Telecommunication customer churn prediction using machine learning methods

Monika Zdanavičiūtė[1,2], Rūta Juozaitienė[1,2,3] and Tomas Krilavičius[1,2]

[1] *Vytautas Magnus University, Faculty of Informatics, Vileikos street 8, LT-44404 Kaunas, Lithuania*
[2] *Centre for Applied Research and Development, Lithuania*
[3] *Vilnius University, Vilnius, Lithuania*

Abstract

These days telecommunication sector has grown significantly due to the use of smart technologies, and it is likely to continue to grow. The main resource of telecommunications companies is customers, but due to the relatively high level of competition in this field, most customers are not tied to a single service company. To understand the key factors contributing to customer churn rate, we have analysed the real data of one telecommunication company. The data from 2020-01-01 to 2022-03-07 consisted of information on 21128 users, 140970 payments and 350379 calls. The main contribution of our work was to develop a churn prediction model which identifies customers who are most likely subject to churn. We performed experiments using k-nearest neighbours, support vector machine, decision trees, random forest, naive Bayes classifiers and the Cox proportional hazard model with time-varying covariates. Results showed that the Cox regression model with time-varying covariates was superior to classical classification methods because it can take into account static user parameters and reflect their changes over time.

Keywords

Churn prediction, telecommunication churn, survival analysis, churn, telecommunications

1. Introduction

These days telecommunication sector has grown significantly due to the use of smart technologies, and it is likely to continue to grow. The main resource of telecommunications companies is customers, but due to the relatively high level of competition in this field, most customers are not tied to a single service company. Therefore, in order to create a successful business in this field, it is necessary to know your client, his needs and opportunities. Data collected by telecommunication companies on a daily basis can be very helpful in gaining a proper understanding of customer behavior. Analysis of such data is needed to understand what factors may be associated with a customer leaving the customer base. The main goal of this research is to develop a churn prediction model which identifies customers who most likely subject to churn.

2. Literature review

The term customers churn can be described as the loss of customers to a company [1]. Due to the specifics of telecommunications companies, this is a common business problem in this area. To retain customers, companies try to predict which consumers are going to leave in a variety of ways.

A genetic algorithm has been proposed to identify customers who intend to change their telecommunications company in the near future [2]. The database used in the study consisted of 5250 customers call data. Each user's profile consisted of information about his behavior and habits (average monthly costs for local and international calls, average amount of internet data, average monthly call time, amount of roaming and special services used). In the developed model, the genetic algorithm, by iteratively adjusting the coordinates of each profile in the plane, creates certain groups of similar elements. The efficiency of the model was evaluated by observing the change in the error function in each iteration. The algorithm grouped customers into four clusters - 1% of very high-spending customers, 9% of high and medium-spending customers, 12% of medium-spending customers, and 78% of low-spending customers.

The study [3] was conducted by analyzing data from 7043 telecommunication customers, 1869 of whom had already left the customer base. Each customer is described by 21 variables, one of which is a binary, which indicates whether the user has already left the company. Using XGBoost (Extreme Gradient Boosting Tree), k-Nearest Neighbors and Random Forest methods, customers are classified

IVUS 2022: 27th International Conference on Information Technology, May 12, 2022, Kaunas, Lithuania
monika.zdanaviciute@vdu.lt (M. Zdanavičiūtė);
ruta.juozaitiene@vdu.lt (R. Juozaitienė);
tomas.krilavicius@vdu.lt (T. Krilavičius)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3611/paper15.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

into two groups according to this variable. Accuracy and F-score measures were used to assess the accuracy of the models, which showed that the XGBoost method was the best classifier. In addition, this method was used to find out which variables most influence customer exit. This study found that customers with higher monthly charges are more likely to churn.

Data from the SyriaTel telecommunications service provider were used for the [4] study. The analyzed period is 9 months (about 10 million users), and the available information includes data about client (age, gender, place of residence, type of contract concluded, services received), his actions (calls, messages, and internet usage), mobile device (device type, brand, model), and telecommunications tower infrastructure. To better describe users, the available data was used to create a social network for all customers and to calculate variables such as degree centrality measures, similarity values, and customer's network connectivity for each customer. For model training and testing, data were separated into training (70%) and testing (30%) sets. Because the data sets were unbalanced (there were significantly more outgoing customers than active ones), the classification was done in two ways: by balancing the sets and applying the data as it is. Using Decision Trees, Random Forest, GBM (Gradient Boosted Machine Tree), and XGBoost algorithms, customers were classified into two classes: churned and existing customers. The AUC (Area under curve) was used to determine accuracy. The obtained results showed that the XGBoost algorithm classifies customers best according to the available data.

The data set for the study [5] consists of call records obtained from the University of California, Department of Information and Computer Science. The data set provides information on the use of the 3333 customer mobile system, which consists of 15 quantitative, 5 categorical variables and a binary variable, describing whether the customer has left the customer base of the telecommunications service provider. In the analysis of the available call data, each user is assigned variables describing his call habits and, using classification methods, these customers are divided into two classes according to said binary variable. The research uses Neural Networks, Support Vector Machine and Bayesian classification methods. The data set is divided into training (80% of all data) and testing (20% of all data) sets so that the training set is balanced. Then 95% of the customers who leave and 5% of the existing ones remain in the testing set. The study revealed that the support vectors method best separates customers into two groups.

The study [6] uses telecommunication users data for customer analysis, which stores basic user information (age, gender, etc.), plan order information (payment method, monthly fee, full-time fee, etc.). It also provides information about the services (telephones, internet, television, insurance, etc.) and information on whether the customer is active or has already churned. Clustering (k-Means, DBSCAN) and classification methods (Multi-Layer Perceptron, Back Propagation algorithm, Decision Trees, Logistic Regression, Support Vector Machine) were used to analyze this data. The classification models were evaluated with several measures of Precision and Accuracy, and the Back Propagation algorithm and Multi-Layer Perceptron best predicted the customer's retraction. When clustering analysis was applied to groups, better active and inactive clients were separated using the DBSCAN algorithm.

The customer loyalty task is usually formulated as a classification task, the data set of which consists of active and churned users. To solve this problem, the literature suggests the use of k–Nearest Neighbors [3], Neural Networks [5][6], Support Vectors [6][5], and Bayesian classifiers [5]. Decision Tree, Random Forest, and XGBoost algorithms can also be used to analyze customer loyalty [3][4]. However, there are also cases when this problem is solved by applying clustering methods, e.g. Genetic [2] or k–means algorithm [6]. This type of task is usually based on user information, as well as payment history and calls data. Research shows that customers who pay more for services tend to change telecommunications operators.

3. Methods

1. k-Nearest Neighbors is an algorithm that stores all available cases and classifies new cases based on a similarity measure (distance functions). Euclidian distance function [7]:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

2. Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes [7]. Hyperplane equation:

$$w^T x + b = 0 \tag{2}$$

To define an optimal hyperplane we need to

maximize the width of the margin (w):

$$\max \frac{2}{\|w\|} \tag{3}$$

3. Decision Tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label [8]. A quantitative measure of randomness, entropy, is used to select a feature in a node. The initial entropy of the set E:

$$H(E) = -\sum_{k_i \in K} P(k_i|E) \log_2 P(k_i|E), \tag{4}$$

where

$$P(k_i|E) = \frac{|e : e \in E, e \in k_i|}{|E|}, \tag{5}$$

mean $E_1, \ldots, E_n$ entropy after division:

$$B(E, p) = \sum_{j=1}^{n} P(v_j|E) H(E_j), \tag{6}$$

where

$$P(v_j|E) = \frac{E_j}{E}. \tag{7}$$

4. Random Forest is an ensemble learning method for classification tasks that operates by constructing a multitude of decision trees at training time. The output of the random forest is the class selected by most trees [9].
5. Naïve Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors [7]. This assumption is called class conditional independence.

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)}. \tag{8}$$

6. Cox proportional hazard model with time varying covariates is method for investigating the effect of several variables upon the time a specified event takes to happen. In a Cox proportional hazards regression model, the measure of effect is the hazard rate. Hazard function for individual i:

$$h_i(t) = h_0 \exp(\beta_1 x_{i_1} + \beta_2 x_{i_2} + \ldots + \beta_n x_{i_n}) \tag{9}$$

where $h_0(t)$ is the baseline hazard function, $x_{i_1}, x_{i_2} \ldots, x_{i_n}$ – covariates, $\beta_1, \beta_2, \ldots, \beta_n$ – regression coefficients.

7. Confusion matrix was used to assess the accuracy of user classification. Elements of the confusion matrix:
   - TP (true positive) - The user is expected not to churn and he remains.
   - TN (true negative) - The user is expected to churn and he churns.
   - FP(false positive) - The user is expected to remain but he churns.
   - FN (false negative) - The user is expected to churn but remains.

4. Data set

The analyzed data consists of three data sets in the range from 2020-01-01 to 2022-03-07:

1. Users data set. Individual users information, which includes demographic and other data provided during registration. This study analyzes 21128 users.
2. Payments data set. 140970 payment records showing when and what type of plan was purchased and how much it cost. There are two types of plans: monthly and yearly.
3. CDR (call detail record) data set. A real-time data records documenting telephone calls or other telecommunications operations (3350379 records).

After the data transformations, a list of variables describing the users was created (Table 1).

Table 1
Created user-defining variables

Created variables used for churn prediction
Total amount of seconds
Total amount of calls
Number of failed calls
Ratio of failed calls to total calls
The amount of not failed calls
Total amount of active days
Mean call duration
Max call duration
Median between calls
Median between active days
Number of contacts called
Total amount of purchased plans
Last plan before (amount of days)
Total amount paid

5. Churn definition

In order to assess the risk of customer's churn, the definition of churn must first be de fined. Since user leaving the customer base can be described in several ways, it is necessary to monitor client behavior and changes in activity and decide which definition best describes churn. In the study, user churn is described in two different w ays. D ifferent problem-solving methods are used for each of these two options.

1. The user is classified as a churned customer if he has not purchased a new plan 35 days after the first plan purchase.
   Figure 1 shows a bar graph showing the distribution of the number of plans purchased by customers. It shows that most customers have only bought one plan.

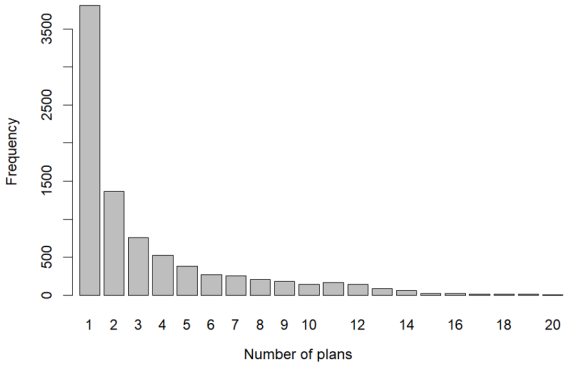

Figure 1: Frequency of number of plans pur-chased by the user

   The distribution of intervals between plan orders for users who have purchased more than one plan is shown in Figure 2. It shows that most plans are ordered every 30 days, in other words, most plans are ordered on a regular monthly basis. There are also some users who order multiple plans on the same day. The data set for the classification models consists of variables describing user behavior (Table 1), calculated on the 25th day after the purchase of the first plan. Class labels indicate whether the customer has purchased a second plan within 35 days after the first plan. Five different methods are used for classification: k-Nearest Neighbors, Sup-port Vectors Machine, Decision Tree, Ran-dom Forest and Naïve Bayes classifier. This definition of churn can only be used to predict consumers purchasing monthly

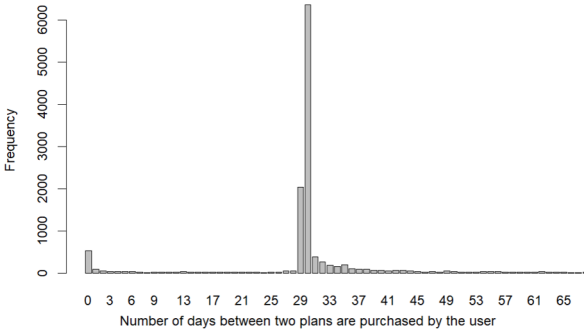

Figure 2: Frequency of number of days between plans

   plans, so it was decided to define churn in another, more universal way.
2. The user is classified as a churned customer if he does not use the services provided by the company for 25 consecutive days (does not call anyone).
   To find the optimal interval of days, after which we could treat the user as leaving, rather than just taking a break between calls, a percentage of users returned to the system after x days of inactivity is calculated. In the graph shown in Figure 3, the abscissa axis reflects the number of inactive days x, and the ordinate axis corresponds to the number of users (in percent). The blue bar then shows the percentage of users who had an x-day interval between calls, and the red bar represents the percentage of users who returned to the system after x days (call again). It can be seen from this graph that almost all users have had a one-day interval ($x = 1$) between calls and only about 60% of them have returned to the system after this interval. Nearly 80% of users have had a thirty-day interval ($x = 30$) between calls, with less than 25% returning to the system. There is no clear break in the number of users who have not returned to the system, but there is a steady decrease in the number of users who have returned to the system. It has been decided that 25 days is a sufficient period of inactivity to consider a user leaving the system.
   The user is monitored from the first day of registration until churn (25 inactive days in a row). In this case, it is not the static variables that are observed, but their change

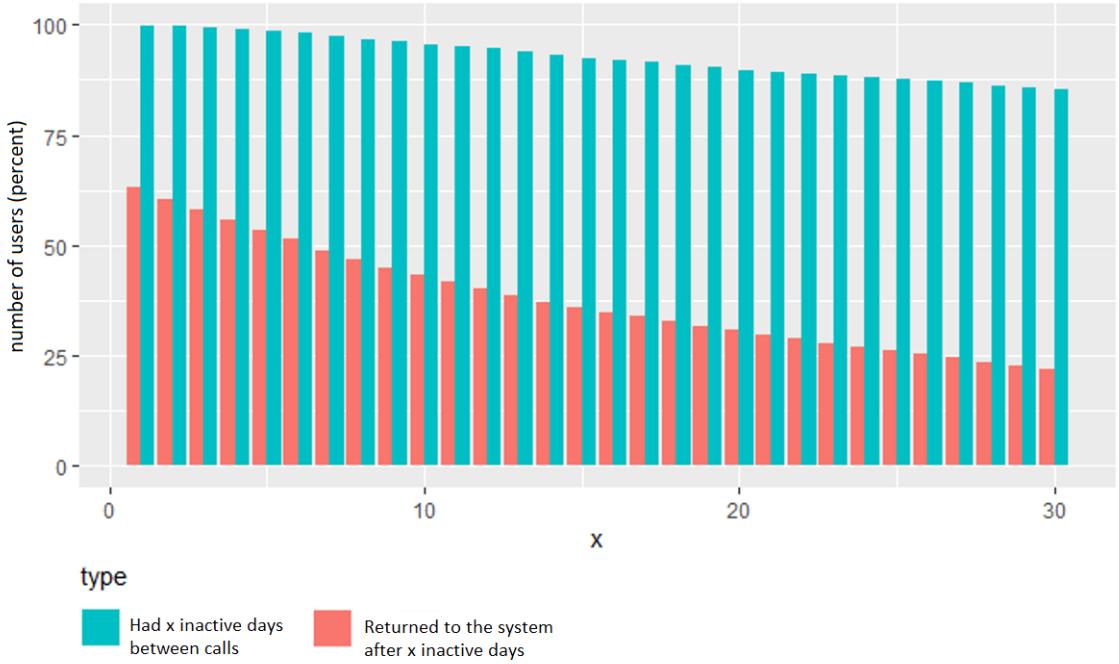

Figure 3: Frequency of users returning to the system after x days of inactivity

over time.
There are times when after a long break (after so called churn) the user returns to the system and starts using the services again. For such cases, the algorithm is designed so that withdrawn customer is still monitored, and when he returns to the system (calls again), he is treated as a newly logged-in user.

6. Experiments

6.1. Evaluating the purchase of a plan

Users classification is performed by dividing customers into these two groups:

- An active user is one who re-orders a plan within 35 days after the first order of the plan.
- Withdrawn - did not order another plan within 35 days after ordering the first plan.

The selected forecast period is 10 days. User data is tracked for 25 days from the first plan purchase. Based on these data, the characteristics describing the user's behavior are calculated on the 25th day after the purchase of the first p lan. An attempt is then made to assign the user to one of the classes (predicted to remain in the system or leave). The characteristics describing user activity are presented in Table 1.

Customers are divided into model training (70% data) and testing (30% data) sets. Five different methods are used for classification: k-Nearest Neighbors, Support Vectors Machine, Decision Tree, Random Forest and Naïve Bayes classifier.

The values of the confusion matrix elements evaluating the accuracy of the listed classification methods and the percentage accuracy of all models is given in Table 2. It can be seen that the SVM model achieves the best accuracy.

Table 2
Evaluation of classification models accuracy

Method	TP	TN	FP	FN	Accuracy
k-Nearest Neighbors	634	3747	650	1308	69.11 %
Support Vector Machine	631	4094	303	1311	74.54 %
Decision Tree	688	4028	369	1254	74.40 %
Random Forest	780	3937	460	1162	74.41 %
Naïve Bayes classifier	629	3931	466	1313	71.94 %

6.2. Estimating inactive time intervals between calls

5000 users were randomly selected from the full list of users. For each of them, the variables in Table 1 are calculated on every day from the time the user registers until he leaves or until the end of the entire data range. This results a data set in which each user is described not by one row, but by as many rows as number of days the user has been in the system for. Having this data set it is possible to track how user activity has changed over time. Any user who leaves the system for more than 25 days (does not call anyone for 25 days) and then returns to it (calls again) is treated as a new user (assigned a new identification n umber). As a result, the creation of such a user data set increases the number of users to a total of 15435. A training set (70 % of these users) is used to create the model.

To select the most appropriate Cox regression model, three different c ombinations o f variables were created and three models were constructed. Table 3 shows the variables for all three models. In each case, only non-correlated, statistically significant variables are included in the model.

Table 3
Variables describing user behavior in Cox models

Variable	M1	M2	M3
active days count	x	x	x
mean call duration	x	x	x
max call duration	x		x
number of contacts	x	x	x
nul call ratio	x	x	x
median days between calls	x	x	x
max days between calls	x	x	
median days between active d	x	x	x
total paid	x	x	x
last plan before	x	x	x

The accuracy of these three models was assessed using a test set (30% of users). Four dates in the analyzed period were selected for model testing. Active users are selected on a specific date and it is predicted that after 10 they will still be active or churned. The same is repeated with four different dates. In this way, the real performance of the model is verified, when both short-term customers and long-term customers are evaluated. Some users may have been analyzed several times at different times. In total, the model evaluated customers 2741 times, of which 1976 users did not quit and 765 when users left the system. The accuracy of the models is presented in Table 4. It can be seen that M1 and M3 models have achieved equal accuracy and are more suitable for predicting churn than M2 model.

Table 4
Evaluation of Cox models accuracy

Model	TP	TN	FP	FN	Accuracy
M1	1059	529	236	917	57.94 %
M2	991	539	226	985	55.82 %
M3	1059	529	236	917	57.94 %

7. Conclusion

Experiments with the telecommunication customer data set show that:

1. After assessing the specifics of the available data, it was decided to define user activity in two ways: according to the plans to be purchased and according to the frequency of calls made.
2. In the case where the customer is considered active as long as he regularly purchases the call plans offered by the supplier, the following classification algorithms were used to segment the users: k-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and Naïve Bayes classifier. Based on experimental studies, it can be stated that the classification of Support Vector Machine outperforms other methods.
3. In the case where user activity is defined by a faster-changing indicator – the frequency of calls, it was decided to use the Cox regresion model with time varying covariates to divide users into groups. This model is superior to classical classification methods in that it can take into account not only static user parameters but also their change over time.

In further research the possibility of combining these two methods to predict the likelihood of customer churn may be considered.

References

[1] B. Huang, M. T. Kechadi, B. Buckley, Customer churn prediction in telecommunications, Expert Systems with Applications 39 (2012) 1414–1425.

[2] H. REN, Y. ZHENG, Y. rong WU, Clustering analysis of telecommunication customers, The Journal of China Universities of Posts and

Telecommunications 16 (2009) 114 – 128. URL: http://www.sciencedirect.com/science/article/pii/S1005888508602149. doi:https://doi.org/10.1016/S1005-8885(08)60214-9.

[3] J. Pamina, B. Raja, S. SathyaBama, M. Sruthi, A. VJ, et al., An effective classifier for predicting churn in telecommunication, Jour of Adv Research in Dynamical & Control Systems 11 (2019).

[4] A. K. Ahmad, A. Jafar, K. Aljoumaa, Customer churn prediction in telecom using machine learning in big data platform, Journal of Big Data 6 (2019) 28.

[5] I. Brânduşoiu, G. Toderean, H. Beleiu, Methods for churn prediction in the pre-paid mobile telecommunications industry, in: 2016 International Conference on Communications (COMM), 2016, pp. 97–100. doi:10.1109/ICComm.2016.7528311.

[6] I. M. Mitkees, S. M. Badr, A. I. B. ElSeddawy, Customer churn prediction model using data mining techniques, in: 2017 13th International Computer Engineering Conference (ICENCO), IEEE, 2017, pp. 262–268.

[7] J. Han, J. Pei, M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[8] G. Norkevičius, G. Raškinis, Lietuvių kalbos garsų trukmės modeliavimas klasifikavimo ir regresijos medžiais, naudojant didelės apimties garsyną, Informacinės technologijos 2007: konferencijos pranešimų medžiaga, Kauno technologijos universitetas, 2007 m. sausio 31 d.-vasario 1 d. Kaunas: Technologija, 2007 (2007).

[9] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.