Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications

Jan Manďák[1] | *VŠB-Technical University of Ostrava, Ostrava, Czech Republic*
Jana Hančlová[2] | *VŠB-Technical University of Ostrava, Ostrava, Czech Republic*

Abstract

Customer churn, loss of customers due to switch to another service provider or non-renewal of commitment, is very common in highly competitive and saturated markets such as telecommunications. Predictive models need to be implemented to identify customers who are at risk of churning and also to discover the key drivers of churn. The aim of this paper is to use demographic and service usage variables to estimate logistic regression model to predict customer churn in European Telecommunications provider and to find the factors influencing customer churn. An interesting findings came out of the estimated model – younger customers who are shorter time with company, who use mobile data and sms more than traditional calls, having occasional problem with paying bills, with students account and ending contract in the near future are typical representatives of customers who tend to leave the company.

An interaction terms added as explanatory variables showed that effect of usage of data and voice vary depending on the year of birth. The quality of the logistic regression model was assessed by Hosmer-Lemeshow test and pseudo R squared measures. An independent testing data set was further used to evaluate the predictive ability of the model by computation of performance metrics such as the area under the ROC curve (AUC), sensitivity and precision. The resulting model was able to catch 94.8% of customers who in fact left the company. Quality of the model was confirmed also by high value of AUC metric equal to 0.9759. Logistic regression represents a very useful tool in prediction of customer churn not only thanks to its interpretability, but also for its predictive power.

Keywords

Customer churn, telecommunications, predictive analytics, logistic regression, sensitivity

JEL code

C35, C38, C53, L96

[1] VŠB-TU Ostrava, Faculty of Economics, Department of Systems Engineering, Sokolská třída 33, 702 00 Ostrava, Czech Republic. E-mail: jan.mandak@gmail.com, phone: (+420)737674192.
[2] VŠB-TU Ostrava, Faculty of Economics, Department of Systems Engineering, Sokolská třída 33, 702 00 Ostrava, Czech Republic. E-mail: jana.hanclova@vsb.cz, phone: (+420)597322285.

INTRODUCTION

With continuously decreasing costs of data storage, telecommunications companies have access to various data sources, which can be beneficial to various types of customer analyses. Traditional transactional data stored in databases can be combined with unstructured data such as complaints or feedback scraped from social networks or call recordings. These data are used to create predictive models with the use of algorithms such as logistic regression, decision trees, random forests or neural networks. Predictive models can help decision makers to identify customers who are likely to churn (Balasubramanian and Selvarani, 2014). Telecommunications companies can then offer customers new incentives to stay. But it is not sufficient to predict who is likely to churn, but also what are the key factors causing the churn. With this knowledge in hands marketing departments can target retention campaigns to the right customer groups and also change the whole range of services.

The goal of this paper is to predict the customer churn in European Telecommunications Company with the use of logistic regression. The estimated model should reveal especially the key factors leading customers to leave the company. Another aim of the paper is also to assess quality of both logistic regression model and its predictions. This paper should also confirm the suitability of the usage of logistic regression for customer churn prediction.

The theoretical background of customer churn, its types, dimensions, consequences and benefits of proper churn management process as well as review of current literature are described in the introductory part of the paper. The methodological part contains definition of logistic regression, maximum likelihood method used for estimation of beta coefficients, statistical tests such as Wald test or Hosmer Lemeshow test, pseudo R-squared measures, ROC curve and performance statistics sensitivity and precision. Then, the variables used for creation of the model are at first introduced, interesting patterns explored during graphical analysis are shown and finally estimation results and results of statistical tests together with performance metrics are listed. The most interesting results as well as comparison with similar studied are content of the last part of the paper.

1 CUSTOMER CHURN

The term *"customer churn"* is used to indicate those customers who are about to leave for a competitor or end their subscription. Customer churn or customer attrition has become an important issue for organizations particularly in subscription based businesses, where customers have a contractual relationship which must be ended. Customer churn in telecommunications industry is really hot topic, because it is saturated market and where it is difficult to attract new customers and because it is relatively easy to switch to another company (Canale and Lunardon, 2014). It is generally accepted that acquiring a new customer costs five to six times more than retaining the existing one. For telecom operators it's preferable to invest into existing customers and renew their trust, rather than attract new ones characterized by a higher churn rate. *Churn management* is the process of identifying customers, who are valuable for company and are likely to churn, and taking proactive steps to retain them. The measurement for the number of customers moving out during a specific period of time is called *Churn rate*. People responsible for the churn management process should take care of these three dimensions: *WHO* (which customers are likely to churn), *WHEN* (will the customers churn in a week, month or year?) and *WHY* (what are the reasons of customer churn).

It is necessary to distinguish between two types of churn (Lazarov and Capota, 2007). In case of *voluntary churn* – the customer decides to cancel his contract and to switch to another provider. For companies it is necessary to know the reason of churn before applying the right retention strategy. Reasons for churn may include dissatisfaction with the quality of the service, too high costs, not competitive price plans, no rewards for customer loyalty, bad support, long time of problem solutions, privacy concerns, etc. *Involuntary churn* is a situation when the company discontinues the contract itself, e.g. because of fraud

or non-payment. This type of churn can be healthy for a company, because company loses non-profitable or problem-causing customers.

Dahiya and Talwar (2015) emphasize that machine learning models work well if there is enough time spent to prepare meaningful features. Thus, having the right features is usually the most important thing. With the still decreasing costs of data storage, telecommunications companies have access to various data sources, which can be beneficial for analysis of customer churn. It is therefore necessary to invest time into feature engineering, because well prepared features can also help us identify the reasons of churn.

Proper churn management can undoubtedly save company a huge amount of money. Van den Poel and Lariviere (2004) summarize the economic value of customer retention, which may have several benefits: satisfied customers can bring new ones and share positive references, long-term customers tend to buy more and are less sensitive to competitors marketing offers and company can also focus on satisfying the existing customer's needs.

There are many approaches applicable to distinguish churners and non-churners, such as association rules, classification models, clustering, or various types of visualization. Logistic regression, together with decision trees and neural networks belongs to the most frequently used classification methods for churn prediction. Several authors have used logistic regression to detect potential churners in Telecommunications.

Olle and Cai (2014) gathered dataset of 2000 subscribers from an Asian Telecommunications operator. Location, Age, Tenure or Tariff were some of their explanatory variables. Logistic regression model estimated in WEKA achieved precision 0.72 and recall 0.75. Gürsoy (2010) analyzed churn in a major Telecommunication firm in Turkey. Logistic regression model was estimated in SPSS Clementine software. The following explanatory variables were confirmed as statistically significant using Wald test: discount package, customer age or average length of call. Ahn et al. (2006) investigated determinants of customer churn in the Korean mobile telecommunications service market using logistic regressions. Their results indicate that call quality-related factors influence customer churn. Also heavy users are more likely to leave the service provider. Sebastian and Wagh (2017) gathered dataset with over 2000 customers described by 22 variables. They achieved accuracy 0.8 by the use of backward stepwise logistic regression. They also made the results more understandable by visualization in Power BI. Modelling telecom customer attrition is important also in African countries, Oghojafor et al. (2012) state that e.g. in Nigeria the annual churn rate come up to 41%. They created a well-structured and compliant questionnaires and obtained 6000 subscribers of Telecom service providers. They distinguish churners from non-churners by questioning respondents whether they would you like to change their current service provider. Stepwise logistic regression model revealed that call expenses, providers' advertisement medium, type of service plan, number of mobile connections and providers' service facilities are the most important factors driving churn.

2 LOGISTIC REGRESSION

Logistic regression is member of a class of models called *generalized linear models* (Zumel and Mount, 2014). The aim of generalized linear models for a binary dependent variable is to estimate a regression equation that relates the expected value of the dependent variable y to one or more predictor variables, denoted by $\boldsymbol{x}$ (Heeringa et al., 2010). A naïve approach is to model y as a linear function of $\boldsymbol{x}$, but linear regression doesn't capture the relationship between y and $\boldsymbol{x}$ and moreover it may produce predictions that are outside the permissible range 0-1. A better alternative is a nonlinear function that yields a regression model that is linear in coefficients and it is possible to transform the resulting predicted values to the range 0-1. These functions are called in the terminology of generalized linear models link functions (Heeringa et al., 2010). The two most common link functions used to model binary survey variables are the *logit* and the *probit*. The logit, natural logarithm of the odds, can be modelled by a linear regression model:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{kp} x_k, \quad (1)$$

where $\pi(\boldsymbol{x})$ is the conditional probability that $y = 1$ given the covariate vector $\boldsymbol{x}$, β_0, β_1, … , β_k are estimated regression coefficients of the logit model and x_1, x_2, … , x_k are the explanatory variables. The left-hand side of the Formula (1) is called the *log-odds* or *logit* and can take values from the interval (-∞; ∞). The expression $\frac{\pi(x)}{1-\pi(x)}$ is called the *odds* and can take on any value between 0 and ∞. Values close to 0 indicate very low and values close to ∞ indicate very high probability.

The usual practice after estimation of the model coefficients is to assess the significance of the explanatory variables (Hosmer and Lemeshow, 2000). *Wald test* can be used to test the statistical significance of the coefficients $\boldsymbol{\beta}$ in the model. Wald test calculates a Z statistic (2), which is for i-th variable computed as:

$$Z = \frac{\widehat{\beta}_i}{\widehat{SE}(\widehat{\beta}_i)}, \quad (2)$$

where $\widehat{SE}(\widehat{\beta}_i)$ is an estimated standard error of the estimated regression coefficient $\widehat{\beta}_i$. This Z value is then squared, yielding a Wald statistic with a chi-square distribution.

An important step after the model building is assessing the fit of the model. It can be done using e.g. the *Hosmer-Lemeshow test*. The Hosmer-Lemeshow test is a goodness of fit test, which tells how well the model fits the data. Specifically, it calculates if the observed event rates match the expected event rates in population subgroups. Data is first regrouped by ordering the predicted probabilities and forming the number of groups, g. The Hosmer-Lemeshow test statistic (3) is calculated with the following formula:

$$G^2_{HL} = \Sigma_{j=1}^{g} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)}, \quad (3)$$

where n_j = number of observations in the j^{th} group, O_j = number of observed cases in the j^{th} group and E_j = number of expected cases in the j^{th} group. This statistics follows a *chi-squared* distribution with $(g - 2)$ degrees of freedom.

There have also been proposed several *pseudo-R^2 measures* for measuring predictive power of logistic regression. The most often used R squared measures in statistical software's appear to be one proposed by McFadden (1974) and Cox and Snell (1989) along with its corrected version known as Nagelkerke's R^2 (Nagelkerke, 1991). All three measures are based on comparison of the value of likelihood function for model with no predictors and model being estimated.

For the classification tasks where the class which we want to predict is much less frequent, the most known performance metric – *accuracy* – is not sufficient. Two other metrics, *precision* and *sensitivity (recall)*, are much more important. *Accuracy*, *precision or sensitivity* can be computed from the information available in *confusion matrix* (Lantz, 2013), which categorizes predictions according to whether they match the actual value in the data. One dimension indicates the possible categories of predicted values while the other dimension indicates the same for actual values. There are four categories in 2 x 2 confusion matrix (Lantz, 2013):

- True Positive (TP): Correctly classified as the class of interest;
- True Negative (TN): Correctly classified as not the class of interest;
- False Positive (FP): Incorrectly classified as the class of interest;
- False Negative (FN): Incorrectly classified as not the class of interest.

From the business point of view the two most important predictive measures are *sensitivity* and *precision*. By *precision* we mean the ratio of correct predictions, where the model predicts that the customers should churn. A company should avoid low values of *precision*, because it means that

the money is spent in retention campaigns to those customers, who would stay regardless beneficial retention offer. It is computed as true positives divided by sum of true positives and false positives:

$$Precision = \frac{TP}{TP + FP}. \tag{4}$$

Sensitivity in our case measures the ratio of customers, who churned, and the model was able to predict them. Formally it is calculated as the ratio of true positives and sum of true positives and false negatives:

$$Sensitivity\ (recall) = \frac{TP}{TP + FN}. \tag{5}$$

Another metric, *Specificity*, is the proportion of correctly classified non-churners. Formally it is calculated as the ratio of true negatives and sum of true negatives and false positives:

$$Specificity = \frac{TN}{TN + FP}. \tag{6}$$

The overall performance of a classifier summarized over all possible thresholds is given by the *area under the (ROC) curve (AUC)*. *AUC* ranges from 0.5 (classifier with no predictive value) to 1 (a perfect classifier).

To sum up, the first and probably the most important and difficult thing is to gather appropriate input variables. Then, an exploratory data analysis is necessary to get an idea of which variables could be useful for classification. Further the model is estimated and its quality is tested e.g. using Wald test, Hosmer-Lemeshow test or pseudo R-squared measures. The final step is the computation of predictive performance measures like sensitivity or AUC with the use of independent test data set.

3 RESULTS AND DISCUSSION

This section of the paper is devoted to application of logistic regression on two real data sets of approximately 50 000 customers from a European Telecommunications. Two independent data sets were available, training data set, whose purpose is to train classification models, and testing data set to evaluate the predictive performance. The input variables divided into demographic group and service usage group are described at the beginning of this part. Then, some interesting patterns uncovered during exploratory data analysis are shown. The following step is to estimate a logistic regression models using training data set. The estimated model is further described, explained and tested. The testing data set is then used to calculate performance statistics to check the behavior of the model on unseen data.

3.1 Description of the data

The input data for modelling were downloaded from relational tables stored in company data warehouse. Two data sets, both with roughly 50 000 different customers, were randomly selected from the population of approx. 1 million customers. As mentioned earlier, the first one was used to estimate logistic regression model while the second one was left for testing of model's performance. The variables which could help reveal leaving customers were thoroughly selected on the basis of cooperation with customer service managers and IT database experts. Customer managers wanted to see mainly two categories of data – demographic data such as age, customer lifetime or type of account (see Table 1) and service usage data such as consumption of mobile data, voice or monthly invoice paid (see Table 2). They had another interesting suggestions for input variables such as number of calls to other networks, but this information was not easily extractable from the data warehouse.

Three *R* packages were used to further process the data in *R* – *plyr* (Wickham, 2011), *dplyr* (Wickham and Francois, 2016) and *reshape2* (Wickham, 2007). Demographic input variables are shown in Table 1.

Majority of the variables are self-explanatory, but account type and city should be more explained. There are 3 various types of accounts – type A corresponds to personal account, type B to account for students and type D to family account. The variable city is used to distinguish behavior of customers in big and small cities. The five biggest cities were selected based on discussion with company representatives. The living standard of these five cities is higher than standard in the smaller ones.

Table 1 Demographic variables

Variable name	Description
birthYear	Customer's year of birth
delinquent	Did the customer have problem with paying bills at least once in last year?
duration	Customer lifetime
accType	Type of account
contractDuration	Days till the end of contract
city	Indication, whether a customer is from the 5 biggest cities in the country

Source: Company database

The input variables in the service usage group are visible in Table 2. It is important to note that all averages are monthly averages.

The dependent binary variable is called *portout* and tells whether the given customer left the company in 45 days after the date when the input variables were calculated.

Table 2 Service usage variables

Variable name	Description
avgInvoice	Average amount of invoice for last year
avgExtra3	Average overpayment for last 3 months
avgExtra6	Average overpayment for last 6 months
avgExtra12	Average overpayment for last 12 months
avgData	Average monthly data consumed (GB's)
avgVoice	Average monthly voice consumed (min's)
avgSMS	Average monthly number of sent SMS
avgMMS	Average monthly number of sent MMS
VAS	Whether the customer has currently value added services
portout	Whether the customer left the company

Source: Company database

You can see descriptive statistics (minimum, first quartile, median, mean, third quartile and maximum) of numeric variables in Table 3. There are clearly visible some outliers, e.g. maximal values for avgSms and avgMms. Outliers were removed and replaced by medians. The negative values of variable contractDuration means that a given customer had specific time limited (e.g. 1 year) contract which ended in past. Now, the customer does not have time limited contract and pays invoice based on actual usage of services.

Table 3 Descriptive statistics of numeric variables

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
birthYear	1 910	1 959	1 968	1 968	1 980	2003
contractDuration	–6 890	–2 583	–1 274	–1 695	–366	4 866
avgInvoice	0.16	331.69	431.37	567.95	674.85	746.64
avgExtra12	0.00	9.35	46.31	92.19	25.16	6 133.47
avgExtra6	0.00	7.00	48.44	100.91	136.43	6 533.47
avgExtra3	0.00	2.00	42.18	101.82	136.16	9 666.17
avgData	0.00	366.70	2 505.60	10 746.80	8 968	60 602.80
avgVoice	0.00	0.00	0.00	49.15	64.00	2 723.72
avgSms	0.00	0.00	0.00	2.85	0.00	594.55
avgMms	0.00	0.00	0.00	0.11	0.00	680
duration	15	1 335	2 196	2 639	3 833	7 070

Source: Company database

Counts of individual levels for categorical variables are visible in Table 4. The majority of customers have account type A and have no problems with paying bills. Roughly three fifths of customers have value added services and only one fifth of customers is from the 5 biggest cities.

Table 4 Frequencies of levels of categorical variables

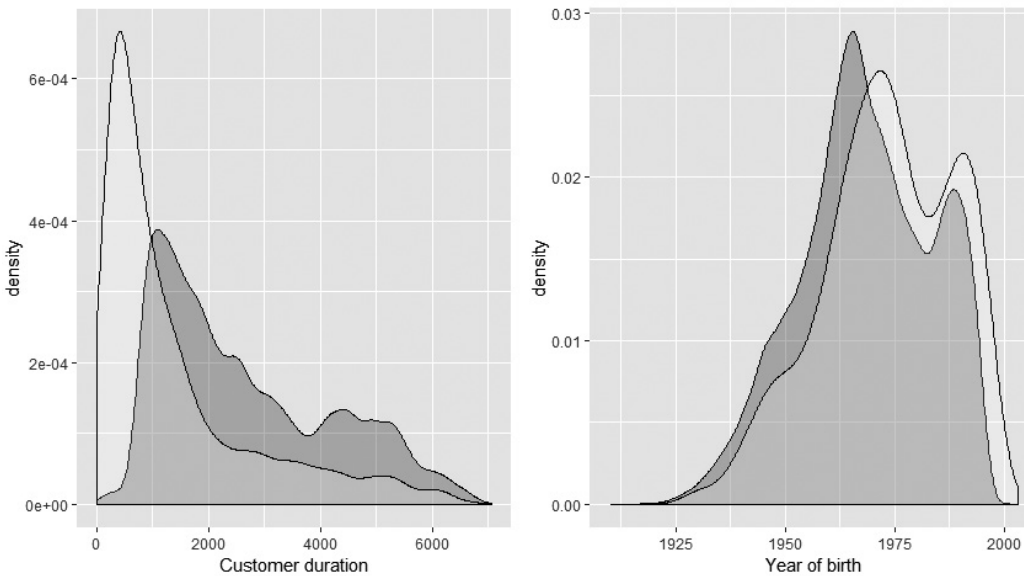
accType		delinquent		vas		city		portout	
A	49 571	*yes*	49 948	*yes*	31 623	*yes*	10 241	*yes*	49 225
B	373	*no*	112	*no*	18 437	*no*	39 819	*no*	835
D	116								

Source: Company database

Two interesting patterns appeared during exploratory data analysis. Overlapping density plots were used for comparison of differences of numeric variables *duration* and *birthYear* grouped by binary dependent variable *portout* (see Figure 1). The density plots were created using the kernel smoothing method with Gaussian kernel. The lighter gray color of area under density curve represents customers,

who left company, the darker gray area under density curve represents customers, who stayed. It is evident that churners are younger and are with company shorter time.

Figure 1 Overlapping density plots for duration and birth year



Source: Own construction

3.2 Description of the estimated model

Estimation results of logistic regression model, which consists of estimated regression coefficient, odds ratio, standard error, value of Wald *z* statistic and corresponding *p-value* are shown in Table 5. According to the regression coefficients, higher values of these numeric variables *increases probability to churn*: birth year, average data and average sms. Also delinquent customers have 2.61 times higher odds to churn than non-delinquent ones and customers with students account type (B) have 2.53 times higher odds to churn than customers with regular personal account. We can infer that younger customers, who use mobile data and sms, with occasional problem with paying their bills and with students account have generally higher tendency to leave the company.

On the other hand, higher values of these numeric variables *decreases probability to churn*: average invoice, days till the end of contract, duration, average overpayment, average voice and average mms. Company should focus retention activities mainly on customers with ending contract in the near future and are shorter time with the company.

Customers with family account in comparison to customers with personal account and customers which have value added services in comparison to those without value added services have minimal odds to churn (in fact, in the training data set there were not a single customer with value added services or with account type D, which left company). Customers from the 5 biggest cities have 0.96 times lower odds to churn than customers from other cities and are therefore less likely to churn.

The behavior of customers is most probably different for various levels of their age and this fact could influence estimated regression coefficients. Interaction terms between birth year of customers and average voice, data and duration were, therefore, embedded into logistic regression model to monitor the effect

of year of birth to regression coefficients of these variables. These variables were selected because it is expected that younger customers use mobile data more and are shorter time with company whereas older customers use voice calls more and are longer time with company.

Table 5 Estimation results

Variable	Estimate	Odds	Std. Error	Z value	p-value
(Intercept)	–33.760	0.000	8.567	–3.941	0.000
birthYear	0.017	1.017	0.004	3.839	0.000
contractDuration	–0.004	0.996	0.000	–29.371	< 2e-16
delinquent true	0.961	2.614	0.262	3.665	0.000
accType B	0.926	2.525	0.488	1.899	0.058
accType D	–15.308	0.000	1 628	–0.009	0.992
avgInvoice	–2.64e-05	1.000	1.32e-04	–0.200	0.841
avgExtra12	–1.18e-04	1.000	1.96e-04	0.601	0.548
avgData	3.05e-04	1.000	1.46e-04	2.093	0.036
avgVoice	–0.084	0.919	0.053	–1.597	0.110
avgSms	0.004	1.004	0.002	1.775	0.076
avgMms	–0.017	0.983	0.005	–0.335	0.737
vas true	–18.983	0.000	185.3	–0.102	0.918
city yes	–0.045	0.956	0.103	–0.439	0.661
duration	-0.013	0.987	0.004	3.008	0.003
avgVoice·birthYear	4.22e-05	1.000	2.68e-05	1.576	0.115
avgData·birthYear	–1.55e-07	1.000	7.42e-08	–2.090	0.037
duration·birthYear	–6.86e-06	1.000	2.22e-06	–3.092	0.002

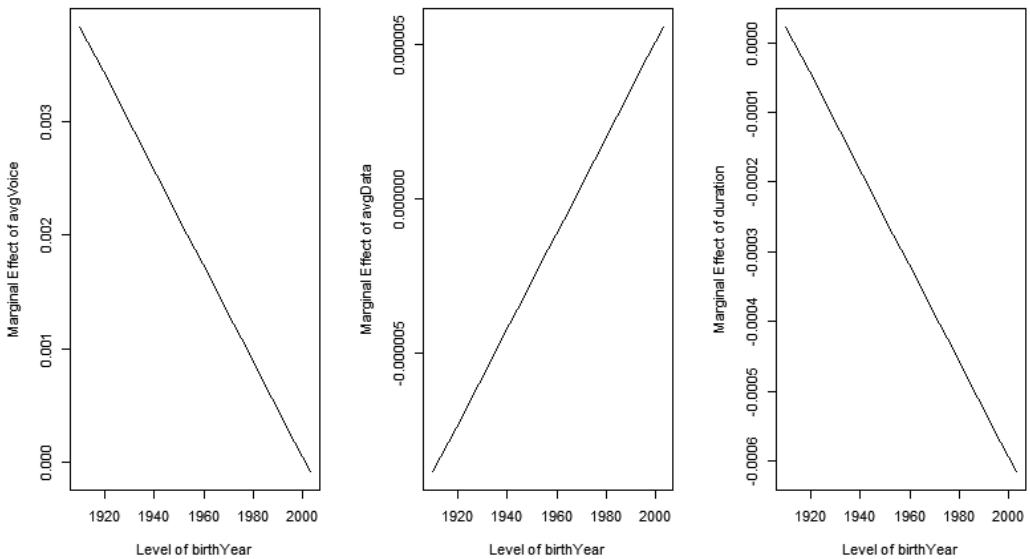
Source: Own construction

The effect of avgVoice, avgData and duration on the probability to churn is not the same depending on the values of birthYear. We can see values of birth year on the x axis and marginal effect of corresponding variable on y axis in Figure 2. The lines in three charts show how the effect of a given variable changes depending on the value of birth year. For example the marginal effects for average voice are calculated as

$$ME_{avgVoice} = \widehat{\beta}_{avgVoice} + \widehat{\beta}_{avgVoice \cdot birthYear} \cdot birthYear. \tag{7}$$

From Figure 2 it is clear that the effect of avgVoice and also duration is higher for older customers while the effect of avgData is higher for younger customers.

Figure 2 Effect of interaction terms



Source: Own construction

The *Hosmer-Lemeshow* test was used as the first one to assess the fit of the model. Hosmer and Lemeshow (2000) recommends to set the parameter g (number of subgroups) as $k + 1$ (number of variables + 1), this parameter was therefore set to 18. The test statistics follows χ^2 distribution with $g - 2 = 16$ degrees of freedom. The *p-value* of test statistic is equal to 2.2e – 16, which is below alpha = 0.05, so the null hypothesis that the observed and expected proportions are the same across all subgroups is rejected. This negative result of the test can be caused by the large data set or by presence of nonlinearities in the model.

Another possibility to assess the fit of the model is by using R^2 measures. R squared measures should be used only as an additional tool for assessing model fit (Hosmer and Lemeshow, 2000). These measures are suitable for comparison of competing models fit to the same set of data. The relatively low values of pseudo R^2 measures (see Table 6.) comparing to the R^2 values of good linear models are the norm and should not be understood as a signal of bad model.

Table 6 Pseudo R squared metrics

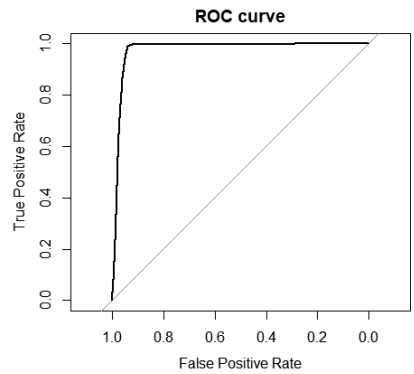
Pseudo R squared	Value
McFadden	0.478
Nagelkerke (Cragg and Uhler)	0.499

Source: Own construction

In the next part the quality of the model is assessed using test data set of 50 060 customers which wasn't used in the process of model training. Complex assessment of the ability of classification model to distinguish between classes is possible with the use of *ROC curve* and

corresponding predictive measure AUC. From Figure 3 it is evident that the *ROC curve* is close to the perfect classifier, which has a curve that passes through the point at 100 percent true positive rate and 0 percent false positive rate. With the AUC value equal to 0.976 the logistic regression classifier is according to the Lantz (2013) outstanding.

Figure 3 ROC curve



Source: Own construction

To compute the values of *sensitivity* and *precision*, it is necessary to determine the probability cutoff (threshold). Hosmer and Lemeshow (2000) recommends choose the threshold in the intersection of sensitivity and specificity. In this case the threshold was set to value 0.065 – customers with predicted probability to churn lower than or equal to 0.065 are predicted to churn and customer with probability higher than 0.065 is predicted to stay.

Table 7 shows the resulting confusion matrix computed for the probability threshold 0.065.

Table 7 Confusion matrix for the threshold 0.065

Prediction / Reference	stay	leave
stay	46 669	43
leave	2 556	792

Source: Own construction

Information in confusion matrix is used to calculate sensitivity and precision. Sensitivity is computed using this formula:

$$Sensitivity = \frac{792}{792 + 43} = 0.948. \tag{8}$$

The value 0.948 tells us that the logistic regression model is able to catch 94.8% of customers, who in reality left company. Another important performance measure is *precision*. *Precision* is calculated as follows:

$$Precision = \frac{792}{2\,556 + 792} = 0.237. \tag{9}$$

The value of *precision* 0.237 means that 23.7% of customers predicted to leave in fact left the company. This ratio is not so big in comparison with the sensitivity, but is acceptable because of the higher importance of *sensitivity* for the company.

It was proved that logistic regression is a useful tool applicable to the area of prediction of customer churn. There are mainly two advantages of using logistic regression – the first one is its interpretability, which is understood as an ability to reveal important factors, their strength and direction. The second one is the predictive power of the algorithm tested and confirmed on the independent testing data set.

CONCLUSIONS

The aim of this paper was to use logistic regression for understanding and prediction of customers who are about to leave Telecommunications provider, which resides in the European Union. Demographic variables such as year of birth, customer lifetime or account type as well as service usage variables such as average monthly voice, data, overpayment or invoice were extracted from the company's data warehouse. Exploratory data analysis revealed that younger customers and customers with shorter lifetime tend to churn more. The estimated logistic regression model confirmed this finding – younger customers who are shorter time with a company, who use mobile data and sms more than traditional calls, with occasional problem with paying bills, with students account and ending contract in near future are typical representatives of customers who tend to leave the company. The interaction terms between year of birth and variables avgVoice, avgData and duration show that the marginal effects of avgVoice and duration are higher for older customers, while the marginal effect of avgData is higher for younger customers. It seems natural that younger customers use their smartphones mainly with connection to the internet whereas older customers use mobile mainly for calling. The quality of the model was confirmed by for logistic regression relatively high value of McFadden and Nagelkerke R-squared measures equal to 0.478 and 0.499, respectively. Three other metrics for assessing quality of the model on independent test data set were calculated – AUC, sensitivity and precision. High values of AUC (0.9759) and sensitivity (0.948) validated that the estimated model is able to predict customers intending to churn. According to the value of sensitivity, the logistic regression model was able to successfully predict 94.8% of customers who in fact actually left the company. To sum up, logistic regression was successfully applied in Telecommunication Company for detection of customers tending to leave the company and also for discovery of the most important drivers of churn.

ACKNOWLEDGEMENTS

This research was supported by the European Social Fund within the project CZ.1.07/2.3.00/20.0296, the Student Grant Competition of Faculty of Economics, VŠB-Technical University of Ostrava within project SP2016/116.

References

AHN, J. H., HAN, S. P., LEE, Y. S. Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, 2006, 30(10–11), pp. 552–568.

BALASUBRAMANIAN, M. AND SELVARANI, M. Churn Prediction in Mobile Telecom System Using Data Mining Techniques. *International Journal of Scientific and Research Publications*, 2014, 4(4), pp. 1–5.

CANALE, A. AND LUNARDON, N. *Churn Prediction in Telecommunications Industry.* A Study Based on Bagging Classifiers, Carlo Alberto Notebooks, 350 p.

COX, D. R. AND SNELL, E. J. *Analysis of binary data.* New York: Chapman and Hall, 1989.

DAHIYA, K. AND TALWAR, K. Customer churn prediction in telecommunication industries using data mining techniques – a review. *International journal of advanced research in computer science and software engineering*, 2015, 5(4), pp. 417–433.

GÜRSOY, S. Customer churn analysis in telecommunication sector. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 2010, 39(1), pp. 35–49.

HOSMER, D. W. AND LEMESHOW, S. *Applied logistic regression.* New York: Wiley, 2000.

HEERINGA, S., WEST, B., BERGLUND, P. *Applied survey data analysis.* Boca Raton, FL: Chapman & Hall/CRC, 2010.

JAMES, G. R. *An introduction to statistical learning: with applications in R.* New York: Springer, 2013.

LANTZ, B. *Machine learning with R.* Birmingham: Packt Publishing, 2013.

LAZAROV, V. AND CAPOTA, M. *Churn Prediction, Business Analytics Course.* TUM Computer Science, 2007.

MCFADDEN, D. Conditional logit analysis of qualitative choice behavior. In: ZAREMBKA, P. eds. *Frontiers in Econometrics.* New York: Academic Press, 1974.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika*, 1991, 78, pp. 691–692.

OGHOJAFOR, B. et al. Discriminant Analysis of Factors Affecting Telecoms Customer Churn. *International Journal of Business Administration*, 2012, 3(2), pp. 59–67.
OLLE OLLE, G. AND CAI, S. A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 2014, 4(1), p. 55–62.
SEBASTIAN, H. AND WAGH, W. Churn Analysis in Telecommunication using Logistic Regression. *Oriental Journal of Computer Science & Technology*, 2017, 10(1), pp. 207–212.
VAN DEN POEL, D. AND LARIVIERE, B. Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. *European Journal of Operational Research*, 2004, 157(1), pp. 196–217.
WICKHAM, H. AND FRANCOIS, R. *dplyr: A Grammar of Data Manipulation. R package version 0.5.0* [online]. 2016. <https://CRAN.R-project.org/package=dplyr>.
WICKHAM, H. Reshaping Data with the reshape package. *Journal of Statistical Software*, 2007, 21(12), pp. 1–20.
ZUMEL, N. AND MOUNT, J. *Practical data science with R.* Shelter Island, NY: Manning Publications Co., 2014.