TELECOMMUNICATION CUSTOMER CHURN PREDICTION USING MACHINE LEARNING METHODS

RELEVANCE

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.

GOAL

Develop a churn prediction model which identifies customers who are most likely subject to churn.

DATA

The one-year data consists of information on

- 18 333 users*,
- 94 350 payments,
- 233 048 calls.

** In the original data set, some users are registered in the system but have not purchased a plan. Therefore only consumers who have purchased plans are analyzed.*

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

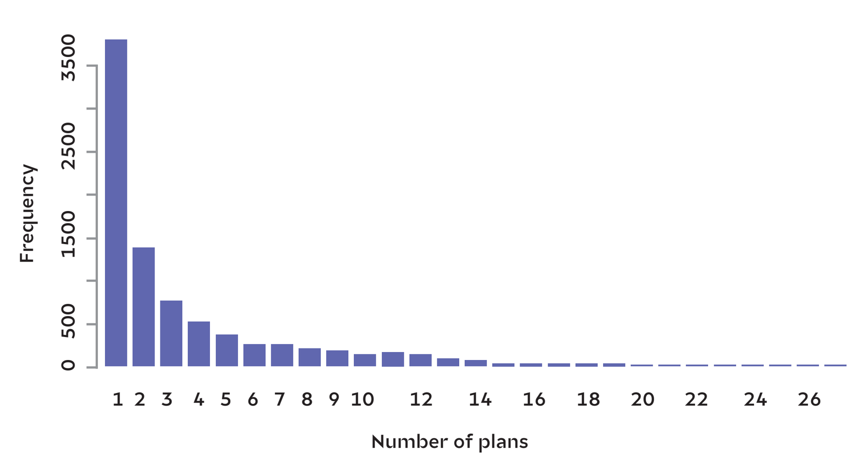
CHURN DEFINITION

In the study, user churn is described in two different ways. Different problem-solving methods are used for each of these two options.

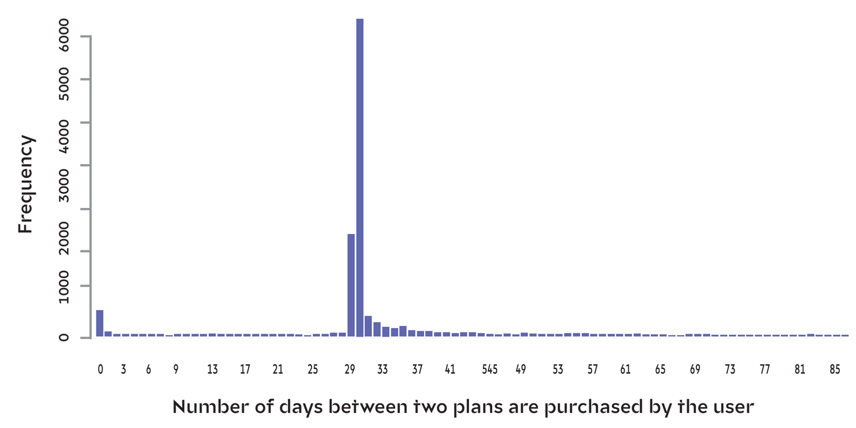
1.

The user is classified as a withdrawing customer if he has not purchased a new plan 35 days after the first plan purchase.

Frequency of number of plans purchased by the user



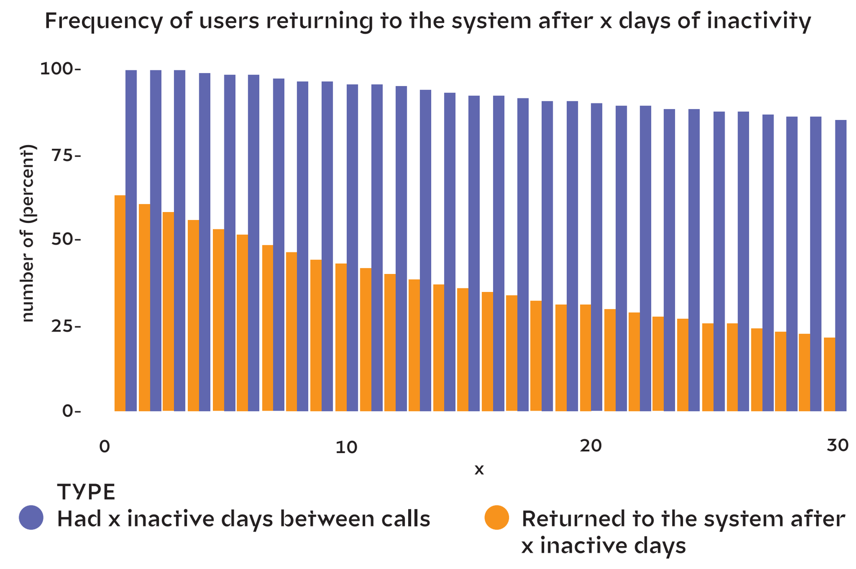
Frequency of number of days between plans



2.

The user is classified as a withdrawing customer if he does not use the services provided by the company for 25 consecutive days *(does not call anyone)*.

Frequency of users returning to the system after x days of inactivity



METHOD

1. Applying the first definition of churn *(evaluating the purchase of a plan)*

- **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:

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

- **Support Vector Machine** *(SVM)* performs classification by finding the hyperplane that maximizes the margin between the two classes. Hyperplane equation:

$$w^T x + b = 0.$$

To define an optimal hyperplane we need to maximize the width of the margin *(w)*:

$$\max \frac{2}{\|w\|}.$$

- **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. A quantitative measure of randomness, entropy, is used to select a feature in a node.
- **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.
- **Naive 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. This assumption is called class conditional independence.

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

2. Applying the second churn definition (estimating inactive time intervals between calls)

- **Cox proportional hazard model with time varying covariates.**

The hazard function for individual $\boldsymbol{i}$:

$$h_i(t) = h_o(t)\exp(\beta_1 x_{i_1}(t) + \beta_2 x_{i_2}(t) + \dots + \beta_n x_{i_n}(t)),$$

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

RESULTS

1. Applying the first definition of churn *(estimation of plan purchase probability)*

METHOD	ACCURACY
k-Nearest Neighbours	70.84%
Support Vector Machine	71.97%
Decision Tree	70.68%
Random Forest	71.15%
Naive Bayes classifier	71.15%

2. Applying the second churn definition *(estimation of the probability of becoming inactive)*

VARIABLE	COEFFICIENT	p-VALUE
Ratio of failed calls to total calls	$4.88 \cdot 10^{-1}$	$<2 \cdot 10^{-16}$
Total amount of active days	$-7.29 \cdot 10^{-3}$	$<2 \cdot 10^{-16}$
Mean call duration	$-2.97 \cdot 10^{-4}$	$<2 \cdot 10^{-16}$
Median between active days	$1.31 \cdot 10^{-1}$	$<2 \cdot 10^{-16}$
Number of contacts called	$-4.54 \cdot 10^{-3}$	$2.58 \cdot 10^{-2}$
Last plan before *(amount of days)*	$6.88 \cdot 10^{-3}$	$<2 \cdot 10^{-16}$
Total amount paid	$-1.88 \cdot 10^{-3}$	$2.53 \cdot 10^{-3}$
	MODEL ACCURACY	76.26%

AUTHORS:

Monika Zdanavičiūtė
monika.zdanaviciute@vdu.lt

Rūta Juozaitienė
ruta.juozaitiene@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

VYTAUTO DIDŽIOJO UNIVERSITETAS
Informatikos fakultetas

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT