

УДК 519.86+330.46

JEL classification: C45, C51, C52, C53

Юрій КЛЕБАН

викладач,
кафедра економіко-математичного
моделювання та інформаційних технологій,
Національний університет «Острозька
академія», Україна
Email: yuriy.kleban@oa.edu.ua
ORCID ID: 0000-0002-7070-5175
Researcher ID: G-0000-2019

Наталія ГОРОШКО

Національний університет «Острозька
академія», Україна
Email: nataliia.horoshko@oa.edu.ua
ORCID ID: 0000-0002-0458-9962

© Юрій Клебан, Наталія Горошко, 2021

Отримано: 05.02.2021 р.
Прорецензовано: 16.02.2021 р.
Рекомендовано до друку: 24.02.2021 р.
Опубліковано: 24.02.2021 р.



Ця стаття розповсюджується на умовах ліцензії Creative Commons Attribution-NonCommercial 4.0, яка дозволяє необмежене повторне використання, розповсюдження та відтворення на будь-якому носії, за умови правильного цитування оригінальної роботи.

Юрій Клебан (Україна)
Наталія Горошко (Україна)

ІДЕНТИФІКАЦІЯ ДЕФОЛТНИХ КЛІЄНТІВ БАНКУ МЕТОДАМИ МАШИННОГО НАВЧАННЯ НА ОСНОВІ БІННІНГУ ПОКАЗНИКІВ

АНОТАЦІЯ

Вступ. В умовах сучасної глобальної кризи проблема якості кредитних портфелів банків є актуальною темою досліджень. Серед методів ефективного управління кредитними ризиками є оцінка кредитоспроможності позичальника. Підвищення якості аналізу сильних та слабких сторін контрагента зменшить виникнення непередбачуваних ризиків у процесі проведення кредитних операцій. З огляду на важливість ролі оцінки кредитоспроможності для прийняття рішень, виникає необхідність покращення та вибору методики, що забезпечить якомога точнішу класифікацію клієнтів банку.

Мета. Метою роботи є вибір найкращого методу прогнозування імовірності дефолту клієнтів комерційного банку на основі аналізу підходів та тестування побудованих моделей.

Метод (методологія). У статті розглянуто методологічні підходи до моделювання неплатоспроможності клієнтів банку та визначення ймовірностей повернення кредитних коштів на основі біннінгу показників. Також у роботі побудовані моделі оцінки кредитного ризику, що ґрунтуються на використанні logit- та probit-регресій, алгоритму екстремального градієнтного бустінгу та штучних нейронних мереж. Проведено порівняльний аналіз ефективності застосування використовуваних підходів.

Результати. Отримані результати продемонстрували високу точність моделей та їх здатність ідентифікувати некредитоспроможних клієнтів. Висновки проведеного дослідження та оцінені математичні підходи можуть бути впроваджені в роботу банківських структур та інших кредитних установ для запобігання збільшенню обсягів проблемних заборгованостей у їхніх кредитних портфелях.

Клебан Ю., Горошко Н. Ідентифікація дефолтних клієнтів банку методами машинного навчання на основі біннінгу показників. *Економічний аналіз*. 2021. Том 31. № 1. С. 133-142.

DOI: <https://doi.org/10.35774/econa2021.01.133>

Ключові слова: математична модель; біннінг; logit-регресія; нейронна мережа; градієнтний бустінг; кредитоспроможність; позичальник.

UDC 519.86+330.46

JEL classification: C45, C51, C52, C53

Yurii KLEBAN

Lecturer,
Department of Economic-Mathematical
Modeling and Information Technologies,
National University of Ostroh Academy, Ukraine
Email: yuriy.kleban@oa.edu.ua
ORCID ID: 0000-0002-7070-5175
Researcher ID: G-0000-2019

Nataliia HOROSHKO

National University of Ostroh Academy, Ukraine
Email: nataliia.horoshko@oa.edu.ua
ORCID ID: 0000-0002-0458-9962

© Yurii Kleban, Nataliia Horoshko, 2021

Received: 05.02.2021

Revised: 16.02.2021

Accepted: 24.02.2021

Online publication date: 24.02.2021



This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 license, which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Yurii Kleban (Ukraine)
Nataliia Horoshko (Ukraine)

IDENTIFICATION OF THE BANK'S DEFAULT CLIENTS BY MACHINE LEARNING METHODS ON THE BASIS OF BINNING

ABSTRACT

Introduction. In the current global crisis, the problem of the quality of banks' loan portfolios is a topical issue. Among the methods of effective credit risk management is the assessment of the borrower's creditworthiness. Improving the quality of analysis of the strengths and weaknesses of the counterparty will reduce the occurrence of unforeseen risks in the process of conducting credit operations. Given the importance of the role of creditworthiness assessment for decision-making, there is a need to improve and choose a methodology that will ensure the most accurate classification of the bank's clients.

Purpose. The aim of the work is to choose the best method for predicting the probability of default of commercial bank customers based on the analysis of approaches and testing of the built models.

Method (methodology). The paper considers methodological approaches to modeling the insolvency of bank customers and determining the probability of repayment of loans based on binning indicators. Also, the credit risk assessment models based on the use of logit and probit regressions, the algorithm of extreme gradient boosting and artificial neural networks are constructed. The comparative analysis of the efficiency of the application of the used approaches is carried out.

Results. The obtained results demonstrated the high accuracy of the models and their ability to identify non-creditworthy customers. The findings of the study and evaluation of mathematical approaches can be implemented in the work of banking structures and other credit institutions to spread the amount of problem fees in their loan portfolios.

Kleban, Y., & Horoshko, N. (2021). Identification of the bank's default clients by machine learning methods on the basis of binning. *Economic analysis*, 31(1), 133-142.

DOI: <https://doi.org/10.35774/econa2021.01.133>

Ключові слова: model; binning; logit regression; neural network; gradient boosting; creditworthiness; individual borrower.

Вступ

У сучасній ринковій економіці кредитні відносини відіграють важливу роль в економічному зростанні країни, забезпечуючи розширення споживчого попиту на ринку товарів та послуг. Однак, надмірність надання банківських позик характеризується підвищенням кредитного ризику для банківських установ. Загроза втратити частину грошових коштів призводить до скорочення прибутків комерційних банків. В нормативному документі Національного банку України [1] цей вид ризиків визначається як «наявний або потенційний ризик для надходжень і капіталу, який виникає через неспроможність сторони, що взяла на себе зобов'язання, виконати умови будь-якої фінансової угоди із банком або в інший спосіб виконати взяті на себе зобов'язання». Тому головним завданням для банків є забезпечення мінімального рівня дефолтів при зростаючому обсязі кредитів.

Мета та завдання статті

Метою роботи є вибір найкращого методу прогнозування ймовірності дефолту клієнтів комерційного банку на основі аналізу підходів та тестування побудованих моделей. Результатом роботи найкращого підходу буде визначення ненадійних клієнтів банку, які не здатні виконати боргові зобов'язання відповідно до умов кредитного договору з мінімальною похибкою класифікації. Основними завданнями є дослідження підходів щодо оцінювання кредитоспроможності позичальників коштів та оцінка ймовірності їх неповернення у випадку надання банком кредиту.

Об'єкт дослідження – кредитні відносини між банківською установою та її клієнтами щодо надання коштів у позику та плати за неї, а предмет – економіко-математичні методи та моделі оцінки спроможності виконати позичальником банківських установ взяті на себе зобов'язання.

Виклад основного матеріалу дослідження

Визначення рівня кредитоспроможності позичальника є першим етапом реалізації кредитних відносин, адже слугує основою для прийняття рішення про надання кредиту. Адекватна діагностика поточного стану позичальника та прогноз можливих напрямків його змін в подальшому дозволяють звести до мінімуму ризик неповернення коштів через виникнення складного фінансового становища позичальника. Провівши інтерпретацію результатів оцінки кредитоспроможності із застосуванням сучасних економіко-математичних методів та моделей, відповідальна особа здатна прийняти раціональне та обґрунтоване рішення стосовно надання або ненадання кредиту потенційному позичальнику.

Проблемам аналізу кредитоспроможності позичальників коштів комерційних банків присвячена значна кількість зарубіжних та вітчизняних наукових праць. Дослідженню кредитної політики банківських установ та її методів займалися В. Вітлінський, В. Вовк, О. Васюренко, О. Пернарівський, С. Мочерний, Е. Гілл, Р. Коттер, М. Гольцберг. Серед актуальних розробок економіко-математичного інструментарію для оцінки кредитоспроможності можна розділити на декілька напрямків: статистичні (лінійні та логістичні регресії, дерева рішень, метод k найближчих сусідів) та нестатистичні методи (нечітка логіка, нейромережі, експертні системи, лінійне та цілочисельне програмування). Огляд таких підходів наведено у працях Г. Великованенко [2], А. Матвійчука [17], О. Новоселецького та О. Якубець [11], Х. Абду [7], С. Аккока [8], Е. Лі [9] та інших. В основі більшості наукових праць покладена скорингова модель, що дозволяє банку оцінити можливість потенційного позичальника вчасно виконати боргові зобов'язання на підставі кількісних показників та якісних його характеристик. З розвитком технологій, їх почали використовувати у поєднанні з методами штучного інтелекту [3,8,9,11,13], що дозволяє оптимізувати та покращити якість процесу аналізу. Однак, в Україні єдиної та чітко розробленої методики оцінки спроможності виконати позичальником взяті на себе зобов'язання немає.

Побудові та використанню економіко-математичної моделі передують розробка її концептуальної схеми. У даному випадку побудова моделі ідентифікації дефолтного клієнта включає ряд кроків, а саме: постановка проблеми; вибір та обґрунтування вхідних факторів моделі; вибір конфігурації моделі та налаштування її параметрів; перевірка моделі на адекватність; експериментальне моделювання; аналіз та інтерпретація результатів.

Оскільки рівень спроможності позичальника виконати взяті зобов'язання обумовлюється як кількісними, так і якісними факторами, то важливо передбачити ефективну обробку цих даних. Один із варіантів є процедура біннінгу (англ. binning), алгоритм якої полягає у розбитті змінних моделі на категорії базуючись на розрахунку показників вагомості ознаки *WOE* й інформаційної значимості *IV*. Таким чином, побудова моделі передбачатиме проведення таких етапів [11]:

- збір даних для дослідження, формування тренувальної та тестової вибірок;
- розбиття значень пояснюючих змінних на категорії за різними алгоритмами біннінгу;
- розрахунок показників *WOE* (англ. *Weight Of Evidence*) та *IV* (англ. *Information Value*) для кожної категорії;

- побудова скорингових моделей на тренувальній вибірці для різних варіантів категоризації вхідних змінних;
- оцінка адекватності побудованих моделей на тестовій вибірці за критерієм AUROC (англ. *area under receiver operating characteristic*);
- аналіз отриманих результатів, формування висновків щодо ефективності алгоритмів біннінгу.

Відповідно до результатів моделювання здатність до виконання взятих зобов'язань фізичної особи може бути:

- високою – ризик невиконання зобов'язань незначний;
- задовільною – ризик невиконання зобов'язань низький;
- нестабільною – ризик невиконання зобов'язань середній;
- незадовільною – ризик невиконання зобов'язань високий;
- критичною – ризик невиконання зобов'язань критичний.

Якщо здатність до виконання зобов'язань є нестабільною, незадовільною чи критичною, банку рекомендовано прийняти рішення про посилення превентивної роботи з такими клієнтами.

Як зазначалось вище, через наявність якісних та кількісних змінних різної розмірності виникає необхідність нормалізації числових значень. Для категоризації факторів розрахуємо показники *WOE* та *IV* за формулами:

$$WOE_i = \ln\left(\frac{B_i}{G_i}\right), i = \overline{1, k} \quad (1)$$

де B_i – відношення кількості ненадійних позичальників у i -й категорії до загального числа ненадійних позичальників у вибірці; G_i – частка надійних угод за i -ю категорією відносно їх загальної кількості; k – кількість категорій змінної.

$$IV = \sum_{i=1}^k (B_i - G_i) \cdot WOE_i \quad (2)$$

Показник інформаційної значимості оцінює ступінь впливу категорії на результуючу змінну. Для побудови та тестування моделей використовувалася кросвалідація з розбиттям на 10 груп [21].

Математична постановка обраної проблеми оцінки фінансово-економічного стану позичальника може зводитись до класичної задачі економетрики, а саме: дослідження зміни значення обраної величини під впливом зміни значень певної множини факторів. Тож, виникає необхідність встановити взаємозв'язок між обраними чинниками впливу та фактом повернення чи неповернення кредиту контрагентом [17].

Для прогнозування поведінки позичальника зручно використати бінарну змінну, що набуватиме лише двох значень: «0» чи «1» (повернення чи неповернення кредиту). Проте, використовуючи звичайну множинну регресію, ми отримаємо недопустимі значення для поставленої задачі, оскільки вони не належатимуть відрізку [0;1]. Таким чином, задача побудови регресії полягає не в передбаченні значень бінарної змінної, а в моделюванні неперервної змінної, що приймає значення з інтервалу [0;1].

Формальна постановка задачі для моделей бінарного вибору (*probit*- та *logit*-моделі) полягає в наступному. Нехай змінна *Status* визначає факт погашення (непогашення) кредитних зобов'язань та набуває значень 0 чи 1 відповідно. Кожен клієнт характеризується m -мірним вектором ознак, що в сукупності утворює масив інформації $X = (x_{ij})$, ($j = \overline{1, m}$ – номер показника; $i = \overline{1, n}$ – номер позичальника). Класичне рівняння лінійної регресії має вигляд:

$$Status_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im} + \varepsilon_i \quad (3)$$

де β_m – шукані коефіцієнти регресії, ε_i – випадкова величина.

Припустимо, що змінна $P_1 = P(Status = 1)$ – ймовірність дефолту позичальника. Тоді рівняння (3) може бути представлено так: $P_1 = P(Status = 1) = X\beta'$.

Як зазначалося вище, лінійна модель ймовірності (3) не дасть потрібних значень залежної змінної на проміжку [0;1]. Для вирішення цієї проблеми використовують *probit*- та *logit*-моделі, в яких припускається, що:

$$P_1 = F(X\beta') \quad (4)$$

де F – функція, область значень якої належить інтервалу [0;1].

В основі функції F найчастіше використовують такі два види функцій розподілу:

- функція стандартного нормального розподілу (для *probit*-моделі):

$$F(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad (5)$$

- функція логістичного розподілу (для *logit*-моделі):

$$F(z) = \frac{e^z}{1 + e^z} \quad (6)$$

Для вибору дослідником кращої моделі ймовірності необхідно оцінити якість кожної з них,

знайшовши коефіцієнти β (4). Найпоширенішим способом оцінки параметрів є метод максимальної правдоподібності.

Варто зазначити, що градієнтний бустинг є одною з потужних технік машинного навчання для вирішення задач класифікації та регресії за допомогою прогнозних моделей. Зокрема, XGBoost – одна з домінуючих та найефективніших реалізацій алгоритму градієнтного бустингу, що базується на дереві рішень та характеризується високою швидкістю обчислень та продуктивністю моделей.

Перейдемо безпосередньо до машинного моделювання на основі методу XGBoost. Даний метод вважається однією з найбільш потужних технік для вирішення задач класифікації та регресії за допомогою прогнозних моделей. Зокрема, XGBoost – одна з домінуючих та найефективніших реалізацій алгоритму градієнтного бустингу, що базується на дереві рішень та характеризується високою швидкістю обчислень та продуктивністю моделей.

Алгоритм за даним методом передбачає формування вхідного потоку даних *Dataset D*, функції втрати *L*, числа ітерацій *M*, швидкості навчання η та кількості закінчених вузлів *T* [13]:

1) Спочатку опишемо ціль алгоритму – мінімізація функції втрати, здійснюючи ініціалізацію:

$$\hat{f}_0(x) = \arg \min_{\theta} \sum_{i=1}^n L(y_i, \theta) \quad (7)$$

2) Знаходимо найшвидші відхилення (ризик) для $m = 1, 2, \dots, M$:

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}^{(m-1)}(x)} \quad (8)$$

3) Визначимо структуру ризику $\{\hat{R}_{jm}\}_{j=1}^T$, відбираючи відхилення, що максимізують коефіцієнт підсилення:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} - \frac{G_{jm}^2}{n_{jm}} \right] \quad (9)$$

4) Визначимо вагу складових елементів $\{\hat{w}_{jm}\}_{j=1}^T$ для вивченої структури:

$$\hat{w}_{jm} = \arg \min_{w_j} \sum_{i \in \hat{I}_{jm}} L(y_i, f^{(m-1)}(x_i)) + w_j. \quad (10)$$

5) Включаємо індикатор швидкості навчання η ($0 < \eta \leq 1$):

$$\hat{f}_m(x) = \eta \sum_{j=1}^T \hat{w}_{jm} I(x_i \in \hat{R}_{jm}) \quad (11)$$

6) Комбінуючи вище зазначене, на кожній ітерації *m* задається «крок»:

$$\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x) \quad (12)$$

7) В результаті отримаємо модель на базі XGBoost:

$$\hat{f}(x) = \hat{f}^{(M)}(x) = \sum_{m=0}^M \hat{f}_m(x) \quad (13)$$

Для визначення ефективності моделі будується ROC-крива, яка оцінює коректність бінарної класифікації. Вісь ординат відображає частку правильно визначених позитивних результатів TPR (True Positive Rate), вісь абсцис – частку помилково діагностованих позитивних результатів FPR (False Positive Rate) при варіюванні порога відсікання [12].

Інколи традиційні методи класифікації об'єктів не завжди можуть точно передати сутність досліджуваного явища, у таких випадках науковці все частіше застосовують методи штучного інтелекту, зокрема, нейромереві технології. Здатність працювати з нелінійними моделями, зашумленими даними та високі адаптивні властивості підтверджують доцільність використання нейронних мереж. У загальному випадку побудова нейронної мережі передбачає такі кроки: вибір початкової конфігурації мережі; навчання мережі та проведення модельних експериментів з оцінкою контрольної помилки; виявлення ефекту перенавчання і коригування структури нейронної мережі; перевірка адекватності навчання із застосуванням тестової вибірки [4].

Задачу ідентифікації дефолтних клієнтів банку вирішуватимемо на базі нейронних мереж з конфігурацією багат шарового перцептрона за допомогою алгоритму зворотного поширення помилки та радіальної базисної функції.

Для оцінки ефективності класифікації використано такі показники: загальна точність класифікації за певного рівня розмежування (*Accuracy*), збалансована точність класифікації, що враховує пропорції позитивних та негативних подій (*Balanced Accuracy*), чутливість (*Sensitivity*), специфічність (*Specificity*) та площа під кривою (*AUROC*) [2020]. Всі розрахунки проводилися в програмному середовищі *R* (див. додатки).

Для побудови математичних моделей було використано набір даних банківської установи з характеристиками клієнтів та інформацією щодо

транзакцій. Масив даних містить 348437 спостережень та 48 факторів, які мають як категоріальний, так і числовий характер. Залежною змінною є бінарний показник *Status*, що вказує на

факт невиконання боргових зобов'язань (значення «1») чи погашення позичальником заборгованості (значення «0»). Визначимо структуру вхідної вибірки:

Таблиця 1. Структура вхідного показника *Status* початкового набору даних для побудови математичних моделей

	Подія « <i>bad</i> »	Подія « <i>good</i> »
Значення бінарної змінної	1	0
Кількість записів, од	40 412	457 353
Частка, %	8,12	91,88

Тренувальна та тестова вибірки розділені у пропорції 70/30, дані яких не повторювалися в

межах моделі. Розподіл даних для нашого дослідження представлений у табл. 2 і табл. 3.

Таблиця 2. Структура вхідного показника *Status* початкового набору даних для побудови математичних моделей у тренувальній вибірці

Найменування	Значення «1»	Значення «0»
Кількість записів, од	28 288	28 288
Частка, %	50	50

Таблиця 3. Структура вхідного показника *Status* початкового набору даних для побудови математичних моделей у тестовій вибірці

Найменування	Значення «1»	Значення «0»
Кількість записів, од	12 124	137 206
Частка, %	8,12	91,88

Окрім того, слід зауважити, що вибірки містили також спостереження, де інформація про позичальників була відсутня. Для забезпечення повноти вхідних масивів даних було використано метод найближчого сусіда (KNNimpute), що забезпечує надійне заповнення пропусків [15].

Здійснюючи попередній аналіз даних для оцінки надійності потенційного позичальника, було виявлено найважливіші чинники для включення в моделі. У таблиці 4 наведено топ-10 факторів за ступенем їх важливості.

З метою визначення найефективнішої моделі ймовірності для подальшого прогнозування, використаємо усі зазначені нами вище методи для

аналізу кредитоспроможності на основі тренувальної вибірки.

У машинному навчанні вимір продуктивності моделі (прогнозної здатності) забезпечує вибір найкращої моделі класифікації, зокрема на основі кривої AUC-ROC (англ. Area Under the Curve of Receiver Operating Characteristic), яка дає змогу перевірити та візуалізувати продуктивність завдання класифікації [19]. Це один з найбільш важливих показників оцінки для перевірки ефективності будь-якої моделі класифікації. Чим вище значення AUC, тим краще модель ідентифікує ненадійного контрагента.

Таблиця 4. Ранги важливості параметрів моделі ідентифікації дефолтних клієнтів для моделі на основі алгоритму *xgBoost*

№	Параметр	Важливість, %
1	Відношення суми витрат до поповнень картки	100,00
2	Вид останнього наданого кредиту	83,26
3	Факт поповнення кредитної картки кожного місяця від 1 грн.	80,87
4	Факт активації кредитного ліміту	65,4
5	Наявність прострочок по попередніх кредитах	58,86
6	Суми на депозитах	54,62
7	Щомісячна заробітна плата	47,48
8	Відношення суми зняття готівки до суми поповнень картки	45,65
9	Середньомісячний приріст суми поповнень	44,58
10	Середньомісячне зростання доходів	43,86

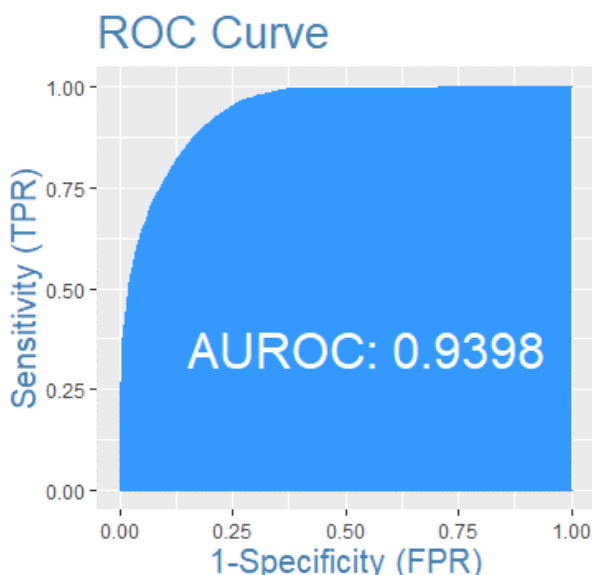


Рис. 1. ROC-крива для тестової вибірки probit-моделі

Зокрема, припускається, що у класифікатора є деякий параметр, варіюючи який ми отримаємо розбиття на два класи. На основі цих порогів або точок відсіку (cut-off value) будується лінія відтинання, що визначає розподіл клієнтів з використанням умов:

- якщо ймовірність непогашення заборгованості більша або рівна значення лінії відтинання, то клієнт вважається «ненадійним» ($\text{predicted_prob} \geq \text{cut_off}$, $p = 1$);

- якщо ж ймовірність невиконання боргових зобов'язань менша лінії відтинання, то клієнт вважається «надійним» ($\text{predicted_prob} < \text{cut_off}$, $p = 0$).

Здійснюючи оцінку загальної точності класифікації, лінію поділу класів (cut-off) встановлено на рівні 0,5, що дало змогу розрахувати наступні показники (табл. 5).

Таблиця 5. Показники ефективності побудованих моделей ідентифікації дефолтного клієнта банку

№	Модель	AUROC	Ginni	Sensitivity	Specificity	Accuracy	Balanced Accuracy
1	Probit	0,9398	0,8796	0,2884	0,9906	0,8099	0,6395
2	Logit	0,9400	0,8800	0,2941	0,9904	0,8152	0,6422
3	xgBoost	0,9464	0,8928	0,3129	0,9911	0,8299	0,6520
4	NeuralNet	0,9115	0,8230	0,2834	0,9901	0,8060	0,6367

Як видно з табл. 5, модель 3, побудована на основі методу градієнтного бустингу xgBoost, має найкращу точність класифікатора. Отримавши найвищі значення показників *Sensitivity* та *Specificity*, вона дозволяє найкраще ідентифікувати як надійних, так і дефолтних клієнтів, що і є метою у вирішенні даної задачі. Саме тому, побудуємо підсумковий розподіл клієнтів із тестової вибірки по персентілям з діапазоном у 10% на основі методу xgBoost (табл. 6).

З таблиці видно, що банку потрібно з обережністю надавати кредити своїм клієнтам: в діапазоні з ймовірністю 0,7-1 (більше 70%) знаходиться 87% від усіх ненадійних позичальників та 57% надійних клієнтів зможуть погасити борг лише з ймовірністю не більше 10%.

Висновки та перспективи подальших розвідок

Провівши аналіз наукових публікацій відомих вчених та ознайомившись із експериментальним підходом до вирішення поставленої задачі, виявлено, що єдиного універсального способу оцінки кредитоспроможності фізичних осіб не існує. В даній статті сформовано три підходи до ідентифікації ненадійного клієнта банку, а саме: ідентифікація за допомогою *logit*- та *probit*-моделей, градієнтного бустингу та нейронної мережі.

Представлені моделі дають змогу не лише отримати відповідь на питання «Чи буде повернуто кредит контрагентом?», а й визначити з якою ймовірністю позичальник здатен виконати чи не виконати зобов'язання, передбачені кредитним договором. Окрім того, банківські установи мають можливість посилювати чи послаблювати умови відбору потенційних позичальників через зміну порогового рівня ризику в моделі.

Результати проведених експериментів засвідчили найвищу точність запропонованого алгоритму градієнтного бустингу, що зараз широко використовується для аналізу великих масивів даних. Перевагами даного методу можна виділити гнучкість та універсальність, оскільки вирішує більшість задач регресії, класифікації та ранжування.

Однак, варто зазначити, що показники ефективності та результати всіх побудованих моделей майже збігаються. Таким чином, зважаючи на високу якість отриманих результатів, запропоновані моделі можуть бути використані для оптимізації прийняття рішень в процесі кредитування клієнтів банківськими установами.

Таблиця 6. Результати класифікації клієнтів методом градієнтного бустингу

Діапазон	Status = 1 кількість Bad клієнтів	Status = 1 частка Bad клієнтів	Status = 0 кількість Good клієнтів	Status = 0 частка Good клієнтів
0.9 - 1	5458	0,45	1632	0,01
0.8 - 0.9	1785	0,15	2681	0,02
0.7 - 0.8	1618	0,13	4885	0,04
0.6 - 0.7	1365	0,11	7277	0,05
0.5 - 0.6	885	0,07	7921	0,06
0.4 - 0.5	531	0,04	8335	0,06
0.3 - 0.4	252	0,02	6588	0,05
0.2 - 0.3	128	0,01	5337	0,04
0.1 - 0.2	59	0	4862	0,04
0 - 0.1	43	0	87688	0,64

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Методичні вказівки з інспектування банків «Система оцінки ризиків»: Постанова Правління Національного банку України від 15.03.2004 № 104. URL: <https://zakon.rada.gov.ua/laws/show/v0104500-04>.
2. Великоіваненко Г. І., Трокоз Л. О. Моделювання кредитоспроможності позичальників комерційного банку. *Наукові записки Національного університету «Острозька академія», серія «Економіка»*. Острог: Вид-во НУ«Острозька академія», 2013. Вип. 22. С. 137-141.
3. Матвійчук А. В. Штучний інтелект в економіці: нейронні мережі, нечітка логіка: монографія. Київ, КНЕУ, 2011. 439 с.
4. Смолева Т. Сучасні методи оцінки кредитоспроможності позичальників банками України. *Фінанси, облік, банкі*. 2014. №1(20). С. 241-245.
5. Thomas L. C. Modelling the credit risk for portfolios of consumer loans: analogies with corporate loan models, *Math. Comput. Simulat.* 79 (8). 2009. 2525–2534.
6. Zenzerović R. Credit scoring models in estimating the creditworthiness of small and medium and big enterprises. *Croatian Operational Research Review*, Vol. 2 No. 1, 2011.
7. Abdou H., Pointon J. Credit scoring, statistical techniques and evaluation criteria: a review of the literature, *Intell. Syst. Account., Finance Manage.* 18 (2–3) (2011) 59–88.
8. Akkoc S. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (AN- FIS) model for credit scoring analysis: the case of Turkish credit card data, *Eur. J. Oper. Res.* 222 (1) (2012) 168–178.
9. Lee T., Chen I. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, *Expert Syst. Appl.* 2005. № 28 (4). 743-752.
10. Crook J., Banasik J. Forecasting and explaining aggregate consumer credit delinquency behaviour, *Int. J. Forecasting* 28 (1) (2012) 145–160.
11. Новоселецький О.М., Якубець О.В. Моделювання кредитоспроможності юридичних осіб на основі дискримінантного аналізу та нейронних мереж. Нейронечіткі технології моделювання в економіці. *Науково-аналітичний журнал*. Київ, 2014. №3. С.120-151
12. Клебан Ю.В. Дослідження способів трансформації даних в контексті підвищення ефективності моделей кредитного скорингу. Нейронечіткі технології моделювання в економіці. *Науково-аналітичний журнал*. Київ, 2019. №8. С.94-123
13. Григорович О.В. Застосування багатосарових перцептронів для класифікації позичальників-юридичних осіб. Нейронечіткі технології моделювання в економіці. *Науково-аналітичний журнал*. Київ, 2019. №8. С.48-64.
14. Nielsen, D. (2016). Tree Boosting with XGBoost. Norwegian University of Science and Technology. URL: https://ntnuopen.ntnu.no/ntnu-mlui/bitstream/handle/11250/2433761/16128_FULLTEXT.pdf?sequence=1&isAllowed=y.

www.econa.org.ua

15. Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman. Missing value estimation methods for DNA microarrays (2001). *Bioinformatics*, Vol. 17 no. 6, P. 520–525, URL: https://www.researchgate.net/publication/220263062_Missing_Value_Estimation_Methods_for_DNA_Microarrays.
16. Stefan Trueck , Svetlozar T.Rachev. Rating Based Modeling of Credit Risk. Academic Press. (2009). 11-30.
17. Матвійчук А.В., Бень В.П. Використання logit- та probit-регресій для оцінки кредитоспроможності позичальника. Фінансовий ринок. Вісник Національного банку України. Київ, 2015. № 5. С.37-41.
18. Вітлінський В. В. Штучний інтелект у системі прийняття управлінських рішень. Нейро-нечіткі технології моделювання в економіці. 2012. №1. С.97-118.
19. Хайкин С. Нейронные сети: полный курс, 2-е издание. М.: Издательский дом «Вильямс», 2006. 1104 с.
20. Fawcett, Tom. An Introduction to ROC Analysis. *Pattern Recognition Letters*. 27 (8) (2011). P. 861–874.
21. Berrar, D. Cross-Validation. Reference Module in Life Sciences. 2018, 1, 542–545. URL: https://www.researchgate.net/publication/324701535_Cross-Validation.

REFERENCES

1. The National Bank of Ukraine. (2004). *Metodychni vkazivky z inspektuvannya bankiv «Systema otsinky ryzykiv»*: Postanova Pravlinnia Natsionalnoho banku Ukrainy vid 15.03.2004 № 104. Retrieved from <https://zakon.rada.gov.ua/laws/show/v0104500-04> [in Ukrainian].
2. Velykoivanenko, G. I., Trokoz, L. O. (2013). *Modelyuvannya kredytopromozhnosti pozychalnykiv komertsiinoho banku* [Modeling of the creditworthiness of commercial bank borrowers]. *Naukovyi zhurnal «Naukovi zapysky Natsionalnoho universytetu«Ostrozka akademiia seriia Ekonomika* [Scientific journal Scientific Notes of the National University of Ostroh Academy, series Economics], 22, 137-141. [in Ukrainian].
3. Matviichuk, A. V. (2011). *Shtuchnyi intelekt v ekonomitsi: neironni merezhi, nechitka lohika*. [Artificial intelligence in economics: neural networks, fuzzy logic: a monograph]. Kyiv, KNEU. [in Ukrainian].
4. Smoleva, T. (2014). *Suchasni metody otsinky kredytopromozhnosti pozychalnykiv bankamy Ukrainy* [Modern methods of assessing the creditworthiness of borrowers by Ukrainian banks.]. *Finansy, oblik, banky* [Finance, accounting, banks], 1(20), 241-245. [in Ukrainian].
5. Thomas, L.C. (2009). Modelling the credit risk for portfolios of consumer loans: analogies with corporate loan models. *Math. Comput. Simulat*, 79 (8), 2525–2534.
6. Zenzerović, R. (2011). Credit scoring models in estimating the creditworthiness of small and medium and big enterprises. *Croatian Operational Research Review*, 2(1).
7. Abdou, H., Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intell. Syst. Account., Finance Manage*, 18 (2-3), 59-88.
8. Akkoc, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (AN-FIS) model for credit scoring analysis: the case of Turkish credit card data, *Eur. J. Oper. Res.*, 222 (1), 168-178.
9. Lee, T., Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, *Expert Syst. Appl.*, 28 (4), 743–752.
10. Crook, J., Banasik, J. (2012). Forecasting and explaining aggregate consumer credit delinquency behaviour, *Int. J. Forecasting*, 28 (1), 145–160.
11. Novoseletskyy, O.M., Yakubets O.V. (2014) *Modeliuvannya kredytopromozhnosti yurydychnykh osib na osovi dyskryminantnoho analisu ta neironnykh merezh* [Legal entities creditworthiness modeling using discriminant analysis and neural networks]. *Nejronechitki tekhnologii modelyuvannya v ekonomitsi. Naukovo-analitychnyi zhurnal* [Neuroreflective modeling technologies in economics. Scientific and analytical journal], 3, 120-151. [in Ukrainian].
12. Kleban, Y. V. (2019). *Doslidzhennia sposobiv transformatsii danykh v konteksti pidvyshchennia efektyvnosti modeley kredytnoho skorynhu* [Research of ways of data transformation in the context of increasing the efficiency of credit scoring models]. *Nejronechitki tekhnologii modelyuvannya v ekonomitsi. Naukovo-analitychnyi zhurnal* [Neuroreflective modeling technologies in economics. Scientific and analytical journal], 8, 94-123. [in Ukrainian].
13. Hryhorovych, O.V. (2019). *Zastosuvannya bagatosharovysh perceptroniv dlya klasyfikatsii posychalnykiv-yurydychnykh osib* [Application of multilayer perceptrons to legal entities borrowers classification]. *Nejronechitki tekhnologii modelyuvannya v ekonomitsi. Naukovo-analitychnyi zhurnal* [Neuroreflective modeling technologies in economics. Scientific and analytical journal], 8, 48-64. [in Ukrainian].
14. Nielsen, D. (2016). *Tree Boosting with XGBoost*. Norwegian University of Science and Technology. Retrieved from <https://ntnuopen.ntnu.no/ntnu-xmloi/handle/11250/2433761> . [in English].
15. Troyanskaya, O., Cantor, M., Sherlock and otherth (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17 (6), 520-525.
16. Trueck, S., Rachev, S.T. (2009). *Rating Based Modeling of Credit Risk*. Academic Press, 11-30.
17. Matviichuk, A. V., Ben, V.P. (2015). The use of logit and probit regressions in borrower's creditworthiness assessment. *The Visnyk of the National Bank of Ukraine*, 5, 37-41.

www.econa.org.ua

18. Vitlinskii, V.V. (2012) Shtuchnyi intellect u systemi pryiniattia upravlinskykh rishen [Artificial intelligence in the management decision-making system]. *Neyronechitki tekhnologii modelyuvannia v ekonomitsi. Naukovo-analitychnyi zhurnal* [Neuroreflective modeling technologies in economics. Scientific and analytical journal], 1, 97-118.
19. Haykin, S. (2006). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Moscow: Williams.
20. Fawcett, T. (2011). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27 (8), 861–874.
21. Berrar, D. (2018). Cross-Validation. Reference Module in Life Sciences. Retrieved from https://www.researchgate.net/publication/324701535_Cross-Validation.