ЗАСТОСУВАННЯ ЛОГІСТИЧНОЇ РЕГРЕСІЇДЛЯ ПРОГНОЗУВАННЯ ВІДТОКУ КЛІЄНТІВ ТЕЛЕКОМУНІКАЦІЙНОЇ КОМПАНІЇ

Земко П. А.
Науковий керівник – к.т.н., доц. Гибкіна Н. В.
Харківський національний університет радіоелектроніки
61166, Харків, просп. Науки, 14, каф. прикладної математики,
тел. (057) 702-14-36, e-mail: polina.zemko@nure.ua

This work is devoted to solving of customer churn prediction problem. The mathematical model of this problem is logistic regression, where inputs are cutomers characteristics and output – prediction if customer is going to refuse company's services. To determine major features the following methods are used: recursive feature elimination, univariate feature selection with chi-square test and mutual information. As a result, a linear classifier was built and estimated on real data.

В історії розвитку кожної компанії, що продає свої послуги, настає момент, коли ринок вже насичений та боротьба за залучення нових клієнтів відходить на другий план. Тоді основною задачею компанії стає утримання вже існуючих клієнтів та запобігання їх відтоку. Сьогодні для вирішення цієї проблеми все частіше застосовують машинне навчання. Найважливішими є методи, що надають можливість отримувати у реальному часі оцінку ймовірності втрати кожного клієнта за його характеристиками. За допомогою отриманих оцінок стає можливим формування групи ризику користувачів, які не є лояльними до компанії, та здійснення заходів щодо запобігання їх відтоку. До таких методів відноситься логістична регресія – лінійна ймовірнісна модель, яка застосовується в задачах класифікації. Не дивлячись на те, що логістична регресія не виконує статистичну класифікацію, її використовують для побудови класифікатора шляхом встановлення порогу дискримінації. У такий спосіб одразу маємо оцінки ймовірності втрати клієнта та можливість класифікувати його як «лояльного» або «нелояльного».

У роботі розглядається задача виділення групи ризику телекомунікаційної компанії для подальшого запобігання відтоку клієнтів, а також виділення ознак, що мають найбільший вплив на оцінку втрати користувача.

Нехай X – множина об'єктів, $Y = \{-1, 1\}$ – множина класів. Для поставленої задачі об'єктами будуть клієнти, для яких відомий список підключених послуг, демографічно-соціальний статус та тривалість користування послугами компанії. Також по кожному користувачу є дані чи користується він послугами зараз ($y = -1$) або розірвав стосунки з компанією ($y = 1$). Кожен об'єкт описується вектором ознак

$\vec{x}_i = \left(f_1(x_i),...,f_m(x_i)\right)$, $i = \overline{1,n}$, де n – кількість об'єктів, m – кількість ознак, а f_j, $j = \overline{1,m}$, – деяка функція-ознака від об'єкта x_i [1].

Потрібно за допомогою алгоритму $a(\vec{x})$ отримати оцінку ймовірності належності об'єкта одному з двох класів $\{-1, 1\}$:

$$a(\vec{x},\vec{w}) = \text{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \text{sign}(\vec{w}^T\vec{x}),$$

де w_j – вагові коефіцієнти, які визначаються в процесі навчання алгоритму; w_0 – поріг прийняття рішення, якому відповідає нульова ознака $f_0(x) = -1$.

Задача навчання алгоритму класифікації полягає у розв'язанні задачі мінімізації наступної функції вартості:

$$Q(\vec{w}) = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \cdot \vec{w}^T\vec{x}_i}\right) \to \min_{\vec{w}}.$$

Побудований таким чином лінійний класифікатор надає можливість за ознаками клієнта оцінити апостеріорну ймовірність того, що він перестане користуватись послугами компанії:

$$P\left(y \mid \vec{x}\right) = \frac{1}{1 + e^{-\vec{w}^T\vec{x}\cdot y}}.$$

Для кращого розуміння природи даних та виявлення закономірностей було проведено первинний аналіз вхідного набору даних. В ході попередньої обробки застосовано стандартизацію до числових та пряме кодування до категоріальних ознак.

Для задач машинного навчання важливим є обґрунтування результатів та визначення характеристик, які є мають найбільший вплив на модель. Для виділення важливих ознак, що є найбільш інформативними для оцінки лояльності користувача, у роботі запропоновано використати наступні методи: метод рекурсивного видалення ознак (RFE), універсальний вибір ознак на основі критерії незалежності хі-квадрат та взаємної інформації [2].

В роботі було побудовано класифікатор логістичної регресії на підмножинах ознак, відібраних зазначеними методами. Найкращий результат отримано методом на основі критерію хі-квадрат зі значенням метрики AUC-ROC 0,7633.

Список використаних джерел:

1. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин). URL: https://bit.ly/1bCmE3Z (дата звернення: 20.02.2021).

2. Advances in Feature Selection with Mutual Information / Verleysen M., Rossi F., François D. [та ін.] // Similarity-Based Clustering. Lecture Notes in Computer Science. 2009. vol 5400. Springer, Berlin, Heidelberg. C. 52-69.