СЕКЦІЯ 4
МАТЕМАТИЧНІ МЕТОДИ, МОДЕЛІ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В ЕКОНОМІЦІ

DOI: https://doi.org/10.32999/ksu2307-8030/2025-54-9

УДК 519.6:336.74

Клебан Ю.В.
старший викладач кафедри інформаційних технологій та аналітики даних
Національного університету «Острозька академія»
ORCID: https://orcid.org/0000-0002-7070-5175
E-mail: yuriy.kleban@oa.edu.ua

Мельниченко Г.М.
студентка
Національного університету «Острозька академія»
ORCID: https://orcid.org/0009-0005-7513-7356
E-mail: halyna.melnychenko@oa.edu.ua

ВПЛИВ МЕТОДІВ БАЛАНСУВАННЯ ДАНИХ НА ЯКІСТЬ ТА ЕКОНОМІЧНУ ЕФЕКТИВНІСТЬ КЛАСИФІКАЦІЇ БАНКІВСЬКИХ КЛІЄНТІВ

Стаття досліджує вплив методів балансування даних (Oversampling, Undersampling, Over&Under, ROSE, SMOTE) на економічну ефективність класифікації ненадійних клієнтів у фінансовому секторі. Основна увага зосереджена на визначенні оптимального порогу класифікації, що максимізує прибуток банку і не шкодить якості прогнозів. Дані зібрані з 700 клієнтів із відсотком ненадійних у 26,3%. Результати показують, що SMOTE демонструє найкращі економічні результати, незважаючи на певні помилки у класифікації. Зокрема, модель SMOTE забезпечує високі показники F1-Score, F2-Score та F-Measure, підвищуючи точність прогнозування кредитних ризиків. Використання ROSE виявилось менш ефективним. Таким чином, застосування збалансованих моделей може бути вигідним для покращення фінансових результатів та зменшення втрат у банківській сфері.

Ключові слова: балансування даних, кредитний ризик, класифікація клієнтів, економічна ефективність, SMOTE, ROSE, оптимізація cutoff-порогів.

Kleban Yurii, Melnychenko Halyna. THE IMPACT OF DATA BALANCING METHODS ON THE QUALITY AND COST-EFFECTIVENESS OF BANK CUSTOMER CLASSIFICATION

The article explores the effectiveness of data balancing methods—Oversampling, Undersampling, Over&Under, ROSE, and SMOTE—in improving the classification of unreliable clients in the financial sector. The primary focus is on determining the optimal classification threshold that maximizes bank profitability while maintaining predictive accuracy. The study uses a dataset of 700 clients, with 26.3% identified as unreliable. Logistic regression models were constructed to evaluate the performance of each balancing technique. The findings indicate that the SMOTE technique exhibits the most substantial economic impact, yielding favorable outcomes in terms of both F1-Score and economic returns. Despite some misclassifications, SMOTE effectively improves the identification of high-risk clients, showcasing enhanced sensitivity and specificity. Conversely, the application of the ROSE method proved to be less effective among the techniques studied, demonstrating lower classification quality and economic efficiency. This suggests that the choice of balancing method significantly influences the financial outcomes and forecast accuracy in banking environments. Additionally, the study emphasizes the necessity of balancing data to resolve class imbalance issues, which is a common challenge in real-world financial applications such as credit risk prediction. The analysis reveals that models benefiting from balanced datasets outperform unbalanced ones, demonstrating improved accuracy and reduced bias towards the dominant class. Overall, the research highlights the importance of selecting appropriate data balancing strategies as a critical factor in enhancing the predictive capacity and economic efficiency of classification models in the banking sector. These findings underscore the potential of employing sophisticated machine learning techniques to minimize credit risk and optimize profitability, providing banks with a strategic advantage in risk management and decision-making processes. The study suggests that adopting advanced data balancing methods like SMOTE can significantly contribute to better financial results and risk mitigation in credit lending practices.

Key words: data balancing, credit risk, customer classification, economic efficiency, SMOTE, ROSE, cutoff threshold optimization.

Постановка проблеми. Однією з основних проблем застосування методів машинного навчання для ідентифікації ненадійних клієнтів у фінансовому секторі є дисбаланс класів у вибірці. Переважання надійних клієнтів над ненадійними зумовлює упередженість алгоритмів на користь більшості, що негативно впливає на точність виявлення клієнтів із підвищеним кредитним ризиком. Така ситуація може призводити до значних фінансових втрат для банківських установ. Для розв'язання цієї проблеми використовуються методи балансування даних, зокрема Oversampling, Undersampling, SMOTE тощо. Однак їх ефективність залежить від особливостей даних та специфіки досліджуваної задачі. Відтак, постає необхідність комплексного аналізу впливу цих методів не лише на якість класифікації, а й на економічну ефективність побудованих моделей, що є важливим аспектом у контексті прийняття фінансових рішень.

Аналіз останніх досліджень і публікацій. Проблема ідентифікації ненадійних клієнтів у фінансовому секторі активно досліджується в останні роки, особливо в контексті застосування методів машинного навчання для оцінки кредитного ризику. Різні дослідження акцентують увагу на важливості вирішення проблеми дисбалансу класів у фінансових даних.

Ahmed Almustfa Hussin Adam Khatir та Marco Bee [1, с. 169] застосовують різні методи машинного навчання для прогнозування кредитних ризиків, використовуючи SMOTE та SMOTEtomek для балансування даних. Їхнє дослідження демонструє, що балансування даних значно покращує точність моделей при класифікації кредитних ризиків.

Migraz Enes Furkan MİLLİ та ін. [2, с. 55] аналізують вплив методів балансування класів на продуктивність технік машинного навчання при оцінці кредитних ризиків, використовуючи дані з Німеччини, Австралії та HMEQ. Результати показали, що гібридні методи, які поєднують передискретизацію та недодискретизацію, демонструють найкращу продуктивність.

У дослідженні Eslam H. S. та ін. [3, с. 1690–1711] використовувались методи машинного та глибокого навчання для прогнозування кредитних ризиків у банківській сфері з застосуванням ансамблевих методів та SMOTE. Дослідження виявило, що DenseNet і ResNet є найбільш точними моделями для прогнозування кредитних ризиків.

Chenyu Yang та ін. [4, с. 30] розглядають різні стратегії балансування вибірки та комбіновані підходи, такі як SMOTE-Tomek Links. Автори дійшли висновку, що гібридні методи, які комбінують oversampling та undersampling, демонструють найкращі результати для класифікації рідкісних подій, таких як дефолти у кредитному ризику.

На основі аналізу останніх досліджень можна зробити висновок, що для коректної оцінки кредитних ризиків необхідно порівняти різні методи балансування вибірки, щоб вибрати найефективніший підхід залежно від характеру даних. Варто зазначити, що більшість наявних досліджень концентрується переважно на ефективності моделей з точки зору статистичних метрик (accuracy, precision, recall, F1-score), приділяючи недостатньо уваги економічному ефекту від їх використання в реальних бізнес-сценаріях. Саме тому існує потреба в дослідженні, яке б оцінювало методи балансування даних не лише за технічними показниками, але й за їхнім впливом на фінансові результати банківських установ.

Метою статті є оцінка ефективності методів балансування даних (Oversampling, Undersampling, Over&Under, ROSE, SMOTE) для класифікації ненадійних клієнтів у фінансовій сфері з урахуванням як точності прогнозів, так і економічного ефекту для банківських установ.

Виклад матеріалу дослідження та його основні результати. У фінансовому секторі критичним завданням є мінімізація кредитних ризиків через точну оцінку надійності клієнтів. Проблема дисбалансу класів виникає, коли кількість надійних клієнтів значно перевищує кількість ненадійних, що призводить до упередженості алгоритмів класифікації на користь більшості та знижує точність виявлення ризикових клієнтів.

Ненадійні клієнти – це особи або організації з високою ймовірністю невиконання фінансових зобов'язань. Їх ідентифікація є важливою частиною управління ризиками, оскільки допомагає знизити фінансові втрати банків.

Незбалансованість даних ускладнює навчання моделей машинного навчання, оскільки алгоритми схильні орієнтуватися на більшість (надійних клієнтів) та ігнорувати меншість (ненадійних клієнтів), що знижує ефективність класифікації високоризикових випадків.

Теоретичною основою дослідження є теорія кредитного ризику та логістична регресія, яка широко застосовується для класифікації та прогнозування ймовірності подій з бінарними значеннями [5]. Ця модель використовується для прогнозування ймовірності повернення кредиту позичальником. Формула логістичної регресії виглядає наступним чином [5]:

$$P(Y=1 \mid X)=\frac{1}{1+e^{-(b0+b1X1+\ldots+bnXn)}}, \qquad (1)$$

Де $P(Y=1|X)$ – ймовірність того, що об'єкт належить до класу 1 (наприклад, «надійний клієнт» або «кредит буде повернутий»), *b0* – вільний коефіцієнт (константа), *b1,b2,...,bn* – коефіцієнти моделі, які визначають вплив кожної змінної *X1,X2,...,Xn* на результат, *X1,X2,...,Xn* – незалежні змінні, які впливають на ймовірність результату (наприклад, фінансові характеристики клієнта).

Ця формула дозволяє обчислити ймовірність того, що подія (наприклад, неповернення кредиту) відбудеться на основі вхідних характеристик (X).

Для побудови моделі передбачено виконання наступних етапів:

1. Збір та підготовка даних про клієнтів.
2. Створення базової моделі логістичної регресії на незбалансованих даних.
3. Застосування методів балансування (Oversampling, Undersampling, Over&Under, ROSE, SMOTE) та побудова відповідних моделей.
4. Визначення cutoff-лінії для класифікації клієнтів з урахуванням моделювання та економічного ефекту.
5. Оцінка моделей за допомогою Confusion Matrix та метрик (F1-Score, F2-Score, F-Measure).
6. Порівняння результатів та формулювання висновків щодо ефективності методів балансування.

У дослідженні використано набір даних про клієнтів банку, що складається з 700 спостережень та 9 змінних [6]. Незалежні змінні включають демографічні та фінансові характеристики клієнтів:

- Вік: переважно 25–45 років, найактивніша група 27–30 років.
- Стаж роботи: до 15 років, значна частка без досвіду.
- Річний дохід: переважно до 120 тис. доларів.
- Борги по кредитній картці: до 5 тис. доларів, багато без боргів.

Залежна змінна "default" – бінарний показник повернення кредиту (0 – повернено, 1 – не повернено).

Набір даних є незбалансованим, так як у вибірці значно переважає клас надійних клієнтів (73,7%), тоді як клас 1 (ненадійні) зустрічається значно рідше (26,3%).

На наступному етапі спостереження розділено на тренувальну та тестову вибірки у пропорції 70/30 зі збереженням дисбалансу класів у початковому наборі даних.

Незбалансованість у тренувальній вибірці може призвести до того, що модель навчання буде схильна передбачати лише найпоширеніший клас, ігноруючи рідкісний клас, що негативно впливає на її продуктивність. У результаті, хоча загальна точність (accuracy) моделі може бути високою, її здатність правильно класифікувати менш представлений клас буде низькою. Для вирішення цієї проблеми використовуються різні методи балансування, ефективність яких порівнюється у даному дослідженні [7, с. 87]. Балансування даних є важливим для покращення результатів моделей машинного навчання, особливо в задачах зі значним дисбалансом класів. У фінансових задачах, таких як прогнозування кредитних ризиків, де переважає один клас (надійні клієнти) над іншим (ненадійні клієнти), методи балансування даних можуть поліпшити продуктивність моделей.

Одним із методів балансування є Oversampling, який збільшує кількість зразків меншості, покращуючи здатність моделі навчатися на цих даних. Undersampling зменшує кількість зразків більшості класу, зменшуючи його домінування, але може призвести до втрати важливої інформації [8]. Метод Over&Under-sampling комбінує створення нових синтетичних зразків меншості та зменшення кількості зразків більшості, покращуючи баланс між класами. Алгоритм ROSE (Random OverSampling Examples) генерує нові зразки меншості на основі статистичних властивостей оригінальних даних [9]. Метод SMOTE (Synthetic Minority Oversampling Technique) створює нові приклади для меншості класу, знаходячи середину між двома схожими прикладами [10].

Після балансування за допомогою вище згаданих методів розподіл класів тренувальної вибірки має вигляд табл. 1.

Після побудови логістичних моделей на незбалансованих та збалансованих різними методами даних перевіряємо точність прогнозування на тестовій вибірці. Однак, для

Таблиця 1

Структура змінної "default" у тренувальній вибірці до та після балансування

Назва методу балансування	Без баланс.		Oversampling		Undersampling		Over&Under		ROSE		SMOTE	
Значення змінної "default"	**0**	**1**	0	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Кількість спостережень, од.	**362**	**129**	362	**362**	**129**	**129**	**342**	**348**	**342**	**366**	**517**	**549**
Частка, %	**73,6**	**26,4**	50	**50**	**50**	**50**	**50,43**	**49,67**	**51,7**	**48,3**	**51,5**	**48,5**

Джерело: створено авторами

класифікації клієнтів недостатньо оцінювати точність моделі за стандартними критеріями. Зокрема, для визначення класу клієнта та оцінки точності використовується cutoff-лінія.

Cutoff (threshold) – це поріг ймовірності, який використовується для прийняття рішення про класифікацію у задачах бінарної класифікації. У логістичній регресії модель прогнозує ймовірність належності спостереження до певного класу. Традиційні методи класифікації орієнтуються на вибір порогу для підвищення якості моделі. У даному дослідженні оптимальний cutoff визначається шляхом аналізу фінансових втрат та вигод, щоб максимізувати прибутковість банку, а не тільки покращити точність моделі [11, с. 505].

Для усіх моделей було обраховано два значення оптимального cutoff для тренувальної вибірки: з точки зору моделювання (рис. 1) та економічної ефективності (рис. 3).

Для оцінки якості моделей з точки зору моделювання одним з основних інструментів є матриця помилок (Confusion Matrix), яка може детально проаналізувати результати роботи моделі, показуючи, скільки елементів було класифіковано правильно для кожного класу, а скільки ні. У матриці є такі значення:

– *True Positives* (TP) – кількість правильних позитивних передбачень;

– *True Negatives* (TN) – кількість правильних негативних передбачень;

– *False Positives* (FP) – кількість неправильних позитивних передбачень;

– *False Negatives* (FN) – кількість неправильних негативних передбачень [12].

Confusion Matrix для незбалансованої моделі виглядає табл. 2.

Ця матриця показує, що модель робить помилки при класифікації позитивних випадків, оскільки маємо 35 помилкових негативи та 19 правильних позитиви. Однак модель добре справляється з класифікацією негативних випадків, оскільки кількість правильних негативів висока (148), і кількість помилкових позитивів низька (1).

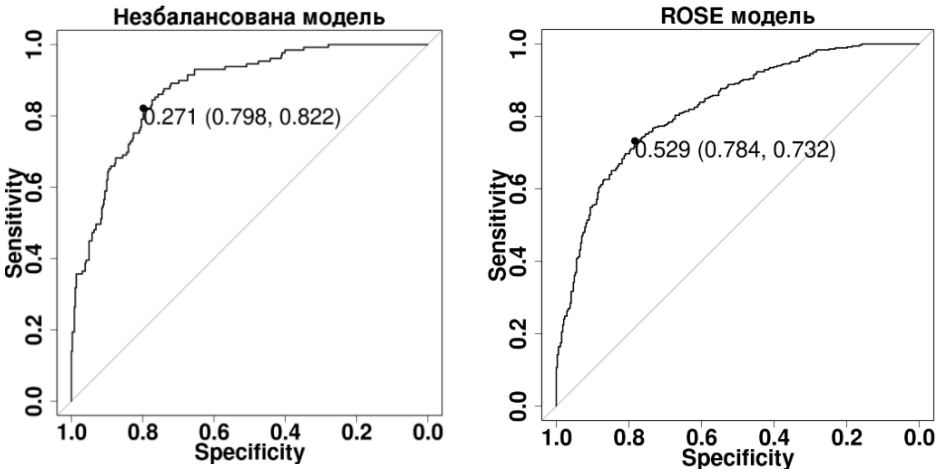

Рис. 1. Графіки ROC-кривих для моделей з незбалансованими даними та даними з балансуванням на основі ROSE з оптимальними значеннями cutoff для моделей

Джерело: побудовано авторами

Таблиця 2

Матриця помилок для незбалансованої моделі

	Negative Prediction	Positive Prediction
Negative Class	148	1
Positive Class	35	19

Джерело: побудовано авторами

Розглянемо також матрицю помилок для найкращої із збалансованих моделей:

Таблиця 3

Матриця помилок для SMOTE моделі

	Negative Prediction	Positive Prediction
Negative Class	139	16
Positive Class	21	33

Джерело: побудовано авторами

Серед збалансованих моделей найкращою є модель SMOTE, оскільки вона має найбільшу кількість правильно класифікованих ненадійних клієнтів (TP = 33) та найменшу кількість FN (21). Найгірше з класифікацією позитивного класу справляється модель ROSE.

Враховуючи значення матриці помилок можна обчислити такі метрики, коли є дисбаланс між класами, як F1-score, F2-score та F-Measure. Вони поєднують чутливість (recall) і точність (precision), щоб дати більш повну оцінку ефективності моделі.

F1-Score використовують, коли важливий баланс між точністю та чутливістю. Він знаходиться в межах від 0 до 1, де 1 означає ідеальну модель. Формула F1-Score має вигляд:

$$F1 - Score = \frac{2 * Precision * Recall}{(Precision + Recall)}. \quad (2)$$

Precision (скільки з передбачених позитивних класів є правильними) та Recall (скільки з реальних позитивних класів було правильно класифіковано) у свою чергу обчислюються за формулами 5–6:

$$Precision = \frac{TP}{(TP + FP)}. \quad (3)$$

$$Recall = \frac{TP}{(TP + FN)}. \quad (4)$$

F2-Score — це модифікація F1-Score, яка більше акцентує на чутливості (recall) та розраховується за формулою:

$$F2 - Score = \frac{5 * Precision * Recall}{4 * Precision + Recall}. \quad (5)$$

F-Measure є загальним терміном для метрик, що поєднують точність і чутливість, де β визначає вагу Recall відносно Precision. Формула має вигляд:

$$F - Measure = \frac{(1 + \beta^2) * Precision * Recall}{\beta^2 * Precision + Recall}. \quad (6)$$

У випадках, коли важливіша чутливість, β може бути більшим, а коли точність – меншим [13].

Зважаючи на результати оцінки моделей (рис. 2), найкращою для класифікації позитивного класу (ненадійних клієнтів) є SMOTE модель, оскільки вона має найвищі значення F1-Score (0.64), F2-Score (0.62) та F-Measure (0.63). Це означає, що дана модель демонструє найкращий баланс між чутливістю і точністю, забезпечуючи високу точність при виявленні клієнтів, які можуть не повернути кредит.

Найгіршою серед збалансованих моделей є ROSE модель, яка має найнижчі показники

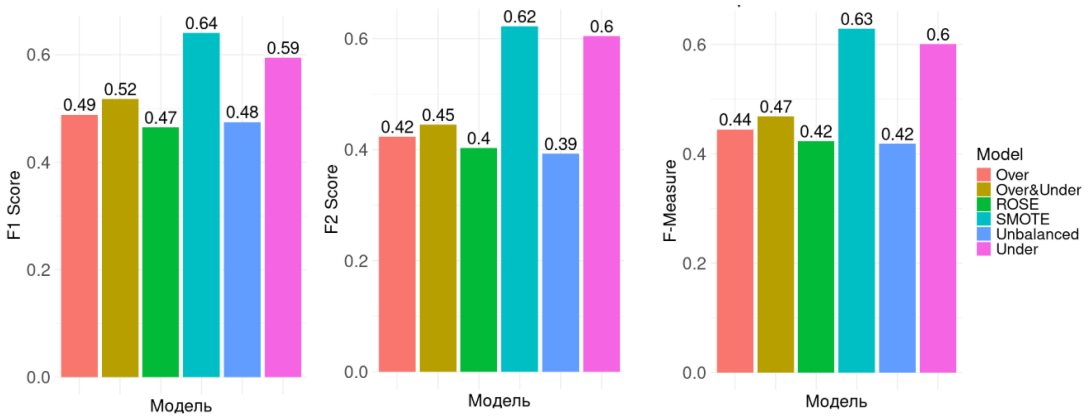

Рис. 2. Графіки порівняння моделей за метриками F1-score, F2-score та F-Measure

Джерело: побудовано авторами

у всіх трьох метриках, що свідчить про її погану здатність точно класифікувати позитивний клас.

З точки зору економічної ефективності загальні втрати банку можна оцінити за формулою:

$$L=P+I+C_{oper}+C_{legal}-C_{res}, \quad (7)$$

де P – це основна сума кредиту, I – втрачений прибуток, тобто упущені відсотки, які банк не отримує, C_{oper} – операційні витрати, витрати на видачу кредиту, такі як аналіз позичальника, обробка заявки тощо, C_{res} резервування, коли банки резервують частину коштів на покриття ризиків неповернення, Cl_{egal} – юридичні витрати, які виникають, якщо банк намагається стягнути борг через суд, Прибуток від поверненого кредиту можна обрахувати за формулою:

$$G=I+C_{fee}-C_{oper}, \quad (8)$$

де I – відсотки, це основний прибуток, який банк отримує від наданого кредиту, $Cfee$ – комісії, плата, яку банк стягує за оформлення кредиту чи за дострокове погашення [14, с. 50].

Обчислення оптимального порогу класифікації для економічної ефективності здійснюється за такими умовами: якщо банк видає кредит клієнту у сумі 10000 гривень на 1 рік під 30% річних, то у разі повернення клієнтом боргу банк отримає прибуток:

I=1000*0.3-500=2500 грн.

а якщо ж клієнт не повертає кредит, то збиток банку складе:

L=10000+300+500-5000=8500 грн.

З точки зору математичної моделі на незбалансованих даних cutoff становить 0.27, що може покращити якість моделі, але не враховує реальні фінансові наслідки неправильних рішень.

Однак при cutoff 0.68 кількість помилкових рішень значно скорочується (з 73 до 5 клієнтів), що дозволить банку мінімізувати збитки від неповернення кредитів і збільшити чистий прибуток. Це показує важливість адаптації математичних моделей до бізнес-реалій.

Проаналізувавши табл. 4 і табл. 5, можна сказати, що у випадку визначення cutoff на основі моделювання більшість моделей демонструють значні фінансові втрати, що сягають 355500 грн у незбалансованій моделі. Лише модель ROSE (38500 грн) та Under (26500 грн) показали позитивний фінансовий результат, тоді як всі інші моделі мають збитки. Це свідчить про те, що математично обґрунтоване значення оптимального порогу не завжди враховує реальні фінансові наслідки. Також, хибно класифікованих надійних клієнтів залишається досить багато (від 31 до 142 осіб), що може призводити до потенційної втрати прибутку банком.

У табл. 5, де cutoff визначався з точки зору економічної ефективності, ситуація суттєво покращується. Усі моделі на тренувальних даних демонструють прибутки, а не збитки, що є головним показником їхньої ефективності в реальному бізнес-середовищі. Найкращі результати отримані для моделі SMOTE (394000 грн), Over&Under (292500 грн) та Over (253000 грн). Навіть

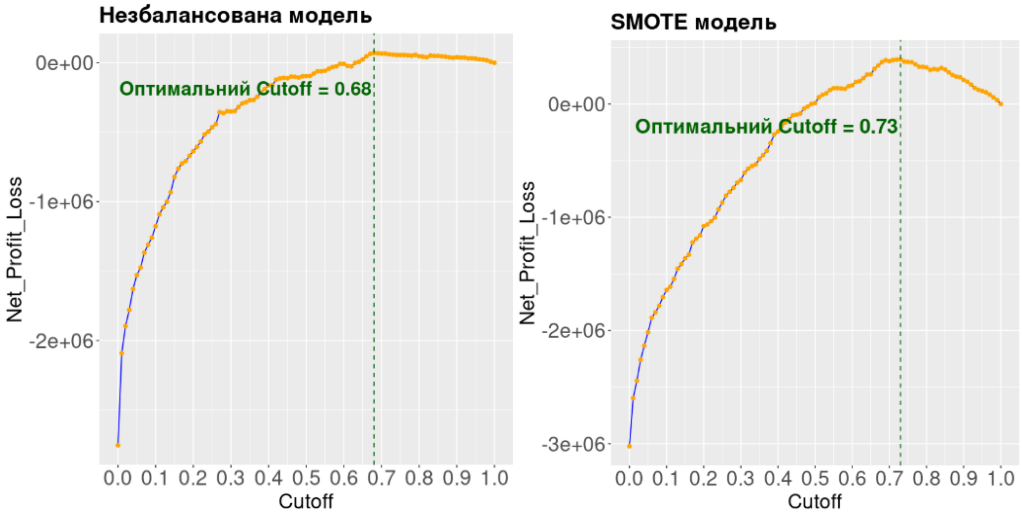

Рис. 3. Графіки залежності Cutoff і прибутку/збитку банку для моделей на незбалансованих даних, даних балансованих зі SMOTE та оптимальні значення Cutoff

Джерело: побудовано авторами

Таблиця 4

Аналіз прибутку/збитку на основі точності класифікації математичної моделі для різних підходів балансування даних на тренувальній вибірці

№	Спосіб балансування даних	Оптимальний cutoff	Прибуток/ збиток, грн	Хибно класифіковані надійні клієнти, ос.
1	Незбалансована модель	0,27	-355500	73
2	Over	0,44	-2500	95
3	**Under**	0,43	26500	31
4	**Over&Under**	0,43	-105500	103
5	ROSE	0,53	38500	74
6	SMOTE	0,48	-22000	142

Джерело: побудовано авторами

Таблиця 5

Аналіз прибутку/збитку на основі економічної ефективності для різних підходів балансування даних на тренувальній вибірці

№	Модель	Оптимальний cutoff	**Прибуток/ збиток, грн**	Хибно класифіковані надійні клієнти, ос.
1	Незбалансована модель	0,68	70000	5
2	Over	0,81	253000	17
3	**Under**	0,71	142000	8
4	**Over&Under**	0,81	292500	10
5	ROSE	0,71	223000	32
6	SMOTE	0,73	394000	46

Джерело: побудовано авторами

найгірший фінансовий результат (70000 грн у незбалансованій моделі) є значно кращим, ніж у табл. 2. Також зменшилася кількість хибно класифікованих надійних клієнтів. У незбалансованій моделі їх кількість зменшилася з 73 до 5 осіб, що значно знижує ризик втрати хороших позичальників. Найкращі результати на тренувальній вибірці в цій категорії показала модель Under (8 осіб), тоді як у SMOTE їх залишилося найбільше (46 осіб), проте це компенсується значним прибутком.

Після застосування оптимального cutoff з точки зору економічного ефекту на тестових даних (табл. 6), Under модель показала найбільший прибуток – 286500 грн, але кількість хибно класифікованих позитивних випадків (False Negatives) в цій моделі була найбільшою (24), що вказує на зниження її ефективності у прогнозуванні проблемних клієнтів у реальних умовах.

Хоча SMOTE модель не продемонструвала найвищого прибутку, вона є однією з найкращих моделей з точки зору балансу між економічним результатом і кількістю неправильно ідентифікованих «1». Таким чином, вона може бути вибрана як хороша альтернатива, коли потрібно забезпечити хороший економічний результат при допустимому рівні похибок.

Таблиця 6

Аналіз прибутку/збитку на основі економічної ефективності для різних підходів балансування даних на тестовій вибірці

№	Модель	Оптимальний cutoff	Прибуток/ збиток, грн	Хибно класифіковані надійні клієнти, ос.	**F1-Score**
1	Незбалансована модель	0,68	107000	7	0,48
2	Over	0,81	146000	11	0,49
3	**Under**	0,71	**286500**	**24**	0,52
4	**Over&Under**	0,81	131500	9	0,59
5	ROSE	0,71	152000	12	0,47
6	SMOTE	0,73	218500	16	**0,64**

Джерело: побудовано авторами

Найгіршою моделлю є незбалансована модель, яка на тестових даних показала найменший прибуток у 107000 грн. Хоча кількість хибно класифікованих надійних клієнтів була низькою (7 осіб), прибуток був значно нижчим у порівнянні з іншими моделями, що вказує на те, що ця модель не була ефективною з економічної точки зору в прогнозах.

Висновки. У результаті аналізу різних моделей класифікації з урахуванням економічної ефективності для прогнозування кредитного ризику було визначено, що моделі, які використовують техніки балансування вибірки, показують кращі результати порівняно з незбалансованою моделлю. Зокрема, SMOTE модель продемонструвала найкращу ефективність у класифікації ненадійних клієнтів, маючи високі значення показників оцінки моделей. Що стосується економічної ефективності, SMOTE модель також показала кращі результати, оскільки її застосування призводить до прибутку у сумі 218500 грн. Це свідчить про те, що хоча вартість помилкових позитивів та негативів у цій моделі дещо вища, вона дозволяє отримати максимальний прибуток завдяки кращій класифікації ненадійних клієнтів. Це робить її найбільш відповідною для бізнесових цілей, оскільки допомагає ефективно виявляти високий ризик неповернення кредиту.

Натомість, ROSE модель виявилася найгіршою серед збалансованих методів, показуючи низькі показники якості класифікації та економічної ефективності. Її використання призводить до менш точних прогнозів і, як наслідок, до меншого прибутку для банку.

Таким чином, для ефективного управління кредитним ризиком доцільно обирати моделі, які зберігають баланс між високою чутливістю та точністю, зокрема SMOTE модель, а також враховувати економічні показники, що дозволяють максимізувати прибуток і знизити збитки.

Результати досліджень показують, що методи машинного навчання для класифікації ненадійних клієнтів надають банкам значні переваги, зокрема можливість покращити точність прогнозування за допомогою балансування даних у випадку дисбалансу класів. Це підтверджує те, що балансування може бути корисним для уникнення помилок класифікації та підвищення економічної ефективності моделі.

У підсумку, використання машинного навчання дозволяє банкам підвищити ефективність, знизити витрати та збільшити прибутковість, а балансування даних залишається важливим інструментом для точнішої класифікації, особливо в умовах дисбалансу.

БІБЛІОГРАФІЧНИЙ СПИСОК:

1. Ahmed Almustfa Hussin Adam Khatir and Marco Bee. Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination? *Risks.* 2022, 10, p. 169–190.
2. Migraç Enes Furkan MİLLİ, İpek DEVECİ KOCAKOÇ, Serkan ARAS. Investigating the Effect of Class Balancing Methods on the Performance of Machine Learning Techniques: Credit Risk Application. *Izmir Journal of Management,* 2024, 5, p. 55–69.
3. Haque, F. M. A., & Hassan, Md. M. (2024). Bank Loan Prediction Using Machine Learning Techniques. *American Journal of Industrial and Business Management*, 14, 1690–1711. DOI: https://doi.org/10.4236/ajibm.2024.1412085
4. Chenyu Yang, Yanjie Dong, Jiachen Lu, Zherui Peng. Solving Imbalanced Data in Credit Risk Prediction: A Comparison of Resampling Strategies for Different Machine Learning Classification Algorithms, Taking Threshold Tuning into Account. *MLMI 2022: 2022 5th International Conference on Machine Learning and Machine Intelligence*. p. 30–40.
5. Logistic Regression in Machine Learning. URL: https://www.geeksforgeeks.org/understanding-logistic-regression/
6. Zolghadr Z. Bank Loan / Credit Scoring for Bank Customers. URL:https://www.kaggle.com/datasets/zahrazolghadr/bank-loan-cleaned-ver1/data
7. Гавриленко С. Ю., Зозуля В. Д., Омельченко В. В. Дослідження методів підвищення якості класифікації на незбалансованих даних. *Системи управління, навігації та зв'язку.* 2023. № 2. С. 87–91.
8. Undersampling, Oversampling and SMOTE, Ensemble Method and Cost Sensitive Learning techniques for dealing with Imbalanced Data. *Medium.* URL: https://medium.com/@abhishekjainindore24/undersampling-oversampling-and-smote-ensemble-mehtod-and-cost-sensitive-learning-techniques-for-08efb557ec68
9. ROSE: Generation of synthetic data by Randomly Over Sampling… URL: https://rdrr.io/cran/ROSE/man/ROSE.html
10. Overcoming Class Imbalance with SMOTE. *Train in Data.* URL: https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote/
11. Verbraken, T., Bravo, C., Weber, R., Baesens, B.. Development and Application of Consumer Credit Scoring Models Using Profit-Based Classification Measures. *European Journal of Operational Research*. 2014, 238(2), p. 505–513.
12. What is a confusion matrix? Jacob Murel Ph.D. IBM/2024. URL: https://www.ibm.com/think/topics/confusion-matrix
13. Jason Brownlee. Tour of Evaluation Metrics for Imbalanced Classification. URL: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
14. Мостовенко Н. А., Коробчук Т. І. Кредитний менеджмент: Навчальний посібник. Луцький національний технічний університет. Луцьк : Волиньполіграф ТМ, 2016. 280 с.

REFERENCES:

1. Khatir, A. A. H. A., & Bee, M. (2022). Machine learning models and data-balancing techniques for credit scoring: What is the best combination? *Risks, 10*, 169–190.
2. MİLLİ, M. E. F., Deveci Kocakoç, İ., & Aras, S. (2024). Investigating the effect of class balancing methods on the performance of machine learning techniques: Credit risk application. *Izmir Journal of Management, 5*, 55–69.
3. Haque, F. M. A., & Hassan, Md. M. (2024). Bank loan prediction using machine learning techniques. *American Journal of Industrial and Business Management, 14*, 1690–1711. DOI: https://doi.org/10.4236/ajibm.2024.1412085
4. Yang, C., Dong, Y., Lu, J., & Peng, Z. (2022). Solving imbalanced data in credit risk prediction: A comparison of resampling strategies for different machine learning classification algorithms, taking threshold tuning into account. *Proceedings of the 5th International Conference on Machine Learning and Machine Intelligence (MLMI 2022)*, 30–40.
5. GeeksforGeeks. Logistic regression in machine learning. Available at: https://www.geeksforgeeks.org/understanding-logistic-regression/
6. Zolghadr, Z. (n.d.). Bank loan / credit scoring for bank customers. Kaggle. Available at: https://www.kaggle.com/datasets/zahrazolghadr/bank-loan-cleaned-ver1/data
7. Havrylenko, S. Yu., Zozulia, V. D., & Omelchenko, V. V. (2023). Doslidzhennia metodiv pidvyshchennia yakosti klasyfikatsii na nezbalansovanykh danykh [Research on methods for improving classification quality on imbalanced data]. *Systemy upravlinnia, navyhatsii ta zv'iazku – Control, Navigation and Communication Systems, 2*, 87–91.
8. Jain, A. Undersampling, oversampling and SMOTE, ensemble method and cost-sensitive learning techniques for dealing with imbalanced data. *Medium*. Available at: https://medium.com/@abhishekjainindore24/undersampling-oversampling-and-smote-ensemble-mehtod-and-cost-sensitive-learning-techniques-for-08efb557ec68
9. ROSE: Generation of synthetic data by randomly over sampling. Available at: https://rdrr.io/cran/ROSE/man/ROSE.html
10. Train in Data. Overcoming class imbalance with SMOTE. Available at: https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote/
11. Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research, 238*(2), 505–513.
12. Murel, J. (2024). What is a confusion matrix? *IBM*. Available at: https://www.ibm.com/think/topics/confusion-matrix
13. Brownlee, J. Tour of evaluation metrics for imbalanced classification. *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
14. Mostovenko, N. A., & Korobchuk, T. I. (2016). *Kredytnyi menedzhment: Navchalnyi posibnyk* [Credit management: A textbook]. Lutsk National Technical University. Volynpoligraf TM.

Стаття надійшла до редакції 27.02.2025.
The article was received 27 February 2025.