ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ ОЦІНЮВАННЯ КРЕДИТНИХ РИЗИКІВ НА ОСНОВІ АНСАМБЛЕВИХ МОДЕЛЕЙ

Халімончук Р.А.[1], Гуськова В.Г.

Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна

[1] halimonchukrostik@gmail.com

У роботі представлено підхід до створення інформаційної системи для оцінювання кредитних ризиків позичальників із використанням методів машинного навчання та ансамблевих моделей. Система реалізує повний цикл обробки даних і дозволяє порівнювати ефективність різних алгоритмів, серед яких логістична регресія, LightGBM, MLP та комбіновані ансамблі. Додатково розроблено зручний веб-інтерфейс, що забезпечує завантаження даних, вибір моделей, перегляд метрик та графічних візуалізацій, спрощуючи проведення аналізу. Отримані експериментальні результати підтверджують, що ансамблеві методи підвищують точність прогнозування дефолтів, а запропонована система може бути корисним інструментом у практичних задачах кредитного скорингу.

Ключові слова: кредитний ризик, скоринг, машинне навчання, LightGBM, MLP, логістична регресія, стекінг, блендинг, гібридна модель.

1. ВСТУП

У сучасній фінансовій сфері задачі оцінювання кредитних ризиків потребують швидких, точних та адаптивних рішень. Традиційні статистичні методи часто не враховують складні взаємозв’язки між ознаками та не забезпечують достатньої точності при великих обсягах даних. У зв’язку з цим дедалі більшої популярності набувають інтелектуальні системи, побудовані на методах машинного навчання та ансамблевого моделювання.

Ансамблеві алгоритми дозволяють поєднати результати різних моделей та підвищити стабільність прогнозування. Це робить їх ефективним інструментом у задачах кредитного скорингу, де помилки класифікації можуть призвести до фінансових втрат або необґрунтованих відмов у кредитуванні.

У роботі створено інформаційну систему, яка забезпечує автоматизоване завантаження даних, налаштування моделей, їх навчання та формування прогнозів, що робить підхід придатним до використання у банківських установах.

2. ПОСТАНОВКА ЗАДАЧІ ТА ВИХІДНІ ДАНІ

Метою дослідження є створення інформаційної системи, що забезпечує автоматизовану оцінку ймовірності дефолту позичальників на основі ансамблевих моделей машинного навчання. Задача передбачає поєднання кількох компонентів: підготовку структурованого набору ознак, розроблення базових моделей, побудову ансамблів, отримання прогнозів та презентацію результатів у зручному для аналізу вигляді.

У роботі значну увагу приділено попередній обробці даних. Було виконано очищення вибірки від пропусків, трансформацію категоріальних ознак, масштабування числових змінних, а також обробку змінних із вкрай нерівномірним розподілом. Для уникнення зміщення моделей під час навчання застосовано стратифікований поділ на тренувальну та тестову вибірки, що дозволяє зберегти пропорцію дефолтних клієнтів.

3. МЕТОДИ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1. Побудова базових моделей

На першому етапі побудовано три базові моделі. Логістична регресія використана як інтерпретований статистичний метод, що дозволяє оцінити значущість ознак та забезпечує базову точність. LightGBM, який реалізує градієнтний бустинг на деревовидних структурах, продемонстрував високу здатність моделювати складні залежності та взаємодії між ознаками. Нейронна мережа MLP забезпечила можливість аналізу нелінійних закономірностей та виявила чутливість до параметрів навчання.

3.2. Реалізація ансамблевих методів

Для підвищення точності класифікації побудовано ансамблеві моделі.

Стекінг реалізовано як дворівневу структуру, де базові моделі формують прогнозні ймовірності, які надалі виступають вхідними ознаками для метамоделі на основі логістичної регресії.

Блендинг забезпечує об’єднання моделей на окремій валідаційній вибірці, що дозволяє спростити процедуру тренування при збереженні високої ефективності.

Гібридна модель GBDT+LR поєднує нелінійність дерев LightGBM із інтерпретованістю лінійної моделі, що досягається шляхом перетворення листків дерев у набір бінарних ознак.

3.3. Інформаційна система та інтерфейс

Розроблена інформаційна система оснащена веб-інтерфейсом, який забезпечує послідовну роботу з даними та моделями. Користувач може завантажити власний CSV-файл або використати демонстраційний датасет, після чого вибрати цільову змінну для моделювання. Інтерфейс дозволяє гнучко обирати одну або кілька базових моделей, а також визначати ансамблевий метод.

У налаштуваннях передбачено вибір параметрів тестового поділу та випадкового генератора, що забезпечує відтворюваність результатів. Після запуску моделювання система автоматично формує таблицю метрик (AUC та Accuracy) та відображає графічні візуалізації. Користувачу доступні стовпчикові діаграми для порівняння моделей та графік розподілу прогнозних ймовірностей, який демонструє поведінку різних алгоритмів на тестовій вибірці. Така функціональність забезпечує швидкий аналіз результатів і дозволяє ефективно порівнювати якість окремих моделей та ансамблю.

На рис. 1 наведено головне вікно інтерфейсу, яке демонструє основні елементи взаємодії з даними та моделями. На рис. 2 показано результати роботи системи: таблицю метрик для обраних алгоритмів, а також графічні візуалізації – порівняння AUC та Accuracy та розподіл прогнозних ймовірностей для кожної моделі.

Представлені візуалізації свідчать про те, що система не лише автоматизує процес моделювання, а й забезпечує наочне порівняння результатів, що є критично важливим для прийняття рішень у задачах кредитного скорингу. Така інтеграція аналітичних інструментів робить інтерфейс практичним та зручним для подальшого застосування у фінансовій сфері.

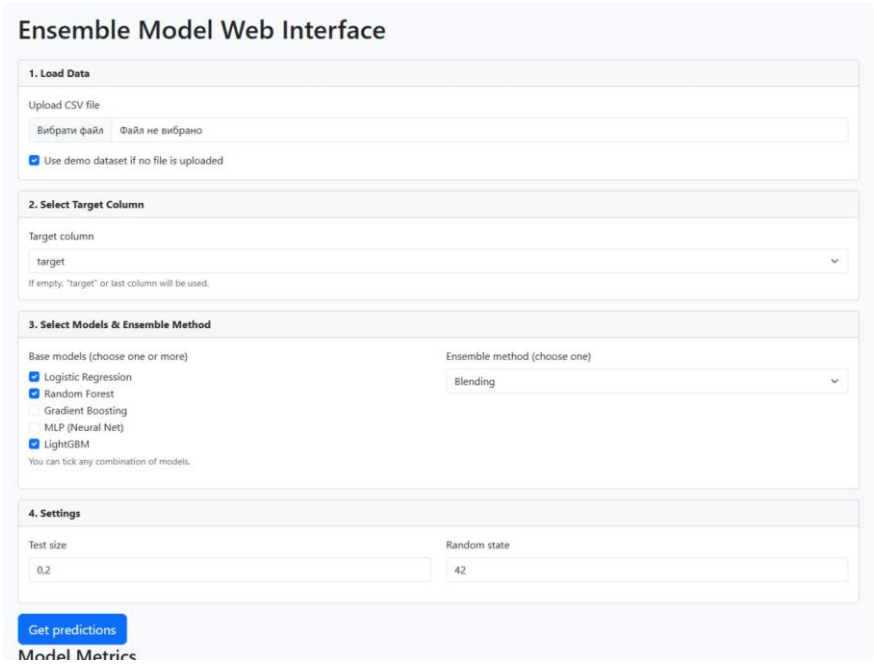

Рисунок 1. Інтерфейс користувача

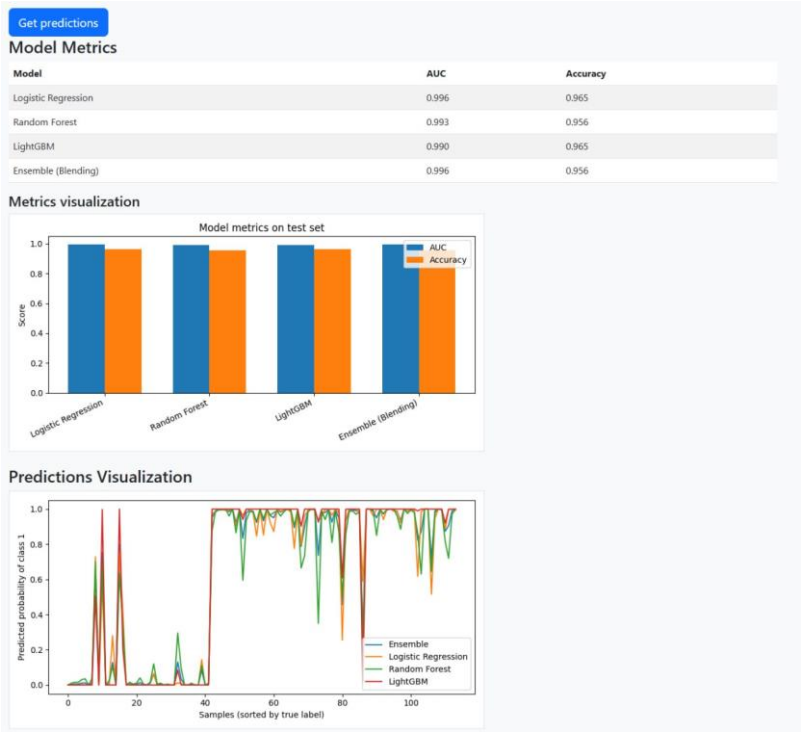

Рисунок 2. Таблиця метрик і графічна візуалізація в інтерфейсі

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

4.1. Логістична модель

Оцінювання якості моделей проводилося на тестовій вибірці, яка була повністю відокремлена від даних, використаних у процесі навчання. Для зменшення впливу випадкових

коливань у розподілі даних додатково застосовано перехресну валідацію. Такий підхід забезпечив отримання усереднених результатів, що коректно відображають загальну надійність і стабільність моделей.

Логістична регресія виступила базовою моделлю для порівняння з іншими алгоритмами (табл. 1). Її лінійний характер забезпечив інтерпретованість та стабільну класифікацію більшості «надійних» клієнтів. Водночас модель показала обмежену здатність розпізнавати складні нелінійні взаємозв'язки, що призвело до пропуску частини дефолтних випадків. Це типова особливість лінійних методів, які не враховують взаємодії ознак і тому слугують радше початковим орієнтиром для оцінювання інших підходів.

Таблиця 1. Метрики логістичної регресії

	Precision	Recall	F1-score	Support
Class 0	0.92	1.00	0.96	70687
Class 1	0.32	0.00	0.00	6191
Accuracy	-	-	0.92	76878
Macro avg	0.62	0.50	0.48	76878
Weighted avg	0.87	0.92	0.88	76878

4.2. LightGBM

Модель LightGBM продемонструвала значно вищу точність завдяки можливості моделювати складні структури та взаємодії між ознаками. Завдяки деревоподібній архітектурі й механізму градієнтного бустингу вона ефективно розпізнавала ризикових позичальників, зменшила кількість хибних негативних прогнозів і показала стабільність навіть за нерівномірних або зашумлених даних. LightGBM став одним із найрезультативніших методів серед протестованих (табл. 2).

Таблиця 2. Метрики LightGBM

	Precision	Recall	F1-score	Support
Class 0	0.92	1.00	0.96	70687
Class 1	0.57	0.02	0.04	6191
Accuracy	-	-	0.92	76878
Macro avg	0.74	0.51	0.50	76878
Weighted avg	0.89	0.92	0.88	76878

4.3. Нейронна мережа (MLP)

Багатошаровий перцептрон добре впорався з нелінійними зв'язками в даних та продемонстрував високу здатність розпізнавати дефолтних клієнтів. Однак модель виявилась чутливою до параметрів навчання, що в окремих випадках призводило до збільшення хибнопозитивних рішень. У загальному результаті MLP показав сильні сторони у задачах із складною структурою ознак, але вимагає точного налаштування гіперпараметрів (табл. 3).

Таблиця 3. Метрики нейронної мережі

	Precision	Recall	F1-score	Support
Class 0	0.93	0.94	0.93	70687
Class 1	0.17	0.15	0.16	6191
Accuracy	-	-	0.87	76878
Macro avg	0.55	0.54	0.55	76878
Weighted avg	0.87	0.87	0.87	76878

4.4. Стекінг

Стекінг, як дворівневий ансамблевий підхід, став однією з найефективніших моделей дослідження (табл. 4). Метамодель об'єднала прогнози логістичної регресії, LightGBM та MLP, навчившись використовувати їх сильні сторони. Це дозволило суттєво знизити кількість помилок класифікації та досягти найвищих значень AUC серед усіх алгоритмів. Стекінг продемонстрував найкраще узагальнення на тестовій вибірці.

Таблиця 4. Метрики стекінгу

	Precision	Recall	F1-score	Support
Class 0	0.92	0.99	0.96	70687
Class 1	0.48	0.06	0.10	6191
Accuracy	-	-	0.92	76878
Macro avg	0.70	0.53	0.53	76878
Weighted avg	0.89	0.92	0.89	76878

4.5. Блендинг

Блендинг показав результати, близькі до стекінгу, використовуючи простішу схему об'єднання прогнозів на окремій валідаційній частині даних (табл. 5). Метод виявився ефективним і менш вимогливим до ресурсів, забезпечивши хороший баланс між точністю та обчислювальною простотою. Він успішно усунув частину недоліків окремих моделей і показав стабільну роботу.

Таблиця 5. Метрики блендингу

	Precision	Recall	F1-score	Support
Class 0	0.93	0.94	0.93	70687
Class 1	0.17	0.15	0.16	6191
Accuracy	-	-	0.87	76878
Macro avg	0.55	0.54	0.55	76878
Weighted avg	0.87	0.87	0.87	76878

4.6. Гібридна модель (GBDT + LR)

Гібридна модель поєднала деревоподібний підхід LightGBM із лінійною логістичною регресією. Використання leaf-фіч дозволило побудувати новий простір ознак, що відображає структуру рішень дерев. Це забезпечило чіткіше розмежування класів та високу точність при збереженні інтерпретованості моделі. За результатами дослідження, гібридний підхід став одним із найстабільніших і найзбалансованіших методів (табл. 6).

Таблиця 6. Метрики гібридної моделі

	Precision	Recall	F1-score	Support
Class 0	0.92	0.99	0.95	70687
Class 1	0.34	0.07	0.12	6191
Accuracy	-	-	0.91	76878
Macro avg	0.63	0.53	0.54	76878
Weighted avg	0.88	0.91	0.89	76878

5. ВИСНОВКИ

У роботі створено інформаційну систему для оцінювання кредитних ризиків позичальників із використанням сучасних методів машинного навчання. Проведені дослідження показали, що окремі моделі забезпечують різний рівень точності: логістична регресія слугує інтерпретованою базою, LightGBM демонструє високу якість за рахунок деревоподібної структури, а нейронна мережа добре працює з нелінійними залежностями.

Найкращих результатів вдалося досягти за допомогою ансамблевих методів. Стекінг та гібридна модель GBDT+LR показали найвищі значення AUC і найменшу кількість помилок класифікації, що підтверджує ефективність комбінування різних алгоритмів. Блендинг також продемонстрував стабільні результати при нижчій складності реалізації.

Розроблений інтерфейс системи забезпечив зручну взаємодію з моделями, можливість налаштовувати параметри та аналізувати результати без необхідності втручання в програмний код.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow. / A. Géron; O’Reilly Media, USA, 2023. 850 p.
2. Brownlee, J. Ensemble Learning Algorithms with Python. / J. Brownlee; Machine Learning Mastery, Australia, 2022. 290 p.
3. Zhang, C.; Ma, Y. Ensemble Machine Learning: Methods and Applications. / C. Zhang, Y. Ma; Springer, London, 2012. 360 p.
4. Friedman, J.; Hastie, T.; Tibshirani, R. The Elements of Statistical Learning. / Springer, New York, 2017. 745 p.
5. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. / Proceedings of the 22nd ACM SIGKDD Conference; San Francisco, USA, 2016. 13 p.
6. Ke, G.; Meng, Q.; Finley, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. / NIPS Proceedings; California, USA, 2017. 15 p.
7. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. / Journal of Machine Learning Research, Vol. 12, 2011. p. 2825-2830.

8. Thomas, L. C.; Edelman, D. B.; Crook, J. N. Credit Scoring and Its Applications. / SIAM, Philadelphia, USA, 2017. 384 p.

9. Abdou, H. A.; Pointon, J. Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review. / Intelligent Systems in Accounting, Finance and Management; Vol. 18, 2011. p. 59–88.

10. 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, Vol. 247, 2015. p. 124–136.

11. Marqués, A. I.; García, V.; Sánchez, J. S. Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles. / Expert Systems with Applications, Vol. 39, 2012. p. 10244–10250.

12. Jadhav, S.; Channe, H. Comparative Study of KNN, Naive Bayes and Decision Tree Classification Techniques. / International Journal of Science and Research, Vol. 5 (1), 2016. p. 1842–1845.