlier. Long (1997.) introduces the adjusted classification indicator, which corrects the previous indicator using the category with the largest frequency.

$$R_{ak}^2 = \frac{(n_{11} + n_{22}) - max_r}{n - max_r(n_{r.})}$$

Where,$n_{r.}$ is the row total of tables row *r.*, the frequency of the dependent variable's. The adjusted classification R^2 can be described, as by how much percentage is the error of the prediction reduced if the independent variables are known, compared to when only the boundary probabilities, that is the estimation would be based only on the probabilities $\mathrm{P(y = 0)}$ and $\mathrm{P(y = 1)}$known from the dependent variables distribution. This indicator is equal to the Goodman and Kurskal λ (Long, 1997.)

An important argument against the usage of the indicators originated from the classification table, is that they are highly determined by the dependent variable's distribution in the sample. Similarly to Greene (2003.) Hosmer and Lemeshow (2000.) also emphasises that the classification indicator's value is influenced by relative ratio of the dependent variable's two values. They discovered, that the group with the larger element number will always have better prediction. This is an aspect, which isn't related to the model's fitting goodness. The classification reduces the continuous result variable probability model to a dichotomous result variable model, which is also considered as a disadvantage. It is shown that there is little difference between the estimated probabilities of 0.48 and 0.52, however the use of the 0.5 threshold value divides the two cases into different (opposing) groups. Against this argument is, that the initial model, the observed dependent variable is also dichotomous, thus it is an acceptable expectation, that the estimated dependent variable should also be dichotomous. For example a products demand can be described with a dichotomous variable, the continuous probability variable is latent, it is "only" a partial results required for the calculations. For the threshold value determinations there are several- partial- practical solutions, for example the exclusion of the estimated probabilities near the 0,5 threshold value, or the adjustment of the threshold according to an optimisation aspect. The threshold adjustment will be detailed later on.

Indeed, in case of a model with two different dependent variables, or in case of differing sample the classification indicators cannot be used, because the difference of the two model rather depend on the distribution of the dependent variables, than referring to the "quality" of the models. However this does not excludes the possibility to compare the models to each other using different model-tests, where the dependent variable's distribution is constant. My opinion is that in practice the classification table cannot be neglected for an investigation. Beside the mythological aspects, it must be taken into account that this is an easy to understand quality indicator of the model.

The customer attrition prediction model based on the logistic regression model

Model specification

The aim is to specify a logistic regression model, which determines the probability of customer attrition:

$$P(Y = 1) = \frac{e^{b_0 + b_1 x_1 + \cdots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + \cdots + b_k x_k}}$$

Where, Y=1 is the event of customer attrition, and P(Y=1) is the probability of this event. The model specification is done by selecting the independent, explanatory variable x_i, so that the model's prediction capability is as high as possible.

The model's possible inputs

From the model layout it can be seen that the explanatory variables can come from several data sources: from personal, telephone, online and mail customer service. As mentioned earlier these data can be:

- Data about the customers: age, gender, type of house (that influences the possibility of substitute products), Address (indicators corresponding to financial conditions can be generated from this).
- Data about the supplier: size of the bill, length of the service being used, reporting faults due to low heat value, frequency of inspection.
- Exogenous data referring to prices, marketing activity of competitors can also be implemented into the model.

The logistic regression model makes it possible to try out all explanatory variables available in the CRM database. The significance of these variables are determined during the model specification.

Methodological-technical suggestions for the model specification

During the specification of binominal logistic regression models, practical statements and suggestions are made, which can only be found in econometrical books very rarely, but may be important for practicing researchers:

1. The model with the best rate of correct results, have many explanatory variables which have no significant effect on the dependent variable, but by keeping them under control and eliminating their indirect effects increases the model's explanatory power. During the model specification every single non significant variable's involvement was investigated, in the final model only those were included which increased the rate of correct results. This question points the attention to method of explanatory variables selection. The statistical, econometrical software, such as SPSS, offer several procedures for the method of involving the explanatory variable into the regression model. Seven such different possibility is given in the SPSS, which –mainly in case of the linear regression model- often results the same model, but in our case the model selection process has significance. One of the differences between them, is that the ENTER method also leaves the non significant variables in the model, whereas the other six method offered by SPSS do not. During the comparison of the model selection methods, it was found that the ENTER method results the best model, but it has its price;

it requires far more time than the other methods. For every single non significant independent variable it needs to be determined, whether the involvement or exclusion improves more the rate of correct results.

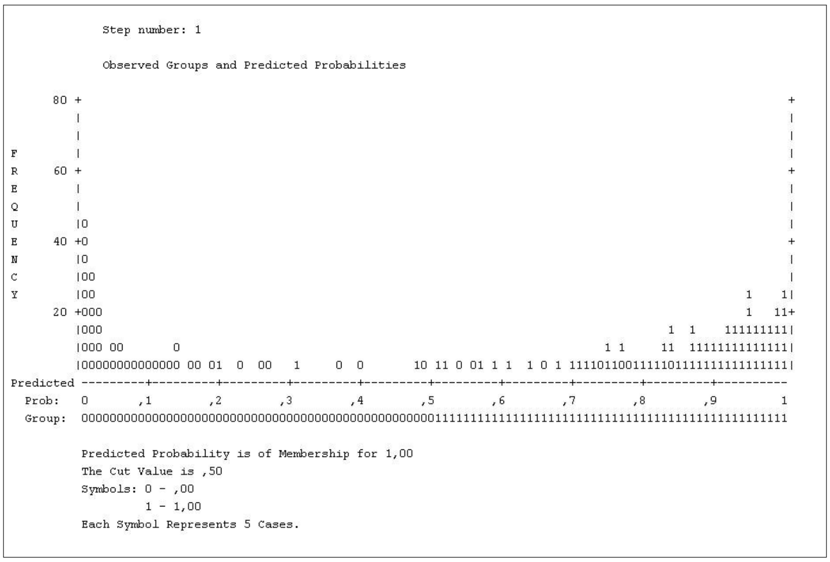
2. Better results can be achieved if using categorical independent variable, rather than using the same criteria's "higher measuring level" numerical variable. In practice it is worth converting the numerical variable to categorical. The reason for this statement, which is unusual in the data analysis, is that the independent variable's (e.g. income) effect is not linear, but there are categories (e.g. income levels), which – in case at a given level of the other explanatory variables – have significant effect, and other categories do not.
3. The model's rate of correct results is increased, if the categorical explanatory variables have more values, categories. During market research, data analysis, it often happens, that the nominal variables that have relatively more categories are recoded to a variable with fewer categories, at the binominal logistic regressive model specification this is not recommended. The reason for this is probably the same as in the previous statement: the nonlinear relationship between the dependent and independent variables can be modelled better, if there are more categories present.

The output of the model

The model output is an indicator with a value can vary between 0 and 1, which is assigned to every customer and shows the probability of attrition. If this probability is greater than 0,5 then the customer is classified into the attrition segment, if the probability is smaller than 0,5 then the customer is classified into the loyal segment. The model's prediction accuracy, the correctness rate of the estimated and actual grouping is showed by the classification table.

The continuous random variable character of the model output, gives a further possibility, the grouping threshold can be adjusted according to our optimisation aims, it is not needed to insist with the generally accepted 0,5 threshold value.

The following histogram compares the grouping, resulted from the actual and the estimated probabilities, based on a previous campus research:



```
             Step number: 1

             Observed Groups and Predicted Probabilities

      80 +                                                                                                    +
         |                                                                                                    |
         |                                                                                                    |
F        |                                                                                                    |
R     60 +                                                                                                    +
E        |                                                                                                    |
Q        |                                                                                                    |
U        |0                                                                                                   |
E     40 +0                                                                                                   +
N        |0                                                                                                   |
C        |00                                                                                                  |
Y        |00                                                                                          1    1|
      20 +000                                                                                         1   11+
         |000                                                                                1  1    111111111|
         |000 00       0                                                            1 1     11  11111111111111|
         |00000000000000 00 01  0  00   1     0  0       10 11 0 01 1 1  1 0 1 1111011001111101111111111111111|
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0       ,1       ,2       ,3       ,4       ,5       ,6       ,7       ,8       ,9         1
  Group:  0000000000000000000000000000000000000000000000000011111111111111111111111111111111111111111111111111

          Predicted Probability is of Membership for 1,00
          The Cut Value is ,50
          Symbols: 0 - ,00
                   1 - 1,00
          Each Symbol Represents 5 Cases.
```

Figure 2. Comparison of the grouping based on the actual and estimated probabilities

Source: Edited by the author based on illustrative data

The abscissa corresponds to the predicted probability with the interval [0-1], The ordinate corresponds to the frequency of the event giving the current value. It can be seen that the model reached high accuracy, only a few cases are listed into the not appropriate domain. The ratio of the two different type of error can be modified by changing the classification threshold from 0,5 .

The most important results of the model is the customer loyalty measure, which directly determines the customer relationship, and furthermore the efficiency of the different promotions, CRM elements and CRM channels, is shown by the analysis of the estimated explanatory variable's parameters. According to this, the customer relationship management can be optimized and reviewed periodically.

Referens:

1. Amemiya T. (1985): Advanced econometrics. : Harvard University Press. 521. p., ISBN 0-674-00560-0
2. Borgulya I. (1998): Neurális hálók és fuzzy rendszerek. Dialog Campus.ISBN: 9789639123274
3. Fajszi B. – Cser L. (2004): Üzleti tudás az adatok mélyén. Budapesti Műszaki és Gazdaságtudományi Egyetem. 260. p., ISBN: 9634215580
4. Franses P.H. – Paap R. (2001): Quantitative models in marketing research. Cambridge: Cambridge University Press. 206. p., ISBN 0-521-80166-4
5. Greene W. (2003): Econometric analysis. Fifth Edition. New Jersey, Upper Saddle River: Prentice Hall. 1083. p., ISBN 9788177586848
6. Hajdu O. (2003): Többváltozós statisztikai számítások. Budapest: Aula Kiadó 457. p., ISBN 963-215-600-5
7. Hosmer D. W. — Lemeshow S. (2000): Applied Logistic Regression, 2nd edition. New York: Wiley, 392p., ISBN 9780470582473
8. Kumar A. - VIthala R. R. - HARSH S. (1995) An Empirical Comparison of Neural Network and Logistic Regression Models, Marketing Letters, Vol. 6. No. 4. 251-264. p., ISSN: 0923-0645 http://dx.doi.org/10.1007/BF00996189
9. Long J. S. (1997): Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks: Sage. 297 p., ISBN 9780803973749 http://dx.doi.org/10.1086/231290
10. Oravecz B. (2007): Credit scoring modellek és teljesítményük értékelése, Hitelintézeti Szemle a Magyar Bankszövetség kiadványa, Vol. 2007. No. 6. 607-627. p., ISSN 1588-6883
11. Sajtos L.—Mitev A. (2007): SPSS kutatási és adatelemzési kézikönyv. Budapest: Alinea Kiadó. 402. p., ISBN 978-963-9659-08-7
12. de SÁ J.P.M. (2007): Applied Statistics Using SPSS, STATISTICA, MATLAB and R.. Heidelberg: Springer. 520. p., ISBN 978-3-540-71971-7 http://dx.doi.org/10.1007/978-3-540-71972-4
13. Schweidel D. A. , Fader P. S., Bradlow E. T. (2008): Understanding Service Retention Within and Across Cohorts Using Limited Information, Journal of Marketing Vol. 72. No. 1. 82-94. p., ISSN 0022-2429 http://dx.doi.org/10.1509/jmkg.72.1.82