УДК 004.85

DOI: 10.31673/2412-9070.2020.042231

В. В. ЖЕБКА, канд. техн. наук, доцент,
Державний університет телекомунікацій, Київ

ДОСЛІДЖЕННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА ЇХ ЗАСТОСУВАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВІДТОКУ КОРИСТУВАЧІВ ТЕЛЕКОМУНІКАЦІЙНИХ ПОСЛУГ

Досліджено проблему відтоку клієнтів телекомунікаційних компаній: закордонних та українських. Визначено, що доречним є використання компаніями інформаційних систем, які дають можливість спрогнозувати поведінку користувача послуг та попередити компанію у разі наявності ризиків. Встановлено, що закордонні компанії більш прогресивні щодо досліджень та утримання клієнтів. В українських компаніях інструменти дослідження неструктурованих даних є застарілими. З метою розширення інструментарію українських компаній було проаналізовано напрацювання закордонних науковців. У процесі оброблення великої кількості інформації для побудови коротко- та середньострокових прогнозів поведінки споживачів доречним є вироблення моделей поведінки користувачів та використання методів машинного навчання. Обґрунтовано, що науковці надають перевагу методу випадкового лісу. Проаналізовано методи машинного навчання: сформульовано їх переваги та недоліки. Досліджено характеристики методу випадкового лісу. Визначено, що подальшими удосконаленнями в даній сфері можуть бути композиції алгоритмів. Досліджено метод композиції алгоритмів — бустинг. Розкрито його переваги та недоліки, а також особливості даного методу. Визначено етапи побудови моделі для створення прогнозу. На основі проаналізованої літератури щодо відтоку користувачів окреслено пропозиції, які можуть зменшити скорочення їх кількості. Запропоновано пріоритетні напрямки подальших досліджень, що стосуються оптимізаційних задач методів машинного навчання, зокрема в умовах невизначених факторів.

Ключові слова: відтік користувачів; моделювання; методи машинного навчання; бустинг; неструктуровані дані.

Вступ

Постановка задачі. Сьогодні для телекомунікаційної галузі характерним є швидкий розвиток: зростає кількість послуг та товарів. Визначальними для галузі є висока конкуренція між компаніями. Саме тому важливим є утримання наявних користувачів та залучення нових (що за сучасних умов є достатньо важким завданням).

З метою збільшення прибутку та утримання користувачів телекомунікаційні компанії вводять нові послуги та відповідні тарифи. Проте кожного разу не зрозуміло, як буде поводити себе користувач у разі введення нового тарифного плану або розширення існуючого. Саме тому доречним є використання компаніями інформаційних систем, що дають можливість спрогнозувати поведінку користувача послугами та дозволять попередити компанію у разі наявності ризиків.

Мета та задачі дослідження. Моделювання поведінки споживача доречно здійснювати, використовуючи методи машинного навчання, проте кожен із цих методів є ефективним лише для своєї конкретної задачі. А тому в даній статті буде проведено аналіз методів машинного навчання з метою визначення тих методів (або їх комбінацій), які дадуть найбільш точний результат за визначених метриках.

Нині існує достатньо велика кількість досліджень, присвячених моделюванню та тестуванню різних видів машинного навчання, зокрема і для прогнозування відтоку користувачів. Виконано дослідження та розроблено рекомендації з попередження відтоку користувачів. Проте за даними Національної комісії, що здійснює державне регулювання у сфері зв'язку та інформатизації, українські телекомунікаційні компанії продовжують використовувати примітивну статистику для оброблення даних та створення прогнозів.

Тому доречним є дослідити закордонні напрацювання з даної тематики та адаптувати їх до українських реалій. Перша частина цього завдання і буде висвітлена в даній статті.

Основна частина

Протягом багатьох років телекомунікаційні компанії по-різному намагалися утримати користувачів: від акцій та знижок до застосування продуманих стратегій. А для цього потрібні глибокі, актуальні знання про сегментацію клієнтів.

Головні дані здобувають з великих наборів неструктурованих даних. Компанії, які обробляють і аналізують дані в режимі реального часу, правильно визначають сегменти і потреби клієнтів — підвищують лояльність своїх клієнтів і отримують нових.

Багато закордонних компаній, де нагромаджується велика кількість даних, застосовують методи машинного навчання та інтелектуального аналізу даних. Саме у великих неструктурованих даних компанії шукають підказки. Але дуже часто інструменти, що використовують компанії, не дають змогу дістати вірогідну інформацію. Іншою проблемою є швидкість отримання інформації.

Алгоритми машинного навчання можна описати як навчання цільової функції f, яка співвідносить вхідні параметри та вихідні змінні.

До методів та моделей машинного навчання належать такі: лінійна регресія, логістична регресія, бінарна регресія, лінійний дискримінантний аналіз, дерева прийняття рішення, наївний байєсів класифікатор, K-найближчих сусідів (KNN), мережі векторного квантування (LVQ), метод опорних векторів (SVM), випадковий ліс (бегінг), бустинг (AdaBoost), кластеризація, пошук асоціативних правил, нейронні мережі тощо.

Розглянемо лише основні з них, які на базі аналізу та в процесі проведених досліджень дають найвищу точність під час моделювання поведінки споживачів, зокрема в разі прогнозування відмови клієнтів.

Досвід застосування аналізу даних у телекомунікаційних компаніях описано в дослідженнях [1–5], які подають переважно прогнозування відтоку користувачів.

В [1] порівнювалися два алгоритми: дерево рішень та нейронні мережі. Автори надали перевагу дереву рішень, хоча в процесі дослідження точність кожного алгоритму становила більш як 90%.

У [2] автори досліджували такі алгоритми: логістичну та лінійну регресію, лінійний дискримінантний аналіз, дерево рішень, K-найближчих сусідів, нейронну мережу.

У [3] було розглянуто такі алгоритми, як K-найближчих сусідів, градієнтний бустинг, наївний байєс, випадковий ліс. Проте саме випадковий ліс дав найкращу точність (91%) та повноту (87%).

У [4] дослідник Ніл А. Акільдірим аналізував статистичні дані телекомунікаційної компанії за допомогою алгоритмів випадкового лісу та методу опорних векторів. Найкращий результат було здобуто під час використання методу випадкового лісу. У результаті проведеного моделювання точність для моделі випадковий ліс, яку було створено для прогнозування стану клієнтів телекомунікаційної компанії, становила 0,89. Для методу опорних векторів цей показник досяг лише 0,647.

У [5] висвітлено результати дослідження методів логістичної регресії (точність 0,87) та випадкового лісу (точність 0,96).

Отже, з огляду на проаналізовані результати цих та багатьох інших досліджень, можна виокремити основні методи, які для визначеної задачі дають найкращий результат: випадковий ліс, дерева прийняття рішення, K-найближчих сусідів, наївний байєсів класифікатор. Проте найкращі результати точності дає метод випадкового лісу.

Однак варто зауважити, що порівнювати результати даних досліджень не завжди доцільно, оскільки показники метрик, які використовуються в дослідженнях, залежать від способу збору даних та їх початкової структури, оброблення даних та добору найкращих параметрів для моделі (таблиця).

Переваги та недоліки методів машинного навчання

Метод	Переваги	Недоліки
Бінарна регресія	Простота. Швидкість. Точність. Добре інтерпретується. Володіє інструментами оцінювання якості моделі	Виникають труднощі за наявності нелінійних зв'язків (наприклад, між відтоком користувачів та факторами, які на це впливають). Параметр, що прогнозується, дуже часто належить неперервному числовому діапазону
K-найближчих сусідів	Простота. Добре інтерпретується	Висока складність прогнозу
Наївний байєсів класифікатор	Простота. Швидкість. Точність. Надійність	Розглядає ознаки незалежно одна від одної, проте не завжди
Дерева рішень	Інтуїтивність. Модель легко інтерпретується користувачем. Універсальні для вирішення задач класифікації та регресії	Нестабільність процесу (невеликі зміни в наборі даних можуть призвести до побудови нового дерева). Складність контролю розміру дерева
Логістична регресія	Відображає гарні результати для завдань бінарної класифікації	Залежність від набору даних. Низька стійкість до помилок
Лінійний дискримінантний аналіз	Простота реалізації. Легка інтерпретація результатів	Чутливість до розподілу вхідних даних
Метод опорних векторів	Найпопулярніший для класичної класифікації. Простота. Швидкість	Застосування можливе лише для розв'язання завдань із двома класами
Випадковий ліс	Ефективно обробляє дані з великою кількістю ознак і класів. Працює з неперервними та дискретними ознаками. Нечутливий до монотонних перетворень	Схильність до перенавчання в разі зашумленості даних. Велика розмірність моделей. Висока обчислювальна складність
Нейронні мережі	Стійкі до шумів. Розв'язують задачі навіть за наявності невідомих закономірностей. Завадостійкі. Швидкі. Переучуються у разі зміни середовища	Можлива незрозумілість причин прийнятого рішення. Відсутність гарантії отримання однозначних повторюваних результатів

Більшість дослідників проаналізованих праць віддали перевагу випадковому лісу, тому доречно висвітлити коротку характеристику даного методу.

Випадковий ліс — метод бегінгу, який містить велику кількість окремих дерев рішень, які діють як ансамбль. Інакше кажучи — відбувається класифікація методом дерева рішень та вибирається результат, який було здобуто найбільшу кількість разів.

В алгоритмі для всіх вибірок будуються дерева рішень, під час побудови яких для кожного вузла вибираються випадкові ознаки.

Загалом усі алгоритми машинного навчання можна поділити на сильні (дають високий рівень надійності, близький до 1) та слабкі (точність яких коливається трішки більш як 0,5).

Жоден алгоритм не працює ідеально. Тому до однієї і тієї самої задачі застосовуються різні алгоритми, які можуть відбивати різні результати з різною точністю. Доречним є їх об'єднання в один алгоритм, щоб компенсувати недоліки і посилити переваги кожного з них. Саме тому вдосконалення в даній сфері можливе через застосування комбінації методів. Наприклад, у [5] висвітлено дослідження, що визначає досить ефективну комбінацію регресії Кокса (модель пропорційних ризиків), яка дає змогу прогнозувати ризик та оцінювати вплив на нього незалежних змінних, з бінарною регресією, яка визначає ймовірність відмови від послуги та визначення факторів, що впливають на неї. У зазначеному дослідженні встановлено, що застосування комбінації цих методів уможливить на основі історичних даних про користувачів визначення причини і ймовірності відтоку клієнтів у встановленому часовому проміжку.

У процесі розгляду різних комбінацій алгоритмів слід зупинитися на понятті бустингу – ансамблевому методі, головною перевагою якого є послідовне використання алгоритму, причому кожен алгоритм звертає увагу на помилки попереднього.

Головна ідея бустингу — створення сильного класифікатора на основі кількох слабких. На кожній ітерації додаються моделі, які будуть виправляти помилки попередніх. Для бустингу важливим є відсутність аномалій у даних.

Ефективність бустингу забезпечується тим, що алгоритм на кожній ітерації будує базовий алгоритм, який дійсно ефективний тільки на частині підвибірки. Об'єднавши техніки градієнтного бустингу та бегінгу, можна підвищити ефективність алгоритмів: на кожному кроці алгоритму градієнтного бустингу обчислення відбувається, ґрунтуючись не на всю навчальну вибірку, а лише на випадкову підвибірку фіксованого розміру.

Алгоритмічну композицію, що лежить в основі бустингу, можна подати у вигляді [7]

$$a(x)=C(F(b_1(x),...,b_T(x)))=\operatorname{sign}\left(\sum_{t=1}^{T}\alpha_t b_t\right),\ x\in X. \tag{1}$$

Функціонал якості композиції алгоритмів визначається як кількість помилок, яких припускаються на навчальній вибірці:

$$Q_T=\sum_{i=1}^{l}\left[y_i\sum_{t=1}^{T}\alpha_i b_t(x_i)<0\right]. \tag{2}$$

Науковець Воронцов у [7] пропонує ввести дві евристики, додавши до яких експоненціальну апроксимацію

$$[y_i b(x_i)<0]\le e^{-y_i b(x_i)}, \tag{3}$$

приводять до алгоритму AdaBoost.

Коротко сформулюємо переваги і недоліки бустінгу.

Переваги бустингу:

- дає можливість розглядати різні функції втрат;
- працює з будь-яким сімейством базових алгоритмів;
- гарантує коректність на навчальній вибірці з достатньо слабкими додатковими обмеженнями;
- уможливлює виокремлення шумових об'єктів;
- дає змогу проведення математичних та алгоритмічних оптимізацій завдяки простоті методу та чіткому математичному обґрунтуванню, що також дозволить прискорити роботу алгоритму.

Недоліки бустингу:

- повільний, трудомісткий процес;
- без додаткових модифікацій алгоритм підлаштовується під дані, зокрема під помилки;
- алгоритм погано застосовується до побудови композиції з достатньо складних та потужних алгоритмів. Це можливий, але трудомісткий процес, який не забезпечує належну якість.

Водночас наступні пропозиції можуть зменшити скорочення клієнта:

1. Опитування клієнтів міжнародного плану, щоб зрозуміти больові точки та виявити першопричини, що стосуються наміру. Потім вживайте заходів для вирішення цих проблем.

2. Ескалація всіх дзвінків після першого дзвінка клієнта, аби гарантувати, що будь-які проблеми, з якими стикається клієнт, виправлені до наступного дзвінка. Активно перевіряйте клієнтів для підтвердження, що їхню проблему виправлено.

3. Опитування клієнтів, чиї тарифи на дзвінки вищі за середні, щоб перевірити наявність намірів. Визначте першопричини, які стосуються наміру збільшити, і вживайте заходів для усунення цих проблем.

Висновки

Здобуті результати дослідження дають можливість скласти картину наявних методів навчання та їх застосування до моделювання поведінки користувачів телекомунікаційних послуг, зокрема у процесі прогнозування відтоку клієнтів.

Подальшими дослідженнями в даному напрямку може бути дослідження оптимізаційних задач машинного навчання, одним із розгалужень якого є розгляд оптимізаційних задач в умовах невизначеностей даних та факторів.

Список використаної літератури

1. ***Condamoor*** *Ravi Building Predictive Models for Customer Churn in Telecom. URL: https://www.experfy.com/blog/building-predictive-models-for-customer-churn-in-telecom.*

2. ***Canale A., Lunardon N.*** *Churn prediction in telecommunications industry. A study based on bagging classifiers telecom // Carlo Alberto Notebooks, 2014. Vol. 350. P. 1–11. URL: https://www.carloalberto.org/assets/working-papers/no.350.pdf.*

3. ***Khan A. A., Sanjay J., Sepehri M. M.*** *Applying data mining to customer churn prediction in an Internet service provider // Int. J. Comput. Appl. 2010. Vol. 9, No. 7. P. 8–14. URL: http:// www.ijcaonline.org/volume9/number7/pxc3871889.pdf.*

4. ***Telecom*** *Churn Analysis. URL: https://towardsdatascience.com/customer-churn-analysis-4f77cc70b3bd*

5. ***Neal A.*** *Akyildirim Brief Overview of Customer Churn Analysis and Prediction with Decision Tree Classifier. URL: http://rstudio-pubs-static.s3.amazonaws.com/277278_427ca6a7ce7c4eb688506efc7a6c2435.html*

6. ***Скоринговое*** *моделирование финансовых потоков от взыскания / Т. И. Григорчук, З. В. Максименко, Л. Ф. Розанова, Г. Р. Бикбулатова // Нефтегазовое дело: електрон. версія журн. 2015. № 5. С. 630–655.*

7. ***Воронцов.*** *Курс «Машинное обучение» 2019 (Школа анализа данных). URL: https://ya-r.ru/2020/05/07/vorontsov-kurs-mashinnoe-obuchenie-2019-shkola-analiza-dannyh/*

В. В. Жебка

ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ИХ ПРИМЕНЕНИЕ ДЛЯ ПРОГНОЗИРОВАНИЯ ОТТОКА ПОЛЬЗОВАТЕЛЕЙ ТЕЛЕКОММУНИКАЦИОННЫХ УСЛУГ

Исследована проблема оттока клиентов телекоммуникационных компаний: иностранных и украинских. Определено, что уместно использование компаниями информационных систем, которые позволят спрогнозировать поведение пользователя услуг и предупредить компанию при наличии рисков. Проанализированы методы машинного обучения: определены их преимущества и недостатки. Исследованы характеристики метода случайного леса. Установлено, что дальнейшими усовершенствованиями в данной области могут быть композиции алгоритмов. Исследован метод композиции алгоритмов — бустинг. Установлены его преимущества и недостатки, определены особенности данного метода. На основе проанализированной литературы по оттоку пользователей определены предложения, которые могут уменьшить сокращения их количества. Предложены приоритетные направления дальнейших исследований, касающихся оптимизационных задач методов машинного обучения, в частности в условиях неопределенных факторов.

Ключевые слова: отток пользователей; моделирование; методы машинного обучения; бустинг; неструктурированные данные.

V. V. Zhebka

RESEARCH OF MACHINE LEARNING METHODS AND THEIR APPLICATION FOR FORECASTING USE OUTFLOW BY TELECOMMUNICATIONS SERVICES

The article examines the problem of outflow of customers of telecommunications companies: foreign and Ukrainian. It is determined that it is appropriate for companies to use information systems that will predict the behavior of users of services, and will warn the company in the presence of risks. It is established that foreign companies are more progressive in research and customer retention. In Ukrainian companies, unstructured data research tools are outdated. In order to expand the tools of Ukrainian companies, the achievements of foreign scientists have been studied. When processing a large amount of information in order to build short- and medium-term forecasts of consumer behavior, it is appropriate to build models of user behavior and use machine learning methods. It is established that scientists prefer the method of random forest. The methods of machine learning are analyzed in the work: their advantages and disadvantages are determined. The characteristics of the random forest method are investigated. It is established that further improvements in this area may be compositions of algorithms. The method of composition of algorithms — boosting is investigated. Its advantages and disadvantages are established, features of this method are defined. The stages of building a model for creating a forecast are defined. Based on the analyzed literature on the outflow of users, proposals have been identified that can reduce the reduction in their number. The priority directions of further researches concerning optimization problems of methods of machine learning, in particular in the conditions of uncertain factors are established.

Keywords: outflow of users; modeling; machine learning methods; boosting; unstructured data.