DOI: https://doi.org/10.35774/econa2024.01.179 **JEL classification: C45, C51, C52, C53**
UDC: 519.85+330.4

Ольга КРИВИЦЬКА
доктор економічних наук, професор,
кафедра економіко-математичного моделювання та інформаційних технологій,
Національний університет «Острозька академія», Україна
E-mail: olha.kryvytska@oa.edu.ua
ORCID ID: 0000-0002-0844-3362
Researcher ID: G-4917-2018

Юрій КЛЕБАН
старший викладач,
кафедра економіко-математичного моделювання та інформаційних технологій,
Національний університет «Острозька академія», Україна
E-mail: yuriy.kleban@oa.edu.ua
ORCID ID: 0000-0002-7070-5175
Researcher ID: AAN-3865-2021

Андрій ЯГОДКА
Національний університет «Острозька академія», Україна
E-mail: andrii.yahodka@oa.edu.ua
ORCID ID: 0009-0006-5676-188X
Researcher ID: KCK-6034-2024

УТРИМАННЯ КЛІЄНТІВ КОМЕРЦІЙНОГО БАНКУ ЯК ЗАДАЧА КЛАСИФІКАЦІЇ У МАШИННОМУ НАВЧАННІ

Анотація

Вступ. Відтік клієнтів є поширеною проблемою для багатьох галузей бізнесу, зокрема банківської сфери. Для розвитку банку необхідно залучати нових клієнтів, адже кожний втрачений клієнт призводить до зменшення прибутку і вимагає витрат часу і зусиль на пошук нового. Відтік клієнтів відбувається, коли клієнт припиняє користуватися продуктом або послугою банку. Утримання інтересу клієнта є більш вигідним і дешевим, ніж спроби залучити нового. У зв'язку з цим, скорочення відтоку клієнтів стає однією з ключових задач для бізнесу. Банки, які можуть зберігати та приваблювати нових клієнтів, мають значно більші шанси на успіх. Саме тому, використання методів машинного навчання стає одним з ключових інструментів для вирішення задачі скорочення відтоку клієнтів. Адже потенційно ці методи можуть допомогти банківським установам оптимізувати свої процеси та підвищити прибуток.

Мета. Метою роботи є оцінка ефективності використання методів машинного навчання для утримання клієнтів банку, їх побудова, тестування та оцінка економічного ефекту.

Метод (методологія). У даній статті досліджується питання утримання клієнтів комерційного банку шляхом визначення ймовірності відтоку клієнтів за допомогою класифікації методами машинного навчання. Для цього будуть побудовані моделі логістичної регресії (GLM), дерев рішень (Decision Trees), випадкового лісу (Random Forest), а також за допомогою методу опорних векторів (SVM), k-

Отримано: 18.01.2024 р. **Рекомендовано до друку:** 03.02.2024 р. **Опубліковано:** 28.02.2024 р.

Як цитувати: Кривицька О., Клебан Ю., Ягодка А. Утримання клієнтів комерційного банку як задача класифікації у машинному навчанні. *Економічний аналіз.* 2024. Том 34. № 1. С. 179-190. DOI: https://doi.org/10.35774/econa2024.01.179

найближчих сусідів (k-NN), та наївного алгоритму Баєса (Naive Bayes). Для оцінки якості побудованих моделей застосовуватиметься матриця помилок (Confusion Matrix).

Результати. Отримані результати виявили високу точність побудованих моделей та їх здатність ефективно виявляти клієнтів банку, які мають схильність до відтоку. Висновки цієї статті можуть бути корисними для розробки стратегій утримання клієнтів не лише для комерційних банків, але й для різних секторів бізнесу, де втрата клієнтів є актуальною проблемою.

Ключові слова: відтік клієнтів; банкінг; мінімізація витрат; вплив моделювання; ефективність моделей; класифікація клієнтів; автоматизація рішень; випадковий ліс.

Вступ

У сучасному економічному середовищі конкуренція в галузі банківських послуг надзвичайно висока. Комерційні банки постійно шукають способи не лише привернути нових клієнтів, але й утримати існуючих, враховуючи, що збереження клієнтів виявляється ключовим фактором для забезпечення стійкого фінансового успіху. У цьому контексті важливим є розробка ефективних стратегій утримання клієнтів, які базуються на аналізі їхньої поведінки та потреб.

За останні десятиліття швидкий розвиток методів машинного навчання та аналізу даних зробив їх необхідним інструментом для вирішення цієї проблеми. Використання алгоритмів класифікації у машинному навчанні стало ефективним способом для ідентифікації клієнтів з різним ступенем вірогідності відтоку [1].

Мета та завдання статті

Метою роботи дослідити та оцінити ефективність наявних методів машинного навчання для класифікації відтоку клієнтів та обрати найкращий для їх подальшого використання при розробці стратегій утримання клієнтів у комерційних банках. Основним завданням є розробка ефективних моделей класифікації, які дозволять ідентифікувати клієнтів з ризиком відтоку, протестувати побудовані моделі та оцінити їх економічний ефект.

Виклад основного матеріалу дослідження

У сучасних умовах, коли конкуренція на ринку зростає, а клієнти стають все більш вимогливими, банківські установи повинні зосередитися на забезпеченні високої якості обслуговування своїх клієнтів, розвитку партнерських відносин та збільшенні рівня лояльності клієнтів. Для цього необхідно використовувати різноманітні інструменти та методи, які дозволяють збільшити ефективність утримання клієнтів та залучення нових [2]. Крім традиційних методів, таких як програми лояльності, спеціальні пропозиції та знижки для постійних клієнтів, індивідуальне обслуговування та вирішення проблем клієнтів, банки можуть використовувати в комбінації з ними сучасні технології, такі як машинне навчання.

Стан проблеми полягає в тому, що зростаючий рівень конкуренції між банками зробив утримання клієнтів однією з ключових задач для бізнесу. Використання методів машинного навчання може допомогти банкам підвищити ефективність взаємодії з клієнтами та забезпечити розвиток свого бізнесу. Виявлення і класифікація клієнтів, які відмовилися від послуги банків, або клієнтів із ризиком відтоку, дозволяє банкам заздалегідь ввести зміни у свою бізнес-стратегію для мінімізації можливих втрат. Використання методів машинного навчання може допомогти банкам прогнозувати, які клієнти можуть покинути їхній банк та вчасно приймати дії щодо їх утримання [3]. За даними останніх досліджень, залучення нових клієнтів в умовах сучасних конкурентних ринків обходиться до 10 разів дорожче, ніж утримання існуючих [4].

На сьогоднішній день існує безліч досліджень про зменшення відтоку клієнтів. У багатьох з них використовуються різні методи та підходи. Багато досліджень зосереджені на використанні методів машинного навчання для прогнозування відтоку клієнтів та розробки стратегій утримання клієнтів. Прикладом такого дослідження є робота Скота Нестліна, Суналі Гупта та ін. про вимірювання та розуміння точності прогнозування моделей відтоку клієнтів [5]. Їх дослідження показало, що моделі прогнозування відтоку клієнтів залишаються ефективними протягом

щонайменше трьох місяців. Також було ідентифіковано п'ять методологічних підходів, де логістична регресія та моделі на основі дерев виявилися найкращими.

Іншим прикладом дослідження скорочення відтоку клієнтів за допомогою моделей машинного навчання є робота А. Тамадоні, С. Стакховича та М. Евінга [6]. В даному дослідженні проведено порівняння різних моделей прогнозування відтоку клієнтів у неумовних налаштуваннях. Застосовано два методи даних (SVM та boosting), ймовірнісну модель Pareto/NBD, логістичну регресію та метрику RFM. Оцінено їхню ефективність через прибуток клієнтів і встановлено, що метод boosting забезпечує найбільший прибуток. Модель Pareto/NBD не перевершує логістичну регресію. Дослідження підтверджує перевагу методу boosting у більшості випадків, за винятком ситуацій з дуже низьким відтоком клієнтів або невеликим обсягом даних, які викликають успіх логістичного методу.

Деякі дослідження також вивчають взаємозв'язок між показниками задоволеності клієнтів та їх лояльністю до бренду. Прикладом цього є дослідження Джонатана Л., Джангхюка Л. та Лоуренса Ф [7]. У роботі детально досліджено вплив лояльності та задоволеності клієнтів на їх відтік. З даного дослідження можна зробити висновок, що задоволеність клієнта має позитивний вплив на лояльність клієнтів і, відповідно, зменшує коефіцієнт відтоку. Проте варто пам’ятати, що як тільки рівень задоволеності послугами впаде, відразу ж з’являється висока ймовірність відмови від послуг банку, оскільки клієнти звикають до певного рівня якості послуг.

З вказаних вище досліджень можна помітити, що використання методів машинного навчання виявилося дієвим інструментом утримання клієнтів у багатьох дослідженнях. Результати підтверджують стійкість ефективності прогнозування відтоку протягом тривалого періоду, а також перевагу деяких методів, таких як boosting. Дослідження також вказують на важливість збереження високого рівня задоволеності клієнтів, який сприяє їхній лояльності та зменшенню відтоку. Однак, низький рівень задоволеності може призвести до збільшення відтоку, оскільки клієнти мають тенденцію звикати до певного рівня обслуговування. Результати цих досліджень підкреслюють важливість використання методів машинного навчання у сучасних стратегіях управління клієнтською базою та розвитку бізнесу [8].

Для ідентифікації того, чи відмовиться клієнт від послуг компанії зазвичай використовується класифікація. Це функція інтелектуального аналізу даних, яка точно передбачає цільовий клас для кожного випадку в даних і розуміє зв’язок між різними групами даних. Класифікаційні моделі дозволяють визначити, чи залишиться клієнт у банку, чи піде. Для цього використовуються різноманітні алгоритми, такі як логістична регресія, дерева рішень, метод опорних векторів тощо. Класифікаційні моделі є важливим інструментом у боротьбі з відтоком клієнтів та підвищенні їх лояльності [9].

Одним з базових методів математичного моделювання є логістична регресія. Вона базується на математично орієнтованому підході до аналізу впливу одних змінних на інші [10]. Прогнозування здійснюється шляхом формування набору рівнянь, що пов'язують вхідні значення (тобто, що впливають на відтік клієнтів) з вихідним полем (ймовірністю відтоку). У логістичному регресійному аналізі прогнозується ймовірність того, що залежна змінна матиме два значення. Крім того, змінні в моделі є безперервними. Через цю особливість цей метод часто використовується для класифікації спостережень. Класичне рівняння логістичної регресії має вигляд [11]:

$$p(y = 1 \mid x1,\ldots,xn) = f(y) \quad (1)$$

$$f(y) = \frac{1}{(1+e^{-y})} \quad (2)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (3)$$

, де:

- y - цільова змінна для кожного індивідуума j (клієнта в моделюванні відтоку), y - бінарна мітка класу (0 або 1);
- β_0 є константою;
- β_1 вага, що надається конкретній змінній, пов'язаній з кожним клієнтом j (j=1,.... m);
- $x_1,\ldots,x_n$ це факторні (незалежні) змінні для кожного клієнта j, на основі яких потрібно спрогнозувати y.

До основних методів моделювання також відносяться дерева рішень, які являють собою деревоподібний граф, що відображає взаємозв'язки між змінними. Використовувані для вирішення проблем класифікації та прогнозування, моделі дерев рішень представляються та оцінюються за принципом "зверху-вниз". Для оцінки набору даних клієнта за допомогою розробки дерева рішень, класифікація виконується шляхом зміни дерева до тих пір, поки не буде досягнутий корінний вузол. При цьому оцінюючи відповідність клієнта значенню churner, якщо ймовірність відтоку висока, або non churner, якщо ймовірність відтоку низька, відповідне значення присвоюється його вузлу [11, 12].

Ще одним методом є метод опорних векторів – один з найпопулярніших алгоритмів керованого навчання, який використовується як для класифікації, так і для регресійних задач. Метою алгоритму SVM є створення найкращої лінії або межі рішення, яка може розділити n-вимірний простір на класи так, щоб ми могли легко віднести нову точку даних до правильної категорії в майбутньому. Ця найкраща межа рішення називається гіперплощиною [13]. Цей метод вибирає екстремальні точки/вектори, які допомагають у створенні гіперплощини. Ці екстремальні випадки називаються опорними векторами, а алгоритм - машиною опорних векторів [14].

Також існує алгоритм випадкового лісу – це алгоритм машинного навчання, який найчастіше використовується для класифікації та регресію вченими, що займаються аналізом даних та машинним навчанням. Цей алгоритм належить до керованого машинного навчання, що широко використовується в задачах класифікації та регресії. Він будує дерева рішень на різних вибірках і бере їх більшість голосів для класифікації та середнє значення у випадку регресії [15].

Також використовується наївний алгоритм Баєса – один з найважливіших алгоритмів машинного навчання, який допомагає вирішувати проблеми класифікації. Він походить від теорії ймовірностей Баєса і використовується для класифікації тексту, де ви навчаєтесь на наборах даних високої розмірності. Одними з найкращих прикладів наївного алгоритму Баєса є аналіз настроїв, класифікація нових статей та фільтрація спаму. Це алгоритм, який вивчає ймовірність кожного об'єкта, його особливості та групи, до яких він належить. Він також відомий як імовірнісний класифікатор. Наївний алгоритм Баєса відноситься до керованого навчання і в основному використовується для вирішення завдань класифікації [16].

Для оцінки якості побудованих моделей найчастіше використовується матриця помилок (Confusion Matrix). Вона є важливим інструментом при оцінці ефективності моделей класифікації, надаючи детальну інформацію про відповідність між прогнозованими і фактичними класами, що дозволяє ідентифікувати, наскільки ефективно модель працює під час визначення до якого класу віднести те чи інше спостереження [17].

Матриця помилок представляється у вигляді таблиці, де кожен ряд відповідає фактичному класу, а кожний стовпець - прогнозованому класу. Зазвичай вважаються два класи: "позитивний" (наприклад, "клієнт, який відмовиться від послуг") та "негативний" (наприклад, "клієнт, який залишається"). Однак, у випадку багатокласової класифікації, матриця помилок може мати більше рядків і стовпців, які відповідатимуть різним класам [18].

У Confusion Matrix використовуються чотири основних показники:

- True Positives (TP): Кількість спостережень, які належать до позитивного класу і були правильно визначені моделлю як позитивні;
- True Negatives (TN): Кількість спостережень, які належать до негативного класу і були правильно визначені моделлю як негативні;
- False Positives (FP): Кількість спостережень, які належать до негативного класу, але були помилково визначені моделлю як позитивні;
- False Negatives (FN): Кількість спостережень, які належать до позитивного класу, але були помилково визначені моделлю як негативні.

За допомогою цих значень можна обчислити показники, що характеризують продуктивність моделі, такі як [19]:

1. F-міра (F1-score): Цей показник є гармонічним середнім між точністю і чутливістю і використовується для оцінки збалансованості моделі. Розраховується за формулою:

$$F1\text{-score} = 2 * (Precision * Recall) / (Precision + Recall) \quad (4)$$

2. Точність (Precision): Вимірює наскільки точно модель класифікує позитивні приклади. Формула має вигляд:

$$Precision = TP / (TP + FP) \quad (5)$$

3. Чутливість (Recall) або Сприйнятливість (Sensitivity): Вимірює наскільки ефективно модель розпізнає позитивні приклади. Розраховується за формулою:

$$Recall = TP / (TP + FN) \quad (6)$$

4. Специфічність (Specificity): Вимірює наскільки ефективно модель розпізнає негативні приклади. Розраховується за формулою:

$$Specificity = TN / (TN + FP) \quad (7)$$

5. Точність прогнозування (Accuracy): Вимірює загальну точність моделі і враховує як правильно класифіковані позитивні, так і негативні приклади. Формула має вигляд:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

6. Збалансована точність (Balanced Accuracy): Цей показник приділяє однакову вагу кожному класу, незалежно від його частки в сукупності спостережень. Він дозволяє оцінити загальну продуктивність моделі, враховуючи незбалансованість класів, коли кількість спостережень різних класів може суттєво відрізнятися. Формула має вигляд:

$$Balanced\ Accuracy = (Sensitivity + Specificity) / 2 \quad (9)$$

Ці показники допомагають оцінити продуктивність моделі класифікації та зрозуміти, наскільки вона ефективно розпізнає кожен з класів. Зазвичай, вибір показника залежить від специфіки задачі та вимог до моделі. Наприклад, якщо важливо мінімізувати FP (False Positives) результати, то використовують точність (Precision). Якщо важливо мінімізувати FN (False Negative) результати, то використовують чутливість (Recall) [19].

Також для оцінки якості бінарного моделювання використовується ROC-крива (англ. Receiver Operating Characteristic curve) – це графік, який використовується для візуалізації продуктивності моделі класифікації на основі її чутливості та специфічності при різних порогах відсічення. Ця крива обчислюється за допомогою розрахунку чутливості та специфічності для різних значень порогу відсічення, які визначаються на основі прогнозів моделі для тестових даних.

Перед розробкою та використанням моделі необхідно створити її концептуальну схему. У цьому випадку процес побудови моделі, яка визначатиме схильних до відтоку клієнтів, включає кілька етапів: постановку проблеми; вибір та обґрунтування вхідних факторів моделі; визначення конфігурації моделі та налаштування її параметрів; експериментальне моделювання на тренувальній вибірці; перевірку моделі на адекватність; аналіз результатів.

Отже, для побудови моделі потрібно виконати наступні кроки:

- зібрати дані для дослідження, оцінити рівень кореляції між числовими показниками, сформувати тренувальну та тестову вибірки;
- побудувати моделі та оцінити їх адекватність на тестовій вибірці за допомогою Confusion Matrix та ROC-кривої (англ. Receiver operating characteristic).
- проаналізувати отримані результати, сформувати висновків щодо результатів моделювання на основі економічного ефекту.

Для побудови моделей було використано набір даних комерційного банку з інформацією про клієнтів та їх рахунки. Датасет складається з 10000 спостережень і 12 змінних. Незалежні змінні містять інформацію про клієнтів, як категоріальну, так і числову. Залежною змінною є бінарний показник «Churn», що вказує, чи клієнт продовжує користуватися послугами банку (значення «0»,) чи відмовився від них (значення «1»). Структура нашої вибірки наступна:

Таблиця 1. Структура залежної ознаки Churn початкового набору даних для подальшого проведення моделювання

Назва	Подія «відтік клієнта»	Подія «клієнт залишиться»
Значення бінарної змінної	1	0
Кількість спостережень, од.	2 037	7 963
Частка, %	20,37	79,63

Тренувальні та тестові вибірки розділені у пропорціях 70/30. Розподіл даних продемонстровано у табл. 2 і табл. 3.

Таблиця 2. Структура залежної ознаки Churn у тренувальній вибірці

Назва	Значення змінної «1»	Значення змінної «0»
Кількість спостережень, од.	1 426	5 574
Частка, %	20,37	79,63

Таблиця 3. Структура залежної ознаки Churn у тестовій вибірці

Назва	Значення змінної «1»	Значення змінної «0»
Кількість спостережень, од.	611	2 389
Частка, %	20,37	79,63

Наш набір даних є незбалансованим, оскільки частка клієнтів, які відмовилися від послуг банку, становить лише 20,37% від загальної кількості спостережень, тоді як решта 79,63% складають клієнти, що залишилися.

Незбалансованість тренувального набору даних у моделюванні відтоку клієнтів може призвести до недооцінки проблеми та порушення об'єктивності моделі. Моделі можуть страждати від перекосу в напрямку більш представленого класу та погано визначати клієнтів, які справді схильні до відтоку. Також незбалансованість може спричинити зміщення оцінки точності моделі та призвести до недооцінки її ефективності. Тому, для вирішення цієї проблеми використовуватиметься балансування вибірки за допомогою методу ROSE (Random Over-Sampling Examples) [20]. Це метод на основі bootstrap, який призначений для вирішення задач бінарної класифікації в умовах наявності рідкісних класів. ROSE обробляє як безперервні, так і категоріальні дані, генеруючи синтетичні приклади з умовної оцінки щільності двох класів. Після балансування тренувальна вибірка має наступний вигляд:

Таблиця 4. Структура залежної ознаки Churn у тренувальній вибірці після балансування

Назва	Значення змінної «1»	Значення змінної «0»
Кількість спостережень, од.	3 458	3 542
Частка, %	49,40	50,60

Для порівняння важливості факторів використаємо Random Forest, який показує для кожної змінної, наскільки вони важливі для включення у модель. Графік середнього зниження точності показує, наскільки модель втрачає точність, якщо виключити кожну змінну. Чим більше страждає точність, тим важливішою. Для обрахунку важливості використовується середнє зменшення коефіцієнта Джині, яке є мірою того, як кожна змінна робить внесок в однорідність вузлів і дерев. Чим вище значення точності середнього спаду або середнього спаду коефіцієнта Джині, тим вища важливість змінної в моделі. У таблиці 5 продемонстровано рівень важливості кожної незалежної ознаки.

Таблиця 5. Важливість факторних ознак для моделювання відтоку клієнтів банку на основі алгоритму Random Forest

№	Факторна ознака	Важливість, Джині
1	Кількість використовуваних послуг	714,30
2	Вік клієнта	708,98
3	Середній баланс картки	349,87
4	Кредитний рейтинг	303,83
5	Приблизна заробітна платня	296,33
6	Час користування послугами банку	296,15
7	Країна проживання	147,94
8	Наявність статусу активного користувача	104,49
9	Стать	60,38
10	Наявність кредитної картки	43,68

Побудуймо усі згадані вище моделі на основі цього датасету для подальшої перевірки їх точності на тестовій вибірці.

Таблиця 6. Результати моделювання відтоку клієнтів комерційного банку

№	Модель	AUROC	F1	Precision	Recall	Specificity	Balanced Accuracy
1	NeuralNet	0.5892	0,3754	0,2467	0,7856	0,3864	0,5860
2	SVM	0.8543	0,5985	0,5206	0,7038	0,8342	0,7690
3	**Random Forest**	**0.8609**	**0,6133**	**0,5444**	**0,7021**	**0,8497**	**0,7759**
4	Decision Trees	0.8089	0,5501	0,4529	0,7005	0,7836	0,7420
5	Logit	0.7721	0,4953	0,3841	0,6972	0,7141	0,7057
6	KNN	0.5600	0,3248	0,2382	0,5106	0,5823	0,5464

З таблиці 6 можна побачити, що третя модель, побудована за допомогою методу Random Forest, має найвищу точність класифікації відтоку клієнтів комерційного банку. Вона має найвищі значення ROC, F1, Precision, Specificity та Balanced Accuracy, тобто здатна найкраще дати відповідь на запитання, чи відмовиться клієнт від послуг банку, чи залишиться надалі.

За допомогою ROC кривої, можна оцінити ефективність моделі. Значення AUC (Area Under the Curve) для моделі побудованої методом випадкового лісу складає 0.8609. Це означає, що модель має високу прогнозовану точність і добре відрізняє клієнтів, які мають високу ймовірність відтоку та тих, які залишаться з компанією. Завдяки цій моделі компанія може знайти схильних до відтоку клієнтів і прийняти відповідні заходи для їх збереження, такі як спеціальні пропозиції, знижки, акції, подарункові сертифікати та ін.

Порівняймо економічний ефект від кожної моделі. Для цього використовуватиметься матриці помилок (англ. Confusion Matrix) кожної з побудованих моделей. Для обчислення відносного економічного ефекту усіх створених моделей припустимо, що внаслідок втрати клієнта компанія втрачає 800 ум. од. Коли ж модель ідентифікує клієнта як «схильного до відтоку», то для проведення дій з його утримання (знижки, спеціальні пропозиції, акції, подарункові сертифікати) в середньому витрачатиметься 100 ум. од.

Якщо розглядати це детальніше, то коли модель робить позитивний прогноз відтоку для позитивного фактичного значення відтоку (True Positive), то компанія нічого не втрачатиме, оскільки правильно ідентифікує схильних до відтоку клієнтів і прийме ряд мір по їх утриманню. Якщо ж модель робить позитивний прогноз відтоку для негативного фактичного значення відтоку (False Positive), то відповідно компанія здійснить дії щодо утримання клієнта,

який не мав наміру покидати компанію і витратить 100 ум. од. Якщо модель зробить негативний прогноз для негативного фактичного значення відтоку (True Negative), то компанія нічого не робитиме і не втратить кошти. Якщо ж модель зробить негативний прогноз для позитивного фактичного значення відтоку (False Negative), то компанія не здійснить ніяких дій щодо утримання клієнта і відповідно втратить 800 ум. од., що є найгіршим сценарієм.

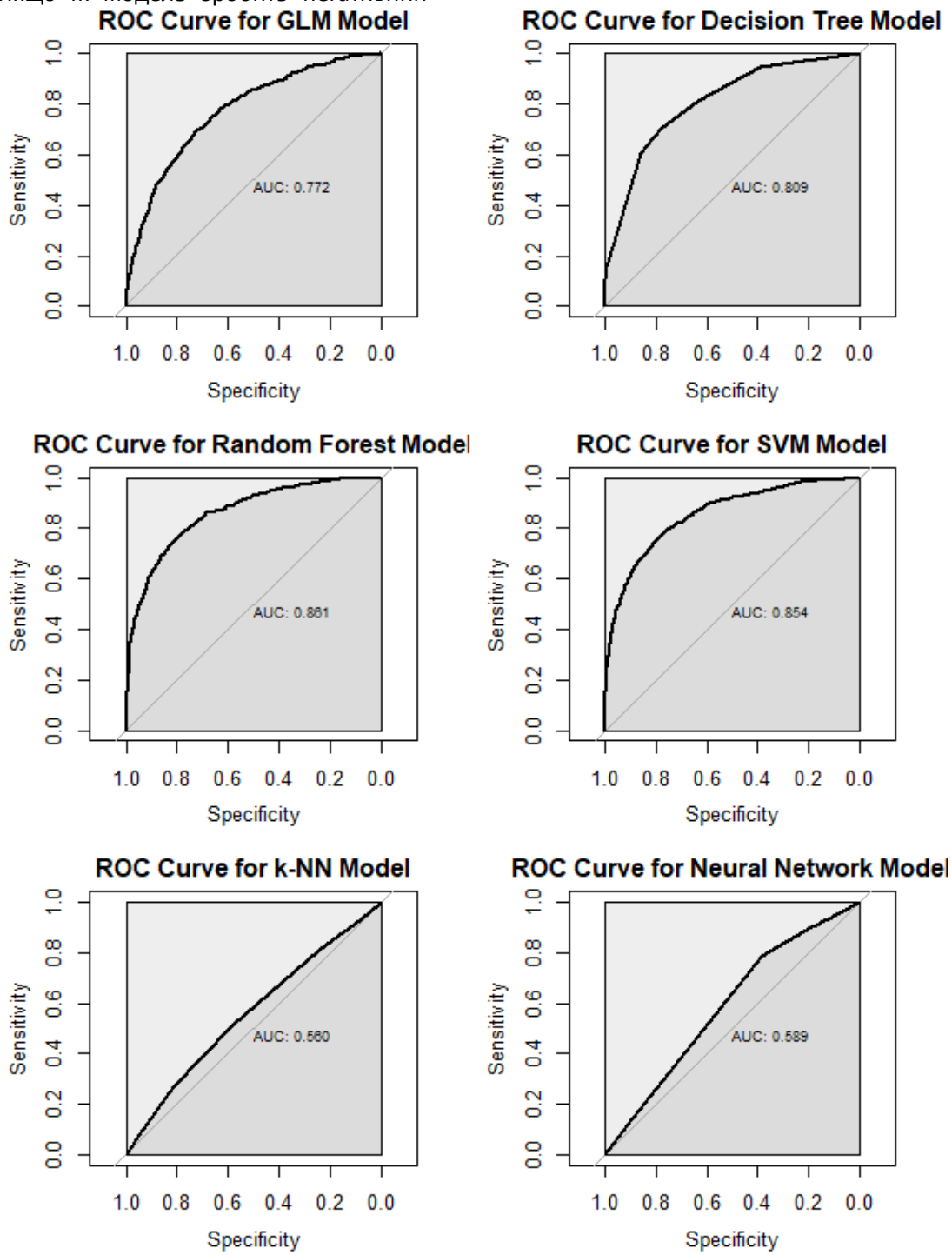

Рис. 1. Графіки ROC-кривих кожної з побудованих моделей

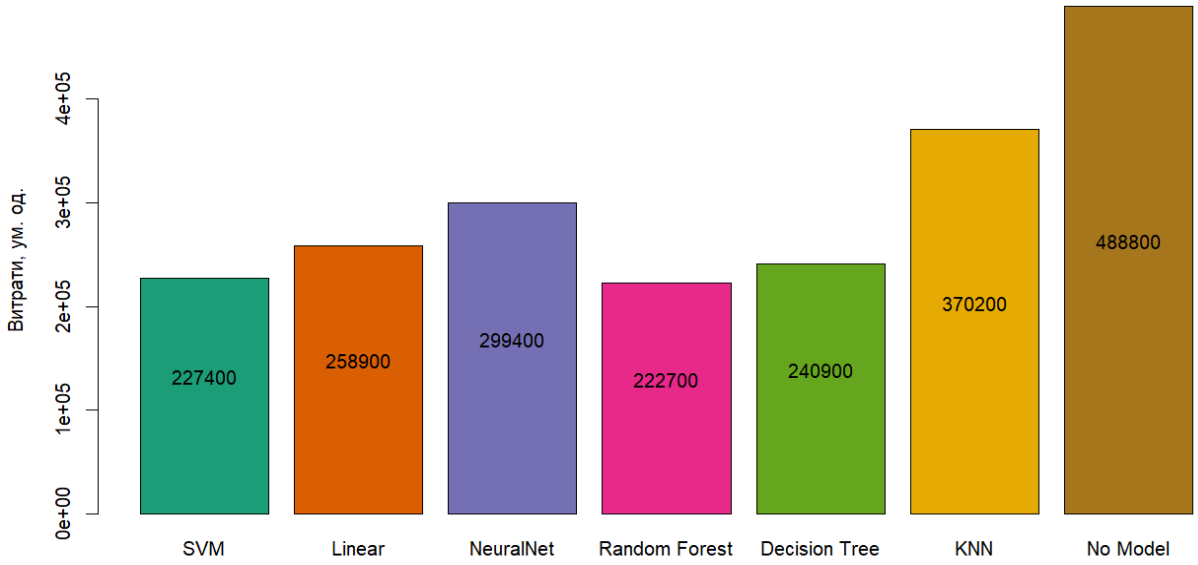

Рис. 1. Графік порівняння втрат внаслідок відтоку клієнтів залежно від використовуваної моделі

Можна дійти висновку, що модель побудована методом випадкового лісу, яка попередньо була обрана найточнішою, максимально мінімізує спричинені відтоком клієнтів умовні витрати. Варто зазначити, що не використовуючи ніяку зі створених моделей витрати компанії будуть більшими, порівняно з ситуацією, за якої компанія використовуватиме одну з розроблених моделей.

Отже, модель побудована методом випадкового лісу здатна правильно визначити 86,09%, що дозволить банку здійснити певні дії щодо їхнього утримання (спеціальні пропозиції, акції та ін). Проте варто пам'ятати, що не існує універсальної моделі машинного навчання, тому потрібно проводити більш детальні дослідження і обирати найкращий метод моделювання виходячи зі специфіки поставленого завдання.

Висновки та перспективи подальших розвідок

Результати проведених досліджень засвідчують, що використання методів машинного навчання для утримання клієнтів відкриває перед банками ряд значних переваг. По-перше, такий підхід дозволяє автоматизувати та оптимізувати процес управління клієнтською базою, що зменшує трудомісткість та витрати на утримання клієнтів. Моделі машинного навчання виявляють себе ефективними в прогнозуванні поведінки клієнтів та ідентифікації тих, хто схильний до відтоку, що дозволяє банкам вчасно реагувати та вживати заходів для їх утримання.

По-друге, використання моделей машинного навчання дозволяє банкам вдосконалити свої стратегії маркетингу та персоналізувати підходи до клієнтів. Аналіз даних та прогнозування поведінки клієнтів дозволяють банкам надавати індивідуальні пропозиції та акції, що збільшує їхню лояльність та знижує ризик відтоку.

Крім того, моделі машинного навчання можуть швидко адаптуватися до змінних умов ринку та поведінки клієнтів, що дозволяє банкам бути гнучкими та реагувати на зміни в реальному часі. В цілому, використання методів машинного навчання в управлінні клієнтською базою дозволяє банкам підвищити ефективність своєї роботи, знизити витрати та збільшити прибутковість.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Customer Churn: Definition, Rate, Analysis and Prediction. 2015. URL: https://www.questionpro.com/blog/customer-churn/.

2. Customer churn 101: What it is, why churn happens, and what you can do about it [Електронний ресурс]. – 2014. URL: https://www.paddle.com/resources/customer-churn.

3. X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009 International Conference on Electronic Commerce and Business Intelligence, Beijing, China, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.86.

4. Kotler, Philip; Keller, Kevin Lane; Adi Maulana; Wibi Hardani.Manajemen Pemasaran Jilid 1 / Philip Kotler, Kevin Lane Keller; Editor: Adi Maulana, Wibi Hardani.2009.

5. Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43(2), 204-211. https://doi.org/10.1509/jmkr.43.2.204

6. Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. Journal of Service Research, 19(2), 123-141. https://doi.org/10.1177/1094670515616376

7. Lee, J., Lee, J. and Feick, L. (2001), "The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in France", Journal of Services Marketing, Vol. 15 No. 1, pp. 35-48. https://doi.org/10.1108/08876040110381463.

8. Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425. URL: http://dx.doi.org/10.1016/j.eswa.2011.08.024.

9. Amin, A., Khan, C., Ali, I., & Anwar, S. (2014). Customer churn prediction in telecommunication industry: With and without counter-example. Nature-Inspired Computation and Machine Learning (pp. 206-218). Springer. URL: http://dx.doi.org/10.1007/978-3-319-13650-9_19.

10. Seema Baghla and Gaurav Gupta. 2022. Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce. Microprocess. Microsyst. 94, C (Oct 2022). https://doi.org/10.1016/j.micpro.2022.104680.

11. ÇELIK, Özer; OSMANOĞLU, Usame Ömer. Comparing to Techniques Used in Customer Churn Analysis. Journal of Multidisciplinary Developments, [S.l.], v. 4, n. 1, p. 30-38, july 2019. ISSN 2564-6095. <http://www.jomude.com/index.php/jomude/article/view/62>.

12. X. Hu, Y. Yang, L. Chen and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 129-132, doi: 10.1109/ICCCBDA49378.2020.9095611.

13. Jinho Kim, Byung-Soo Kim, and Silvio Savarese. 2012. Comparing image classification methods: K-nearest-neighbor and support-vector-machines. In Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics (AMERICAN-MATH'12/CEA'12). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 133–138.

14. Support Vector Machine Algorithm [Електронний ресурс]. – 2016. URL: https://www.javatpoint.com/machinelearning-support-vector-machine-algorithm

15. Introduction to Random Forest in Machine Learning. 2017. URL: https://www.section.io/engineering-education/introduction-to-random-forest-in-machinelearning/.

16. An Introduction to Naive Bayes Algorithm for Beginners. 2017. URL: https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners#whatis-the-naive-bayes-algorithm?

17. Patro, V. M., & Patra, M. R. (2014). Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy. Transactions on Engineering and Computing Sciences, 2(4), 77–91. https://doi.org/10.14738/tmlai.24.328.

18. Sokolova, Marina & Lapalme, Guy. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management. 45. 427-437. 10.1016/j.ipm.2009.03.002.

19. Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.

20. Hayati, Mardhiya & Mutmainah, Siti & Ghufran, Syed. (2021). Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification. International Journal of Artificial Intelligence Research. 4. 86. 10.29099/ijair.v4i2.152.

REFERENCES

1. *Customer Churn: Definition, Rate, Analysis and Prediction*. (2015). https://www.questionpro.com/blog/customer-churn/.

2. Customer churn 101: What it is, why churn happens, and what you can do about it. (2014). https://www.paddle.com/resources/customer-churn

3. X. Zhang, G. Feng and H. Hui, "Customer-Churn Research Based on Customer Segmentation," 2009 International Conference on Electronic Commerce and Business Intelligence, Beijing, China, 2009, pp. 443-446, doi: 10.1109/ECBI.2009.86.

4. Kotler, Philip ; Keller, Kevin Lane ; Adi Maulana ; Wibi Hardani.Manajemen Pemasaran Jilid 1 / Philip Kotler, Kevin Lane Keller; Editor: Adi Maulana, Wibi Hardani.2009

5. Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 43(2), 204-211. https://doi.org/10.1509/jmkr.43.2.204

6. Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing Churn Prediction Techniques and Assessing Their Performance: A Contingent Perspective. *Journal of Service Research,* 19(2), 123-141. https://doi.org/10.1177/1094670515616376

7. Lee, J., Lee, J. and Feick, L. (2001), "The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in France", *Journal of Services Marketing*, 15(1), 35-48. https://doi.org/10.1108/08876040110381463.

8. Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1), 1414-1425.

9. Amin, A., Khan, C., Ali, I., & Anwar, S. (2014). Customer churn prediction in the telecommunication industry: With and without counter-example. In Nature-Inspired Computation and Machine Learning (pp. 206-218). Springer.

10. Seema Baghla and Gaurav Gupta. 2022. Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce. Microprocess. Microsyst. 94, C (Oct 2022). https://doi.org/10.1016/j.micpro.2022.104680

11. ÇELIK, Özer; OSMANOĞLU, Usame Ömer. Comparing to Techniques Used in Customer Churn Analysis. Journal of Multidisciplinary Developments, [S.l.], v. 4, n. 1, p. 30-38, july 2019. ISSN 2564-6095. http://www.jomude.com/index.php/jomude/article/view/62.

12. X. Hu, Y. Yang, L. Chen and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 129-132, doi: 10.1109/ICCCBDA49378.2020.9095611.
13. Jinho Kim, Byung-Soo Kim, and Silvio Savarese. (2012). Comparing image classification methods: K-nearest-neighbor and support-vector-machines. In Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics (AMERICAN-MATH'12/CEA'12). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 133–138.
14. Support Vector Machine Algorithm. (2016). Retrieved from https://www.javatpoint.com/machinelearning-support-vector-machine-algorithm
15. Introduction to Random Forest in Machine Learning. (2017). Retrieved from https://www.section.io/engineering-education/introduction-to-random-forest-in-machinelearning/
16. An Introduction to Naive Bayes Algorithm for Beginners. (2017). Retrieved from https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners#whatis-the-naive-bayes-algorithm?
17. Patro, V. M., & Patra, M. R. (2014). Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy. Transactions on Engineering and Computing Sciences, 2(4), 77–91. https://doi.org/10.14738/tmlai.24.328
18. Sokolova, Marina & Lapalme, Guy. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45, 427-437. 10.1016/j.ipm.2009.03.002.
19. Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.
20. Hayati, Mardhiya & Mutmainah, Siti & Ghufran, Syed. (2021). Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification. International Journal of Artificial Intelligence Research, 4. 86. 10.29099/ijair.v4i2.152.

Olha Kryvytska, Doctor of Sciences (Economics), Professor, Department of Economic-Mathematical Modeling and Information Technologies, National University of Ostroh Academy, Ukraine
Yurii Kleban, Senior Lecturer, Department of Economic-Mathematical Modeling and Information Technologies, National University of Ostroh Academy, Ukraine
Andrii Yahodka, National University of Ostroh Academy, Ukraine

CUSTOMER RETENTION IN COMMERCIAL BANKING AS A CLASSIFICATION TASK IN MACHINE LEARNING

Abstract
Introduction. Customer churn is a common problem for many industries, particularly the banking sector. To thrive, banks need to attract new customers, as each lost customer leads to a decrease in profit and requires time and effort to acquire a new one. Customer churn occurs when a client ceases to use a bank's product or service. Retaining customer interest is more beneficial and cost-effective than attempting to attract new ones. Therefore, reducing customer churn becomes one of the key tasks for businesses. Banks that can retain and attract new customers have significantly higher chances of success. Hence, the use of machine learning methods becomes one of the key tools for addressing the task of reducing customer churn. These methods have the potential to help banking institutions optimize their processes and increase profitability.
Purpose. The aim of the study is to assess the effectiveness of using machine learning methods for customer retention in a bank, including their construction, testing, and evaluation of the economic impact.
Method (methodology). This article investigates the issue of retaining customers of a commercial bank by determining the probability of customer churn using classification methods of machine learning. Logistic regression models (GLM), decision trees (Decision Trees), random forests (Random Forest), as well as support vector machines (SVM), k-nearest neighbors (k-NN), and naive Bayes algorithm will be constructed for this purpose. The quality of the constructed models will be evaluated using a confusion matrix.
Results. The obtained results revealed high accuracy of the constructed models and their ability to effectively identify bank customers prone to churn. The conclusions of this article may be valuable for developing customer retention strategies not only for commercial banks but also for various business sectors where customer attrition is a relevant issue.

Keywords: customer churn; banking; cost minimization; modeling impact; model effectiveness; customer classification; decision automation; random forest.
Cite as: Kryvytska, O., Kleban, Y., and Yahodka, A. (2024). Customer retention in commercial banking as a classification task in machine learning. *Economic analysis,* 34 (1), 179-190. DOI: https://doi.org/10.35774/econa2024.01.179