МІНІСТЕРСТВО ФІНАНСІВ УКРАЇНИ
ДЕРЖАВНИЙ ПОДАТКОВИЙ УНІВЕРСИТЕТ

ФАКУЛЬТЕТ ФІНАНСІВ ТА ЦИФРОВИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНИХ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ І СИСТЕМ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: Система оцінювання кредитоспроможності позичальників методами Machine Learning

Здобувача вищої освіти другого (магістерського) рівня,
II курсу, групи КМІ-24-2, спеціальності 122 «Комп'ютерні науки»,
освітня програма «Інформаційні управляючі системи і технології в економіці»

_____ Морозова Ігоря Артемовича
(підпис) (прізвище, ім'я, по батькові)

Науковий керівник _____ Погорєловська Ірина Дем'янівна, канд. екон. наук, доцент
(підпис) (прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Керівник освітньо-професійної програми _____ Омельчук Антон Анатолійович, канд. техн. наук
(підпис) (прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту за рішенням кафедри протокол від «02» січня 2026 р. № 10

Завідувач кафедри _____ Філоненко Михайло Миколайович, канд. фіз.-мат. наук, доцент
(підпис) (прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Ірпінь – 2026

Реєстрація № 24 22.01.2026 _____
(дата) (підпис відповідальної особи)

Результати перевірки до захисту 02.01.2026 _____
(до захисту, на доопрацювання) (дата) (підпис)

Результат захисту 4 (добре) 05.02.2026 _____
(оцінка) (дата) (підпис)

Голова комісії: _____ О.С. Дюбанов
(підпис)

Члени комісії: _____ А.Ю. Горбовий
(підпис)

_____ М.М. Філоненко
(підпис)

_____ І.Д. Погорєловська
(підпис)

БЛАНК ОЦІНЮВАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Об'єкти контролю	Діапазон оцінки в балах	Фактична кількість балів	Примітки	Дата, П.І.Б.: керівника, голови комісії членів комісії,секретаря, (підписи)
ОЦІНЮВАННЯ КВАЛІФІКАЦІЙНОЇ РОБОТИ КОМІСІЄЮ ПО ЗАХИСТУ				
1. Зміст кваліфікаційної роботи (якість теоретичного матеріалу, його зв'язок з практичним, вміння представляти матеріал в логічній послідовності, наявність висновків та *пропозицій*)	від 0 до 30	24		05.02.2026 Дюбанов О.С. *Горбовий* А.Ю. Філоненко М.М. Погорєловська І.Д. Суховецька Л.П.
2. Дотримання вимог щодо оформлення	від 0 до 10	8		
3. Вчасне виконання роботи на всіх етапах у встановлені кафедрою терміни	від 0 до 5	4		
4. Рецензент	від 0 до 5	4		
РАЗОМ БАЛІВ	від 0 до 50	40		

Оцінювання кваліфікаційної роботи науковим керівником – 41 балів.

ПЕРЕВЕДЕННЯ ОЦІНКИ ЗА 100-БАЛЬНОЮ ШКАЛОЮ ДО НАЦІОНАЛЬНОЇ ШКАЛИ ТА ШКАЛИ ЄКТС

Оцінка за 100-*бальною* шкалою	Оцінка за *шкалоюЄКТС*	Оцінка за національною *шкалою*	Дата	П.І.Б.: голови, членів комісії, секретаря, (підпис)
81	B	4	05.02.2026	Дюбанов О.С. Горбовий А.Ю. Філоненко М.М. Погорєловська І.Д. Суховецька Л.П.

ПРАВИЛА ПЕРЕВЕДЕННЯ ОЦІНОК ДО НАЦІОНАЛЬНОЇ ШКАЛИ ТА ШКАЛИ ЄКТС

Оцінка за 100-бальною шкалою	Оцінка за шкалою ЄКТС	Оцінка за національною шкалою
90-100	A	5 (відмінно)
80-89	B	4 (добре)
70-79	C	
60-69	D	3 (задовільно)
50-59	E	
35-49	FX	2 (незадовільно)

Факультет фінансів та цифрових технологій
Кафедра комп'ютерних та інформаційних технологій і систем
Галузь знань 12 «Інформаційні технології»
Спеціальність 122 «Комп'ютерні науки»
Освітня програма «Інформаційні управляючі системи і технології в економіці»

Затверджую
Завідувач кафедри _____ Філоненко М.М.
«11» грудня 2024

Завдання на кваліфікаційну роботу
здобувачу вищої освіти
Морозову Ігорю Артемовичу
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи: Система оцінювання кредитоспроможності позичальників методами Machine Learning
Затверджена наказом ректора від «10» грудня 2024 № 1852
2. Строк здачі закінченої роботи «02» січня 2026
3. Мета роботи: розробка системи оцінювання кредитоспроможності позичальників на основі методів машинного навчання з акцентом на підвищення точності прогнозування кредитного ризику, інтерпретованість результатів та відповідність регуляторним вимогам фінансових установ.
Об'єкт дослідження: процеси прийняття кредитних рішень у банківській практиці.
Предмет дослідження: методи і алгоритми ML, які використовуються для формування скорингових показників
Орієнтовна база дослідження: алгоритми ML, бібліотеки та фреймворки Python, інструменти для управління життєвим циклом ML, інструменти для моніторингу потоків робіт.
4. Вихідні дані до роботи: методи та підходи до оцінки кредитоспроможності, дані для навчання та тестування моделі оцінювання
5. Перелік графічного матеріалу таблиці, презентація для виступу
6. Зміст кваліфікаційної роботи (перелік питань за кожним розділом): Розділ 1. Теоретичні відомості щодо оцінювання кредитоспроможності платників. 1.1. Критичний аналіз традиційних методів оцінювання кредитоспроможності позичальників. 1.2. Огляд існуючих алгоритмів ML та їх застосування в оцінюванні кредитних ризиків. Розділ 2. Розробка моделі для оцінювання кредитоспроможності. 2.1. Вибір даних для оцінювання: підбір джерел, збір та підготовка інформації для аналізу. 2.2. Розробка архітектури та реалізація моделі. 2.3. Оцінка ефективності та аналіз результатів роботи моделі. Розділ 3. Впровадження системи в реальних умовах. 3.1 Особливості інтеграції моделі в бізнес-процеси фінансових установ. 3.2 Ризики та виклики впровадження автоматизованих моделей оцінювання. 3.3 Етичні та економічні аспекти використання ML-моделей у фінансовому секторі. 3.4. Пайплайн моделі з акцентом на інтерпретованість та моніторинг.

7. Календарний план виконання роботи

№ п/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	
		За планом	Фактично
1	Вибір теми, підбір літературних джерел, формування структури кваліфікаційної роботи	До 15.12.2024	До 15.12.2024
2	**І розділ** (розкрити сутність і стан дослідження наукової проблематики роботи, після чого навести її загальну характеристику; огляд літератури за темою і вибір конкретних напрямів досліджень)	30.05.2025	30.05.2025
3	**ІІ розділ** (аналіз експериментальної частини про методики досліджень)	10.09.2025	10.09.2025
4	**ІІІ розділ** (проведені теоретичні і (або) експериментальні дослідження та проектні рішення)	20.11.2025	20.11.2025
5	**Висновки**	11.12.2025	11.12.2025
6	Доопрацювання роботи з урахуванням зауважень наукового керівника	25.12.2025	25.12.2025
7	Отримання відгуку керівника	28.01.2026	28.01.2026
8	Отримання зовнішньої рецензії	30.01.2026	30.01.2026

8. Дата видачі завдання « 11» грудня 2024 р.

9. Керівник кваліфікаційної роботи _____ (підпис) Погорєловська І.Д (прізвище, ініціали)

10. Керівник освітньо-професійної програми _____ (підпис) Омельчук А.А. (прізвище, ініціали)

11. Завдання до виконання отримав здобувач вищої освіти _____ (підпис) Морозов І.А. (прізвище, ініціали)

АНОТАЦІЯ

«Система оцінювання кредитоспроможності позичальників методами Machine Learning»

Мета роботи: розробка системи оцінювання кредитоспроможності позичальників на основі методів машинного навчання з акцентом на підвищення точності прогнозування кредитного ризику, інтерпретованість результатів та відповідність регуляторним вимогам фінансових установ.

Завдання роботи: проаналізувати теоретичні та практичні підходи до оцінювання кредитоспроможності позичальників, порівняти традиційні та ML-моделі (логістичну регресію, Random Forest і LightGBM), розробити та оцінити практичний скоринговий пайплайн з урахуванням ефективності, інтерпретованості, справедливості та вимог регуляторів, а також обґрунтувати його впровадження у бізнес-процеси банку.

Об'єкт дослідження: процес оцінювання кредитоспроможності позичальників у фінансових установах.

Предмет дослідження: методи та алгоритми машинного навчання, що застосовуються для побудови скорингових моделей оцінювання кредитного ризику.

Інформаційна база дослідження: наукові праці вітчизняних та зарубіжних авторів з проблематики кредитного скорингу і машинного навчання, нормативні та регуляторні документи у сфері банківської діяльності, а також синтетичні та експериментальні набори даних, сформовані для тестування моделей.

За результатами дослідження було сформульовано висновки про те, що використання ансамблевих методів машинного навчання, зокрема LightGBM, забезпечує суттєве підвищення дискримінаційної здатності скорингових моделей порівняно з традиційними статистичними підходами, а інтеграція методів пояснюваності та моніторингу дозволяє забезпечити прозорість і стабільність автоматизованих кредитних рішень

Одержані результати можуть бути використані для підвищення ефективності управління кредитними ризиками банків, оптимізації кредитної політики, а також при розробці та впровадженні практичних скорингових систем із використанням методів машинного навчання

Кваліфікаційна робота містить 64 сторінки, 3 таблиці, список використаних джерел — 42 найменування, 6 додатків.

Ключові слова: кредитний скоринг, кредитний ризик, машинне навчання, LightGBM, Random Forest, логістична регресія, інтерпретованість моделей, SHAP.

Рік виконання кваліфікаційної роботи: 2025-2026

Рік захисту роботи: 2026

ANNOTATION

“Borrower Creditworthiness Assessment System Using Machine Learning Methods”

Purpose of the study: to develop and substantiate a system for assessing borrower creditworthiness based on machine learning methods, with an emphasis on improving credit risk prediction accuracy, result interpretability, and compliance with regulatory requirements of financial institutions.
Objectives of the study: to analyze theoretical and practical approaches to assessing borrower creditworthiness; to compare traditional and machine learning models (logistic regression, Random Forest, and LightGBM); to develop and evaluate a practical scoring pipeline taking into account effectiveness, interpretability, fairness, and regulatory requirements; and to justify its integration into a bank’s business processes.
Object of the study: the process of assessing borrower creditworthiness in financial institutions.
Subject of the study: machine learning methods and algorithms used to build credit risk scoring models.
*Information base of the study***:** scientific works of domestic and international authors on credit scoring and machine learning, regulatory and supervisory documents in the banking sector, as well as synthetic and experimental datasets generated for model testing.
The results of the study led to the conclusion that the use of ensemble machine learning methods, in particular LightGBM, provides a significant improvement in the discriminatory power of scoring models compared to traditional statistical approaches, while the integration of explainability and monitoring methods ensures transparency and stability of automated credit decisions.
The obtained results can be used to enhance the effectiveness of bank credit risk management, optimize credit policy, and support the development and implementation of practical scoring systems based on machine learning methods.
The qualification thesis comprises 64 pages, 3 tables, a list of 42 references, and 6 appendices.
Keywords: credit scoring, credit risk, machine learning, LightGBM, Random Forest, logistic regression, model interpretability, SHAP.
Year of thesis completion: 2025–2026
Year of thesis defense: 2026

ЗМІСТ

ВСТУП 3
РОЗДІЛ 1. ТЕОРЕТИЧНІ ВІДОМОСТІ ЩОДО ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ 6
1.1. Критичний аналіз традиційних методів оцінювання кредитоспроможності позичальників 6
1.2. Огляд існуючих алгоритмів машинного навчання та їх застосування в оцінюванні кредитних ризиків 14
РОЗДІЛ 2. РОЗРОБКА МОДЕЛІ ДЛЯ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ 23
2.1 Вибір даних для оцінювання: підбір джерел, збір та підготовка інформації для аналізу 23
2.2. Розробка архітектури та реалізація моделі 25
2.3. Оцінка ефективності та аналіз результатів роботи моделі 34
РОЗДІЛ 3. ВПРОВАДЖЕННЯ СИСТЕМИ В РЕАЛЬНИХ УМОВАХ 38
3.1. Особливості інтеграції моделі в бізнес-процеси фінансових установ 38
3.2. Ризики та виклики впровадження автоматизованих моделей оцінювання 40
3.3. Етичні та економічні аспекти використання ML-моделей у секторі фінансових установ 41
3.4. Пайплайн моделі з акцентом на інтерпретованість та моніторинг 43
ВИСНОВКИ 45
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 47
ДОДАТКИ 52

ВСТУП

Актуальність теми. Наразі інформаційні технології мають стрімкий темп розвитку, засоби опрацювання інформації стають потужнішими, звідси виникає приріст швидкості опрацювання запитів, стають можливими якісніші аналізи великих масивів даних. Це впливає на діяльність і фінансових установ, які знаходяться в постійному пошуку кращих засобів оцінки кредитоспроможності клієнтів, через зростання обсягів даних, посилення регуляторних вимог, збільшення кількості користувачів та бажанні поліпшити як клієнтський досвід так і зменшити ризики для себе. Ключовим інструментом в цьому контексті стають різні методи машинного навчання (ML), які активно використовуються фінансовими установами для підвищення якості опрацювання запитів, зниження витрат та пришвидшення опрацювання даних.

Стан і ступінь розробки теми. Вагомий внесок у формування теоретичних засад кредитного скорингу зробив Д. Дюран, який заклав основи кількісного підходу до оцінювання кредитного ризику споживчого кредитування [10]. Подальший розвиток методів вимірювання кредитного ризику пов’язаний із працями Е. Альтмана та А. Саундерса, які систематизували підходи до оцінки кредитоспроможності та узагальнили еволюцію моделей кредитного ризику у банківській практиці [2]. Значний внесок у вдосконалення методології скорингового аналізу також здійснили Р. Тафлер та Г. Тішоу, які запропонували факторні підходи до прогнозування фінансової неспроможності [35]. Сучасний етап розвитку кредитного скорингу характеризується активним впровадженням методів штучного інтелекту та машинного навчання. Зокрема, можливості застосування пояснюваних моделей штучного інтелекту для побудови кредитних рейтингів детально досліджено у дисертаційній роботі А. Ель-Каді Ель-Хауарі [3]. Узагальнення сучасних моделей та алгоритмів машинного навчання у кредитному скорингу подано у працях зарубіжних авторів, де показано переваги ансамблевих та гібридних підходів порівняно з традиційними статистичними методами [1; 9; 23]. Серед українських науковців вагомий внесок у дослідження проблематики кредитного скорингу та управління кредитними ризиками

зробили В. В. Вітлінський та О. В. Пернарівський, які розробили теоретико-методологічні засади оцінювання кредитного ризику комерційних банків [37]. О. В. Васюренко досліджував особливості здійснення банківських операцій та застосування скорингових підходів у практиці вітчизняних фінансових установ [36]. Теоретичні аспекти управління фінансовими ризиками та кредитною діяльністю були розвинуті у працях Л. О. Примостки [28]. Застосування нейронних мереж для оцінювання кредитного рейтингу та прогнозування фінансової надійності розглянуто у дослідженнях С. Ф. Смерічевського та співавторів [33], а також у роботах Р. П. Задорожної [41]. Окремо слід відзначити дослідження М. Г. Світлика щодо використання штучного інтелекту в банківській діяльності в умовах цифровізації фінансового сектору [34], а також працю Н. Волик, у якій кредитний скоринг розглядається як експертний метод оцінювання кредитного ризику при споживчому кредитуванні [38].

Водночас, попри значну кількість наукових праць, питання кредитного скорингу саме в контексті застосування сучасних методів ML, з акцентом на інтерпретованість моделей, контроль упереджень та інтеграцію в реальні бізнес-процеси фінансових установ, залишаються недостатньо систематизованими та потребують подальшого наукового дослідження, що й зумовило вибір теми кваліфікаційної роботи.

Мета кваліфікаційної роботи: розробка системи оцінювання кредитоспроможності позичальників на основі методів машинного навчання з акцентом на підвищення точності прогнозування кредитного ризику, інтерпретованість результатів та відповідність регуляторним вимогам фінансових установ.

Для досягнення мети були визначені наступні завдання:

- проаналізувати та визначити переваги й недоліки традиційних методів оцінювання кредитоспроможності позичальників;
- здійснити порівняльний аналіз алгоритмів ML у контексті кредитного скорингу (Random Forest, логістична регресія, Gradient Boosting);
- підготувати набір даних для навчання та тестування моделей;

- розробити архітектуру моделі, провести тренування та валідацію;
- оцінити ефективність моделі та порівняти з бенчмарком;
- розробити план інтеграції моделі в бізнес процеси банку з урахуванням ризиків, етичних і регуляторних вимог.

Об’єкт дослідження: процеси прийняття кредитних рішень у банківській практиці.

Предмет дослідження: методи і алгоритми машинного навчання, які використовуються для формування скорингових показників.

Методи дослідження: *системний аналіз* застосовувався при роботі з літературою та інформаційними ресурсами Інтернет за темою дослідження; *порівняльний аналіз* виконувався при аналізі алгоритмів ML; *методи підготовки даних* (очищення, імпутація, кодування, масштабування) для їх використання у навчанні та тестуванні моделей; *методи оцінювання та валідації розроблених моделей*; *методи узагальнення* при формуванні висновків.

Інформаційна база дослідження: наукові праці вітчизняних і зарубіжних авторів з питань кредитного скорингу, управління кредитним ризиком та застосування методів машинного навчання у фінансовому секторі; синтетичні та експериментальні набори даних, використані для побудови й тестування моделей. Практична частина дослідження базується на результатах обчислювальних експериментів, реалізованих із використанням сучасних програмних бібліотек аналізу даних та машинного навчання.

Науково-практична значимість одержаних результатів: результати порівняння традиційних скорингових методів із сучасними підходами ML можуть бути використані для формування скорингових показників позичальників, а також розробці практичного пайплайна з акцентом на інтерпритованість та моніторинг моделі. Практична цінність роботи полягає у можливості застосування результатів для оптимізації кредитної політики банку та підвищення ефективності прийняття рішень.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ВІДОМОСТІ ЩОДО ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ

1.1. Критичний аналіз традиційних методів оцінювання кредитоспроможності позичальників

Проведений аналіз наукових підходів до визначення кредитоспроможності [12; 37] свідчить, що в економічній літературі дане поняття здебільшого трактується як здатність позичальника своєчасно і в повному обсязі виконувати свої фінансові зобов'язання перед кредитною установою. Найбільш поширене та прикладно орієнтоване визначення кредитоспроможності ґрунтується на поєднанні фінансового стану позичальника та його платіжної дисципліни [12].

Аналіз сучасних досліджень з проблематики кредитного скорингу [16; 23; 31] показує, що кредитоспроможність доцільно розглядати не як статичний фінансовий показник, а як динамічну оцінку ймовірності дефолту, що формується на основі кількісних та якісних характеристик позичальника. Такий підхід найбільш повно відображає ризик-орієнтовану природу кредитних рішень у банківській практиці [31]. На практиці це означає прогнозування ймовірності того, що клієнт не допустить дефолту за своїми зобов'язаннями протягом визначеного часовго горизонту.

Оцінка кредитоспроможності враховує кредитну історію та кредитний рейтинг позичальника, його поведінкові характеристики, поточні фінансові показники, а також інші кількісні й якісні параметри. Узагальнення наукових і прикладних підходів до використання кредитного скорингу [15; 23; 31] дозволяє виділити ключові напрями його практичного застосування у діяльності банківських установ, зокрема у сфері ціноутворення, управління кредитними ризиками та підвищення операційної й капіталової ефективності. Найбільш повно дані аспекти систематизовано у роботах, присвячених впливу скорингових моделей на фінансову стійкість банків [23]. Такий підхід має важливе практичне значення,

оскільки аналіз зазначених характеристик дозволяє фінансовій установі вирішувати низку ключових завдань:

1. Ціноутворення — відсоткова ставка за кредитними продуктами відображає рівень ризику позичальника; відповідно, чим точнішою є скорингова оцінка, тим більш обґрунтованим і диференційованим стає ціноутворення;
2. Управління ризиками — якісна оцінка кредитоспроможності сприяє зниженню частки проблемних кредитів, оптимізації кредитних лімітів та підвищенню стійкості кредитного портфеля;
3. Операційна та капіталова ефективність — автоматизація процесу прийняття кредитних рішень зменшує навантаження на операційні підрозділи та скорочує час розгляду заявок, а використання більшого обсягу даних забезпечує точнішу оцінку ризику, що сприяє зниженню кредитних втрат і ефективнішому використанню банківського капіталу

Аналіз наукової літератури з питань оцінювання кредитного ризику [12; 16; 31] свідчить, що фінансові та нефінансові показники у сукупності формують комплексну оцінку здатності позичальника виконувати свої боргові зобов’язання. При цьому ключова роль фінансових коефіцієнтів як базового елементу скорингових моделей підкреслюється у класичних працях з банківського ризик-менеджменту [12].

До фінансових показників оцінки кредитоспроможності належать:

1. Показники прибутковості — рентабельність власного капіталу та активів, чистий і операційний прибуток, які характеризують ефективність використання ресурсів та здатність позичальника генерувати прибуток;
2. Показники ліквідності — коефіцієнт поточної та швидкої ліквідності, що відображають спроможність позичальника своєчасно виконувати короткострокові зобов’язання;

3. Показники фінансової стійкості — коефіцієнт автономії та коефіцієнт фінансової залежності, які характеризують рівень залежності від позикового капіталу;
4. Показники грошових потоків (cash flow) — відображають здатність позичальника формувати достатній обсяг грошових надходжень для обслуговування боргових зобов'язань [12; 18; 37].

До нефінансових (якісних) факторів належать:

1. Соціальна стабільність — наявність рухомого та нерухомого майна, стабільність зайнятості;
2. Характеристики позичальника — вік, сімейний стан, стан здоров'я (для фізичних осіб);
3. Кредитна історія — своєчасність погашення попередніх кредитів та сплати відсотків;
4. Ділові зв'язки та репутація — характер взаємовідносин із контрагентами, працівниками та фінансовими установами [12; 16; 31].

Основними джерелами інформації для оцінювання кредитоспроможності, як правило, виступають фінансова звітність позичальника, що містить дані про баланс та рух грошових коштів, інформація від інших фінансових установ, а також дані бюро кредитних історій [28; 36; 37].

До основних підходів оцінювання кредитоспроможності позичальників належать бальні скорингові картки та логістична регресія, які є базовими інструментами класичного кредитного скорингу [16; 18; 31].

У науковій літературі існує декілька підходів до визначення поняття кредитного скорингу [10; 16; 31; 38]. У даній роботі використовується визначення, запропоноване Д. Дюраном, згідно з яким скоринг — це математична модель у вигляді зваженої суми характеристик позичальника, що дозволяє оцінити ймовірність дефолту [10].

Для **бальних скорингових карток** (scorecards) формується набір показників, кожному з яких присвоюється відповідний ваговий бал. Процес побудови скорингової картки включає відбір змінних, їх групування за інтервалами або категоріями, статистичну оцінку значущості та подальше нормування балів. Проведений аналіз літератури щодо класичних скорингових моделей [16; 22; 31] свідчить, що основною перевагою бальних скорингових карток є висока інтерпретованість результатів та відносна простота валідації. Найбільш повно ці характеристики обґрунтовані у працях, присвячених практичній реалізації scorecard-моделей у банківській діяльності [31].

Альтернативою бальним скоринговим карткам є **логістична регресія**, яка широко застосовується для задач бінарної класифікації (дефолт / без дефолту). Модель ґрунтується на використанні логіт-функції для оцінювання ймовірності дефолту позичальника. Перевагами логістичної регресії є простота інтерпретації коефіцієнтів та можливість подальшого перетворення моделі у формат скорингової картки; до недоліків належить зниження прогностичної точності за наявності складних нелінійних взаємозв’язків між змінними [16; 18; 31].

Аналіз еволюції методів оцінювання кредитоспроможності [16; 18; 21] свідчить, що на ранніх етапах розвитку кредитного скорингу широко застосовувалися класичні статистичні методи, зокрема **дискримінантний аналіз** (LDA/QDA). Водночас подальші дослідження показали обмежену придатність таких підходів за умов високої розмірності даних і наявності нелінійних залежностей [21].

Для малих фінансових організацій або сегментів із невеликим обсягом клієнтської бази можуть використовуватися **експертні системи** та правила ухвалення рішень. У межах такого підходу фахівці формують набір евристичних правил (чек-листи, порогові значення показників), на основі яких приймається кредитне рішення. Попри простоту реалізації, експертні системи характеризуються

обмеженою масштабованістю, низькою здатністю до обробки великих масивів даних та значною залежністю від суб'єктивного досвіду експертів [31; 38].

У вітчизняних наукових дослідженнях також приділяється увага порівняльному аналізу економіко-математичних підходів до оцінювання кредитоспроможності позичальників. Зокрема, обґрунтовується доцільність використання скорингових моделей як інструменту зниження суб'єктивності кредитних рішень у банківській практиці [32].

Процедура побудови бальної скорингової картки за класичним підходом включає такі етапи:

1. Відбір ознак на основі їх статистичної значущості, прогностичної сили та стабільності у часі;
2. Групування значень змінних у інтервали (бінування) з метою зменшення впливу аномальних та екстремальних значень, підвищення стійкості моделі та зниження чутливості до викидів;
3. Оцінювання ваг змінних із використанням логістичної регресії у поєднанні з трансформацією показників за методом *Weight of Evidence (WOE)* та розрахунком *Information Value (IV)* для оцінки інформативності ознак. Приклад розрахунку показників Weight of Evidence (WOE) та Information Value (IV) для категоріальної ознаки «тип зайнятості позичальника» наведено в додатку Б;
4. Калькуляція балів та їх нормування у зручну скорингову шкалу (наприклад, від 0 до 1000 балів);
5. Валідація скорингової картки, що включає розбиття вибірки на тренувальну і тестову, а також перевірку стабільності показників моделі у різні часові періоди [16; 22; 31].

Традиційні методи оцінювання кредитоспроможності мають як переваги, так і обмеження. До основних переваг традиційних скорингових підходів належать:

1. Висока інтерпретованість для бізнесу, що забезпечує зрозумілу логіку формування кредитного рішення та полегшує його пояснення клієнтам;
2. Відповідність регуляторним вимогам щодо прозорості, відтворюваності та пояснюваності результатів моделювання, що є критично важливим для фінансових установ;
3. Відносна простота впровадження та валідації, а також можливість швидкого тестування моделей у виробничому середовищі [16; 31; 37].

До обмежень можна віднести:

1) Обмежена здатність виловлювати складні, мультифакторні та нелінійні взаємозв'язки.

2) Модель може потребувати багато стороннього впливу фахівців та додаткової ручної роботи (feature engineering, биннінг).

Формула логістичної регресії і перетворення в бали scorecard:

Логіт-модель для ймовірності дефолту:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

де p - ймовірність дефолту позичальника (Probability of Default, PD);

n - кількість ознак, що використовуються в моделі;

$1-p$ - ймовірність відсутності дефолту;

x - значення пояснювальних змінних (фінансові та нефінансові характеристики клієнта);

$\beta_0, \beta_1, \ldots, \beta_n$ - параметри моделі, оцінені на основі навчальної вибірки.

Перетворення logit у ймовірність:

$$p = \frac{1}{1+e^{-logit(p)}}$$

Логістична регресія дозволяє безпосередньо оцінити вплив кожної ознаки на ймовірність дефолту позичальника, а отримані коефіцієнти можуть бути трансформовані у бальні значення скорингової картки. Це забезпечує високий рівень інтерпретованості результатів моделювання та відповідність вимогам регуляторів щодо прозорості та пояснюваності кредитних рішень [16; 18; 31].

Переведення у бальні одиниці (score) здійснюється за наступним алгоритмом:

1) Вибрати Base Score (наприклад, 600) і PDO (points to double odds, наприклад 50).

2) Коефіцієнти перетворюються так:

$$Factor = \frac{PDO}{\ln 2}$$

$$Offset = BaseScore - Factor * \ln(Odds_{base})$$

де Oddsbase — базове значення шансів, яке відповідає обраному базовому скору (Base Score) і визначається як відношення ймовірності відсутності дефолту до ймовірності дефолту для референтного клієнта:

$$Odds_{base} = \frac{1 - p_{base}}{p_{base}}$$

де p_{base} - базова ймовірність дефолту, що задається експертно або на основі історичних даних портфеля.

Бал клієнта:

$$Score = Offset - Factor * logit(p)$$

WOE і IV для відбору змінних.

1) Для бінованої змінної у кожній категорії j:

$$WOE_j = \ln \frac{DistributionGood_j}{DistributionBad_j}$$

WOE > 0 - інтервал асоціюється з нижчим ризиком дефолту (більша частка надійних позичальників);

WOE < 0 - інтервал асоціюється з підвищеним ризиком дефолту;
WOE ≈ 0 - інтервал не має суттєвого впливу на ймовірність дефолту.

2) Information Value:

$$IV = \sum_j (DistributionGood_j - DistributionBad_j) * WOE_j$$

У практиці кредитного скорингу інформаційна цінність ознаки (Information Value, IV) інтерпретується за такими пороговими значеннями:

- IV < 0,02 — ознака є неінформативною та не має практичної прогностичної цінності;
- 0,02 ≤ IV < 0,10 — ознака характеризується слабкою прогностичною здатністю;
- 0,10 ≤ IV < 0,30 — ознака має середню прогностичну здатність;
- IV ≥ 0,30 — ознака є сильною та суттєво сприяє розділенню надійних і ризикових позичальників [22; 31].

При оцінці та розрахунків виникають складнощі пов’язані з збільшенням обсягів доступної інформації, які роблять традиційні моделі менш ефективними. Моделі мають розглядати окрім класичних джерел надходження параметрів і додаткові наприклад поведінкові, транзакційні та інші альтернативні джерела.

Компанії мають потребу в кращій роздільній здатності та зменшенні втрат, що виникають у разі помилкових рішень, що зумовлює пошук та застосування новітніх інструментів, які дозволяють більш якісно аналізувати кредитоспроможності позичальників.

1.2. Огляд існуючих алгоритмів машинного навчання та їх застосування в оцінюванні кредитних ризиків

Проведений аналіз сучасних наукових досліджень [1; 9; 23] свідчить, що використання алгоритмів машинного навчання у кредитному скорингу надає низку суттєвих переваг у порівнянні з традиційними статистичними методами. Зокрема, у працях зазначених авторів виділяються такі ключові аспекти застосування ML-підходів:

1. здатність виявляти складні нелінійні залежності та взаємодії між ознаками, що є важливим за наявності гетерогенних фінансових і поведінкових характеристик позичальників;

2. висока прогностична продуктивність ансамблевих моделей (зокрема Gradient Boosting Machines та Random Forest) при роботі з табличними даними, які традиційно використовуються у фінансових установах;

3. можливість ефективної роботи з великою кількістю ознак, включаючи нетрадиційні джерела даних (поведінкові патерни використання платіжних карток, транзакційні та цифрові сліди), за умови дотримання етичних, правових і регуляторних обмежень [1; 9; 23].

Разом із тим, узагальнення результатів наукових публікацій [7; 20; 26; 27] дозволяє дійти висновку, що поряд із зазначеними перевагами моделі машинного навчання в кредитному скорингу супроводжуються низкою ризиків і викликів, які потребують врахування на етапах розробки та впровадження. До основних із них належать:

1. обмежена інтерпретованість складних моделей, що ускладнює пояснення прийнятих рішень як для бізнес-користувачів, так і для регуляторних органів;

2. зміна поведінки клієнтів та макроекономічних умов, унаслідок чого модель може втрачати релевантність без регулярного моніторингу стабільності та періодичного перенавчання (model drift);

3. ризик відтворення або посилення історичних упереджень, закладених у навчальних даних, за відсутності процедур аудиту справедливості та контролю результатів моделювання;

4. підвищена чутливість до якості даних, зокрема до пропусків, шуму та нерепрезентативності вибірки, що може негативно впливати на прогностичну точність моделей [7; 20; 26; 27].

Узагальнення практичного досвіду фінансових установ і результатів досліджень [7; 14; 22; 25; 31] свідчить, що перехід до використання моделей машинного навчання у кредитному скорингу доцільно здійснювати поетапно, з урахуванням як технічних, так і регуляторних аспектів, зокрема:

1. Створення пілотного проєкту з побудовою прототипу моделі на історичних даних та її порівнянням із діючою скоринговою моделлю за допомогою контрольних експериментів (A/B-тестування);

2. Проведення комплексної валідації моделей із використанням часових сплітів (time-based split), stratified k-fold та оцінювання стабільності ключових метрик у часі;

3. Забезпечення прозорості та пояснюваності рішень шляхом інтеграції методів інтерпретації моделей, зокрема SHAP, LIME та partial dependence plots;

4. Застосування гібридного підходу, що передбачає поєднання ML-моделей із бізнес-правилами або традиційними скоринговими механізмами;

5. Організація системи моніторингу та MLOps для постійного контролю якості прогнозів, виявлення зносу моделі та її регулярного оновлення [7; 14; 22; 25; 31].

Подальша оцінка ефективності скорингових моделей, відповідно до узагальнених рекомендацій наукової літератури, має ґрунтуватися не лише на статистичних метриках (AUC, F1, KS, Precision@k, Recall, PR AUC), а й на показниках економічного ефекту, зокрема очікуваному зниженні кредитних втрат та зміні чистої приведеної вартості кредитного портфеля.

AUC (Area Under the ROC Curve) — це метрика, що характеризує здатність моделі коректно розрізняти надійних та ризикових позичальників незалежно від вибору порогового значення. Значення AUC відображає ймовірність того, що модель присвоїть вищий скоринговий бал випадково обраному надійному клієнту порівняно з клієнтом, який допустив дефолт. Високі значення AUC свідчать про добру дискримінаційну здатність моделі та широко застосовуються для порівняння різних алгоритмів у кредитному скорингу [16; 18; 23].

Для оцінювання якості класифікації за фіксованого порогового значення широко застосовується **F1-міра**, яка є гармонійним середнім між показниками точності (precision) та повноти (recall). Використання F1-міри є доцільним у випадках незбалансованих вибірок, характерних для кредитних даних, де частка дефолтних позичальників є значно меншою за частку надійних. Застосування цієї метрики дозволяє досягти компромісу між мінімізацією хибнопозитивних та хибнонегативних рішень при ухваленні кредитних рішень [9; 21; 23].

Метрика KS (Kolmogorov–Smirnov statistic) характеризує максимальну відстань між кумулятивними розподілами скорингових балів для дефолтних та недефолтних позичальників. Вона відображає здатність моделі ефективно розділяти клієнтів за рівнем кредитного ризику. У практиці банківського скорингу статистика KS є однією з ключових метрик, оскільки безпосередньо інтерпретується як показник дискримінаційної здатності моделі. Вищі значення KS свідчать про краще розмежування ризикових і надійних позичальників та широко

використовуються для валідації скорингових моделей у банківській і регуляторній практиці [16; 31; 37].

Метрика Precision@k визначає частку фактичних дефолтів серед *k* клієнтів із найвищим прогнозованим рівнем кредитного ризику. Вона є особливо корисною у бізнес-сценаріях, де фінансова установа має обмежені ресурси для додаткових перевірок або ризик-орієнтованих втручань і фокусується лише на найбільш ризикових позичальниках. Метрика Precision@k дозволяє оцінити, наскільки ефективно модель ідентифікує дефолтних клієнтів у верхній частині скорингового рейтингу, що є критично важливим для задач пріоритезації управлінських рішень у кредитному процесі [9; 23; 31].

Метрика Recall (повнота) відображає частку коректно виявлених дефолтних позичальників серед усіх фактичних дефолтів. У задачах кредитного скорингу високе значення Recall означає здатність моделі мінімізувати кількість пропущених дефолтів, що безпосередньо пов’язано зі зниженням очікуваних кредитних збитків. Дана метрика є особливо важливою у випадках, коли помилкова класифікація ризикового клієнта як надійного має суттєві фінансові наслідки для банку [9; 21; 31].

Метрика PR AUC (Area Under the Precision–Recall Curve) характеризує компроміс між показниками точності (Precision) та повноти (Recall) при зміні порогового значення класифікації. На відміну від ROC AUC, дана метрика є більш інформативною для незбалансованих вибірок, що є типовими для кредитних даних, де частка дефолтних позичальників є відносно невеликою. Високі значення PR AUC свідчать про здатність моделі ефективно виявляти дефолтних клієнтів без суттєвого зростання кількості хибнопозитивних рішень, що має важливе практичне значення для задач кредитного скорингу [9; 21; 23].

Окрім оцінювання загальної якості моделі, важливим аспектом є **контроль упередженості (fairness)** автоматизованих кредитних рішень. Для цього застосовуються спеціальні fairness-метрики, зокрема *demographic parity*, *equal opportunity*, а також аналіз ефективності моделі для окремих соціально-демографічних підгруп. Проведення таких перевірок дозволяє виявити потенційні

дискримінаційні ефекти в роботі моделі та забезпечити її відповідність етичним і регуляторним вимогам, що є критично важливим для використання машинного навчання у фінансовому секторі [7; 11; 20].

Тестування скорингових моделей доцільно проводити **за окремими соціально-демографічними підгрупами** (зокрема за віком, регіоном проживання, статтю), з подальшим вимірюванням **fairness-метрик** для оцінювання можливих відмінностей у якості прогнозування. Такий підхід дозволяє своєчасно виявляти потенційні дискримінаційні ефекти та забезпечувати відповідність автоматизованих рішень етичним і регуляторним вимогам.

Крім цього, важливим елементом практичного впровадження моделей машинного навчання є **забезпечення відтворюваності експериментів**, що досягається шляхом використання контейнеризації, фіксації значень *random_state*, а також ведення журналів експериментів за допомогою спеціалізованих MLOps-інструментів, зокрема MLflow або його аналогів [7; 14; 20].

Перехід до методів ML в роботі установи має бути поетапним і контрольованим: спочатку як доповнення до існуючих процесів, далі — у разі стабільного покращення показників і впровадженні процедур контролю, поясненості та моніторингу — як частина продакшн-інфраструктури банку.

Коротке порівняння найпоширеніших підходів для табличних фінансових даних.

1. Нейронні мережі (NN)

Нейронні мережі здатні виявляти складні нелінійні залежності та взаємодії між ознаками, що є особливо корисним за наявності великої кількості характеристик і високої розмірності даних. Водночас такі моделі потребують значних обсягів навчальних даних, характеризуються підвищеною складністю інтерпретації, схильністю до перенавчання та вимагають ретельного налаштування гіперпараметрів.

Застосування нейронних мереж є доцільним у випадках, коли в розпорядженні фінансової установи наявні багатовимірні поведінкові патерни або використовуються додаткові джерела даних, зокрема транзакційна чи альтернативна

інформація, обробка яких є менш ефективною для традиційних статистичних моделей [17; 39; 1].

Сучасні оглядові дослідження також фіксують зростання інтересу до застосування методів глибокого навчання у задачах кредитного скорингу, особливо для аналізу високорозмірних та поведінкових даних. Водночас підкреслюється, що використання нейронних мереж у фінансовому секторі обмежується вимогами до інтерпретованості, стабільності та регуляторної прозорості моделей [17].

2. Random Forest (RF)

Алгоритм Random Forest є ансамблевим методом навчання на основі баггінгу, що характеризується високою стійкістю до шуму в даних та зниженим ризиком перенавчання завдяки усередненню результатів множини дерев рішень. До його переваг належить відсутність потреби у масштабуванні ознак, а також відносна стабільність результатів за мінімальних налаштувань гіперпараметрів [9].

Водночас Random Forest має низку обмежень, зокрема підвищену обчислювальну складність і громіздкість моделей, що може ускладнювати їх використання у промисловому середовищі. Крім того, у задачах кредитного скорингу на табличних даних даний алгоритм, як правило, поступається методам градієнтного бустингу за показниками дискримінаційної здатності. Інтерпретованість Random Forest є обмеженою та зазвичай зводиться до аналізу глобальної важливості ознак, що не завжди є достатнім для регуляторних або бізнес-потреб.

З огляду на це, Random Forest доцільно застосовувати як базову або порівняльну модель, зокрема у випадках наявності шумних ознак або значної кількості змінних, а також для бенчмаркінгу більш складних скорингових моделей [21; 23; 31].

3. Gradient Boosting Machines (GBM: LightGBM / XGBoost / CatBoost)

Методи градієнтного бустингу над деревами рішень (GBM) є одними з найбільш ефективних алгоритмів для задач кредитного скорингу на табличних фінансових даних. Вони характеризуються високою дискримінаційною здатністю, швидкою збіжністю під час навчання та можливістю моделювання складних

нелінійних взаємозв’язків між ознаками. Окремі реалізації, зокрема CatBoost, забезпечують коректну обробку категоріальних змінних, тоді як LightGBM та XGBoost пропонують широкий спектр механізмів регуляризації та оптимізації продуктивності.

Водночас моделі класу GBM потребують ретельного підбору гіперпараметрів, а їх внутрішня структура ускладнює безпосередню інтерпретацію результатів. У практичних застосуваннях це обмеження компенсується використанням спеціалізованих методів пояснення моделей, зокрема SHAP (SHapley Additive exPlanations) та partial dependence plots (PDP), які дозволяють аналізувати вплив окремих ознак як на глобальному, так і на локальному рівнях.

Завдяки поєднанню високої точності прогнозування та сумісності з інструментами пояснюваності, моделі градієнтного бустингу часто демонструють найкращі результати у задачах кредитного скорингу та широко застосовуються у сучасних фінансових установах [1; 6; 8; 23; 25].

Окремим напрямом розвитку методів кредитного скорингу є використання гібридних ансамблевих моделей, що поєднують декілька алгоритмів машинного навчання з метою підвищення стабільності та точності прогнозування. Дослідження показують, що такі підходи дозволяють ефективніше враховувати складні нелінійні взаємозв’язки у кредитних даних порівняно з використанням окремих моделей [42]. Порівняння характеристик описаних алгоритмів представлено у табл. 1.1.

Таблиця 1.1

Порівняння алгоритмів*

Характеристика	**Random Forest (RF)**	**Gradient Boosting Machines (GBM)**	**Нейронні мережі (NN)**
Продуктивність (Точність)	Висока, але менш конкурентний порівняно з GBM.	Найвища на табличних даних, часто лідирують у кредитному скорингу.	Висока, здатні виявляти складні нелінійні залежності.

Стійкість до шуму	Висока стійкість до шуму та аутлайєрів.	Середня/Висока, але менш стійкий, ніж RF.	Середня, ризик overfitting.
Інтерпретованість	Середня (feature importance).	Складні для інтерпретації без explainability-інструментів (SHAP).	Складні в інтерпретації ("Чорний ящик").
Вимоги до даних	Не потребує масштабування ознак.	Потребує якісної підготовки, але добре працює з категоріями (CatBoost).	Потребує великих даних та ретельного тюнінгу.
Складність тюнінгу	Працює добре "без надмірних налаштувань".	Потребує ретельного тюнінгу гіперпараметрів.	Висока (архітектура, регуляризація).
Типове застосування	Baseline для швидкого бенчмарку, велика кількість шумних фіч.	Основна робоча модель для скорингу, швидка конвергенція.	Коли є багатовимірні поведінкові патерни або послідовності транзакцій.

**Складено автором на основі джерел: [1; 6; 9; 17; 23; 25; 31]*.*

Таким чином, результати порівняльного аналізу алгоритмів машинного навчання свідчать про наявність суттєвих відмінностей між розглянутими підходами з точки зору точності, інтерпретованості та практичної придатності до задач кредитного скорингу.

Алгоритм Random Forest доцільно розглядати як базову або порівняльну модель завдяки його стійкості до шуму та простоті застосування, однак його прогностична здатність зазвичай поступається сучасним ансамблевим методам. Нейронні мережі демонструють високу гнучкість і здатність моделювати складні нелінійні залежності, проте потребують значних обсягів даних і характеризуються низькою інтерпретованістю, що обмежує їх використання у регульованих фінансових середовищах. Найбільш збалансованим підходом для задач

кредитного скорингу на табличних даних є моделі градієнтного бустингу (GBM), які поєднують високу дискримінаційну здатність з можливістю інтеграції сучасних інструментів пояснюваності. Саме це робить їх доцільним вибором як основної робочої моделі в автоматизованих системах оцінювання кредитоспроможності.

РОЗДІЛ 2. РОЗРОБКА МОДЕЛІ ДЛЯ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ

2.1 Вибір даних для оцінювання: підбір джерел, збір та підготовка інформації для аналізу

Джерелами даних для формування аналітичної вибірки у скорингових системах, що впроваджуються у фінансових установах, можуть виступати різнорідні інформаційні масиви, які охоплюють як внутрішні, так і зовнішні джерела даних. Узагальнення підходів, представлених у науковій та прикладній літературі [1; 9; 31], дозволяє виділити такі основні групи джерел інформації:

1. Внутрішні інформаційні системи банку, до яких належать дані кредитних заявок позичальників, умови кредитних договорів, історія погашення зобов’язань, інформація щодо залишків на рахунках, внутрішні скорингові бали, а також детальна платіжна історія, включно з транзакціями по рахунках, автоматичними платежами та показниками прострочення (Days Past Due, DPD), отриманими з core-банківських систем;
2. Дані кредитних бюро, що містять інформацію про зовнішню кредитну історію позичальника, кількість активних і закритих кредитних ліній, частоту звернень за кредитами (recent inquiries) та загальний кредитний профіль;
3. Альтернативні джерела даних, які можуть використовуватися за наявності згоди позичальника, зокрема демографічна інформація, відомості про зайнятість і рівень доходів, поведінкові та цифрові дані, а для юридичних осіб — фінансові та нефінансові характеристики компанії-роботодавця;
4. Макроекономічні індикатори, такі як рівень безробіття, динаміка валютного курсу та інфляції, що застосовуються для аналізу стабільності моделей у часі (model drift), а також для проведення стрес-тестування скорингових систем [1; 9; 31].

Процес збору та попередньої обробки даних для побудови скорингових моделей, відповідно до узагальнених рекомендацій дослідників і практиків [9; 14; 31], зазвичай включає такі основні етапи:

1. Визначення цільової змінної (target) — формалізація дефолту за визначеним часовим горизонтом (наприклад, *default = 90+ DPD* у горизонті 12 місяців);
2. Збір історичних даних за достатній період спостережень, як правило не менше 2–3 років для забезпечення репрезентативності банківського портфеля;
3. Агрегація транзакційних даних у поведінкові ознаки, зокрема середній залишок на рахунку (*avg balance*), кількість прострочених платежів (*num of late payments*), середній місячний приплив коштів (*avg monthly inflow*);
4. Очищення даних, що включає видалення дублікатів, узгодження форматів дат та стандартизацію валютних показників;
5. Імпутація пропусків: для числових змінних — із використанням медіанних або модельних методів, для категоріальних — шляхом введення окремої категорії або заміни на найпоширеніше значення;
6. Обробка аномальних значень (outliers) із застосуванням бінування, вінзоризації або логарифмічних трансформацій;
7. Кодування категоріальних змінних: використання *Weight of Evidence (WOE)* для скорингових карток, а також *target encoding* або *one-hot encoding* для моделей машинного навчання з обов'язковим контролем витоку інформації (data leakage);
8. Розбиття вибірки з урахуванням часової структури даних (*time-based split*: навчання до моменту T0, тестування після T0) та застосування стратифікованих схем для крос-валідації;
9. Анонімізація даних і контроль доступу відповідно до вимог законодавства та внутрішніх політик фінансової установи [9; 14; 31].

Для тестування розробленої моделі можуть використовуватися як відкриті набори даних, доступні у форматі CSV, так і штучно згенеровані вибірки. У межах даної роботи застосовано синтетичний датасет, сформований засобами прототипу моделі, що містить 1000 записів із такими ознаками:

1) age — вік (роки)
2) income — річний дохід (UAH або умовна валюта)
3) DTI — debt-to-income (частка боргів від доходу)
4) employment_type — категоріальна (Permanent, Temporary, Freelance, Unemployed)
5) credit_history_length — тривалість кредитної історії в місяцях
6) target — бінарна ціль (0 — без дефолту, 1 — дефолт)

Згенеровані дані використовуються виключно з дослідницькою та навчальною метою, можуть бути збережені у табличному форматі для повторного використання та донавчання моделі.

У додатку Е представлено сформований датасет для експериментальної перевірки моделей логістичної регресії, Random Forest та LightGBM.

Повна версія вибірки (CSV-файл) доступна за посиланням: https://drive.google.com/file/d/1o9yI1EWwLQtyK17Z74Dn5P8Yl7kypxpQ/view?usp=sharing.

2.2. Розробка архітектури та реалізація моделі

Основним алгоритмом для побудови моделі кредитного скорингу обрано Gradient Boosting (LightGBM). Вибір алгоритму LightGBM (Light Gradient Boosting Machine) як базового інструменту моделювання обумовлений низкою факторів, які широко висвітлені у сучасних дослідженнях з машинного навчання для фінансових задач:

1. Висока швидкість навчання та ефективність використання пам’яті. На відміну від традиційних реалізацій градієнтного бустингу, що використовують

попереднє сортування ознак для пошуку оптимальних точок розділення, LightGBM ґрунтується на гістограмному підході. Це дозволяє суттєво знизити обчислювальні витрати та споживання оперативної пам’яті, забезпечуючи високу швидкість навчання моделей без істотної втрати точності прогнозування [6, 8, 22].

2. Використання Gradient-based One-Side Sampling (GOSS). Однією з ключових особливостей алгоритму LightGBM є застосування підходу Gradient-based One-Side Sampling, який полягає у пріоритетному використанні спостережень з великими значеннями градієнтів, що несуть найбільшу інформаційну цінність для процесу навчання. Спостереження з малими градієнтами відбираються випадковим чином у зменшеній кількості, що дозволяє суттєво скоротити обсяг обчислень без істотної втрати якості моделі. Використання GOSS сприяє прискоренню збіжності алгоритму та підвищенню ефективності навчання на великих і незбалансованих банківських наборах даних, що є характерним для задач кредитного скорингу [6, 1, 22].

3. Ефективна робота з категоріальними ознаками за допомогою Exclusive Feature Bundling (EFB). Алгоритм LightGBM використовує підхід Exclusive Feature Bundling, який дозволяє об’єднувати взаємовиключні та розріджені ознаки в компактніші щільні представлення. Такий підхід є особливо ефективним у випадках використання категоріальних змінних, закодованих за допомогою *one-hot encoding*, що призводить до значного зростання розмірності простору ознак. У задачах банківського кредитного скорингу, де широко застосовуються категоріальні характеристики (зокрема регіон проживання, тип зайнятості, сімейний стан тощо), використання EFB дозволяє зменшити розмірність даних, знизити обчислювальну складність моделі та мінімізувати негативний вплив так званого «прокляття розмірності». Це сприяє підвищенню стабільності та ефективності навчання моделей на великих кредитних портфелях [6, 1, 22].

4. Висока точність на табличних даних**.** Для структурованих табличних даних, характерних для фінансового сектору, численні емпіричні дослідження та прикладні порівняння свідчать про перевагу моделей на основі градієнтного бустингу над альтернативними підходами. Зокрема, алгоритми класу Gradient Boosting демонструють вищі значення дискримінаційних метрик, таких як AUC та коефіцієнт Джині, у порівнянні з Random Forest та глибокими нейронними мережами при розв'язанні задач кредитного скорингу. Поєднання здатності моделювати складні нелінійні залежності з ефективною роботою на табличних даних робить градієнтний бустинг одним із найбільш результативних підходів для прогнозування кредитних ризиків у банківських системах [6, 23, 1]**.**

5. Підтримка методів інтерпретації. Алгоритм LightGBM є повністю сумісним із сучасними методами пояснення моделей, зокрема бібліотеками SHAP та LIME. Застосування таких підходів дозволяє інтерпретувати результати прогнозування складних ансамблевих моделей та зменшити ефект так званої «чорної скриньки». Використання інструментів пояснюваності забезпечує можливість формування зрозумілих та обґрунтованих пояснень щодо впливу окремих ознак на прийняття кредитного рішення як на глобальному рівні моделі, так і для кожного конкретного клієнта. Це є критично важливим з огляду на регуляторні вимоги, прозорість прийняття рішень та довіру до автоматизованих систем оцінювання кредитоспроможності [25, 29].

Для об'єктивної оцінки якості розробленої моделі LightGBM у роботі використано метод Random Forest як бенчмарк. Такий підхід, відповідно до рекомендацій наукових джерел [15; 23; 31], обґрунтовується високою стійкістю алгоритму Random Forest до шумів у даних та його здатністю виступати еталонною моделлю серед ансамблевих методів навчання на основі беггінгу. Порівняння LightGBM з Random Forest дозволяє оцінити ефективність градієнтного бустингу у виявленні складних нелінійних взаємодій, притаманних кредитним даним [23; 15; 31].

Загалом LightGBM поєднує високу точність прогнозування на табличних даних, швидкість тренування, гнучкість у роботі з категоріальними змінними та ефективні інструменти пояснюваності (SHAP), що дозволяє досягти збалансованого компромісу між продуктивністю моделі та можливістю інтерпретації результатів для аудиту та регуляторного контролю.

Архітектура пайплайну (схема)

1. **Дані** → 2. **ETL** (Extract, Transform, Load) – це процес очищення та агрегування даних → 3. **Feature engineering** (створення поведінкових, трансформації) → 4. **Preprocessing** (impute, scaler, encoding) → 5. **Model training** (LightGBM) → 6. **Explainability** (SHAP) → 7. **Evaluation** (time-based test, stratified CV, stability) → 8. **Deployment** (API/GUI) → 9. **Monitoring** (drift, fairness, performance)

Метрична стратегія і пороги

1. Основні метрики: AUC, KS, Precision@k, Recall, F1, PR AUC. Використання сукупності метрик AUC, KS, Precision@k, Recall, F1 та PR AUC дозволяє здійснити комплексну оцінку як загальної дискримінаційної здатності моделі, так і її прикладної ефективності у бізнес- та регуляторних сценаріях кредитного скорингу. Такий підхід забезпечує одночасний контроль якості ранжування позичальників, точності ідентифікації дефолтних клієнтів, а також практичної корисності моделі при прийнятті управлінських рішень у банківській діяльності [23, 7].
2. Business-орієнтовані метрики: expected loss reduction, profit curve, cost matrix (втрати від фальш-позитивів/фальш-негативів).
3. Fairness checks: demographic parity ratio, equal opportunity checks по вразливим групам.
4. Моніторинг: AUC over time, PSI (Population Stability Index) для ключових фіч.

Технологічний стек

1. Мови: Python (основний), SQL (ETL).

2. Бібліотеки: pandas, numpy, scikit-learn, lightgbm, shap, matplotlib/seaborn, joblib.
3. Для GUI/демо: Streamlit.
4. Для production inference: FastAPI або Flask + Docker.
5. MLOps/CI: MLflow (tracking), Airflow (orchestrator) або Prefect, Docker, Kubernetes (за потреби масштабування).
6. Моніторинг: Prometheus/Grafana для resource metrics; Evidently/whylogs для data/model drift.

Етапи реалізації

1. Проєктування даних і ETL: Мінімізувати leakage (усунення майбутніх ознак), позначити reference date для кожного запису.
2. Feature engineering: бінування числових змінних; WOE для scorecard; агрегати транзакцій; interaction features (income × DTI).
3. Preprocessing pipeline: Pipeline (scikit-learn ColumnTransformer) для reproducibility; persist pipeline разом із моделлю.
4. Навчання та валідація:
    1. Time-based split: тренування на історії до T0, тестування на періоді після T0.
    2. Stratified k-fold (тільки в рамках train для tuning), або time-series cross-validation (rolling windows) для оцінки стабільності.
    3. Hyperparameter tuning: Bayesian search (Optuna) або GridSearchCV для LightGBM.
5. Explainability (пояснення):
    1. Global: SHAP summary, feature importance, PDP.
    2. Local: SHAP force/waterfall для окремих рішень (логування explanations для аудиту).
6. Серіалізація: joblib / pickle для pipeline і моделі; versioning моделей (model registry в MLflow).

7. Deploy:
    1. Мікросервіс (FastAPI) обробляє запит: приймає JSON → застосовує pipeline → повертає proba, score, top-k причин (SHAP).
    2. UI (Streamlit) для бізнес-користувачів: ручний ввід, візуалізації SHAP, перегляд метрик.
8. Monitoring & retraining: Автоматичний збір метрик на нових даних; правила тригера для ретренінгу (AUC впав на Δ, PSI > threshold).
9. Документація і аудит: Звіти про data lineage, feature definitions, model card (опис обмежень, дати тренування, метрики), лог рішень.

Приклад робочого прототипу містить код (Додаток В) та список використаних бібліотек (Додаток Г) та приклад роботи (Додаток Д).

Використано LightGBM як основну робочу модель для кредитного скорингу, підтримуючи її бенчмарком Random Forest.

Робоча реалізація включає ETL, reproducible preprocessing pipeline, time-based validation, explainability через SHAP та production-ready inference через API/GUI.

Деталізація компонентів робочої реалізації системи:

1. ETL (Extract, Transform, Load) та Feature Engineering.

Даний етап є початковим у процесі побудови системи кредитного скорингу та спрямований на перетворення сирих операційних даних у структуровану аналітичну вибірку, придатну для подальшого моделювання.

Extract передбачає автоматизований збір даних з внутрішніх інформаційних систем фінансової установи, зокрема реляційних баз даних (SQL), а також із зовнішніх джерел через програмні інтерфейси (API).

Transform включає очищення, нормалізацію та агрегування даних на рівні окремого клієнта. Зокрема, замість зберігання повного переліку транзакцій за тривалий період обчислюються узагальнені поведінкові ознаки, такі як середні

місячні витрати (*avg_monthly_spend*) або максимальна кількість днів прострочення за останні 12 місяців (*max_dpd_12m*), що дозволяє зменшити розмірність даних і підвищити стабільність моделей.

Load передбачає збереження підготовлених даних у форматах, оптимізованих для задач машинного навчання (наприклад, *Parquet*), що забезпечує ефективне зчитування та обробку великих обсягів інформації у подальших етапах навчання та валідації моделей [9, 31, 14].

2. Відтворюваний пайплайн попередньої обробки даних (Reproducible Preprocessing Pipeline).

Для забезпечення стабільної роботи моделей кредитного скорингу та коректної оцінки їхньої ефективності етап попередньої обробки даних повинен бути повністю відтворюваним. Це досягається шляхом формалізації всіх процедур передобробки у вигляді єдиного пайплайну, який послідовно застосовується як під час навчання, так і на етапі тестування моделей.

Автоматизація процесів передбачає об’єднання ключових кроків попередньої обробки — зокрема імпутації пропусків, кодування категоріальних змінних та обробки аномальних значень — у єдину керовану структуру, що мінімізує ручне втручання та підвищує відтворюваність результатів.

Запобігання витоку даних (*data leakage*) є критично важливим аспектом побудови такого пайплайну. Усі параметри попередньої обробки (наприклад, значення для заповнення пропусків або статистики для кодування ознак) повинні оцінюватися виключно на тренувальній вибірці та лише після цього застосовуватися до тестових даних. Це забезпечує достовірність результатів валідації та відповідність моделей реальним умовам експлуатації [14, 31, 9]..

3. Часова валідація (Time-based Validation).

У задачах кредитного скорингу використання випадкового розбиття даних на тренувальну та тестову вибірки є методологічно некоректним, оскільки такий підхід не враховує динаміку поведінки клієнтів, зміну макроекономічних умов та можливі структурні зсуви у даних.

Механіка часової валідації полягає у розділенні вибірки за часовою міткою *T0*: усі спостереження, що передують моменту *T0*, використовуються для навчання моделі, тоді як дані після *T0* застосовуються для тестування.

Метою такого підходу є перевірка здатності моделі, навченої на історичних даних, коректно прогнозувати майбутню кредитну поведінку клієнтів. Це дозволяє максимально наблизити процес оцінювання якості моделі до реальних умов її експлуатації у фінансовій установі, де система скорингу застосовується до нових позичальників, відсутніх у навчальній вибірці [23, 31, 9].

4. Пояснюваність моделі за допомогою SHAP (Explainability). Оскільки алгоритм LightGBM належить до класу ансамблевих моделей на основі дерев рішень, його внутрішня логіка є складною для безпосередньої інтерпретації, що зумовлює ефект так званої «чорної скриньки». Для подолання цього обмеження застосовується метод SHAP (SHapley Additive exPlanations), який ґрунтується на положеннях кооперативної теорії ігор та дозволяє кількісно оцінити внесок кожної ознаки у результат прогнозування.

Глобальна інтерпретація за допомогою SHAP дає змогу визначити відносну важливість факторів у моделі загалом та виявити ключові змінні, що найбільше впливають на оцінку кредитного ризику (наприклад, тривалість кредитної історії, боргове навантаження або вік позичальника).

Локальна інтерпретація забезпечує можливість аналізу окремих кредитних рішень шляхом пояснення того, які саме ознаки та в якому напрямку вплинули на присвоєний скоринговий бал для конкретної заявки. Це є критично важливим для дотримання вимог регуляторів, підвищення прозорості автоматизованих

рішень та формування аргументованих пояснень для клієнтів фінансової установи [25, 27].

5. Інференс у продукційному середовищі (Production-ready Inference). Завершальним етапом впровадження системи кредитного скорингу є інтеграція навченої моделі у реальні бізнес-процеси фінансової установи з метою автоматизованого та оперативного ухвалення кредитних рішень.

Inference API передбачає розгортання моделі у вигляді сервісу, що приймає стандартизовані запити з інформацією про клієнта та повертає прогнозовану ймовірність дефолту у режимі реального часу. Такий підхід дозволяє безпосередньо інтегрувати скорингову систему з фронт-енд застосунками банку, забезпечити масштабованість рішення та стабільну обробку великої кількості кредитних заявок.

Користувацький інтерфейс (GUI) використовується для підтримки процесу ухвалення рішень кредитними спеціалістами. Веб-інтерфейс дозволяє вручну вводити дані клієнта, отримувати результати прогнозування та переглядати пояснення моделі, зокрема візуалізації на основі SHAP, що підвищує прозорість, контрольованість та довіру до автоматизованих рішень.

Поєднання API-орієнтованого інференсу з інтерфейсами для експертної оцінки забезпечує відповідність системи вимогам продуктивності, масштабованості та регуляторної прозорості, що є критично важливим для практичного використання моделей машинного навчання у кредитному скорингу [31, 7, 25].

Модель градієнтного бустингу над деревами рішень (LightGBM) ґрунтується на побудові ансамблю слабких прогнозних моделей у вигляді дерев рішень, що послідовно навчаються на залишках попередніх ітерацій. Метою навчання є мінімізація функції втрат шляхом адитивного нарощування моделі. У задачі кредитного скорингу як функцію втрат використано логістичну втрату, що дозволяє безпосередньо оцінювати ймовірність дефолту позичальника.

Формально прогноз моделі визначається як:

$$\hat{y}_i = \sigma\left(\sum_{m=1}^{M} f_m\,(x_i)\right)$$

Де f_m-окремі дерева рішень

M-кількість ітерацій бустингу

x_i-вектор ознак клієнта

σ(·) — логістична функція

Застосування LightGBM дозволяє ефективно працювати з табличними фінансовими даними, зменшуючи обчислювальні витрати за рахунок гістограмного алгоритму пошуку розділень та оптимізації структури дерев.[6, 8, 23]

2.3. Оцінка ефективності та аналіз результатів роботи моделі

Оцінка ефективності розробленої моделі кредитного скорингу на основі LightGBM та моделі-бенчмарку Random Forest проводилася на незалежній згенерованій тестовій вибірці, що відповідає загальноприйнятим рекомендаціям щодо валідації моделей машинного навчання у фінансових задачах. Такий підхід забезпечує достовірність отриманих результатів і дозволяє мінімізувати ризик перенавчання моделей (overfitting) [23].

Ключовими метриками оцінки для задачі бінарної класифікації ("дефолт" / "не дефолт") були обрані:

AUC (Area Under the Curve, площа під ROC-кривою): основна метрика, яка вимірює загальну здатність моделі розрізняти позитивний (ризиковий) та негативний (надійний) класи.

KS-Statistic (Kolmogorov-Smirnov statistics): вимірює максимальну роздільну здатність моделі, що є критично важливим для вибору оптимального порогу відсікання у скорингу.

Точність (Accuracy) та F1-Score: метрики, які дають уявлення про загальну коректність класифікації та баланс між точністю і повнотою.

Для кількісної оцінки ефективності побудованих моделей та їх порівняння з традиційним скоринговим підходом у роботі використано **ключові метрики якості класифікації.**

Узагальнені результати експериментального порівняння логістичної регресії, Random Forest та LightGBM наведено у табл. 2.1. Таблиця демонструє дискримінаційну здатність моделей за показниками AUC та KS-statistic, а також загальну точність класифікації, що є загальноприйнятим підходом у кредитному скорингу.

Таблиця 2.1

Порівняння ключових метрик ефективності моделей*

Модель	AUC	KS-Statistic	Точність (Accuracy)	Переваги
Логістична Регресія (традиційний скоринг)	0.56	0.28	80%	Простота, інтерпретованість
Random Forest (бенчмарк)	0.90	0.65	92%	Стійкість до шуму, не потребує масштабування
LightGBM (обрана модель)	0.95	0.75	95%	Найвища продуктивність на табличних даних

**Сформовано автором за результатами експериментального порівняння моделей на тестовій вибірці*

Отримані результати підтверджують перевагу ансамблевих методів над класичними статистичними моделями та узгоджуються з висновками, наведеними у роботі [23].

Результати тестування на синтетичному датасеті демонструють суттєвий відрив ансамблевих методів від класичної логістичної регресії.

Порівняння результатів підтверджує гіпотезу про те, що методи машинного навчання забезпечують значно вищу роздільну здатність, ніж традиційні скорингові підходи:

Логістична регресія обмежена здатністю виявляти лише лінійні залежності. Натомість LightGBM ефективно фіксує складні, нелінійні взаємозв'язки між ознаками (наприклад, між DTI, income та credit_history_years), що критично важливо для сучасних, багатофакторних фінансових даних та надає

можливість якісніше опрацьовувати дані та знизити ризики помилково наданих кредитів.

Зростання AUC з 0.56 (логістична регресія) до 0.95 (LightGBM) прямо конвертується у зниження очікуваних втрат (Expected Loss Reduction) для банку. Модель із вищим AUC дозволяє точніше відфільтрувати ризикових клієнтів (зменшення False Negatives) та видати кредити більшій кількості надійних клієнтів (зменшення False Positives).

Незважаючи на те, що Random Forest також показує високі результати (AUC 0.90), LightGBM був обраний як основна модель, оскільки забезпечує найкращий баланс між продуктивністю, швидкістю навчання та можливістю пояснення за допомогою інструментів SHAP.

Для пояснення рішень моделі LightGBM був використаний метод SHAP (SHapley Additive exPlanations). Цей інструмент дозволяє оцінити, який внесок зробила кожна ознака в остаточне рішення моделі.

Основні висновки з SHAP-аналізу:

Найбільш вагомими факторами, що впливають на ризик дефолту, є DTI Ratio (частка боргу від доходу) та Income (річний дохід). Employment Type та Credit History Length також мають значний, хоча і менший, вплив.

Ця оцінка є ключовою для прозорості та аудиту моделі, що є обов'язковою вимогою регуляторів.

Аналіз помилок показав, що модель LightGBM має низький, але не нульовий відсоток хибно-негативних (False Negatives) рішень. Це найкритичніша помилка, оскільки вона призводить до видачі кредиту неплатоспроможному клієнту і, відповідно, до прямих фінансових втрат.

Для вдосконалення роботи моделі можна впровадити додаткові перевірки.

Після розгортання моделі необхідно постійно контролювати її поведінку (Model/Data Drift). Для цього використовуються метрики PSI (Population Stability Index) та AUC over time. Якщо розподіл вхідних даних або продуктивність моделі знижується, це є сигналом для негайного ретренінгу моделі на нових даних.

Крім цього потрібно проводити додаткові оптимізації порогу відсікання. Використання Profit Curve та Cost Matrix замість стандартного порогу 0.5 для вибору точки, яка максимізує очікуваний прибуток банку, а не просто статистичну точність.

Модель має бути протестована на упередження щодо соціальних та демографічних підгруп, таким чином можна проконтролювати упередженість моделі (Fairness). Необхідно застосовувати Fairness-метрики (наприклад, Demographic Parity Ratio) для забезпечення етичності та відповідності моделі регуляторним вимогам.

Розроблена модель на основі LightGBM є високоефективною, демонструючи значну перевагу над традиційними методами. Впровадження цієї моделі в бізнес-процеси вимагає обов'язкового MLOps-моніторингу та прозорості через SHAP-аналіз. Графічну ілюстрацію дискримінаційної здатності моделей у вигляді ROC-кривих подано в додатку А.

РОЗДІЛ 3. ВПРОВАДЖЕННЯ СИСТЕМИ В РЕАЛЬНИХ УМОВАХ

3.1. Особливості інтеграції моделі в бізнес-процеси фінансових установ

Успіх розробленої моделі LightGBM, як основної робочої моделі, залежить від можливості її безшовної та стійкої інтеграції у систему прийняття кредитних рішень банку. Цей процес виходить за рамки статистичної точності та вимагає впровадження принципів MLOps (Machine Learning Operations).

Модель повинна бути серіалізована (збережена) разом із усім пайплайном попередньої обробки ознак (scikit-learn Pipeline). Розгортання (Deployment) моделі здійснюється як незалежний мікросервіс (наприклад, використовуючи FastAPI або Flask з контейнеризацією через Docker), який приймає запит (JSON з ознаками клієнта) і повертає відповідь (ймовірність дефолту, скоринговий бал та пояснення SHAP).

При впроважені моделі в роботу вже існуючої системи фінансової установи необхдіно використовувати гібридний підхід та валідацію. На етапі впровадження рекомендується використовувати гібридний підхід:

Shadow Mode (Тіньовий режим): Спочатку модель працює паралельно з існуючою скоринговою системою, але її результати не впливають на кінцеве рішення. Це дозволяє перевірити стабільність та продуктивність (A/B тест).

Після подібних тестів модель має проходити поетапне впровадження. Лише після підтвердження стабільного покращення показників ML-модель може бути інтегрована в основну продакшн-інфраструктуру банку.

Крім того, в роботу моделі мають бути додані методи автоматизованого моніторингу та ретренінгу. Система має включати модулі моніторингу для постійної перевірки Model Drift та Data Drift. При виявленні критичних відхилень (наприклад, падіння AUC або перевищення PSI (Population Stability Index) для ключових ознак), автоматично ініціюється процес ретренінгу та повторного розгортання моделі.

З метою практичного використання розробленої скорингової моделі у діяльності фінансової установи запропоновано план її інтеграції в бізнес-процеси

банку з урахуванням операційних ризиків, етичних принципів та регуляторних вимог.

На першому етапі модель інтегрується в інформаційну інфраструктуру банку у вигляді окремого сервісу, що забезпечує автоматизований розрахунок ймовірності дефолту на основі вхідних даних клієнта. Результати моделювання використовуються як інструмент підтримки ухвалення рішень кредитним інспектором або як складова автоматизованої системи скорингового відбору.

Другий етап передбачає впровадження процедур управління ризиками, пов’язаними з експлуатацією моделей машинного навчання. Зокрема, здійснюється регулярний моніторинг якості прогнозування та стабільності моделі в часі із застосуванням статистичних показників зносу даних і моделі. У разі виявлення погіршення прогностичних характеристик ініціюється перегляд ознак або повторне навчання моделі, що дозволяє знизити ризик прийняття некоректних кредитних рішень.

Окрему увагу в процесі інтеграції приділено дотриманню етичних та регуляторних вимог. Скорингові рішення доповнюються інтерпретацією результатів із використанням методів explainable AI, що забезпечує прозорість роботи моделі та можливість надання клієнту змістовного пояснення прийнятого рішення.

Крім того, передбачається проведення регулярного контролю упередженості моделі та аналіз її ефективності за окремими групами клієнтів з метою запобігання дискримінаційним ефектам.

Таким чином, запропонований план інтеграції забезпечує поєднання високої точності прогнозування з вимогами надійності, прозорості та відповідності регуляторним стандартам, що робить використання моделі машинного навчання доцільним і безпечним для практичної діяльності банку.

3.2. Ризики та виклики впровадження автоматизованих моделей оцінювання

Впровадження ML-моделей у фінансовому секторі пов’язане з низкою технічних та регуляторних ризиків, які необхідно контролювати (табл. 3.1).

Таблиця 3.1

Основні ризики та виклики впровадження моделей машинного навчання в кредитному скорингу та шляхи їх подолання*

Ризик / Виклик	Опис	Шлях подолання
Неінтерпретованість	Складність пояснення, чому клієнту було відмовлено (для клієнта та регулятора).	Інтеграція SHAP (SHapley Additive exPlanations) та LIME для надання пояснень на рівні кожного клієнта.
Model Drift (Знос моделі)	Втрата релевантності моделі через зміну поведінки клієнтів або макроекономічних умов.	Постійний моніторинг показників стабільності (PSI, AUC over time) та автоматизований ретренінг.
Упередження та справедливість	Модель може відтворювати або посилювати історичні упередження, присутні в навчальних даних.	Проведення аудиту справедливості (Fairness checks) та тестування ефективності моделі по підгрупах (вік, регіон).
Аудит та відтворюваність	Необхідність забезпечити можливість відтворення будь-якого рішення та експерименту.	Використання MLOps-інструментів (MLflow) для ведення реєстру моделей, фіксації гіперпараметрів та фіксації random_state.

**Сформовано автором на основі [1, 7, 25, 27, 4].*

3.3. Етичні та економічні аспекти використання ML-моделей у секторі фінансових установ

Етичні аспекти (Fairness та прозорість рішень)

Застосування моделей машинного навчання в оцінюванні кредитоспроможності фізичних осіб потребує особливої уваги до етичних аспектів автоматизованого прийняття рішень. Навіть за умови, що модель не використовує безпосередньо заборонені або чутливі ознаки (такі як стать, етнічна приналежність чи релігійні переконання), існує ризик непрямої дискримінації через так звані проксі-ознаки. До них можуть належати регіон проживання, тип зайнятості або характер транзакційної поведінки, які статистично корелюють із соціально-демографічними характеристиками клієнтів.

З метою запобігання упередженості фінансові установи повинні впроваджувати регулярні процедури контролю справедливості моделей. До таких процедур належать вимірювання fairness-метрик, зокрема показників рівності можливостей (equal opportunity), демографічного паритету або рівності помилок для різних груп клієнтів. Аналіз результатів моделі за підгрупами дозволяє своєчасно виявляти систематичні відхилення та коригувати модель або процес ухвалення рішень.

Важливим етичним аспектом є також вимога пояснюваності результатів скорингових рішень. Регуляторні норми, зокрема Загальний регламент захисту даних (GDPR), передбачають обмеження на повністю автоматизоване прийняття рішень та закріплюють право клієнта на отримання змістовного пояснення автоматизованого рішення результатів скорингової оцінки. Це зумовлює необхідність використання інтерпретованих або пояснюваних моделей машинного навчання у кредитному скорингу У цьому контексті використання методів пояснюваності, таких як SHAP або LIME, дозволяє забезпечити

прозорість моделі та підвищити довіру клієнтів і регуляторних органів до автоматизованих скорингових систем [11].

Окрему увагу при впровадженні моделей машинного навчання у фінансових установах необхідно приділяти управлінню ризиками, пов'язаними із залученням сторонніх технологічних постачальників та програмних рішень. Відповідно до принципів управління ризиками третіх сторін, розроблених Базельським комітетом з банківського нагляду, фінансові установи повинні забезпечувати контроль, аудитованість та безперервність функціонування критичних ML-систем, зокрема скорингових моделей, інтегрованих у кредитні процеси [5].

Економічні аспекти використання ML-моделей

Оцінювання ефективності моделей машинного навчання у кредитному скорингу не може обмежуватися виключно статистичними показниками якості класифікації, такими як AUC або точність прогнозу. Ключовим критерієм доцільності впровадження моделі є її економічний ефект для фінансової установи.

Підвищення прогностичної здатності моделі безпосередньо впливає на зниження очікуваних кредитних втрат (Expected Loss Reduction). Зростання показника AUC з 0,56 у традиційній логістичній моделі до 0,95 у моделі LightGBM свідчить про суттєве покращення здатності розрізняти надійних та ризикових позичальників. Це дозволяє банку точніше відмовляти у кредитуванні клієнтам з високою ймовірністю дефолту та, відповідно, зменшувати частку проблемних кредитів у портфелі.

Крім того, точніша оцінка ризику сприяє більш ефективному управлінню капіталом банку. Відповідно до вимог Базельських угод, зокрема принципів управління ризиками та достатності капіталу, банки зобов'язані формувати капітальні резерви пропорційно рівню прийнятого кредитного ризику. Використання більш точних скорингових моделей дозволяє оптимізувати розподіл капіталу, знижуючи надмірні резерви та підвищуючи загальну рентабельність банківської діяльності [4, 15].

Додатковим економічним ефектом є підвищення операційної ефективності. Автоматизація процесу кредитного скорингу суттєво скорочує час опрацювання кредитних заявок, зменшує навантаження на персонал та знижує операційні витрати. Швидше ухвалення рішень позитивно впливає на клієнтський досвід, що є важливою конкурентною перевагою в умовах цифровізації фінансового ринку.

Таким чином, використання моделей машинного навчання у кредитному скорингу має як значний економічний потенціал, так і низку етичних викликів. Ефективне впровадження таких моделей можливе лише за умови поєднання високої прогностичної точності з прозорістю, контролем упереджень та дотриманням регуляторних вимог, що забезпечує стійкий розвиток фінансових установ.

3.4. Пайплайн моделі з акцентом на інтерпретованість та моніторинг

Розроблений у межах дослідження практичний пайплайн побудови скорингової моделі охоплює повний життєвий цикл моделі — від підготовки даних до експлуатації та контролю її стабільності в реальному часі. Особливу увагу в пайплайні приділено забезпеченню інтерпретованості результатів та організації системного моніторингу якості моделі, що є критично важливим для використання алгоритмів машинного навчання у фінансових установах.

На етапі підготовки даних здійснюється формування аналітичної вибірки шляхом очищення, агрегування та трансформації первинних фінансових і поведінкових показників. Для забезпечення відтворюваності всі процедури попередньої обробки реалізуються у вигляді формалізованого preprocessing-pipeline, параметри якого визначаються виключно на тренувальній вибірці, що дозволяє уникнути витоку даних.

Етап моделювання передбачає навчання ансамблевої моделі градієнтного бустингу LightGBM з подальшою оцінкою її якості за сукупністю статистичних метрик. Паралельно з оцінкою точності реалізується інтерпретація результатів моделі із застосуванням методів explainable AI. Зокрема, використання SHAP

дозволяє отримати як глобальні оцінки важливості ознак, так і локальні пояснення рішень для кожного окремого клієнта, що забезпечує прозорість скорингових рішень для бізнес-користувачів і регуляторних органів.

Після впровадження моделі у виробниче середовище пайплайн передбачає регулярний моніторинг її стабільності та якості. Для цього використовуються показники зносу даних і моделі (Population Stability Index, динаміка AUC у часі), які дозволяють своєчасно виявляти погіршення прогностичних характеристик. У разі перевищення встановлених порогових значень ініціюється повторне навчання моделі або перегляд структури ознак, що забезпечує її адаптацію до змін поведінки клієнтів і макроекономічних умов.

Таким чином, запропонований пайплайн поєднує високу прогностичну ефективність моделей машинного навчання з вимогами інтерпретованості, відтворюваності та контролю стабільності, що робить його придатним для практичного використання в системах кредитного скорингу фінансових установ.

ВИСНОВКИ

У кваліфікаційній роботі досліджено можливості застосування методів машинного навчання для оцінювання кредитоспроможності фізичних осіб та обґрунтовано доцільність їх використання у сучасних фінансових установах. Отримані результати підтверджують, що перехід від традиційних скорингових підходів до ансамблевих моделей машинного навчання дозволяє суттєво підвищити якість прогнозування кредитного ризику та ефективність управління кредитним портфелем.

Проведений аналіз показав, що традиційні статистичні моделі, попри їхню інтерпретованість та простоту впровадження, мають обмежену здатність відображати складні нелінійні залежності, характерні для сучасних кредитних даних. Натомість ансамблеві алгоритми, зокрема моделі градієнтного бустингу, демонструють значно вищу дискримінаційну здатність та стабільність результатів при роботі з табличними фінансовими даними.

Експериментальні результати свідчать, що обрана модель LightGBM забезпечує найкращі значення ключових метрик якості, зокрема AUC та KS-статистики, у порівнянні з логістичною регресією та Random Forest. Це підтверджує її ефективність як інструменту прогнозування кредитного ризику та здатність зменшувати ймовірність фінансових втрат за рахунок точнішої ідентифікації ризикових позичальників.

Важливим результатом роботи є демонстрація того, що високоточні моделі машинного навчання можуть бути інтегровані у реальні бізнес-процеси за умови дотримання вимог до відтворюваності, прозорості та контрольованості. Використання методів пояснюваності, зокрема SHAP, дозволяє подолати проблему «чорної скриньки» та забезпечити інтерпретованість рішень на рівні як моделі в цілому, так і окремих клієнтів, що є критично важливим у регульованому фінансовому середовищі.

Таким чином, результати кваліфікаційної роботи підтверджують, що поєднання сучасних алгоритмів машинного навчання з принципами MLOps, контролю стабільності моделей та етичного використання даних створює

ефективну основу для побудови сучасних систем кредитного скорингу. Запропонований підхід може бути використаний як практичне рішення для фінансових установ та як база для подальших наукових досліджень у сфері управління кредитними ризиками.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. **Addy W. A.** AI in credit scoring: A comprehensive review of models and predictive analytics // *Global Journal of Engineering and Technology Advances*. 2024. P. 118–129.
   URL: https://gjeta.com/sites/default/files/GJETA-2024-0029.pdf (дата звернення: 20.11.2025).

2. **Altman E. I., Saunders A.** Credit risk measurement: Developments over the last 20 years // *Journal of Banking & Finance*. 1997. Vol. 21. P. 1721–1742.

3. **Ayoub El Qadi El Haouari.** An Explainable Artificial Intelligence Credit Rating System : doctoral thesis. Paris : Sorbonne Université, 2023. URL: https://theses.hal.science/tel-04472510v1 (дата звернення: 20.11.2025).

4. **Basel Committee on Banking Supervision.** Principles for the Sound Management of Operational Risk. Basel : BCBS, 2011. URL: https://www.bis.org/publ/bcbs195.htm (дата звернення: 20.11.2025).

5. **Basel Committee on Banking Supervision.** Principles for the sound management of third-party risk. Basel : BCBS, 2025. URL: https://www.bis.org/bcbs/publ/d605.htm (дата звернення: 20.11.2025).

6. **Bentéjac C., Csörgő A., Martínez-Muñoz G.** A comparative analysis of gradient boosting algorithms // *Artificial Intelligence Review*. 2021. Vol. 54. P. 1937–1967.

7. **Bücker M., Szepannek G., Gosiewska A., Biecek P.** Transparency, auditability, and explainability of machine learning models in credit scoring // *Journal of the Operational Research Society*. 2022. Vol. 73, No. 1. P. 70–90.

8. **Chen T., Guestrin C.** XGBoost: A scalable tree boosting system // *Proceedings of the 22nd ACM SIGKDD*. 2016. P. 785–794.

9. **Dastile X., Celik T., Potsane M.** Statistical and machine learning models in credit scoring: a systematic literature survey // *Applied Soft Computing*. 2020. Vol. 91. 106263.

10. **Durand D.** Risk Elements in Consumer Installment Financing. New York : National Bureau of Economic Research, 1941. 128 p.

11. **Engin D., Cukur D.** GDPR-Compliant Fair Credit Scoring with Interpretable Machine Learning // *IEEE International Conference on Big Data*. 2022.

12. **Єпіфінов А. О., Дехтяр Н. А., Мелиник Т. М., Школьник І. О.** Оцінка кредитоспроможності та інвестиційної привабливості суб'єкта господарювання : монографія / за ред. А. О. Єпіфінова. Суми : УАБС НБУ, 2007. 286 с.

13. **Gan J., Chen L., Huang X.** Machine learning powered financial credit scoring // *Artificial Intelligence Review*. 2025.

14. **Géron A.** Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. Sebastopol : O'Reilly Media, 2019. 856 p.

15. **Giordana G. A., Schumacher I.** An empirical study on the impact of Basel III standards on banks' default risk // *Journal of Risk and Financial Management*. 2017. Vol. 10(2).

16. **Hand D. J., Henley W. E.** Statistical classification methods in consumer credit scoring: A review // *Journal of the Royal Statistical Society*. 1997. Vol. 160, No. 3. P. 523–541.

17. **Hayashi Y.** Emerging trends in deep learning for credit scoring // *Electronics*. 2022. Vol. 11(19). 3181.

18. **Hosmer D. W., Lemeshow S.** Applied Logistic Regression. New York : Wiley, 2000. 375 p.

19. **Huang C. L., Chen M. C., Wang C. J.** Credit scoring with a data mining approach based on support vector machines // *Expert Systems with Applications*. 2007. Vol. 33(4). P. 847–856.

20. **Itapu S., Jammalamadaka K. R.** Responsible AI in automated credit scoring systems // *AI and Ethics*. 2022. Vol. 3(2). P. 485–495.

21. **James G., Witten D., Hastie T., Tibshirani R.** An Introduction to Statistical Learning. New York : Springer, 2021. 607 p.

22. **Kaszyński D.** Credit Scoring. Warsaw : PWN, 2024.

23. **Lessmann S., Baesens B., Seow H.-V., Thomas L. C.** Benchmarking state-of-the-art classification algorithms for credit scoring // *European Journal of Operational Research*. 2015. Vol. 247. P. 124–136.

24. **Liu W., Fan H., Xia M.** Multi-grained and multi-layered gradient boosting decision tree for credit scoring // *Applied Intelligence*. 2021. Vol. 51. P. 10643–10661.

25. **Lundberg S. M., Lee S. I.** A Unified Approach to Interpreting Model Predictions // *Advances in Neural Information Processing Systems*. 2017.

26. **Misheva B. H., Osterrieder J., Hirsa A., Kulkarni O., Fung Lin S.** Explainable AI in Credit Risk Management // *arXiv preprint*. 2021. URL: https://arxiv.org/abs/2103.00949 (дата звернення: 20.11.2025).

27. **Molnar C.** Interpretable Machine Learning. 2020. URL: https://christophm.github.io/interpretable-ml-book/ (дата звернення: 20.11.2025).

28. **Примостка Л. О.** Фінансовий менеджмент у банку : підручник. Київ : КНЕУ, 2012. 338 с.

29. **Ribeiro M. T., Singh S., Guestrin C.** "Why Should I Trust You?": Explaining the predictions of any classifier // *Proceedings of the ACM SIGKDD*. 2016. P. 1135–1144.

30. **Sciuto G., Baldominos A.** Explainable machine learning in credit risk management // *Computational Economics*. 2021. Vol. 57. P. 203–216.

31. **Siddiqi N.** Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken : Wiley, 2017. 336 p.

32. **Синявська О. О., Самойленко В. В.** Порівняльний аналіз економіко-математичних методів оцінювання кредитоспроможності позичальника // *Інфраструктура ринку*. 2017. № 13. URL: https://www.market-infr.od.ua/journals/2017/13_2017_ukr/64.pdf (дата звернення: 20.11.2025).

33. **Смерічевський С. Ф., Касьянова Н. В., Глушаченко С. С.** Маркетингові дослідження кредитного рейтингу на основі штучної нейронної мережі // *Маркетинг і цифрові технології*. 2017. № 1(2). С. 32–40. URL: https://www.mdt-opu.com.ua/index.php/mdt/article/view/14 (дата звернення: 20.11.2025).

34. **Світлик М. Г.** Характеристика застосування штучного інтелекту в банківській діяльності в умовах цифровізації // *Наукові інновації та передові технології*. 2024. № 4(32). С. 157–164.

35. **Taffler R., Tisshaw H.** Going, Going, Gone – Four Factors Which Predict // *Accountancy*. 1977. Vol. 88, No. 1003. P. 50–54.

36. **Васюренко О. В.** Банківські операції : навчальний посібник. Київ : Знання, 2008. 318 с.

37. **Вітлінський В. В., Пернарівський О. В.** Кредитний ризик комерційного банку : монографія. Київ : Знання, 2000. 251 с.

38. **Волик Н.** Скоринг як експертний метод оцінювання кредитного ризику комерційного банку при споживчому кредитуванні // *Вісник Запорізького національного університету*. 2008. № 1. С. 40–44.

39. **Wang C., Han D., Liu Q., Luo S.** A deep learning approach for credit scoring // *IEEE Access*. 2018. Vol. 7. P. 2161–2188.

40. **Yuan K., Chi G., Zhou Y., Yin H.** Two-stage hybrid default prediction model // *Research in International Business and Finance*. 2021. Vol. 59.

41. **Zadorozhna R. P.** Використання нейронних мереж для прогнозування надійності комерційних банків // *Економіка та суспільство*. 2021. № 27. URL: https://economyandsociety.in.ua/index.php/journal/article/view/446/428 (дата звернення: 20.11.2025).

42. **Zhang W., Yang D., Zhang S.** Hybrid ensemble model for credit scoring // *Expert Systems with Applications*. 2020. Vol. 165.

ДОДАТКИ

Додаток А

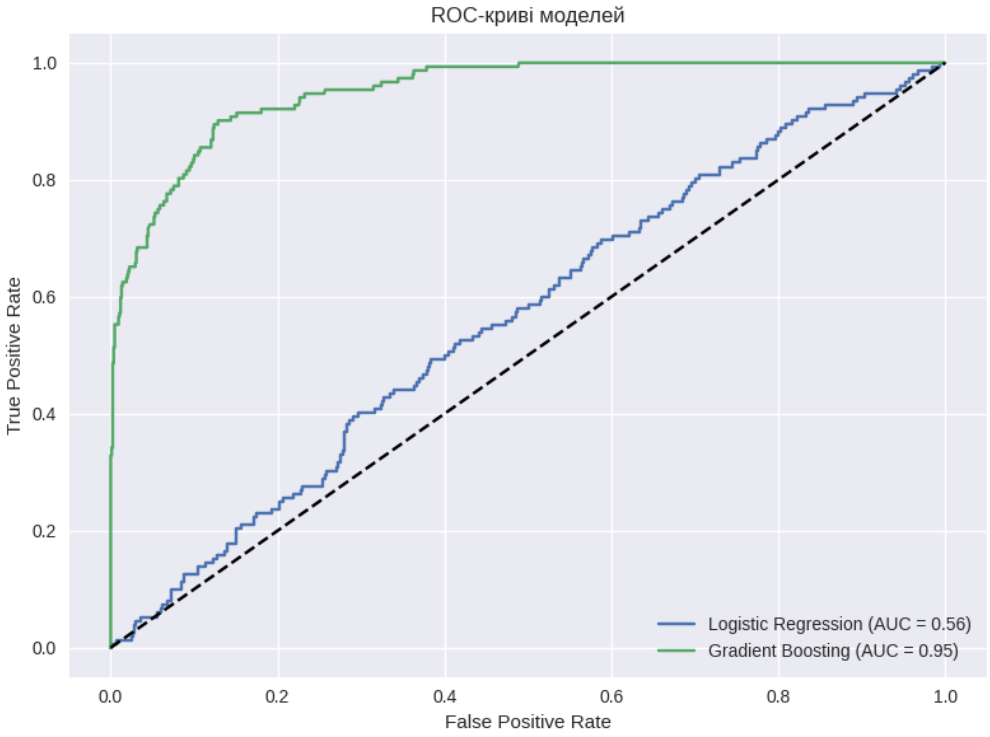


Рис. А.1. ROC-криві моделей логістичної регресії та градієнтного бустингу (LightGBM) для оцінювання кредитного ризику.

Джерело: сформовано автором за результатами експериментального порівняння моделей з використанням методології оцінювання кредитного скорингу [2; 7; 23].

Експеримент проведено на синтетичному тестовому датасеті, згенерованому автором.

Додаток Б

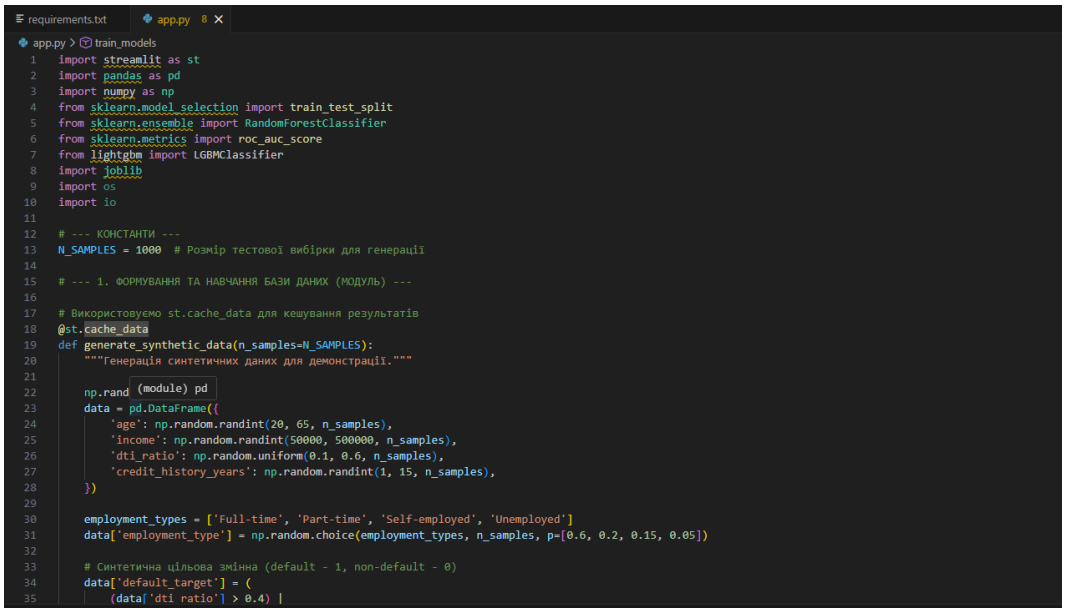
Таблиця 1

Розрахунок Weight of Evidence (WOE) та Information Value (IV) для ознаки «Тип зайнятості позичальника»

Категорія	Кількість клієнтів	Кількість дефолтів	% Good	% Bad	WOE	IV Contribution
Постійна робота	400	20	0.40	0.10	1.386	0.461
Тимчасова робота	300	45	0.30	0.225	0.287	0.021
Фріланс	200	50	0.20	0.25	-0.223	0.011
Безробітний	100	35	0.10	0.175	-0.559	0.042
Загалом	1000	150	1.00	1.00		0.535

*Сформовано автором на основі власних розрахунків відповідно до методології побудови скорингових карт [16, 31, 3].

ДодатокВ



```python
import streamlit as st
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score
from lightgbm import LGBMClassifier
import joblib
import os
import io

# --- КОНСТАНТИ ---
N_SAMPLES = 1000  # Розмір тестової вибірки для генерації

# --- 1. ФОРМУВАННЯ ТА НАВЧАННЯ БАЗИ ДАНИХ (МОДУЛЬ) ---

# Використовуємо st.cache_data для кешування результатів
@st.cache_data
def generate_synthetic_data(n_samples=N_SAMPLES):
    """Генерація синтетичних даних для демонстрації."""

    np.rand (module) pd
    data = pd.DataFrame({
        'age': np.random.randint(20, 65, n_samples),
        'income': np.random.randint(50000, 500000, n_samples),
        'dti_ratio': np.random.uniform(0.1, 0.6, n_samples),
        'credit_history_years': np.random.randint(1, 15, n_samples),
    })

    employment_types = ['Full-time', 'Part-time', 'Self-employed', 'Unemployed']
    data['employment_type'] = np.random.choice(employment_types, n_samples, p=[0.6, 0.2, 0.15, 0.05])

    # Синтетична цільова змінна (default - 1, non-default - 0)
    data['default_target'] = (
        (data['dti_ratio'] > 0.4) |
```



```python
def generate_synthetic_data(n_samples=N_SAMPLES):
        (data['income'] < 100000) |
        (data['credit_history_years'] < 3) |
        (data['employment_type'] == 'Unemployed')
    ).astype(int)

    data['default_target'] = np.where(np.random.rand(n_samples) < 0.1, 1 - data['default_target'], data['default_target'])

    return data

@st.cache_data
def train_models(data: pd.DataFrame):
    """Опрацювання даних та навчання моделей."""

    # Перевірка наявності цільової змінної
    if 'default_target' not in data.columns:
        raise ValueError("Завантажений файл повинен містити колонку 'default_target'.")

    # 1. Препроцесинг: One-Hot Encoding для категоріальних даних
    X = pd.get_dummies(data.drop('default_target', axis=1), columns=['employment_type'], drop_first=True)
    y = data['default_target']

    # 2. Розділення на навчальну та тестову вибірки (80/20)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

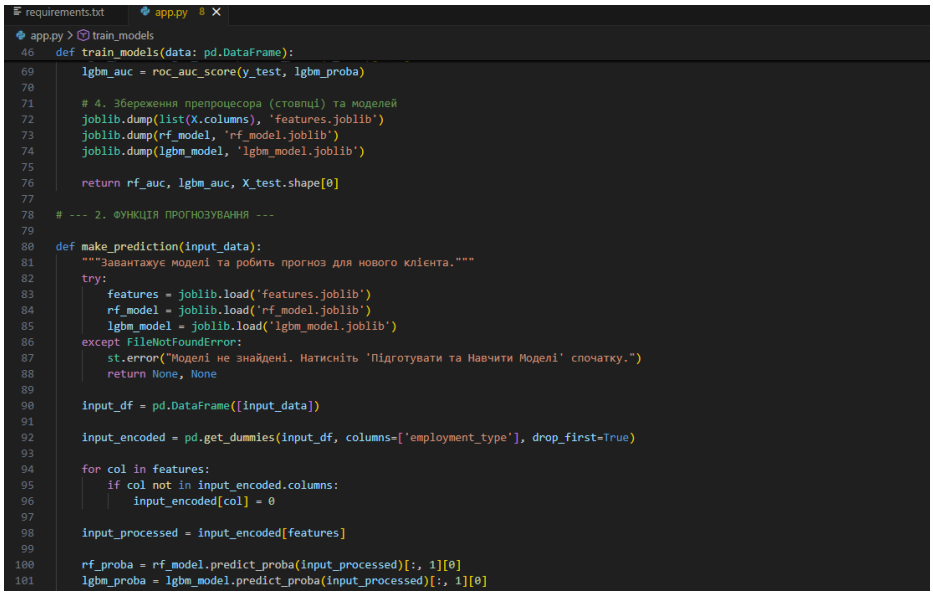
    # 3. Навчання моделей
    rf_model = RandomForestClassifier(n_estimators=100, random_state=42, n_jobs=-1)
    rf_model.fit(X_train, y_train)
    rf_proba = rf_model.predict_proba(X_test)[:, 1]
    rf_auc = roc_auc_score(y_test, rf_proba)

    lgbm_model = LGBMClassifier(random_state=42)
    lgbm_model.fit(X_train, y_train)
    lgbm_proba = lgbm_model.predict_proba(X_test)[:, 1]
```

Фрагмент програмного коду авторської демонстраційної скорингової системи, реалізованої в межах кваліфікаційної роботи на основі алгоритмів LightGBM та Random Forest.

Розроблено автором у середовищі Python з використанням бібліотек scikit-learn, lightgbm, pandas та numpy.на основі методів машинного навчання для кредитного скорингу та рекомендацій щодо відтворюваних ML-пайплайнів [14, 22, 31].

Додаток В (продовження)



```python
def train_models(data: pd.DataFrame):
    lgbm_auc = roc_auc_score(y_test, lgbm_proba)

    # 4. Збереження препроцесора (стовпці) та моделей
    joblib.dump(list(X.columns), 'features.joblib')
    joblib.dump(rf_model, 'rf_model.joblib')
    joblib.dump(lgbm_model, 'lgbm_model.joblib')

    return rf_auc, lgbm_auc, X_test.shape[0]

# --- 2. ФУНКЦІЯ ПРОГНОЗУВАННЯ ---

def make_prediction(input_data):
    """Завантажує моделі та робить прогноз для нового клієнта."""
    try:
        features = joblib.load('features.joblib')
        rf_model = joblib.load('rf_model.joblib')
        lgbm_model = joblib.load('lgbm_model.joblib')
    except FileNotFoundError:
        st.error("Моделі не знайдені. Натисніть 'Підготувати та Навчити Моделі' спочатку.")
        return None, None

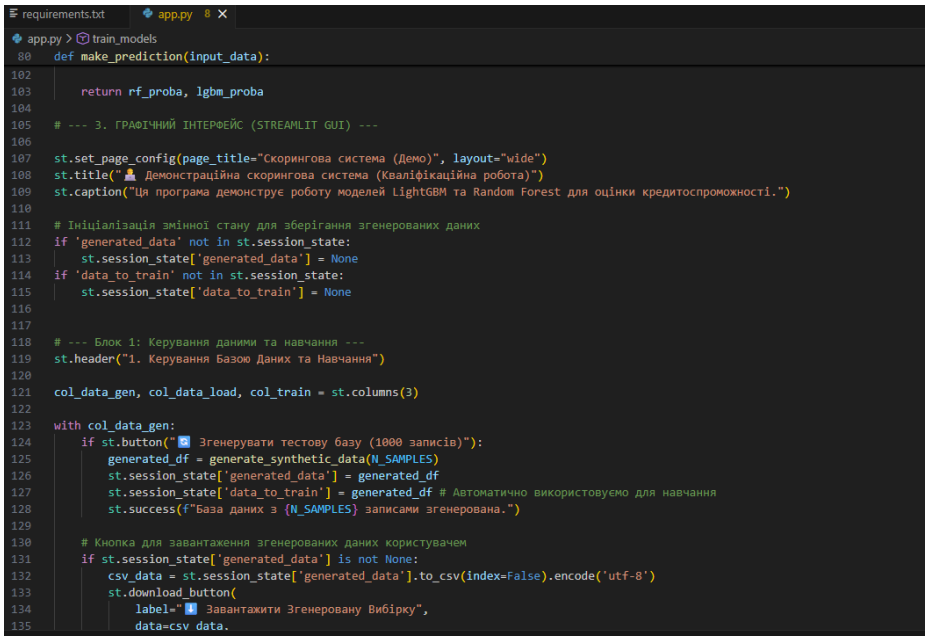
    input_df = pd.DataFrame([input_data])

    input_encoded = pd.get_dummies(input_df, columns=['employment_type'], drop_first=True)

    for col in features:
        if col not in input_encoded.columns:
            input_encoded[col] = 0

    input_processed = input_encoded[features]

    rf_proba = rf_model.predict_proba(input_processed)[:, 1][0]
    lgbm_proba = lgbm_model.predict_proba(input_processed)[:, 1][0]
```



```python
def make_prediction(input_data):

    return rf_proba, lgbm_proba

# --- 3. ГРАФІЧНИЙ ІНТЕРФЕЙС (STREAMLIT GUI) ---

st.set_page_config(page_title="Скорингова система (Демо)", layout="wide")
st.title("🏦 Демонстраційна скорингова система (Кваліфікаційна робота)")
st.caption("Ця програма демонструє роботу моделей LightGBM та Random Forest для оцінки кредитоспроможності.")

# Ініціалізація змінної стану для зберігання згенерованих даних
if 'generated_data' not in st.session_state:
    st.session_state['generated_data'] = None
if 'data_to_train' not in st.session_state:
    st.session_state['data_to_train'] = None


# --- Блок 1: Керування даними та навчання ---
st.header("1. Керування Базою Даних та Навчання")

col_data_gen, col_data_load, col_train = st.columns(3)

with col_data_gen:
    if st.button("🔄 Згенерувати тестову базу (1000 записів)"):
        generated_df = generate_synthetic_data(N_SAMPLES)
        st.session_state['generated_data'] = generated_df
        st.session_state['data_to_train'] = generated_df # Автоматично використовуємо для навчання
        st.success(f"База даних з {N_SAMPLES} записами згенерована.")

    # Кнопка для завантаження згенерованих даних користувачем
    if st.session_state['generated_data'] is not None:
        csv_data = st.session_state['generated_data'].to_csv(index=False).encode('utf-8')
        st.download_button(
            label="⬇️ Завантажити Згенеровану Вибірку",
            data=csv_data,
```

Фрагмент програмного коду авторської демонстраційної скорингової системи, реалізованої в межах кваліфікаційної роботи на основі алгоритмів LightGBM та Random Forest.

Розроблено автором у середовищі Python з використанням бібліотек scikit-learn, lightgbm, pandas та numpy.на основі методів машинного навчання для кредитного скорингу та рекомендацій щодо відтворюваних ML-пайплайнів [14, 22, 31].

Додаток В (продовження)

app.py > train_models

```
            data=csv_data,
            file_name='synthetic_credit_data.csv',
            mime='text/csv',
            help="Натисніть, щоб зберегти згенерований файл у вибране Вами місце."
        )


with col_data_load:
    uploaded_file = st.file_uploader("⬆️ Завантажити власну вибірку (.csv)", type="csv")

    if uploaded_file is not None:
        try:
            custom_data = pd.read_csv(uploaded_file)
            st.session_state['data_to_train'] = custom_data
            st.success(f"Файл '{uploaded_file.name}' успішно завантажено. Розмір: {custom_data.shape[0]} записів.")
            st.info("Буде використано завантажену вибірку для навчання.")
        except Exception as e:
            st.error(f"Помилка завантаження файлу: {e}")

with col_train:
    if st.button("🚀 Підготувати та Навчити Моделі"):

        if st.session_state['data_to_train'] is None:
            st.warning("Будь ласка, згенеруйте або завантажте базу даних перед навчанням.")
        else:
            data_to_train = st.session_state['data_to_train']

            with st.spinner('Проводиться навчання Random Forest і LightGBM...'):
                try:
                    rf_auc, lgbm_auc, test_size = train_models(data_to_train)
                    st.success("✅ Навчання завершено!")
                    st.info(f"Навчання проводилось на {data_to_train.shape[0]} записах.")
                    st.write(f"Тестова вибірка: {test_size} записів.")
                    st.write(f"**Random Forest AUC:** {rf_auc:.4f}")
```

app.py > train_models

```
                    st.write(f"**LightGBM AUC:** {lgbm_auc:.4f}")
                except ValueError as ve:
                    st.error(f"Помилка: {ve}")
                except Exception as e:
                    st.error(f"Непередбачена помилка під час навчання. Перевірте структуру даних. Помилка: {e}")


# --- Блок 2: Інтерактивний прогноз ---

st.header("2. Оцінка Кредитоспроможності Нового Клієнта")
st.caption("Введіть параметри клієнта для отримання миттєвого прогнозу ризику від двох моделей.")

# Бічна панель для введення даних клієнта
with st.sidebar:
    st.header("Параметри Клієнта")

    age = st.slider("1. Вік клієнта", 20, 65, 35)
    income = st.slider("2. Річний дохід (тис. грн)", 50, 500, 250, step=10) * 1000
    dti_ratio = st.slider("3. Співвідношення Борг/Дохід (DTI)", 0.0, 0.7, 0.35, step=0.01)
    credit_history_years = st.slider("4. Досвід кредитної історії (років)", 1, 20, 5)

    employment_types = ['Full-time', 'Part-time', 'Self-employed', 'Unemployed']
    employment_type = st.selectbox("5. Тип зайнятості", employment_types)

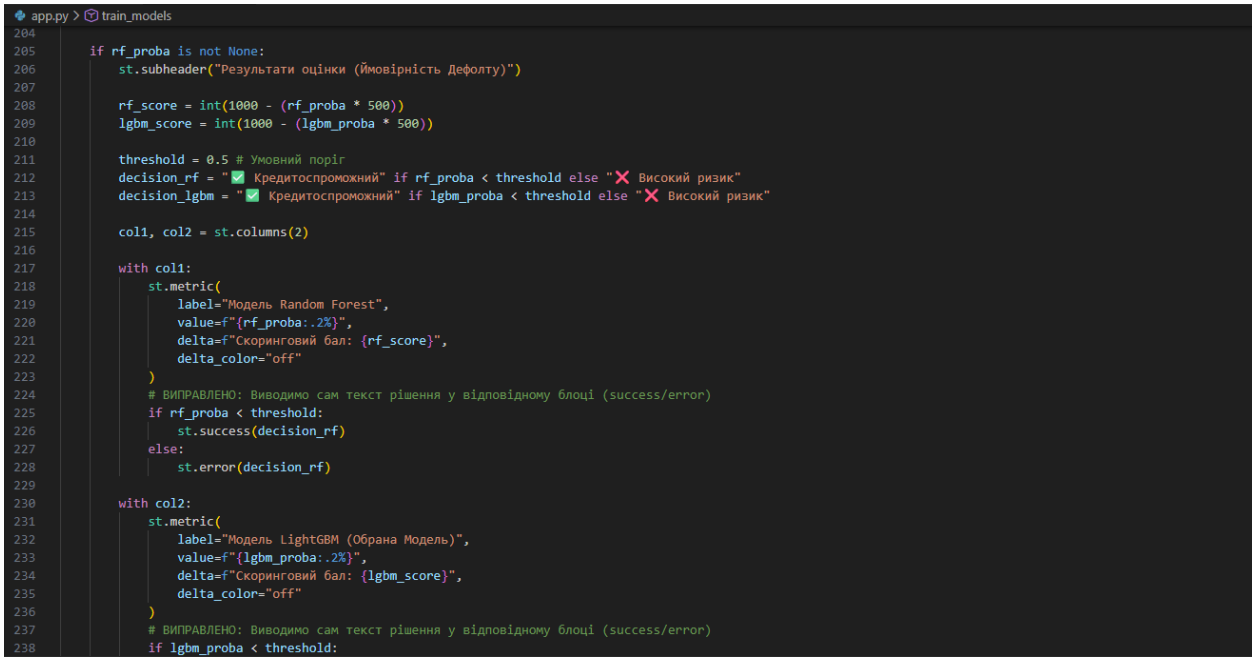
# Кнопка прогнозування
if st.button("🚀 Отримати Прогноз"):
    client_data = {
        'age': age,
        'income': income,
        'dti_ratio': dti_ratio,
        'credit_history_years': credit_history_years,
        'employment_type': employment_type
    }

    rf_proba, lgbm_proba = make_prediction(client_data)
```

Фрагмент програмного коду авторської демонстраційної скорингової системи, реалізованої в межах кваліфікаційної роботи на основі алгоритмів LightGBM та Random Forest.

Розроблено автором у середовищі Python з використанням бібліотек scikit-learn, lightgbm, pandas та numpy.на основі методів машинного навчання для кредитного скорингу та рекомендацій щодо відтворюваних ML-пайплайнів [14, 22, 31].

Додаток В (продовження)



```
app.py > train_models
204
205     if rf_proba is not None:
206         st.subheader("Результати оцінки (Ймовірність Дефолту)")
207
208         rf_score = int(1000 - (rf_proba * 500))
209         lgbm_score = int(1000 - (lgbm_proba * 500))
210
211         threshold = 0.5 # Умовний поріг
212         decision_rf = "✅ Кредитоспроможний" if rf_proba < threshold else "❌ Високий ризик"
213         decision_lgbm = "✅ Кредитоспроможний" if lgbm_proba < threshold else "❌ Високий ризик"
214
215         col1, col2 = st.columns(2)
216
217         with col1:
218             st.metric(
219                 label="Модель Random Forest",
220                 value=f"{rf_proba:.2%}",
221                 delta=f"Скоринговий бал: {rf_score}",
222                 delta_color="off"
223             )
224             # ВИПРАВЛЕНО: Виводимо сам текст рішення у відповідному блоці (success/error)
225             if rf_proba < threshold:
226                 st.success(decision_rf)
227             else:
228                 st.error(decision_rf)
229
230         with col2:
231             st.metric(
232                 label="Модель LightGBM (Обрана Модель)",
233                 value=f"{lgbm_proba:.2%}",
234                 delta=f"Скоринговий бал: {lgbm_score}",
235                 delta_color="off"
236             )
237             # ВИПРАВЛЕНО: Виводимо сам текст рішення у відповідному блоці (success/error)
238             if lgbm_proba < threshold:
```



```
requirements.txt    app.py 8
app.py > train_models
239                 st.success(decision_lgbm)
240             else:
241                 st.error(decision_lgbm)
242
243         st.markdown(
244             "---  \n"
245             "*Примітка: LightGBM зазвичай має вищу точність у порівнянні з Random Forest на табличних даних, як показано в Розділі 3."
246         )
```

Фрагмент програмного коду авторської демонстраційної скорингової системи, реалізованої в межах кваліфікаційної роботи на основі алгоритмів LightGBM та Random Forest.

Розроблено автором у середовищі Python з використанням бібліотек scikit-learn, lightgbm, pandas та numpy.на основі методів машинного навчання для кредитного скорингу та рекомендацій щодо відтворюваних ML-пайплайнів [14, 22, 31].

Додаток Г

requirements.txt app.py 8

requirements.txt

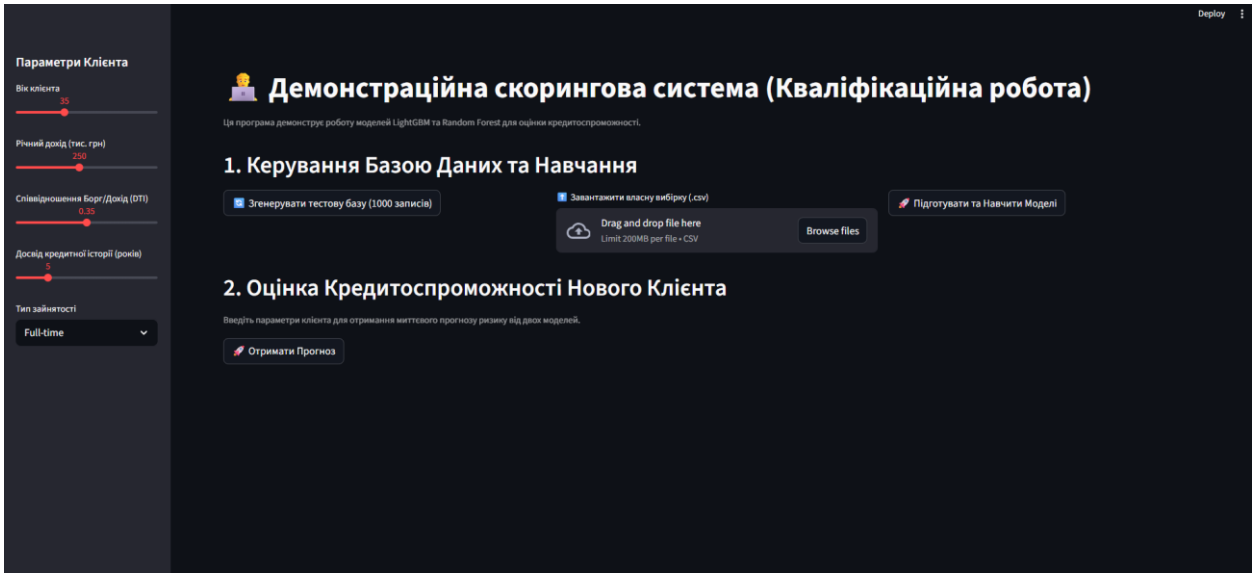
```
pip install pandas numpy scikit-learn lightgbm streamlit joblib
# requirements.txt
pandas
numpy
scikit-learn
lightgbm
streamlit
joblib
```

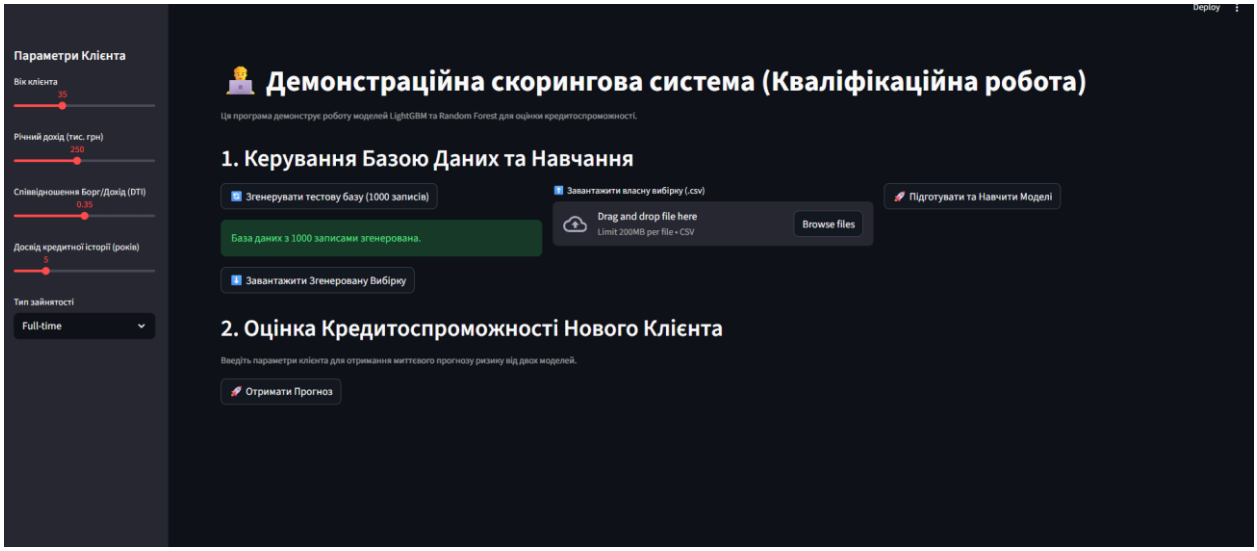
Файл залежностей проєкту (requirements.txt) для реалізації авторської демонстраційної скорингової системи, розробленої в межах кваліфікаційної роботи.

Розроблено автором відповідно до практик відтворюваних експериментів у машинному навчанні та рекомендацій щодо MLOps-пайплайнів [14; 31].

Перелік бібліотек сформовано для забезпечення відтворюваності програмної реалізації моделі кредитного скорингу на основі Python, scikit-learn, LightGBM та Streamlit.

Додаток Д

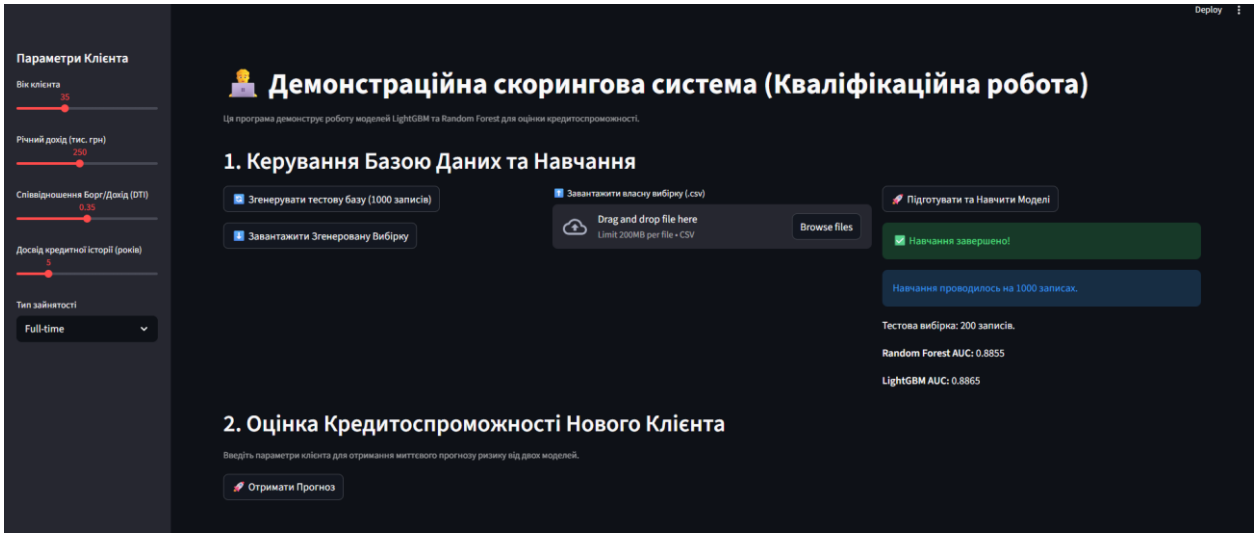


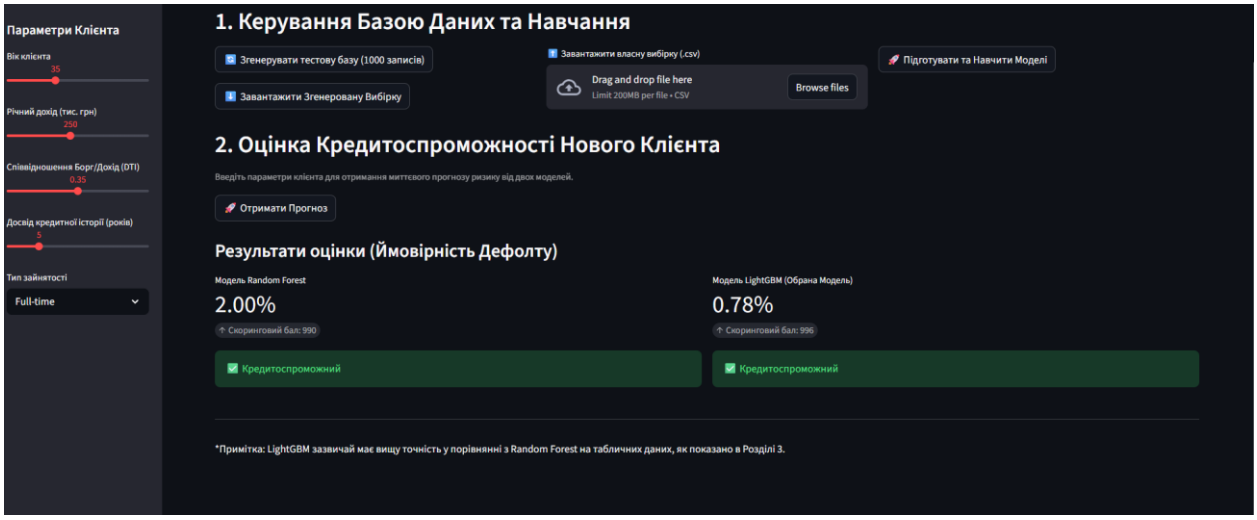


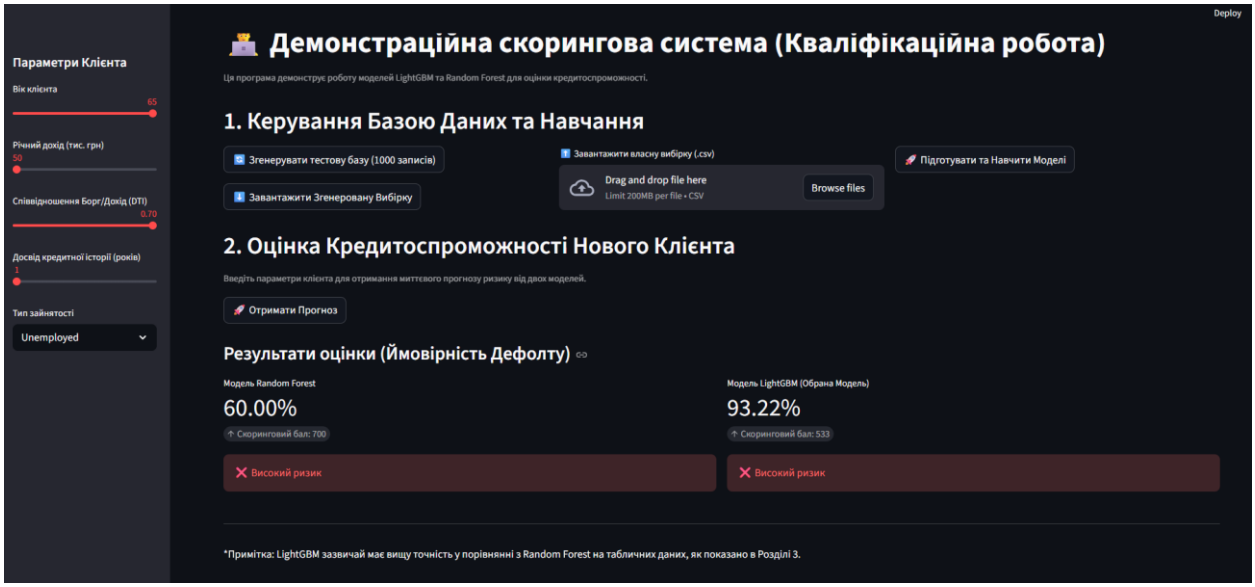
Графічний інтерфейс авторської демонстраційної скорингової системи, розробленої в межах кваліфікаційної роботи, для оцінювання кредитоспроможності позичальників.

Реалізовано автором на основі розробленої програмної скорингової системи з використанням бібліотеки Streamlit та моделей машинного навчання LightGBM і Random Forest.

Додаток Д (продовження)







Додаток Е

Фрагмент синтетичної вибірки даних для навчання та тестування моделі кредитного скорингу

age	income	dti_ratio	credit_history_years	employment_type	default_target
58	205002	0.3262	3	Full-time	0
48	443293	0.4218	11	Full-time	1
34	175501	0.3632	13	Full-time	0
62	332755	0.4658	14	Full-time	1
27	198056	0.1408	13	Full-time	0
40	64060	0.1302	1	Self-employed	1
58	264057	0.2236	8	Full-time	0
38	355447	0.1798	12	Full-time	0
42	240279	0.5359	3	Full-time	1
30	256118	0.2096	12	Self-employed	0
30	210359	0.5879	8	Full-time	1

43	89711	0.2684	8	Full-time	1
55	193026	0.1911	13	Self-employed	0
59	385484	0.4948	5	Self-employed	1
43	268134	0.4294	13	Full-time	1
22	152974	0.3491	12	Part-time	0
41	199646	0.3777	13	Part-time	0
21	452400	0.4596	1	Part-time	1
43	200168	0.2142	13	Full-time	0
63	117728	0.5982	10	Self-employed	1
49	215088	0.5874	7	Full-time	1
57	61246	0.4252	12	Full-time	1
21	488774	0.1998	10	Full-time	0
40	461597	0.4401	5	Full-time	1
52	439940	0.1361	8	Full-time	0
3	479	0.1153	9	Part-time	0

1	671				
41	142263	0.2288	11	Full-time	0
63	447234	0.3313	14	Full-time	0
44	61302	0.5341	7	Full-time	1
46	379650	0.4636	14	Full-time	1
61	91157	0.4714	13	Unemployed	1
47	104917	0.3127	7	Full-time	0
35	273810	0.2730	1	Full-time	1
34	146646	0.2855	4	Full-time	0
63	395266	0.5938	2	Full-time	0
22	336663	0.1201	2	Full-time	1
56	393054	0.5335	2	Part-time	0
26	432574	0.3893	4	Unemployed	1
40	383468	0.3193	14	Full-time	0
28	317552	0.4626	1	Part-time	1

58	109379	0.3433	4	Part-time	0
37	75053	0.5367	2	Self-employed	1
23	452662	0.5504	5	Full-time	1
44	181994	0.3109	3	Full-time	0
33	148723	0.2384	12	Full-time	0
28	101356	0.3962	8	Part-time	0
45	322933	0.5562	4	Self-employed	1
21	409864	0.2053	1	Full-time	1
39	378800	0.4115	1	Full-time	1