Systemy Logistyczne Wojsk
Zeszyt 59 (2023)
ISSN 1508-5430, s. 169-188
DOI: 10.37055/slw/186379

Instytut Logistyki
Wydział Bezpieczeństwa, Logistyki i Zarządzania
Wojskowa Akademia Techniczna
w Warszawie

Military Logistics Systems
Volume 59 (2023)
ISSN 1508-5430, pp. 169-188
DOI: 10.37055/slw/186379

Institute of Logistics
Faculty of Security, Logistics and Management
Military University of Technology
in Warsaw

Application of machine learning methods to analyze customer migration risk in terms of corporate financial security

Zastosowanie metod uczenia maszynowego do analizy ryzyka migracji klientów w kontekście bezpieczeństwa finansowego firmy

Ramona-Monica Stoica
ramona.stoica@mta.ro; ORCID: 0009-0002-2539-0415
Faculty of Aircraft and Military Vehicles, Military Technical Academy „Ferdinand I", Romania

Radu Vilău
radu.vilau@mta.ro; ORCID: 0000-0002-7724-0067
Faculty of Aircraft and Military Vehicles, Military Technical Academy „Ferdinand I", Romania

Daniela Voicu
daniela.voicu@mta.ro; ORCID: 0009-0009-4839-8162
Faculty of Aircraft and Military Vehicles, Military Technical Academy „Ferdinand I", Romania

Małgorzata Grzelak
malgorzata.grzelak@wat.edu.pl; ORCID: 0000-0001-6296-7098
Faculty of Security, Logistics and Management, Military University of Technology, Poland

Abstract. Effective prediction of customer migration is only possible through knowledge of the customer life cycle, which is characterized by the length of the relationship between buyer and provider, i.e. customer retention. A concept of opposite importance is customer migration, defined as the partial or total abandonment of the products or services offered by a company. Its knowledge and ability to predict it is crucial in terms of ensuring the continued financial security of target companies. The primary objective of this article was to present a method for assessing the risk of telecom industry customer migration using machine learning methods. The main research problem was defined in the form of a question: is it possible to effectively support decision-making and marketing strategy development by using machine learning methods to minimise customer migration? The hypothesis of the research conducted was also defined: It

is possible to effectively predict the risk of customer migration in the telecommunications industry based on machine learning models and using available databases.The objective was achieved through the use of research methods, theoretical deductions such as and induction, system analysis and synthesis, and mathematical modelling, which additionally allowed for a practical analysis of the migration of customers of the telecommunications industry. Predictors with the greatest impact on the phenomenon under study were selected. It should be noted that the gain chart indicates that, in the case of contacting the 20% of customers selected by the models, the target coverage would be at the following levels, respectively: 70% for the boosted tree model and the decision tree based on the CART algorithm, and 75% for the random forest model. The research niche addressed in the article is the development of methods for assessing migration risk using machine learning techniques. The tool developed in the article can support decision--making in the creation of marketing campaigns aimed at retaining the largest number of customers.
Keywords: machine learning, churn rate, economic security, telcom services, prediction

Abstrakt. Skuteczne przewidywanie migracji klientów możliwe jest jedynie dzięki znajomości cyklu życia klienta, który charakteryzuje się długością relacji pomiędzy kupującym a dostawcą, czyli utrzymaniem klienta. Pojęciem o przeciwnym znaczeniu jest migracja klientów, rozumiana jako częściowa lub całkowita rezygnacja z oferowanych przez firmę produktów lub usług. Jej wiedza i umiejętność jej przewidywania jest kluczowa z punktu widzenia zapewnienia ciągłego bezpieczeństwa finansowego przejmowanych spółek. Podstawowym celem artykułu było przedstawienie metody oceny ryzyka migracji klientów branży telekomunikacyjnej z wykorzystaniem metod uczenia maszynowego. Główny problem badawczy został zdefiniowany w formie pytania: czy można skutecznie wspierać podejmowanie decyzji i rozwój strategii marketingowej poprzez wykorzystanie metod uczenia maszynowego w celu minimalizacji migracji klientów? Postawiono także hipotezę przeprowadzonych badań: Można skutecznie przewidzieć ryzyko migracji klientów w branży telekomunikacyjnej w oparciu o modele uczenia maszynowego i wykorzystując dostępne bazy danych. Cel został osiągnięty poprzez zastosowanie metod badawczych, wniosków teoretycznych takich jak indukcja, analiza i synteza systemowa oraz modelowanie matematyczne, które dodatkowo pozwoliło na praktyczną analizę migracji klientów branży telekomunikacyjnej. Wybrano predyktory mające największy wpływ na badane zjawisko. Należy zauważyć, że wykres zysków wskazuje, że w przypadku kontaktu z 20% klientów wybranych przez modele docelowe pokrycie kształtowałoby się odpowiednio na następujących poziomach: 70% dla modelu drzewa wzmocnionego i modelu opartego na drzewie decyzyjnym na algorytmie CART i 75% na losowym modelu lasu. Niszą badawczą poruszoną w artykule jest rozwój metod oceny ryzyka migracji z wykorzystaniem technik uczenia maszynowego. Opracowane w artykule narzędzie może wspomagać podejmowanie decyzji przy tworzeniu kampanii marketingowych mających na celu utrzymanie jak największej liczby klientów.
Słowa kluczowe: uczenie maszynowe, wskaźnik migracji, bezpieczeństwo finansowe, usługi telekomunikacyjne, prognozowanie

Introduction

The telecoms market in Poland is characterised by maturity and high saturation, and its value in 2022 was estimated at PLN 40.63 billion (Włoszczyzna, 2022). The year was notable for being the first time that the value of revenues fell (by 0.4%, year-on-year), with a comparable number of subscribers. Rising energy prices, inflation, the effects of the Covid-19 pandemic and the ongoing war in Eastern Europe continue to strongly influence the increase in the price of communication services, as well as increased customer migration, mainly in the area of switching from prepaid to postpaid offers. Increasing competition in the market is forcing telecommunications companies to take measures to consolidate their position in the market. One of the important elements of the competitive struggle is the stability

and predictability of the revenue stream (subscription fees). In the context of this, it is therefore reasonable to develop models that support customer migration analysis, support higher customer retention and lead to a minimisation of the customer churn rate (Alomar, Abdallah, 2022).

The above was set as the scientific objective of the study. Effective prediction of customer migration is only possible through knowledge of the customer life cycle, which is defined in the literature as a marketing concept that indicates that both the needs and the set of values of customers change at different stages. It consists of the transition of the customer between successive phases of the cycle, accompanied by a change in its relationship with its environment (Hajduk et. al, 2022). There are three basic phases: customer acquisition, development and retention. Relationship management itself, on the other hand, can be carried out using CRM (customer relationship management) concepts, and in particular analytical CRM, which supports the process of collecting and analysing data on individual consumer behaviour.

A fundamental concept characterising the length of the buyer-supplier relationship is customer retention. It manifests itself through repeat purchases by buyers of selected products and services from a selected supplier (Hajduk, Poliak, 2023). Its main measure (and at the same time the main measure of customer retention) is the retention rate, which determines what proportion of all customers renewed their purchases in a given period. This rate is determined on the basis of the actual recorded purchasing behaviour and not solely on declarations collected from customers (Urbanek, 2011). A concept with the opposite meaning is customer migration, defined as the partial or total abandonment of products or services offered by a company. Its basic measure is the migration rate, which in the literature, with reference to telecommunications companies, is called the churn rate. It is considered a key measure of customer loyalty. It indicates what percentage of a company's customers have decided to abandon the services offered or products purchased. It is most often calculated on an annual, quarterly or monthly basis. This indicator is analysed in detail in the following study.

The aim of this article is to present a method for assessing the risk of customer migration in the telecommunications industry using machine learning methods. The main research problem was defined in the form of a question: is it possible to effectively support decision-making and marketing strategy development by using machine learning methods to minimise customer migration? The hypothesis of the research conducted was also defined: It is possible to effectively predict the risk of customer migration in the telecommunications industry based on machine learning models and using available databases. On the basis of a database characterising customer behaviour in the telecommunications industry, three models were built using machine learning algorithms, i.e. the CART algorithm, random forest and boosted trees. On this basis, the factors with the greatest impact on the phenomenon

under study were selected and their individual parameters necessary for the creation of marketing strategies providing the greatest return were determined.

Telcom customer migration - a literature review

The literature review has shown that the problem of customer migration in the telecommunications industry is one of the most important aspects of customer service. The specifics of the industry are characterised by a wide variety of availability of voice and data services, as well as high availability of providers of the above services (Staniewska, 2022). In addition, the above sector is characterised by a high degree of freedom for customers to switch from one service provider to another. The rapid increase in competition in the market makes it necessary for companies to increasingly compete for customers, mainly by providing personalised services at low prices. It is therefore important to develop decision-support models for the design of marketing campaigns in order to retain as many customers as possible (Żółtowski et. al, 2022). There are many approaches available in the literature presenting different methods for modelling customer migration risk. The most common categories of independent variables include (Łapczyński, 2016; Berry, Linoff, 2004):

- migration rates recorded in the past, including mobile phone make and model, customer demographic characteristics and place or channel of sale, customer address,
- customer characteristics, i.e. age, gender, place of residence and other available characteristics,
- variables characterising the type of service provided, i.e. handset make and model, purchase price, date of service activation, type of subscription or additional services held by customers,
- history of service use and charges incurred, e.g. total customer receipts, late charges, number and duration of calls, number of text messages sent.

The most commonly used methods include (Ahmed, Linen 2017):

- preprocessing, class imbalance problem and migration prediction based on sampling (Burez, Van den Poel, 2009; Effendy, Abdurahman Baizal, 2014; Wu, Meng 2016; Sundarkumar et al., 2016),
- function-based methods, such as the support vector method SVM (Xia, Wei, 2014; Maldonado et al., 2015),
- composite methods (Backiel et al., 2015; Backiel et al., 2016; Quihua et al. 2014),
- migration prediction based on Big Data (Huang et al., 2015; Young, Yonguri, 2015),

- machine learning methods (Guo-en, Jin 2008; Xie et al., 2009; Sharma et al., 2013; Kisioglu, Topcu, 2011; Brandusoiu et al., 2016; Preeti et al. 2016; Lalwani et al. 2022; Geiler et al. 2022; AL-Najjar et al. 2022),
- Metaheuristic methods (Ning et al. 2011; Lewis et al. 2023),
- hybrid migration prediction models (Sumathi, 2016; Łapczyński, 2016; Dalvi et al., 2016).

The analysis of customer migration has many benefits, such as:

- less investment in customer acquisition,
- higher profitability for existing customers,
- an increase in sales of ancillary services to long-stay customers,
- greater investor confidence.

The literature review also revealed a number of impediments that can accompany the analysis of customer migration in the telecommunications industry:

- use by business customers of diversified services from different operators,
- the possibility for individual clients to diversify their choice of company according to the type of service in question,
- the difficulty of profiling consumers in relation to their needs,
- changing the terms of the contract at the time of renewal,
- change of personal details or place of residence.

Research methods – decision tree models

In this paper, machine learning models based on three types of decision trees are used to predict the migration risk of industry customers. Decision trees are tools that enable the construction of predictive and descriptive models, providing a more accurate analysis of the functional relationship of variables relative to more simple regression models (Owczarek et. al., 2022). They are a method of multivariate analysis that allows the study of the relationship between the dependent variable and the independent variables measured on a weak scale, i.e. nominal or ordinal, and a strong scale, i.e. interval and quotient (Kozłowski et al., 2021). It is said about classification trees when the dependent variable, expressed on a nominal or ordinal scale, while in the case of regression trees the variable under study is quantitative, at least interval (Łapczyński, 2003).

A decision tree is a graphical model that arises as a consequence of the recursive division of a set of observations A into n disjoint subsets *A1, A2, A3, ..., An*. The above makes it possible to obtain such subsets, which will be characterised by maximum homogeneity from the point of view of the value of the dependent variable. The process of building the model is a multi-step one, and in each successive step a different explanatory variable can be used to obtain the optimal subdivision. According to the procedure, in each successive step, the predictor is selected that

guarantees the best node partitioning, separating the most homogeneous subsets (Łapczyński, 2003). The partitioning and classification of individual observations starts at the root of the tree and ends when one of the terminal classes, i.e. leaf is reached.

Thus, decision trees are a visual representation of a character model (Timofeev, 2004):

$$Y = f\left(x_i\right) = \sum_{k=1}^{K} \alpha_k I\left(x_i \in R_k\right) \tag{1}$$

where:
Y – dependent variable,
X^L - independent variables,
L – number of independent variables,
K – number of segments,
R_k *(k=1, ..., K)* – segment of the space of independent variables,
x_i – observations from the analysed set,
α_k – model parameters,
I – indicator function.

Depending on the type and nature of the explanatory variables, the way in which the indicator function is defined is chosen. In the case of variables of a metric nature, segments R_k are defined by their boundaries in space X^L according to the function below:

$$I\left(x_i \in R_k\right) = \prod_{l=1}^{L} I\left(v_{kl}^{(d)} \le x_{il} \le v_{kl}^{(g)}\right) \tag{2}$$

where the values $v_{kl}^{(d)}$ I $v_{kl}^{(g)}$ denote the upper and lower limits of a segment in the *l-th* dimension of space.

On the other hand, in the case of a non-metric form of the explanatory variables, the subspace R_k should be described by a functional relationship:

$$I\left(x_i \in R_k\right) = \prod_{l=1}^{L} I\left(x_{il} \in B_{kl}\right) \tag{3}$$

where:
B_{kl} - a subset of the variable category X_l.

Furthermore, if the dependent variable Y under study in the model is a nominal variable, a discriminant model represented by a classification tree is used to analyse it, for which the model parameters α_k are determined using the formula:

$$\alpha_k = \arg\max\, p(C_j / x_i \in R_k) \tag{4}$$

where:
$p\left(C_j / x_i \in R_k\right)$ - *the a posteriori* probability that an observation from a segment R_k belongs to the class C_j.

If, on the other hand, the explanatory variable is measured on numerical scales, from the group of strong scales, then a regression model should be used, the visualisation of which will be a regression tree. The parameters of the model are calculated according to the relationship:

$$\alpha_k = \frac{1}{N(k)} \sum_{x_i \in R_k} y_i \tag{5}$$

where:
$N(k)$ - number of observations in the segment R_k,
y_i - values of the dependent variable in the segment R_k.

The final step in building the model is to evaluate it, which is done by analysing the quality of the partitioning of the explanatory variable space, by using two main evaluation measures:

- classification error, Gini coefficient, entropy measure for qualitative variables,
- the variance of the dependent variable in the case of quantitative variables (Berk, 2008).

In the remainder of this article, an interaction decision tree model with the CART (Classification and Regression Trees) algorithm, a random forest model and a boosted tree model were used to build a model for analysing the risk of telecoms customer migration.

Results of the decision trees models

A study of the risk of telecom customer migration was carried out using Orange Telecom's churn database (https://www.kaggle.com/mnassrib/telecom-churn-datasets). It has been cleaned of outliers and records with missing data. It presents the activity of individual customers with the assigned value of the dependent variable indicating whether a customer has cancelled a contract. The database was divided into two subsets in a ratio of 80% (2666 observations) - 20% (667 observations). The first dataset was used as a learning dataset to build and cross-validate the model, while the second dataset was used as a test dataset to assess the predictive ability of the model and calculate classification errors.

Whether a customer migrated to a competitor was indicated as the dependent variable, which was labelled *churn*. The variable under study is a binary variable and takes the values *true* if the customer left and *false* if the contract was renewed. Three qualitative independent variables were used to build the model, i.e. *state name, ownership of a roaming service* and *voicemail,* and sixteen quantitative dependent variables, i.e. *length of contract held (in months), number of voice messages received, total call minutes per day (in min.), total calls per day (in pcs.), total daily charges (in $), total call minutes per evening (in min.), total calls during the evening (in pcs.), total charges during the evening (in $), total call minutes during the night (in min.), total calls during the night (in pcs.), total charges during the night (in $), total international call minutes (in min.), total international calls (in pcs.), total international charges (in $), number of customer service calls (in pcs.).* Due to the qualitative nature of the phenomenon under study, classification issues based on decision trees with the CART algorithm, random forest and boosted trees were used to investigate the risk of customer migration. The study was carried out using *Statictica 13.1* computer software.

Analysis of customer migration using the CART classification tree model

The CART decision tree model was built first. Before starting the analysis, the boundary settings of the presented solution were defined:

- sharing rule - Gini index,
- a priori probability - estimated from a learning sample,
- misclassification costs - equal,
- stopping criterion - with misclassification,
- minimum number of terminal nodes n=66,
- maximum number of levels (depth) of tree n =10,
- maximum number of nodes n=1000,
- error estimation by means of 10-fold test validation.

On the basis of the results obtained, it can be concluded that the highest risk of migration is characterised by customers characterised by leaf number 12, i.e. those for whom the sum of minutes made during the day is less than or equal to 173.5 and those who made more than 3 calls to the customer service during the contract period. The corresponding risk of leaving, for the above customers, is 79.05%.

In the next step, a ranking of the importance of the predictors was constructed to assess the impact of the individual explanatory variables on the phenomenon under study (Table 1). According to the results obtained, it should be concluded that *the sum of minutes of calls made per day* and *the sum of daily charges have the*

greatest impact. The *postal code variable* and the *length of account ownership* are the least important from the point of view of the dependent variable.

Table 1. Predictor importance ranking for the developed CART decision tree

Explanatory variable	Validity
Total minutes per day	100
Total daily charges	100
Place of residence (State)	54
Total minutes per evening	33
Total charges during the evening	33
Roaming services	31
Customer service calls	31
Total international calls	14
Total minutes per night	9
Total charges at night	9
Total international minutes	8
Sum of international charges	8
Total calls per day	6
Total calls at night	6
Number of voice messages	5
Voicemail	5
Total calls during the evening	3
Length of account ownership [month]	3
Postcode	2

Source: Own study

The model's correctness of classification was then assessed. For this purpose, a confusion matrix (misclassifications) and evaluation metrics, i.e. sensitivity, precision and accuracy, were used. The classification matrix is shown in 2. There are four types of classified observations:

- truly positive (TP),
- false positives (FP),
- false negatives (FN),
- truly negative (TN).

Table 2. Decision tree classification matrix based on the CART algorithm

	Observed	Predicted FALSE	Predicted TRUE	Total in line
Number	FALSE	2 191	87	2 278
Percentage from column		91.25%	32.83%	
Percentage of row		96.18%	3.82%	
Percentage of total		82.18%	3.26%	85.45%
Number	TRUE	210	178	388
Percentage from column		8.75%	67.17%	
Percentage of row		54.12%	45.88%	
Percentage of total		7.88%	6.68%	14.55%
Number	Total groups	2 401	265	2 666
Percentage total		90.06%	9.94%	

Source: Own study

Based on the table, the value of the following indicators was calculated:

- sensitivity, which indicates the proportion of correctly predicted positive cases (TP) among all positive cases (which also includes those incorrectly classified as negative (FN), this parameter should take the highest possible value.

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} = \frac{178}{178 + 210} \cdot 100\% = 45.88\%$$

- precision, which indicates how many of the positively predicted cases are true positives; this parameter should take the highest possible value.

$$Precision = \frac{TP}{TP + FP} = \frac{178}{178 + 87} \cdot 100\% = 67.16\%$$

- accuracy indicates the proportion of all forecast values that have been correctly classified; this parameter should take the highest possible value.

$$Accuracy = \frac{TP + TN}{P + N} = \frac{178 + 2191}{2666} \cdot 100\% = 88.85\%$$

In the final step, the developed model was saved as PMML code and then implemented to conduct the study based on a subset of the test sample data.

Analysis of customer migration using a random forest model

The second model developed was a random forest model. Prior to the analysis, the boundary settings of the presented solution were defined:

- a priori probability - estimated from a learning sample,
- misclassification costs - equal,
- number of predictors drawn - 5,
- proportion of random test sample - 20%,
- proportion for sub-samples - 50%,
- number of trees - 100,
- stop learning - percentage decrease in error 5%,
- minimum number of descendants - 5,
- maximum number of levels (depth) of tree n =10,
- maximum number of nodes n=1000.

In the next step, a ranking of the importance of the predictors was constructed to assess the influence of the individual explanatory variables on the phenomenon under study (Table 3). According to the results obtained, it should be concluded that the *place of residence (state in the USA)*, *the sum of minutes made per day of calls* and *the sum of daily charges have the* greatest influence. The variable *postal code* and the *number of voicemail messages* are the least significant from the point of view of the dependent variable.

Table 3. Importance ranking of predictors for the random forest model

Explanatory variable	Validity
Place of residence (state)	100
Total minutes per day	66
Total daily charges	62
Customer service calls	51
Total international calls	40
Roaming services	37
Total minutes per evening	36
Total international minutes	35
Total charges during the evening	33
Sum of international charges	33
Number of voice messages	28
Total calls per day	25
Total calls at night	25
Total minutes per night	23

Explanatory variable	Validity
Length of account ownership [month]	22
Total charges at night	21
Total calls during the evening	18
Voicemail	12
Postcode	6

Source: Own study

The correctness of classification of the model was then assessed (Table 4).

Table 4. Classification matrix of the random forest model

	Observed	Predicted FALSE	Predicted TRUE	Total in line
Number	FALSE	1 787	30	1 817
Percentage from column		90.34%	18.87%	
Percentage of row		98.35%	1.65%	
Percentage of total		83.62%	1.40%	85.03%
Number	TRUE	191	129	320
Percentage from column		9.66%	81.13%	
Percentage of row		59.69%	40.31%	
Percentage of total		8.94%	6.04%	14.97%
Number	Total groups	1 978	159	2 137
Percentage total		92.56%	7.44%	

Source: Own study

Based on the table, the value of the following indicators was calculated: *sensitivity*:

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} = \frac{129}{129 + 191} \cdot 100\% = 40.31\%$$

precision:

$$Precision = \frac{TP}{TP + FP} = \frac{129}{129 + 30} \cdot 100\% = 81.13\%$$

accuracy:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{129 + 1787}{2137} \cdot 100\% = 89.65\%$$

In the final step, the developed model was saved as PMML code and then implemented to conduct the study based on a subset of the test sample data.

Analysis of customer migration using the boosted tree model

The model of boosted trees was proposed as the last one. Before proceeding with the analysis, the boundary settings of the presented solution were defined:

- a priori probability - estimated from a learning sample,
- misclassification costs - equal,
- learning rate - 0.1,
- number of trees - 200,
- proportion of random test sample - 20%
- proportion for sub-samples - 50%,
- stop learning - percentage decrease in error 5%,
- minimum descendant count - 66,
- minimum n offspring - 1,
- maximum number of levels (depth) of tree n =10,
- maximum number of nodes n=7.

In the next step, a ranking of the predictors was carried out, which makes it possible to assess the influence of the individual explanatory variables on the phenomenon under study (Table 5). According to the results obtained, it should be concluded that *the sum of minutes of calls made per day, the place of residence (state in the USA)* and *the sum of daily charges have the* greatest impact. The variable *postal code* and the sum of minutes for calls made during the *evening* are the least significant from the point of view of the dependent variable.

Table 5. Importance ranking of predictors for the boosted tree model

Explanatory variable	Validity
Total minutes per day	100
Total daily charges	100
Status	95
Customer service calls	63
Total international minutes	60

Explanatory variable	Validity
Sum of international charges	60
Total calls per day	58
Total minutes per evening	57
Total charges during the evening	57
Number of voice messages	51
Total international calls	49
Total minutes per night	48
Total charges at night	48
Voicemail	42
Length of account ownership [month]	40
Total calls at night	38
Roaming services	33
Total calls during the evening	30
Postcode	9

Source: Own study

The correctness of classification of the model was then assessed (Table 6).

Table 6. Classification matrix of the boosted tree model

	Observed	Predicted FALSE	Predicted TRUE	Total on the line
Number	FALSE	1 557	273	1 830
Percentage from column		96.95%	51.70%	
Percentage of line		85.08%	14.92%	
Percentage of total		72.96%	12.79%	85.75%
Number	TRUE	49	255	304
Percentage from column		3.05%	48.30%	
Percentage of row		16.12%	83.88%	
Percentage of total		2.30%	11.95%	14.25%
Number	Total groups	1 606	528	2 134
Percentage total		75.26%	24.74%	

Source: Own study

Based on the table, the value of the following indicators was calculated:

- sensitivity:

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} = \frac{255}{255 + 49} \cdot 100\% = 83.88\%$$

- precision:

$$precision = \frac{TP}{TP + FP} = \frac{255}{255 + 273} \cdot 100\% = 48.29\%$$

- accuracy:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{255 + 1557}{2134} \cdot 100\% = 84.91\%$$

In the final step, the developed model was saved as PMML code and then implemented to conduct the study based on a subset of the test sample data.

Comparison of the predictive capabilities of the developed models

In the final stage of the study carried out, a comparison was made between the predictive capabilities of the developed models. For this purpose, based on the *Rapid Implementation* module available in *Statistica 13.1,* the models developed and saved as PMML code were implemented. Then, a subset of the test sample was used to assess their quality, which was an extracted random and hitherto unaccounted for 20% of all observations from the available database. In a first step, the predicted class size was assessed against the total observed size for the *churn* variable. Based on the matrix, it should be noted that the best overall relevance is that of the random forest model (relevance of 90.10%), while the least relevance is that of the boosted tree model (74.74%). From the point of view of customer migration risk analysis, the sensitivity of the model is important, i.e. its ability to correctly identify customers who want to terminate their contract with the company. The results obtained show that the boosted trees model has the highest sensitivity (74.74% correct indications), while the random forest model has the lowest sensitivity (43.16%).

In the next step, the *Receiver Operating Characteristic* (ROC*)* curve plots were analysed to assess the predictive performance of the individual models (Figure 1). The most important parameter for evaluating the *ROC curve* is the *AUC* (*Area Under*

the ROC Curve). It takes values from 0 to 1. The interpretation of the result is based on the classification, according to Kleinbaum and Klein (Table 7).

Table 7. Classification matrix of the boosted tree model

AUC value	Evaluation
0.9<AUC<1.0	Excellent discrimination
0.8<AUC<0.9	Good discrimination
0.7<AUC<0.8	Sufficient discrimination
0.6<AUC<0.7	Weak discrimination
0.5<AUC<0.6	Insufficient discrimination

Source: Kleinbaum and Klein 2010

The results obtained show that all of the developed models have good discrimination, with an area under the curve in the range of 0.8-0.9. As a result, they have a good efficiency in predicting the studied dependent variable, while at the same time not giving grounds to consider them as overfitted models. The random forest model has the best efficiency.

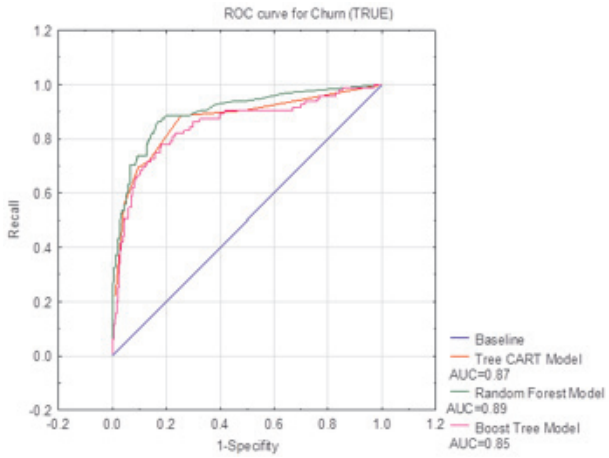

Fig. 1. ROC curve for individual models
Source: Own study

In the final step, the gain chart for the developed models was evaluated (Fig. 2). This graph provides important support for marketing campaigns, as it indicates with what percentage of the customers who have been selected by the model need to be contacted in order to reach the company's target. The blue line is the baseline, indicating what percentage of interested customers the company would reach with

an additional offer if it sent information and enquiries to random customers, e.g. if 20% of customers were contacted, the target coverage would also be 20%. On the other hand, in the case of contacting 20% of the customers selected by the models, the target coverage would be at the following levels respectively: 70% for the boosted tree model and the CART decision tree and 75% for the random forest model.

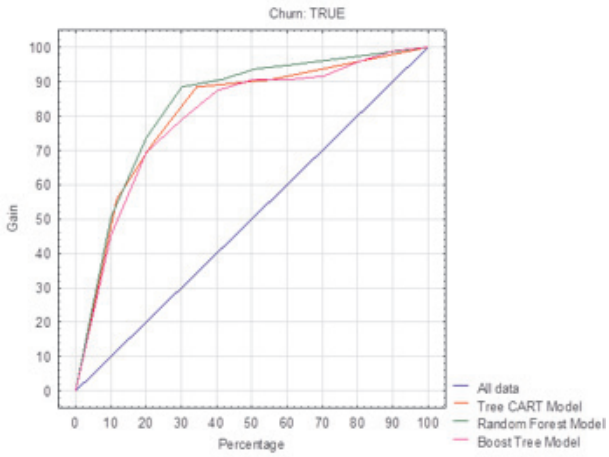

Fig. 2. Gain chart for the developed models
Source: Own study

Conclusion

The market for telecommunications services is one of the elementary and fundamental factors for the proper functioning of society, business and the country as a whole. Increased competition in the market and the outbreak of the COVID-19 pandemic have made it extremely important from a company's point of view to ensure a steady stream of revenue at the lowest possible cost, resulting from customer service. Therefore, it is extremely important to know the customer life cycle and to use tools that support customer relationship management. One of the key elements in ensuring a proper and long-lasting customer relationship is the ability to retain customers.

This paper presents a method for assessing customer migration risk based on decision tree models with the CART algorithm, random forest and boosted tree models. Publicly available customer migration data from Orange Telecom was used to conduct the study. Three prediction models using decision trees were developed,

which were then implemented and the effectiveness of the prediction was evaluated against a test dataset. The developed tool can support the decision-making process when developing marketing campaigns to retain as many profitable customers as possible.

The primary aim of the article was to present a method for assessing the risk of telecoms customer migration using machine learning methods. The stated objective was achieved through a literature analysis, as well as through the application of research methods such as theoretical deduction and induction, system analysis and synthesis, and mathematical modelling, which additionally allowed for a practical analysis of telecoms customer migration.

In summary, the research methods mentioned in the introduction and the research carried out were used to draw the following conclusions about the selected machine learning methods for investigating telecoms customer migration risk:

- The models developed on the basis of the learning dataset make it possible to assess the validity of the predictors and their impact on the likelihood of telecoms customer migration,
- The most significant predictors in terms of migration are *total minutes made per day, total daily charges* and *place of residence,* while the least significant were *postcode, length of account ownership* and *number of voice messages held,*
- The best overall relevance is that of the random forest model (relevance 90.10%), while the least is that of the boosted tree model (74.74%),
- The results obtained show that the boosted tree model has the highest sensitivity (74.74% of correct indications), while the random forest model has the lowest (43.16%),
- *The gain chart* indicates that in the case of contact with the 20% of customers selected by the models, target coverage would be at the following levels respectively: 70% for the boosted tree model and the decision tree based on the CART algorithm and 75% for the random forest model,
- All of the models developed have good discrimination, with an area under the AUC curve in the range 0.8-0.9, so that they have good predictive performance for the dependent variable under study, while at the same time not warranting recognition as overfitted models,
- The random forest model has the best performance, with an AUC value=0.89.
- The results obtained from the research and analysis carried out encourage the continuation of research work towards issues related to:
- analysing the applicability of the developed machine learning methods to the analysis of customer migration in other industries, including the banking sector or the insurance market,
- an analysis of the possibility of using other forecasting methods to assess the phenomenon under investigation, including the use of hybrid forecasting models.

The tool developed in the article can support decision-making in the creation of marketing campaigns aimed at retaining the largest number of customers.

BIBLIOGRAPHY

[1] Ahmed A., Linen D. M., 2017. A review and analysis of churn prediction methods for customer retention in telecom industries. 4th International Conference on Advanced Computing and Communication Systems (ICACCS).

[2] AL-Najjar D., Al-Rousan N., AL-Najjar H., 2022. Machine Learning to Develop Credit Card Customer Churn Prediction. Journal of Theoretical and Applied Electronic Commerce Research, 17(4), https://doi.org/10.3390/jtaer17040077.

[3] Alomar, I., Abdallah, N., 2022. The challenges facing the aviation industry managers in crisis. Logistics management strategies and teamwork management. Military Logistics Systems, 56(1). https://doi.org/10.37055/slw/155067.

[4] Backiel A., Baesens B., Claeskens G., 2016. Predicting time-to-churn of prepaid mobile phone customers using social network analysis. Journal of the Operational Research Society, 1-11.

[5] Backiel A., Verbinnen Y., Baesens B., Claeskens G., 2015. Combining local and social network classifiers to improve churn prediction. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[6] Berry M.J.A., Linoff G.S., 2004. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management, 2nd ed., New York: Wiley and Sons.

[7] Burez J., Van den Poel D., 2009. Handling class imbalance in customer churn prediction. Expert Systems with Applications. 36(3).

[8] Dalvi P. K., Khandge S. K., Deomore A., Bankar A., Kanade V. A., 2016. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. Symposium on Colossal Data Analysis and Networking (CDAN).

[9] Effendy V., Abdurahman Baizal ZK, 2014. Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. 2nd International Conference on Information and Communication Technology (ICoICT).

[10] Hajduk, I. E., Poliak, M., 2023. Proposal of methodology for customers relationships establishing in terms of transport services. The Archives of Automotive Engineering – Archiwum Motoryzacji, 101(3). https://doi.org/10.14669/AM/172917.

[11] Hajduk, I. E., Poliak, M., Gašparík, J., 2022. Quality of transport services and customer satisfaction measurement. The Archives of Automotive Engineering – Archiwum Motoryzacji, 96(2). https://doi.org/10.14669/AM/151707.

[12] https://www.kaggle.com/mnassrib/telecom-churn-datasets [20 June 2023].

[13] Huang Y., Fangzhou Z., Mingxuan Y., Ke D., Yanhua L., Bing N., Wenyuan D., Qiang Y., Jia Z., 2015. Telco churn prediction with big data. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data.

[14] Kleinbaum D. G., Klein, M., 2010. Logistic Regression A Self-Learning Text. Berlin: Springer.

[15] Kozłowski E, Borucka A, Świderski A, Skoczyński P., 2021. Classification Trees in the Assessment of the Road–Railway Accidents Mortality. Energies, 14(12), https://doi.org/10.3390/en14123462.

[16] Lalwani P., Mishra M.K., Chadha J.S. et al., 2022. Customer churn prediction system: a machine learning approach. Computing, 104, https://doi.org/10.1007/s00607-021-00908-y.

[17] Geiler L., Affeldt S., Nadif M., 2022. A survey on machine learning methods for churn prediction. International Journal of Data Science Anal, 14. https://doi.org/10.1007/s41060-022-00312-5.

[18] Łapczyński M., 2003. Classification trees in customer satisfaction and loyalty research, Kraków: Statsoft.

[19] Łapczyński M., 2016. Hybrid predictive models in relationship marketing, Zeszyty Naukowe/ University of Economics.

[20] Lewlisa S., Hrudaya K. T., Tarek G., Hatem E., El-Sayed M. E., 2023. Deep Churn Prediction Method for Telecommunication Industry. Sustainability, 15(5).

[21] Maldonado S., Álvaro F., Verbraken T., Baesens B., Weber R., 2015. Profitbased feature selection using support vector machines-General framework and an application for customer retention. Applied Soft Computing, 35.

[22] Ning L., Hua L., Jie L., Guangquan Z., 2014. A customer churn prediction model in telecom industry using boosting. IEEE Transactions on Industrial Informatics, 10(2).

[23] Owczarek, P., Brzeziński, M., Zelkowski, J., 2022. Evaluation of light commercial vehicles operation process in a transport company using the regression modelling method. Eksploatacja i Niezawodność – Maintenance and Reliability, 24(3), 13. https://doi.org/10.17531/ein.2022.3.13

[24] Pınar K., Topcu Y. I., 2011. Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. Expert Systems with Applications, 38(6).

[25] Qiuhua S., Hong L., Qin L., Wei Z., Kone K., 2014. Improving churn prediction in telecommunications using complementary fusion of multilayer features based on factorization and construction. The 26th Chinese Control and Decision Conference.

[26] Staniewska, E., 2022. Identification and assessment of risk components of enterprise cooperation in the supply chains. Military Logistics Systems, 56(1). https://doi.org/10.37055/slw/155070.

[27] Sumathi T., 2016. Churn Prediction on Huge Sparse Telecom Data Using Meta-heuristic. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 5(7).

[28] Sundarkumar G. G., Vadlamani R., Siddeshwar V., 2015. One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

[29] Timofeev R., 2004. Classification and regression trees (CART) theory and applications, Berlin: Humboldt University.

[30] Urbanek G., 2011. Competencies and enterprise value. Intangible resources in the new economy, Warsaw: Oficyna Wolters Kluwer business.

[31] Włoszczyzna, M., 2022. Corporate social responsibility in the aspect of building corporate image on the example of Orange Polska SA (Master's thesis).

[32] Wu X., Sufang M., 2016. E-commerce customer churn prediction based on improved SMOTE and AdaBoost. 13th International Conference on Service Systems and Service Management (ICSSSM).

[33] Xia G., Wei-dong J., 2008. Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice, 28(1).

[34] Yong L., Yongrui Z., 2015. Research Model of Churn Prediction Based on Customer Segmentation and Misclassification Cost in the Context of Big Data. Journal of Computer and Communications. 3(6).

[35] Żółtowski, B., Simiński, P. and Kosiuczenko, K., 2022. Statistical procedures for determining of parameters for the evaluation of the condition and safety in logistic of military vehicles. Military Logistics Systems, 57(2). https://doi.org/10.37055/slw/163231