“Building the ensembles of credit scoring models”

AUTHORS	Halyna Velykoivanenko https://orcid.org/0000-0001-6326-3965 Svitlana Savina https://orcid.org/0000-0003-0227-7081 http://www.researcherid.com/rid/B-3367-2019 Dmitriy Kolechko https://orcid.org/0000-0002-3515-0514 Vladyslav Ben’

ARTICLE INFO	Halyna Velykoivanenko, Svitlana Savina, Dmitriy Kolechko and Vladyslav Ben’ (2018). Building the ensembles of credit scoring models. *Neuro-Fuzzy Modeling Techniques in Economics*, *7*(1), 21-43. doi:10.21511/nfmte.7.2018.02
DOI	http://dx.doi.org/10.21511/nfmte.7.2018.02
RELEASED ON	Wednesday, 10 April 2019
RECEIVED ON	Wednesday, 04 April 2018
ACCEPTED ON	Wednesday, 19 September 2018

JOURNAL	"Neuro-Fuzzy Modeling Techniques in Economics"
ISSN PRINT	2306-3289
ISSN ONLINE	2415-3516
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	State Higher Educational Establishment "Kyiv National Economic University named after Vadym Hetman"

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
15	6	11

BUSINESS PERSPECTIVES

Publisher:

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine

www.businessperspectives.org

V. HETMAN KNEU

Founder:

State Higher Educational Establishment "Kyiv National Economic University named after Vadym Hetman",
Prospect Peremogy, 54/1,
Kyiv, 03057, Ukraine

https://kneu.edu.ua/

Received on: 4th of April, 2018
Accepted on: 19th of September, 2018

Halyna Velykoivanenko, Ph.D, Professor, Head of Department of Economic and Mathematical Modeling, State Higher Educational Establishment "Kyiv National Economic University named after Vadym Hetman", Ukraine.

Svitlana Savina, Ph.D, Docent, Associate Professor of Department of Economic and Mathematical Modeling, V. Hetman State Higher Educational Establishment "Kyiv National Economic University named after Vadym Hetman", Ukraine.

Dmitriy Kolechko, Member of Management Board, Chief Risk Officer, VPBank, Vietnam.

Vladyslav Ben', Leading Specialist, "MOTOR SICH" JSC, Ukraine.

Halyna Velykoivanenko (Ukraine), **Svitlana Savina** (Ukraine), **Dmitriy Kolechko** (Vietnam), **Vladyslav Ben'** (Ukraine)

BUILDING THE ENSEMBLES OF CREDIT SCORING MODELS

Abstract

The article is devoted to solving the actual problem of increasing the efficiency of assessing the credit risks of individual borrowers by finding the optimal combination of the results of calculations of specific scoring models. The principles of the formation of an ensemble of models are given and the existing approaches to the construction of ensemble structures are analyzed. In the process of experimental research has been applied one of the modifications of the boosting algorithm and implemented the author's algorithm for constructing an ensemble of models based on the specialization of experts. The radial-basis function neural networks were used as specific expert models. As a result of a comparative analysis of the efficiency of the used ensemble technologies it was confirmed that the algorithm for constructing an ensemble based on the specialization of experts proposed by the authors is the most adapted for the task of assessing credit risk.

Keywords credit risk, scoring model, radial-basis function (RBF) network, ensemble (committee) of models, boosting

JEL Classification C45, D81, G21

Г.І. Великоіваненко (Україна), **С.С. Савіна** (Україна), **Д.В. Колечко** (В'єтнам), **В.П. Бень** (Україна)

ПОБУДОВА АНСАМБЛІВ МОДЕЛЕЙ КРЕДИТНОГО СКОРИНГУ

Анотація

Стаття присвячена вирішенню актуального завдання підвищення ефективності оцінювання кредитних ризиків позичальників-фізичних осіб шляхом пошуку оптимального поєднання результатів розрахунків окремих скорингових моделей. Наведено принципи формування ансамблів моделей та проаналізовано існуючі підходи побудови ансамблевих структур. У процесі експериментального дослідження застосовано одну з модифікацій алгоритму бустінгу та реалізовано авторський алгоритм побудови ансамблю моделей на основі спеціалізації експертів. У якості окремих моделей-експертів використовувались нейромережі радіально-базисної архітектури. В результаті проведення порівняльного аналізу ефективності використаних ансамблевих технологій визначено, що найбільш адаптованою для задачі оцінювання кредитного ризику є запропонований авторами алгоритм побудови ансамблю на основі спеціалізації експертів.

Ключові слова кредитний ризик, скорингова модель, радіально-базисна нейронна мережа, ансамбль (комітет) моделей, бустінг

Класифікація JEL C45, D81, G21

ВСТУП

У процесі провадження банківської діяльності генеруються значні обсяги даних. Так, банки зберігають по кожному клієнту анкетні дані, кредитну історію, історію спілкування, фото та відеоматеріали тощо. Крім власної інформації банківські установи часто отримують дані від бюро кредитних історій, операторів мобільного зв'язку, проводять моніторинг інтернет-активності клієнтів. Накопичення настільки різноманітних даних неодмінно призводить до проблем, пов'язаних з Big Data, коли виникає

потреба врахування значної кількості суттєво неоднорідної інформації. Зокрема, при вирішенні одної з найактуальніших задач банківської діяльності – розробки системи скорингової оцінки кредитоспроможності позичальників – кількість характеристичних показників часто може сягати сотень і тисяч.

Проте, класичні рекомендації до побудови математичних моделей збігаються у висновку щодо доцільності застосування обмеженої кількості вхідних факторів. Міллер обґрунтовує оптимальну кількість входів моделей 7 ± 2 [8]. Так, якщо будується економетрична модель (а стандартом при побудові скорингових моделей у банківській сфері є логістичні регресії), то перевищення кількості вхідних факторів понад десять майже напевно призведе до прояву негативного явища мультиколінеарності. І хоча для нелінійних моделей (наприклад, нейронних мереж, які будуть використані у даній статті) це явище не несе прямих загроз, зростання кількості пояснюючих показників зменшує вплив кожного з них окремо на значення результативної змінної. За приблизно однакових результатів моделювання (точності відтворення вихідної статистики, прогнозування на незалежній тестовій вибірці та стійкості результатів прогнозування на різних вибірках) необхідно для подальшого використання брати ту економіко-математичну модель, яка має більш просту структуру.

Якщо ж виникає необхідність врахування значної кількості факторів, то доречно модель зробити ієрархічною і пояснюючі змінні розподілити за різними узагальнюючими групами показників або джерелами інформації. Узагальнення результатів розрахунків отриманих моделей доцільно здійснити за допомогою ансамблевих технологій – одного з напрямків машинного навчання, що є ефективним засобом дослідження *Big Data* [2]. За такого підходу кожна з моделей може бути налаштована на масивах однорідних даних і їх поєднання забезпечить коректне врахування всієї необхідної інформації.

Аналогічні міркування справедливі і стосовно обсягів спостережень у початковій сукупності даних. Так, складно розраховувати на ефективне налаштування однієї моделі на навчальній вибірці, що складається із сотень тисяч записів. Така модель не зможе відслідковувати усі наявні закономірності, оскільки вибірка міститиме інформацію, яка напевно буде поєднувати протилежні тенденції. В результаті цього зростатиме похибка агрегування, адже розрахунок такої моделі буде надто усередненим. Для уникнення подібних пасток моделювання доцільно розбити початковий масив даних на підвибірки, які утворюватимуться з прикладів із однотипними тенденціями поведінки. Для кожної такої вибірки будується окрема модель, узагальнення розрахунків яких реалізується шляхом використання ансамблів. На доцільності створення ансамблів наголошує і Тєрєхов [14], який теоретично обґрунтовує та експериментально демонструє вищу точність вирішення задач класифікації на великих обсягах даних при об'єднанні розрахунків кількох моделей, побудованих для окремих сегментів даних, порівняно із застосуванням одного глобального класифікатора.

1. ЛІТЕРАТУРНИЙ ОГЛЯД

Історично першим дослідженням, присвяченим розробці ансамблів моделей, вважається праця Шапайра 1990 року [13]. У цій статті викладено ідею бустінгу – алгоритму, що дає змогу підвищити точність класифікації однієї моделі. Подальші розробки в цьому напрямі Фрюнда та Шапайра 1997 року [3] привели до винаходу більш ефективної реалізації даного алгоритму, який отримав назву AdaBoost.

Надалі дослідження з розвитку ансамблевих технологій проводились у напрямках пошуку можливостей підвищення ефективності ансамблю. Зокрема, досліджувалась доцільність використання окремих видів моделей, наприклад, різних типів нейромереж. Розглядались різноманітні методи узагальнення результатів розрахунків моделей з метою мінімізації похибки роботи ансамблю на навчальних і тестових даних. Один із найпопулярніших на сьогодні алгоритмів побудови ансамблю моделей, що застосовується для розв'язання задач класифікації, запропоновано Брейманом у праці [1]. Однак, у подібних дослідженнях розглядаються загальні випадки вирішення задач моделювання (безвідносно моделей кредитного скорингу), тому вони мають більш теоретичне спрямування з огляду на мету нашого дослідження.

http://dx.doi.org/10.21511/nfmte.7.2018.02

Лише останнім часом ансамблі моделей почали використовуватись для проведення скорингових оцінок, через що кількість публікацій за даною тематикою досить обмежена. Зокрема, у праці Кузнєцова та Кірєєва [5] описано процедуру розробки ансамблю для розв'язання задачі поведінкового скорингу фізичних осіб, наведено основні аспекти, що впливають на підвищення точності ансамблевих структур, та розглянуто реалізацію одного з них, а саме – метод узагальнення результатів розрахунків окремих моделей комітету. Відкритими для подальших досліджень залишаються інші питання конструювання ансамблевих архітектур, зокрема формування навчальних вибірок для побудови усіх моделей ансамблю.

2. МЕТА ДОСЛІДЖЕННЯ

Метою роботи є вирішення актуального завдання підвищення ефективності оцінювання кредитних ризиків позичальників-фізичних осіб за рахунок пошуку механізму оптимального поєднання результатів розрахунків окремих скорингових моделей.

Досягнення цієї мети зумовлює необхідність вирішення таких завдань дослідження:

- проведення критичного аналізу основних підходів до оцінювання кредитоспроможності позичальників банків, виявлення їх недоліків та переваг,
- збір та аналіз інформаційної бази для розробки математичних моделей оцінки кредитного ризику позичальників банку,
- побудова математичних моделей оцінки кредитного ризику позичальників,
- дослідження різних підходів до формування ансамблів моделей та обґрунтування вибору раціональної ансамблевої структури для оцінювання кредитоспроможності позичальників,
- здійснення порівняльного аналізу результатів класифікації на основі застосування різних варіантів ансамблевих структур моделей.

3. ОБҐРУНТУВАННЯ ДОЦІЛЬНОСТІ ФОРМУВАННЯ АНСАМБЛІВ МОДЕЛЕЙ

При застосуванні ансамблів моделей одночасно з основною задачею розв'язується кілька більш простих задач. Це обумовлюється тим, що досліджуваний повний масив інформації поділяється на сегменти, для кожного з яких розробляється окрема модель, а результати розрахунків моделей об'єднуються визначеним способом.

Однак не очевидним є факт, що результат об'єднання кількох класифікаторів дасть кращу якість, ніж окрема модель. Для дослідження даного питання розглянемо описаний Тереховим спрощений випадок роботи комітету з трьох моделей (які далі називатимемо експертами), що розв'язують задачу бінарної класифікації [14]. Позначимо ймовірності коректної класифікації (точність класифікації) кожним із них через p_1, p_2, p_3. Ймовірності є незалежними. Нехай результат роботи комітету буде визначатись простим голосуванням, тобто певний приклад буде віднесено до того класу, який встановлено трьома чи двома експертами. Ймовірність правильної класифікації комітетом позначимо .

У результаті роботи такого комітету можливі такі варіанти: всі три експерти провели класифікацію коректно, один з трьох помилився, помилились два з трьох та помилились всі три експерти.

Правильне рішення буде прийняте комітетом у двох перших варіантах. Деталізуючи їх отримаємо чотири можливі ситуації для правильного рішення:

- усі три експерти не помилились – імовірність ситуації $\theta_1 = p_1.p_2.p_3$,
- помилився перший експерт, два інші провели класифікацію коректно – ймовірність ситуації $\theta_2 = (1-p_1).p_2.p_3$,

http://dx.doi.org/10.21511/nfmte.7.2018.02

- помилився другий експерт, два інші провели класифікацію правильно – ймовірність ситуації $\theta_3 = p_1.(1-p_2).p_3$,
- помилився третій експерт, два інші правильно провели класифікацію – ймовірність ситуації $\theta_4 = p_1.p_2.(1-p_3)$.

Ймовірність правильної класифікації комітетом буде складатись із сум імовірностей чотирьох описаних ситуацій:

$$P_K = \theta_1 + \theta_2 + \theta_3 + \theta_4 = p_1.p_2.p_3 + (1-p_1).p_2.p_3 + p_1.(1-p_2).p_3 + p_1.p_2.(1-p_3). \quad (1)$$

Якщо зафіксувати на певному рівні одну з імовірностей, наприклад $p_3 = const = C$, то функція (1) перетворюється у функцію двох змінних. Змінюючи значення константи, можна наочно відобразити поведінку функції P_K, досліджуючи таким чином результат роботи комітету моделей за різних значень якості класифікації окремими експертами.

На Рисунку 1 зображено вигляд поверхні, що відповідає виду функції P_K для випадку, коли $p_3 = C = 0.3$, а $0 \le p_1, p_2 \le 0.3$. За таких умов всі три експерти характеризуються низькою точністю класифікації, яка є гіршою, ніж можна було б отримати в результаті випадкового вибору (наприклад, при підкиданні монети).

На Рисунку 1 на поверхні виділено зони поступового зростання значення величини , починаючи від 0 до максимального значення 0.216. Таке максимальне значення функції відповідає ймовірності правильної класифікації комітетом моделей, коли значення ймовірностей усіх трьох експертів окремо дорівнюють 0.3 (але, як бачимо, загальна точність є нижчою, ніж у кожного окремого експерта).

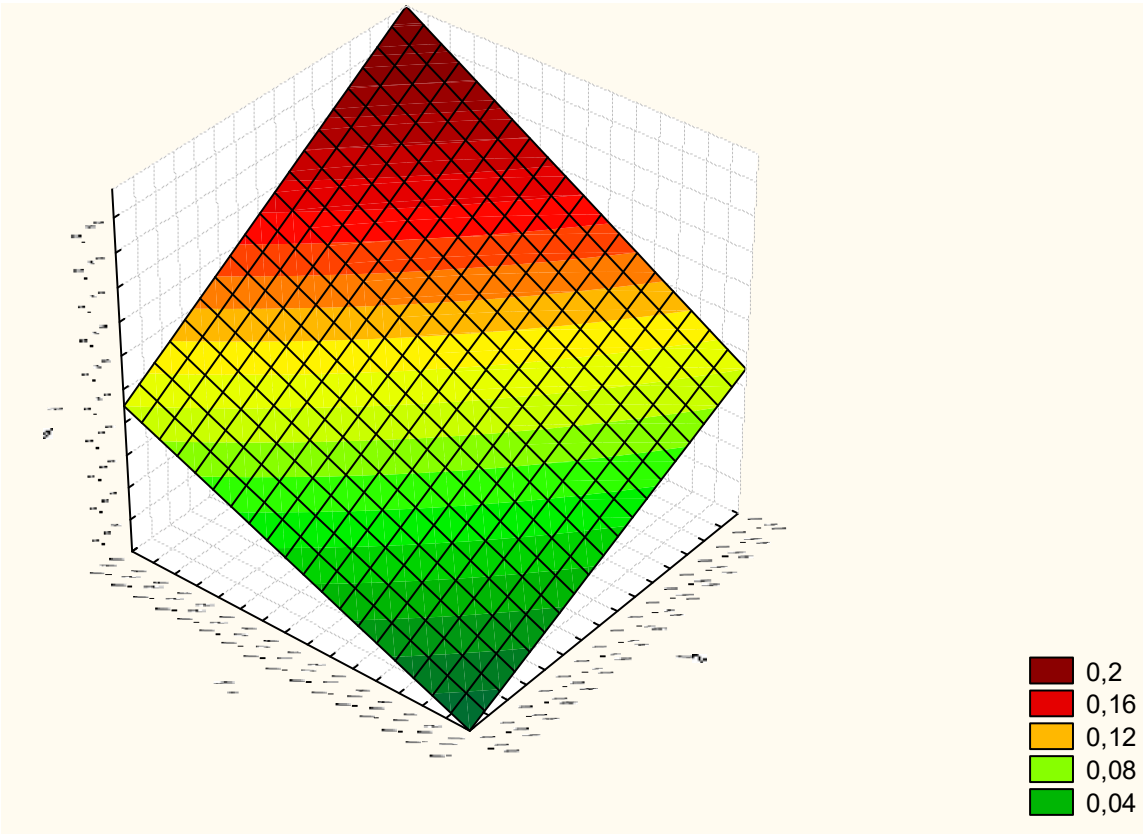

Рисунок 1. Вигляд поверхні при $0 \le p_1, p_2 \le 0.3; p_3 = 0.3$

http://dx.doi.org/10.21511/nfmte.7.2018.02

Дослідимо залежність імовірності правильної класифікації комітетом від точності класифікації двох моделей на повній множині значень імовірностей від 0 до 1 при фіксованій низькій точності третьої моделі на рівні 0.3. Так, у Таблиці 1 наведено значення, що приймає функція у точках з дискретними координатами при зміні від 0 до 1 з кроком 0.1 при $p_3 = 0.3$.

Таблиця 1. Значення ймовірності коректної класифікації комітетом моделей для випадку фіксованої ймовірності коректної класифікації одного з експертів

	Імовірність коректної класифікації другого експерта, p2											
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0	0	0.03	0.06	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.30
	0.1	0.03	0.06	0.10	0.13	0.17	0.20	0.23	0.27	0.30	0.34	0.37
	0.2	0.06	0.10	0.14	0.17	0.21	0.25	0.29	0.33	0.36	0.40	0.44
	0.3	0.09	0.13	0.17	0.22	0.26	0.30	0.34	0.38	0.43	0.47	0.51
Імовір-ність корек-тної класифі-кації першого експер-та, p1	0.4	0.12	0.17	0.21	0.26	0.30	0.35	0.40	0.44	0.49	0.53	0.58
	0.5	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65
	0.6	0.18	0.23	0.29	0.34	0.40	0.45	0.50	0.56	0.61	0.67	0.72
	0.7	0.21	0.27	0.33	0.38	0.44	0.50	0.56	0.62	0.67	0.73	0.79
	0.8	0.24	0.30	0.36	0.43	0.49	0.55	0.61	0.67	0.74	0.80	0.86
	0.9	0.27	0.34	0.40	0.47	0.53	0.60	0.67	0.73	0.80	0.86	0.93
	1.0	0.30	0.37	0.44	0.51	0.58	0.65	0.72	0.79	0.86	0.93	1.00

За вказаних умов поверхня функції P_K є опуклою вгору, що можна бачити за лініями переходу кольорів на Рисунку 1, а також за розрахунками, наведеними в Таблиці 1. Найбільших значень поверхня набуває вздовж лінії $p_1 = p_2$, але при цьому значення функції $P_K < p_1 = p_2$. Отже, комітет у цьому випадку дає нижчу точність оцінювання, ніж окремий експерт, що можна бачити за даними Таблиці 1. Така ситуація завжди характерна для роботи комітету, в якому хоча б одна з моделей має низьку точність класифікації. Перевищення точності оцінювання понад $P_K = 0.5$ такий комітет досягає лише за умови, що ймовірність коректної класифікації двома експертами з комітету буде не менше 0.6 (коли точність оцінювання третім експертом становить 0.3). А таку точність класифікації, яку має перший експерт, наприклад, на рівні 0.7, комітет моделей може продемонструвати лише при точності другого експерта 0.9.

На Рисунку 2 і 3 наведено випадки, коли ймовірність правильної класифікації третім експертом є вищою (p_3=0.5 та 0.7, відповідно).

На Рисунку 2 $p_3 = 0.5$ при $0 \le p_1, p_2 \le 0.5$. Вздовж лінії $p_1 = -p_2 + const$ при $p_3 = 0.5$ значення функції P_K є однаковими, як можемо бачити з даних, наведених у Таблиці 2. Крім того, рівні поверхні функції P_K зростають лінійно, що можна бачити за градієнтом переходу кольорів на Рисунку 2, тобто вона являє собою рівну площину. У даному випадку комітет лише не погіршує роботи окремих експертів.

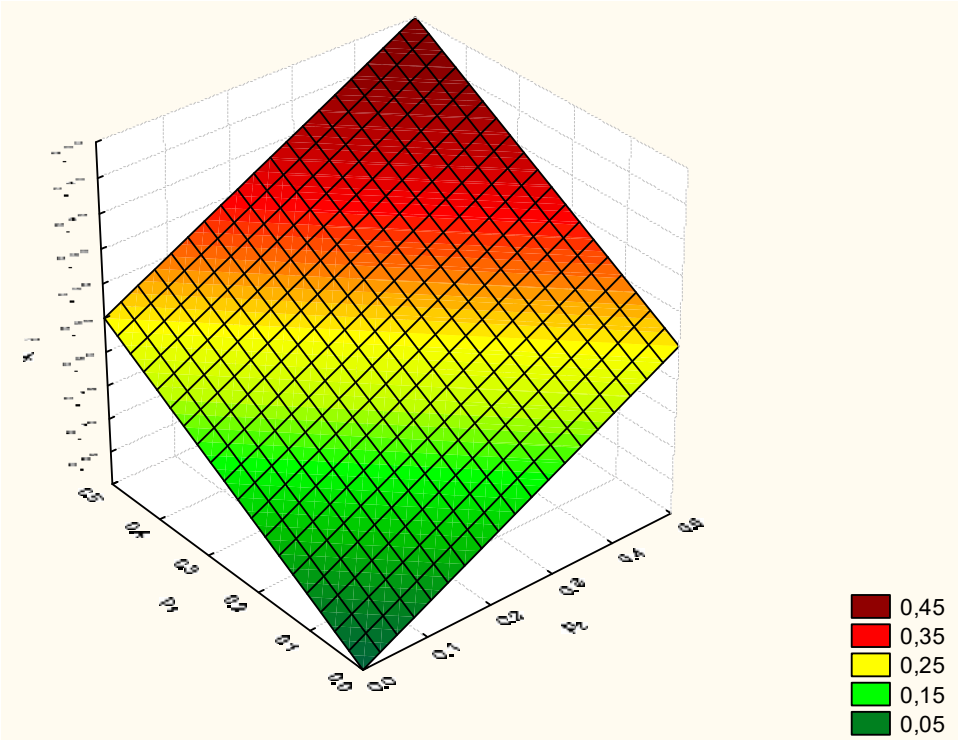

Рисунок 2. Вигляд поверхні при $0 \le p_1, p_2 \le 0.5; p_3 = 0.5$

Таблиця 2. Значення ймовірності коректної класифікації комітетом моделей для випадку ймовірності коректної класифікації третього експерта $p_3 = 0.5$

	Імовірність коректної класифікації другого експерта, p2											
Імовірність коректної класи-фікації першого експерта, p1		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	0	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
	0.1	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
	0.2	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
	0.3	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65
	0.4	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
	0.5	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75
	0.6	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80
	0.7	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
	0.8	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
	0.9	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
	1.0	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00

Рисунок 3, де $p_3 = 0.7$ при $0.5 \le p_1, p_2 \le 1$, відображає випадок, коли поверхня P_K є увігнутою. В даному випадку при $p_1 = p_2$ маємо $P_K > p_1 = p_2$. Отже, точність такого комітету дає результат кращий, ніж його окремі моделі.

http://dx.doi.org/10.21511/nfmte.7.2018.02

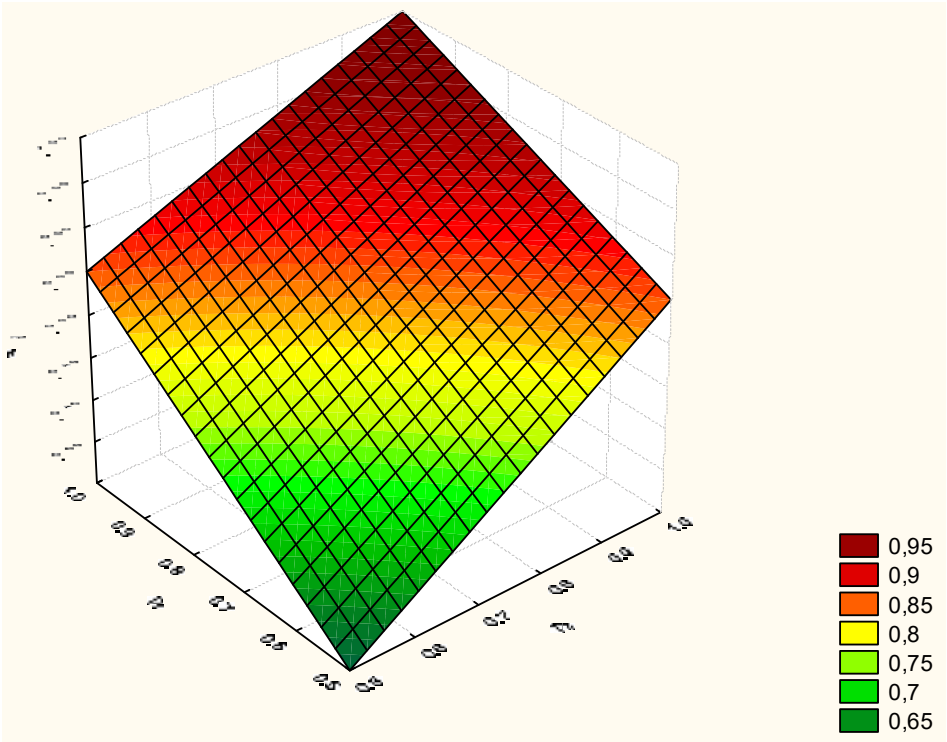

Рисунок 3. Вигляд поверхні P_K при $0,5 \le p_1, p_2 \le 1; p_3 = 0,7$

Рисунок 2 та 3 ілюструють головну умову доцільності створення комітету моделей: якщо точність хоч одного окремого експерта є нижчою за 0.5, то комітет за його участі стає менш ефективним у порівнянні з окремими моделями. Підвищення точності оцінювання комітетом моделей слід очікувати лише у випадку, коли ймовірність правильної класифікації є вищою за 0.5 для всіх експертів комітету [14].

Логіка представлених у тривимірному просторі поверхонь (Рисунок 1-3) може бути відображена в одному рисунку на площині (Рисунок 4).

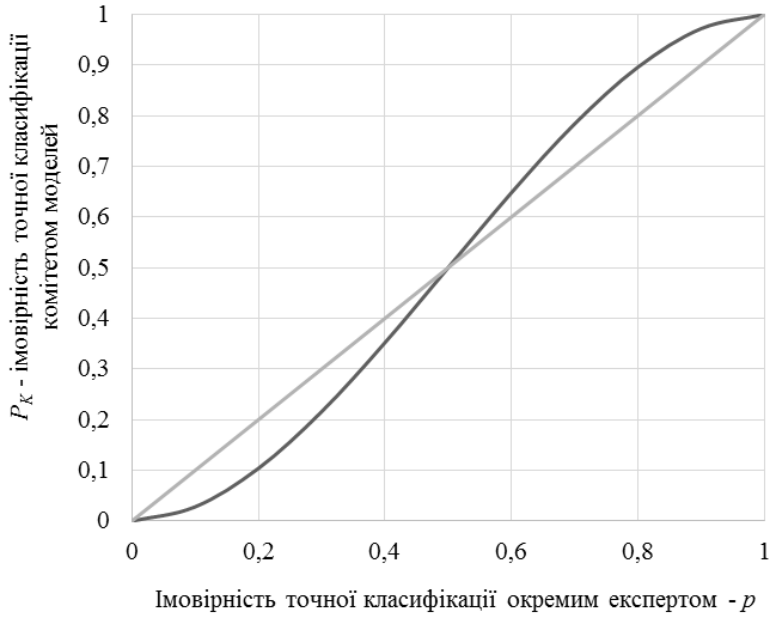

Рисунок 4. Вигляд залежності ймовірності точної класифікації комітетом моделей від точної класифікації окремим експертом p (побудовано з використанням [14])

http://dx.doi.org/10.21511/nfmte.7.2018.02

Крива на Рисунку 4 відображає ймовірність точної класифікації всім комітетом. За горизонтальною віссю відкладається точність класифікації одним експертом (для спрощення приймається, що у всіх експертів однакова ймовірність p вгадування класу кредиту позичальників). Як бачимо на Рисунку 4, при значеннях імовірності $p < 0.5$ точність комітету нижче бісектриси (тобто менше, ніж точність окремого експерта). У точці 0.5 крива лінія перетинає бісектрису, що вказує на рівність ймовірності точної класифікації комітетом та усіх трьох експертів. Далі крива проходить над бісектрисою – комітет демонструє вищу точність класифікації, ніж окремий експерт.

У праці [14] зазначається, що логічним розвитком ідеї створення комітетів для підвищення точності класифікації є формування комітетів із комітетів, однак це не приводить до суттєвого зростання точності таких систем. Комітети лише адаптуються до інформаційного шуму в даних навчальної вибірки, однак точність на нових прикладах спадає (проявляється ефект перенавчання). Тєрєхов у цій праці визначає дві умови зростання точності класифікації ансамблем моделей [14]:

- необхідно підвищити точність класифікації кожної окремої моделі ансамблю,
- потрібно досягти статистичної незалежності похибок моделей ансамблю.

4. ОСНОВНІ ВИДИ АНСАМБЛІВ

На сьогодні розроблено та описано значну кількість різноманітних видів ансамблів [1; 3-6; 9; 13; 14], які різняться за алгоритмами побудови. Деякі з них, наприклад бустінг, мають кілька модифікацій процесу побудови ансамблів та вже перетворились у окремі сімейства алгоритмів. Основні технології формування ансамблів та їх характеристики наведено в Таблиці 3.

Таблиця 3. Алгоритми формування ансамблів моделей та їх основні характеристики

Категорії ансамблів	**Статична інтеграція моделей ансамблю**				**Динамічна інтеграція моделей ансамблю**	
Наявність впливу на початкову щільність розподілу даних	Збереження початкової щільності розподілу даних		Зміна початкової щільності розподілу даних		Зміна початкової щільності розподілу даних	
Назва алгоритму формування ансамблю	Усереднення	Стекінг	Бустінг	Беггінг	Суміш думок експертів	Ієрархічне поєднання думок експертів
Технологія формування навчальної вибірки	Навчальна вибірка єдина для всіх моделей ансамблю, навчання проходить паралельно	Навчальна вибірка єдина для всіх моделей ансамблю, обов'язкове використання різних типів моделей	Для кожної наступної моделі ансамблю використовується нова навчальна вибірка, у якій змінюється розподіл даних, виходячи з результатів попередніх моделей	З однієї навчальної вибірки за допомогою спеціальної процедури випадкового вибору з поверненням штучно утворюються кілька навчальних вибірок для побудови моделей	Початкова вибірка поділяється на кластери, які є навчальними вибірками для всіх моделей ансамблю, зберігаються також оцінки помилок експертів на кожній навчальній вибірці	Початкова вибірка поділяється на кластери, які є навчальними вибірками для всіх моделей ансамблю, зберігаються також оцінки помилок експертів на кожній навчальній вибірці
Схема об'єднання результатів окремих моделей	Усереднення результатів по ансамблю	Застосування метамоделі	Усереднення результатів по ансамблю	Усереднення результатів або зважене усереднення по ансамблю	При класифікації нового об'єкта спочатку визначається найближчий до нього кластер, а потім використовуються ті моделі, які мали найменші помилки в даному кластері (найпростіший спосіб), або поєднання результатів моделей проводиться з вагами, які є виходом деякої керуючої метамоделі	Застосування керуючих метамоделей на двох рівнях ієрархії

http://dx.doi.org/10.21511/nfmte.7.2018.02

Як бачимо за даними цієї таблиці, взагалі виділяють дві категорії ансамблів: зі статичною та динамічною інтеграцією моделей. При статичній інтеграції дані прикладу, для якого проводиться класифікація, не впливають на процедуру об'єднання результатів розрахунків моделей ансамблю. А при динамічній інтеграції процедура об'єднання результатів кожен раз коригується для проведення класифікації кожного конкретного прикладу. Крім того, алгоритми побудови ансамблів розрізняють також за способами формування навчальних вибірок для окремих моделей та схемою узагальнення результатів їх роботи.

Технології, розроблені для категорії статичних ансамблів, доцільно використовувати при дослідженні однорідних даних. До цих технологій відносяться усереднення, стекінг, бустінг, беггінг.

Технологія усереднення результатів роботи моделей ансамблю є найпростішою. Вона передбачає, що всі моделі було побудовано на одній навчальній вибірці, а результат роботи комітету визначається простим голосуванням, тобто більшістю голосів.

Розташування алгоритму стекінгу в Таблиці 3 слід вважати дещо умовним, оскільки Паклін і Орєшков [9] зазначають, що загальна концепція використання даного методу відсутня, а його головна ідея знаходить свою реалізацію в різноманітних варіантах. Сутність даного алгоритму полягає в наступному:

1. для комбінування в ансамбль застосовуються моделі не одного, а різних типів – наприклад, нейромережа, дерево рішень та логістична регресія;
2. замість алгоритму голосування вводиться концепція метанавчання.

На вхід деякої метамоделі, яку називають моделлю першого рівня, подаються результати класифікації кожним експертом, які відносять до нульового рівня. Далі результати розрахунків моделі першого рівня передаються на модель другого рівня і так далі, доки не буде досягнуто певної умови зупинки процесу.

Найбільш розвиненими та поширеними алгоритмами формування комітетів моделей є бустінг та беггінг.

За алгоритмом бустінгу моделі ансамблю будуються послідовно таким чином, щоб кожна наступна модель проводила класифікацію тих прикладів, які не були коректно ідентифікованими моделями на попередніх кроках.

На відміну від бустінгу, за алгоритмом беггінгу моделі будуються паралельно та незалежно одна від одної [9]. Так, на даних вихідного масиву шляхом випадкового відбору формується декілька вибірок. Вони матимуть той самий розмір, що і початковий масив, однак за рахунок випадкового відбору набір прикладів у кожній вибірці буде різним – один і той самий приклад може потрапляти в одну вибірку як кілька разів, так і жодного разу. Далі для кожної вибірки будується окрема модель-експерт. Результати розрахунків узагальнюються голосуванням.

Для масивів неоднорідних даних окремі їх підмножини можуть краще описуватись різними моделями, тобто побудову моделей-експертів слід здійснювати не для всієї вибірки в цілому, а для окремих її частин. У таких випадках мова йде про спеціалізацію експертів. Для врахування спеціалізації при формуванні ансамблів використовуються технології динамічної інтеграції моделей.

Алгоритм побудови ансамблю для неоднорідних даних пропонується у праці [14]. За цим алгоритмом необхідно провести кластеризацію даних із застосуванням карт Кохонена і для кожного кластера обрати той набір моделей, який демонстрував для нього найвищу ефективність. Тоді для класифікації нового прикладу потрібно буде визначити найближчий кластер та застосувати відповідні моделі.

Авторами цієї статті було проведено дослідження [10; 11] з використанням ансамблю моделей для оцінки кредитоспроможності позичальників-фізичних осіб, який побудовано за алгоритмом усереднення результатів моделей. Для розрахунків використовувалась база даних з 2.175 спостережень, яка містить широкий набір характеристичних показників позичальників комерційного банку та відомостей про виконання ними зобов'язань за отриманими кредитами.

З метою відбору найбільш значущих чинників для оцінювання кредитоспроможності позичальників

було використано підхід, що заснований на поєднанні роботи ймовірнісної нейромережі та генетичного алгоритму. Також було застосовано два більш спрощених підходи на основі покрокового включення та покрокового виключення чинників з моделі. Узагальнення результатів моделювання за трьома підходами дозволило виділити з множини початкових пояснюючих чинників шість кількісних (вік, стаж на останньому місці роботи, загальний стаж, наявність депозитів, наявність виплачених у минулому кредитів і кількість дітей у сім'ї) та два якісних (рівень освіти та статус працюючого), які було обрано як вхідні змінні моделей оцінювання кредитоспроможності позичальників.

На основі цих факторів у праці [11] було сформовано ансамбль з трьох нейромереж: дві мережі з радіально-базисною архітектурою (82 та 124 нейрони на прихованому шарі) та тришаровий персептрон з 6 нейронами проміжного шару. Всі моделі мали єдину навчальну вибірку, їх навчання проходило незалежно, загальний розрахунок здійснювався за принципом простого голосування. Межі зміни точності класифікації отриманим ансамблем та його окремими моделями в залежності від умов навчання наведено в Таблиці 4.

Таблиця 4. Межі зміни точності класифікації ансамблем та його окремими моделями залежно від умов навчання

Вид класифікатора	Відсоток правильно класифікованих спостережень	
	у навчальній вибірці, %	у тестовій вибірці, %
Ансамбль моделей	60.7-68.5	47.5-52.3
Окремі нейромережі	55.4-71.9	43.7-55.2

Як бачимо з даних, наведених у Таблиці 4, точність класифікації комітетом моделей не перевищує точності найбільш ефективних за даних умов моделей. Однак зміна умов експерименту приводить до більших коливань в ефективності окремих моделей, ніж сформованого з них комітету.

Зазначимо, що невисокі показники точності класифікації обумовлюються специфікою застосованого у дослідженні пакету *STATISTICA*, в якому границя поділу між класами встановлена на рівні 0.5 (вищі значення розрахунку моделі інтерпретуються як такі, що належать до класу «1», а нижчі значення – до «0»). Однак, проведені нами численні експериментальні розрахунки із різними типами моделей свідчать, що границя раціонального поділу між класами (коли альфа- і бета-помилки моделювання близькі між собою та загальна точність класифікації набуває максимального значення) є дещо вищою – в середньому на рівні 0.56. Тобто, у разі оптимізації рівня поділу класів ці самі моделі демонстрували б більш високі показники ефективності. Однак, у даному дослідженні не було завдання максимального підвищення точності класифікації окремими моделями – необхідно дослідити можливість збільшення ефективності моделювання за рахунок поєднання моделей в ансамблі. І, враховуючи, що зміщення границі поділу між класами має систематичний характер, то висновки, отримані для таких комітетів, будуть релевантними і при їх утворенні з моделей із оптимізованими рівнями поділу класів.

Зауважимо, що база даних, використана у попередніх дослідженнях, та сформований у праці [10] набір показників застосовуватиметься і у цій роботі з метою забезпечення порівнянності результатів моделювання та розвитком авторського підходу до створення ансамблів скорингових моделей оцінювання кредитних ризиків позичальників-фізичних осіб.

5. ЗАСТОСУВАННЯ АЛГОРИТМУ БУСТІНГУ

Як зазначалось вище, однією з умов підвищення ефективності роботи комітету є незалежність похибок його окремих моделей. Для забезпечення виконання цієї умови розроблено алгоритм бустінгу (підсилення). В його основі лежить ідея послідовного навчання експертів, кожен з яких не зможе повторити помилки попереднього, оскільки буде навчатись на іншому масиві даних. Однаковим для всіх експертів є лише обсяг навчальної вибірки.

http://dx.doi.org/10.21511/nfmte.7.2018.02

Для реалізації бустінгу використовувалась схема алгоритму, запропонована у праці Царєгородцева [15]. Процедура застосування бустінгу полягає в послідовному навчанні обраної базової моделі з метою підвищення її якості. Базовою називається модель певної конфігурації, що реалізується у трьох варіантах, які відрізняються між собою за рахунок оптимізації на різних навчальних вибірках. Отримані варіанти базової моделі називаються експертами. Процедура формування навчальних вибірок та навчання експертів проходять послідовно.

Позначимо розмірність загального масиву початкових даних *M*. Для першого експерта з масиву даних *M* обирається випадковим чином вибірка розмірності *N*. На отриманих даних проводиться навчання базової моделі, яка і є втіленням першого експерта. За прикладами, які не потрапили до першої навчальної вибірки, здійснюється розрахунок імовірності дефолту першим експертом, за результатами якої формується навчальна вибірка для другого експерта. Ця вибірка також матиме розмірність *N* і складатиметься наполовину з елементів, які були правильно класифіковані першим експертом, а на другу половину – з елементів, які були класифіковані першим експертом некоректно.

Після навчання другого експерта дані, які не використовувались для формування двох перших навчальних вибірок, використовуються для формування третьої навчальної вибірки. Це здійснюється наступним чином. Перші два експерти проводять класифікацію нових прикладів і до третьої навчальної вибірки включають ті *N* прикладів, за якими два перші експерти дають протилежні результати класифікації. Як можемо бачити з описаної процедури, алгоритм бустінгу, викладений у праці [14], полягає у «послідовній фільтрації прикладів попередніми класифікаторами таким чином, що задача для кожного наступного класифікатора стає складнішою». Після навчання третього експерта комітет можна використовувати для проведення класифікації нових прикладів, узагальнюючи результати всіх експертів простим голосуванням.

З огляду на описану вище процедуру формування навчальних вибірок їх обсяг *N* не може встановлюватись за класичними підходами поділу даних для навчання та тестування моделей. Наприклад, неможливо під навчальну вибірку виділити 70% від загального масиву даних, залишивши 30% для тестування, оскільки після навчання кожного експерта його навчальна вибірка взагалі вилучається з дослідження. Крім того, слід враховувати необхідність отримати потрібну кількість даних з розбіжностями в класифікації першими двома експертами для формування навчальної вибірки третього експерта. Оскільки всі експерти представлені моделями одного типу, то, як правило, розбіжностей в класифікації на невеликих масивах даних є небагато. Тому обсяг навчальної вибірки для невеликих масивів початкових даних у кожному окремому випадку встановлюється емпірично.

У нашому дослідженні повна база даних складалась із 2.075 спостережень. Для вибору максимально можливого обсягу навчальної вибірки було проведено ряд експериментальних розрахунків. Реалізація алгоритму здійснювалась для навчальних вибірок у чотирьох варіантах: 100, 300, 400 та 500 спостережень. В останньому випадку (500 кредитів у навчальній вибірці) було вичерпано всі значення з розбіжностями в класифікації перших двох експертів, тобто збільшити навчальну вибірку понад 500 спостережень було неможливо. Отже, для наявного початкового масиву даних навчальна вибірка може містити не більше 24%.

Деталізовану схему алгоритму бустінгу для досліджуваних даних проілюстровано на Рисунку 5.

За базову модель для всіх варіантів розрахунків було обрано багатошарову нейромережу, оскільки в попередніх дослідженнях моделі такої архітектури демонстрували вищі показники ефективності на наявних даних. Питання вибору типу та конфігурації нейромережі для кожного варіанту навчальної вибірки вирішувалось в результаті проведення експериментальних розрахунків. Розглядалось кілька варіантів конфігурацій нейромереж, які демонстрували найвищі показники точності класифікації. Перевага надавалась моделі, яка мала меншу кількість нейронів на прихованому шарі та показувала вищу точність класифікації на тестовому масиві даних. Було виявлено, що при невеликих розмірах навчальних вибірок доцільно використовувати радіально-базисні мережі з меншою кількістю нейронів проміжного шару. В Таблиці 5 наведено показники ефективності трьох конфігурацій нейромереж, які

продемонстрували найвищу точність класифікації для навчальної вибірки з 500 спостережень.Таблиця 5. Розрахунки ефективності нейромереж при виборі параметрів базової моделі для навчальної вибірки 500 спостережень

Архітектура мережі, кількість входів, кількість нейронів проміжного шару	Відсоток правильно класифікованих спостережень у навчальній вибірці, %	Відсоток правильно класифікованих спостережень у тестовій вибірці, %	Узагальнені дані по всьому масиву (без поділу на навчальну та тестову вибірки)		Правильно класифікованих, %
			Клас	Всього	
Радіально-базисна, 6, 33	75.6	52.8	0	1.131	55.1
			1	1.044	61.3
Радіально-базисна, 6, 50	77.4	51.4	0	1.131	54.5
			1	1.044	60.5
Радіально-базисна, 6, 76	79.2	52.0	0	1.131	55.3
			1	1.044	61.5

На основі даних, наведених у Таблиці 5 можна зробити висновок, що обраний обсяг навчальної вибірки не надає можливості якісного навчання моделі. Про це свідчить значна розбіжність між показниками точності класифікації для навчальної (понад 75%) та тестової (не перевищує 53%) вибірок. Тобто модель, налаштована на невеликій навчальній вибірці, не узагальнює в достатній мірі закономірності поведінки позичальників для ефективного моделювання усього різноманіття варіантів із тестової вибірки. Для підвищення точності класифікації доцільно було б збільшити обсяг навчальної вибірки. Однак, як

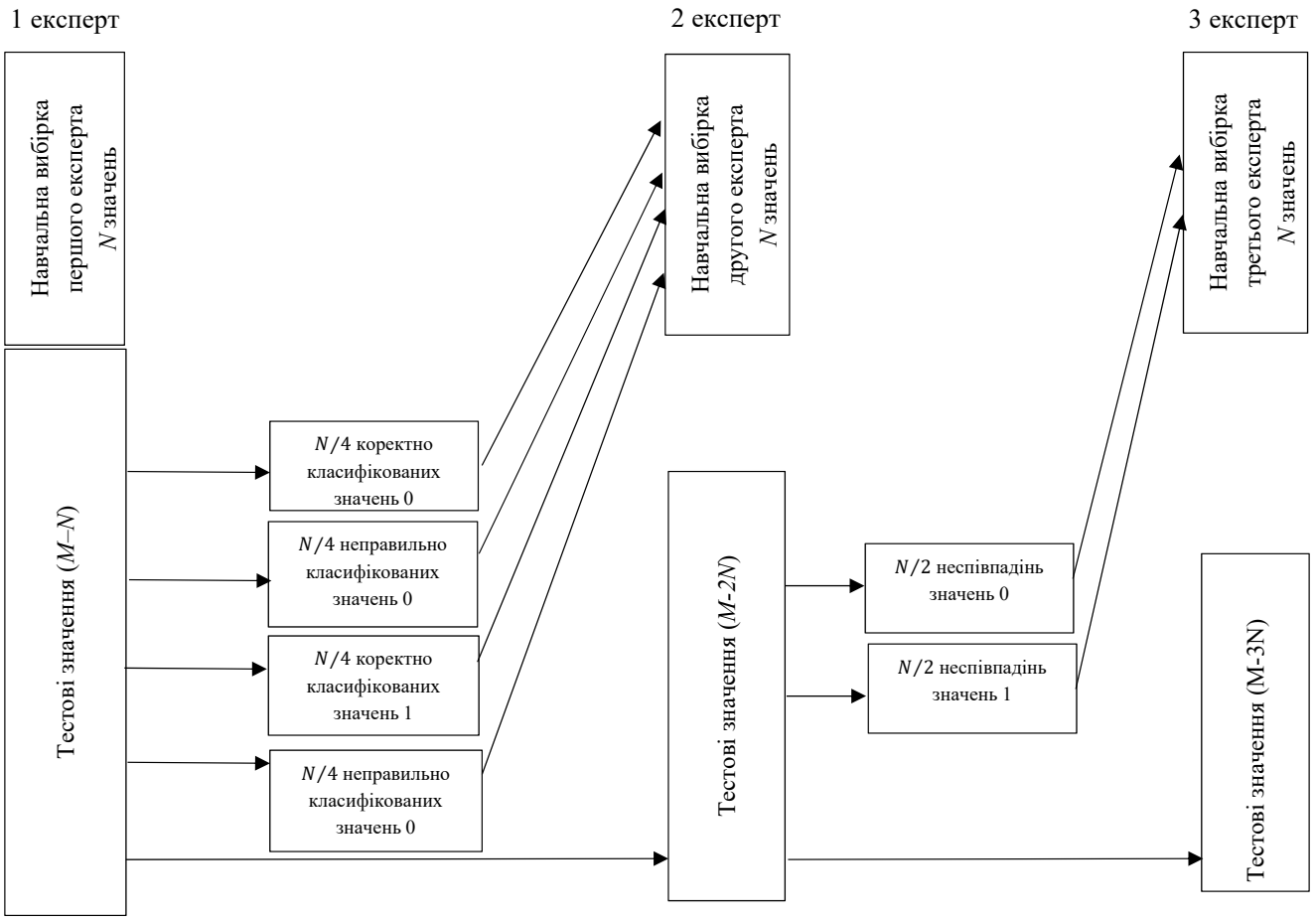

Рисунок 5. Схема алгоритму бустінгу для навчальної вибірки з N значень

http://dx.doi.org/10.21511/nfmte.7.2018.02

наголошувалось вище, у цьому дослідженні такої можливості не було через незначний обсяг наявних даних.

На основі даних, викладених у Таблиці 5, базовою моделлю для навчальної вибірки з 500 спостережень було обрано радіально-базисну нейромережу з 33 нейронами проміжного шару. Така мережа є кращою за показниками ефективності на тестовий вибірці та за узагальненими даними по всьому масиву наявних прикладів.

Аналогічним чином було обрано конфігурацію мереж, які реалізують першого експерта, для трьох інших варіантів навчальних вибірок. Параметри обраних нейромереж та показники їх ефективності наведено в Таблиці 6.

Таблиця 6. Параметри нейромереж та показники ефективності базових моделей першого експерта для трьох варіантів навчальної вибірки

Обсяг навчальної вибірки	Архітектура мережі, кількість входів, кількість нейронів проміжного шару	Відсоток правильно класифікованих спостережень у навчальній вибірці, %	Відсоток правильно класифікованих спостережень у тестовій вибірці, %
100	Радіально-базисна, 6, 25	69.0	51.5
300	Радіально-базисна, 6, 35	82.7	52.8
400	Радіально-базисна, 6, 38	76.8	53.9
500	Радіально-базисна, 6, 33	75.6	52.8

Аналіз даних, наведених у Таблиці 6, вказує на значну розбіжність між показниками правильно класифікованих спостережень на навчальній та тестовій вибірці. Збільшення обсягу навчальної вибірки від 100 до 500 прикладів майже не вплинуло на ефективність класифікації на тестових даних. У даному випадку всі варіанти формування навчальної вибірки не дають змоги здійснити побудову ефективної базової моделі.

Подальша процедура реалізації алгоритму бустінгу використовує обрані базові моделі на нових навчальних вибірках для формування двох наступних експертів ансамблю. Зміни ефективності нових експертів для кожного варіанту базової моделі, що відповідають різним обсягам навчальної вибірки, наведено в Таблиці 7.

Таблиця 7. Зміни ефективності базових моделей для другого та третього експертів при трьох варіантах навчальної вибірки

Архітектура мережі, кількість входів, кількість нейронів проміжного шару	Обсяг навчальної вибірки	Другий експерт			Третій експерт		
		Відсоток правильно класифікованих спостережень у навчальній вибірці, %	Відсоток правильно класифікованих спостережень у тестовій вибірці, %	Обсяг тестових вибірок	Відсоток правильно класифікованих спостережень у навчальній вибірці, %	Відсоток правильно класифікованих спостережень у тестовій вибірці, %	Обсяг тестових вибірок
Радіально-базисна, 6, 25	100	62.0	55.5	1.975	62.0	56.6	1.875
Радіально-базисна, 6, 35	300	61.7	54.4	1.575	58.3	57.2	1.275
Радіально-базисна, 6, 38	400	57.3	55.5	1.375	56.5	58.9	975
Радіально-базисна, 6, 33	500	62.0	55.1	1.175	55.8	58.8	675

Завдання класифікації для наступних двох експертів є складнішим, ніж для першого. Тому всі базові

моделі демонструють зниження точності класифікації на навчальній вибірці, як можемо бачити в Таблиці 7. Проте на тестових даних спостерігається зростання ефективності моделювання. Отже, в процесі реалізації алгоритму бустінгу навчальна вибірка стає різноманітнішою та набуває прикладів з відмінними між собою характеристиками (у тому числі такими, що погано ідентифікувались попередніми експертами), що приводить до більш ефективного узагальнення властивостей та покращення здатності експертів розрізняти класи на нових даних.

Алгоритм бустінгу було винайдено для можливостей підсилення слабких моделей. Отже, показником ефективності застосування даного алгоритму є порівняння точності класифікації трьох експертів (базової моделі, побудованої на трьох різних вибірках) та комітету, сформованого на їх основі. Оскільки всі приклади, що входили до навчальних вибірок, вилучались із подальших досліджень, то порівняння ефективності усіх експертів та комітету здійснювалось лише на даних останньої тестової вибірки для своєї базової моделі. Відповідні розрахунки показників ефективності у розрізі класів надійних та ненадійних позичальників наведено в Таблиці 8.

Таблиця 8. Порівняння ефективності першого, другого, третього експертів та комітету моделей при трьох варіантах навчальної та тестової вибірок

Базова модель	Відсоток правильно класифікованих спостережень першим експертом, %		Відсоток правильно класифікованих спостережень другим експертом, %		Відсоток правильно класифікованих спостережень третім експертом, %		Відсоток правильно класифікованих спостережень комітетом експертів, %	
	надійні позичальники	ненадійні позичальники	надійні п.	ненадійні п.	надійні п.	ненадійні п.	надійні п.	ненадійні п.
Радіально-базисна, 6, 25 (навчальна вибірка – 100, тестова вибірка – 1.875 спостережень)	49.3	54.2	53.3	58.4	53.6	60	53.6	60.8
Загальний відсоток правильно класифікованих спостережень, %	51.7		55.8		56.6		57.1	
Радіально-базисна, 6, 35 (навчальна вибірка – 300, тестова вибірка – 1.275 спостережень)	50.5	58.2	53.4	57.6	54.2	60.1	52.7	62.1
Загальний відсоток правильно класифікованих спостережень, %	54.1		55.4		57.2		57.1	
Радіально-базисна, 6, 38 (навчальна вибірка – 400, тестова вибірка – 975 спостережень)	49.5	66.2	55.2	60.8	56.1	62.2	53.3	63.1
Загальний відсоток правильно класифікованих спостережень, %	57.1		57.7		58.8		57.7	
Радіально-базисна, 6, 33 (навчальна вибірка – 500, тестова вибірка – 675 спостережень)	45.4	71.2	53.4	70.0	45.9	77.7	51.9	70.4
Загальний відсоток правильно класифікованих спостережень, %	55.9		60.1		58.8		59.4	

Як можна бачити з Таблиці 8, майже у всіх випадках експерти після перенавчання на нових вибірках підвищують точність класифікації. Виключення складає лише третій експерт за базовою моделлю із навчальною вибіркою з 500 спостережень. Таку невідповідність можна пояснити ще суттєвішим зміщенням границі раціонального поділу між класами від встановленого у пакеті *STATISTICA* рівня 0.5, що можна бачити за значно збільшеної різниці між альфа- та бета-помилками моделювання, яка, в свою

http://dx.doi.org/10.21511/nfmte.7.2018.02

чергу, впливає на зниження загальної точності класифікації. Крім того, на відміну від попередніх випадків, при формуванні навчальної вибірки для цього експерта було включено всі приклади із розбіжностями в моделюванні ненадійних позичальників першими двома експертами. Відповідно, у тестовій вибірці для перевірки адекватності четвертого комітету моделей не залишається прикладів ненадійних позичальників для здійснення розрахунків за третім експертом. Через це висока ефективність ідентифікації цією моделлю ненадійних клієнтів не враховується при їх розпізнаванні комітетом (саме тому в Таблиці 8 точність класифікації ненадійних позичальників комітетом за четвертою базовою моделлю знаходиться посередині між першими двома експертами), що дещо знижує загальний показник ефективності цього комітету. Однак, навіть незважаючи на це, всі чотири варіанти утворених комітетів демонструють підсилення ефективності базових моделей – точність класифікації комітетом або перевищує кожного з експертів за відповідною базовою моделлю, або є близькою до найкращої з них.

Різновиди алгоритмів бустінгу (наприклад, адаптивного бустінгу AdaBoost) надають можливість створювати більш ефективні комітети моделей. В AdaBoost використовується спеціальна процедура збільшення в навчальних вибірках кількості прикладів, на яких були допущені помилки попередніми експертами. Також при узагальненні результатів за цим алгоритмі враховуються відносні ваги експертів комітету, які визначаються в залежності від їх помилок (вага зменшується при збільшенні похибки класифікації). Для такого варіанту алгоритму Шапайра [12] доведено, що похибка комітету буде зменшуватись експоненційно, якщо похибки окремих експертів менші 0.5.

6. АНСАМБЛЬ НА ОСНОВІ СПЕЦІАЛІЗАЦІЇ ЕКСПЕРТІВ

При моделюванні кредитних ризиків важливо враховувати суттєву неоднорідність досліджуваних даних. Так, клієнти кредитних установ мають значні відмінності за показниками, які використовуються для оцінювання кредитоспроможності. Наприклад, вікова категорія позичальників може змінюватись в межах від 20 до 60 і більше років, рівень освіти – від середньої до наявності двох чи більше вищих освіт тощо. І різні категорії позичальників характеризуються різним рівнем кредитного ризику. Тому виділення з усього наявного масиву даних більш однорідної вибірки, наприклад із позичальників лише з вищою освітою, дасть змогу досліджувати таку їх групу, яка характеризується більш-менш подібними соціально-економічними умовами існування, що дозволить ефективніше виявляти закономірності поведінки позичальників.

Для обробки таких даних доцільно застосовувати підходи до утворення ансамблів, що враховують спеціалізацію експертів. Такі ансамблі відносяться до категорії динамічного об'єднання моделей. У Тєрєхова [14] було описано одну з ідей створення ансамблю даної категорії, яка базується на застосуванні карт Кохонена. Такий підхід вимагає додаткових досліджень щодо вибору параметрів карти, які впливатимуть на результат кластеризації. Універсальних рекомендацій з даного питання немає, тож розмір карт самоорганізації визначається індивідуально для кожної задачі, що обґрунтовано у праці Матвійчука [7]. Крім того, поділ масиву даних на кластери призведе до ще більшого зменшення обсягів навчальних вибірок, що дасть негативний вплив на результати класифікації та може зменшити ефективність моделювання (у випадку невеликої початкової вибірки даних).

Однак реалізація ідеї спеціалізації експертів комітету є дуже привабливою для розв'язання задачі моделювання кредитних ризиків, тому для її втілення було вирішено випробувати підхід, за яким поділ наявних даних на окремі групи не буде базуватись на процедурі кластеризації.

Пропонується такий алгоритм формування ансамблю моделей. Із масиву початкових даних виділяється К якісних показників, які будуть використані при оцінюванні кредитоспроможності позичальників. Для кожного якісного показника визначається кількість його категорій N_j, $j = 1, K$, за якими із загального масиву спостережень виділяються підгрупи, які складатимуть навчальні вибірки для відповідних моделей-експертів $L_{ij}, i = 1, N_j, j = 1, K$. Подібне формування вибірок для оптимізації окремих моделей робить можливою реалізацію ідеї спеціалізації експертів. За потреби, кожен експерт може бути

представлений моделями різних типів – логістичні регресії, нейромережі, дерева рішень тощо.

Загальний результат роботи комітету можна визначати трьома способами: простим чи зваженим голосуванням або за допомогою метамоделі. Для випадку простого голосування використовуються рейтинги моделей-експертів R_{ij}, $i=\overline{1,N_j}$, $j=\overline{1,K}$. Ці рейтинги встановлюються на основі рівня значущості (зокрема, для логістичних регресій) або за рівнем точності класифікації, що демонструє модель на тестових даних (для випадку застосування нейромереж).

Для аналізу кредитоспроможності нового потенційного клієнта обирається група експертів M, які можуть бути задіяні для його класифікації (для застосування яких за цим позичальником є всі необхідні дані та навчальні вибірки яких формувались на значеннях якісних показників, відповідних даному клієнту). Отже, оцінювання кредитного ризику цього позичальника здійснюється комітетом, утвореним з моделей $L_{ij} \in M$. Для проведення розрахунків моделі впорядковуються за їх рейтингами R_{ij}. Аналіз позичальника починається з моделей, які мають найбільш високий рейтинг. Коли більше половини задіяних експертів віднесли аналізованого позичальника до одного класу, процес оцінювання кредитоспроможності припиняється. Якщо виникає ситуація, коли кількість експертів є парною та голоси розділяються порівну, то процедуру узагальнення результатів слід замінити на зважене голосування.

У випадку зваженого голосування для кожного експерта фіксується відсоток правильно класифікованих на тестовій вибірці прикладів t_{ij}, $i=\overline{1,N_j}$, $j=\overline{1,K}$. При обробці даних потенційного клієнта визначається група відповідних йому експертів M. На основі значень t_{ij} обчислюються коефіцієнти компетентності для обраних експертів $L_{ij} \in M$:

$$k_{ij}=\frac{t_{ij}}{\sum_{L_{ij}\in M} t_{ij}}.$$

Загальний результат розрахунків ансамблю є сумою добутків виходів окремих моделей на їх коефіцієнти компетентності. Основою для визначення коефіцієнтів компетентності експерта може також слугувати будь-який інший показник точності моделі, наприклад коефіцієнт Джині.

Третім способом узагальнення результатів роботи комітету є використання метамоделі – моделі верхнього рівня, входами якої будуть розрахунки окремих експертів.

Схематичне зображення описаного вище алгоритму побудови ансамблю моделей наведено на Рисунку 6.

Так, у авторському дослідженні [10] для оцінювання рівня кредитоспроможності фізичної особи запропоновану методику побудови ансамблю реалізовано для випадку використання *logit*-моделей. Враховано вплив таких якісних показників, як рівень освіти та статус працюючого. За кожним з цих показників із початкового масиву даних утворюється по декілька підгруп. Наприклад, показник «рівень освіти» поділяється на два підрівні: «наявність однієї чи більше вищих освіт» та «наявність середньої та середньої спеціальної освіти». З усього масиву даних обираються лише ті записи, які відповідають позичальникам із певним підрівнем даного якісного показника, утворюючи таким чином два окремі масиви більш однорідних даних. Аналогічна процедура здійснюється за іншим показником (статусом працюючого). Додатково було здійснено поділ початкової вибірки на три підгрупи за такою важливою з точки зору оцінки кредитоспроможності фізичних осіб характеристикою, що визначає наявність утриманців. Сформовані таким чином масиви даних реалізують ідею спеціалізації експертів.

За запропонованим підходом початковий масив розбивається на підгрупи, які перетинаються між собою. В такому вигляді формування масивів даних для навчання окремих моделей-експертів схоже на відповідну процедуру за технологією беггінгу. Однак за алгоритмом беггінгу вибірки утворюються випадковим чином, а в запропонованому варіанті формування вибірок є логічно обумовленим та приводить до отримання більш однорідних масивів даних.

Очевидно, що можна будувати моделі, використовуючи будь-яку потрібну кількість якісних показників.

http://dx.doi.org/10.21511/nfmte.7.2018.02

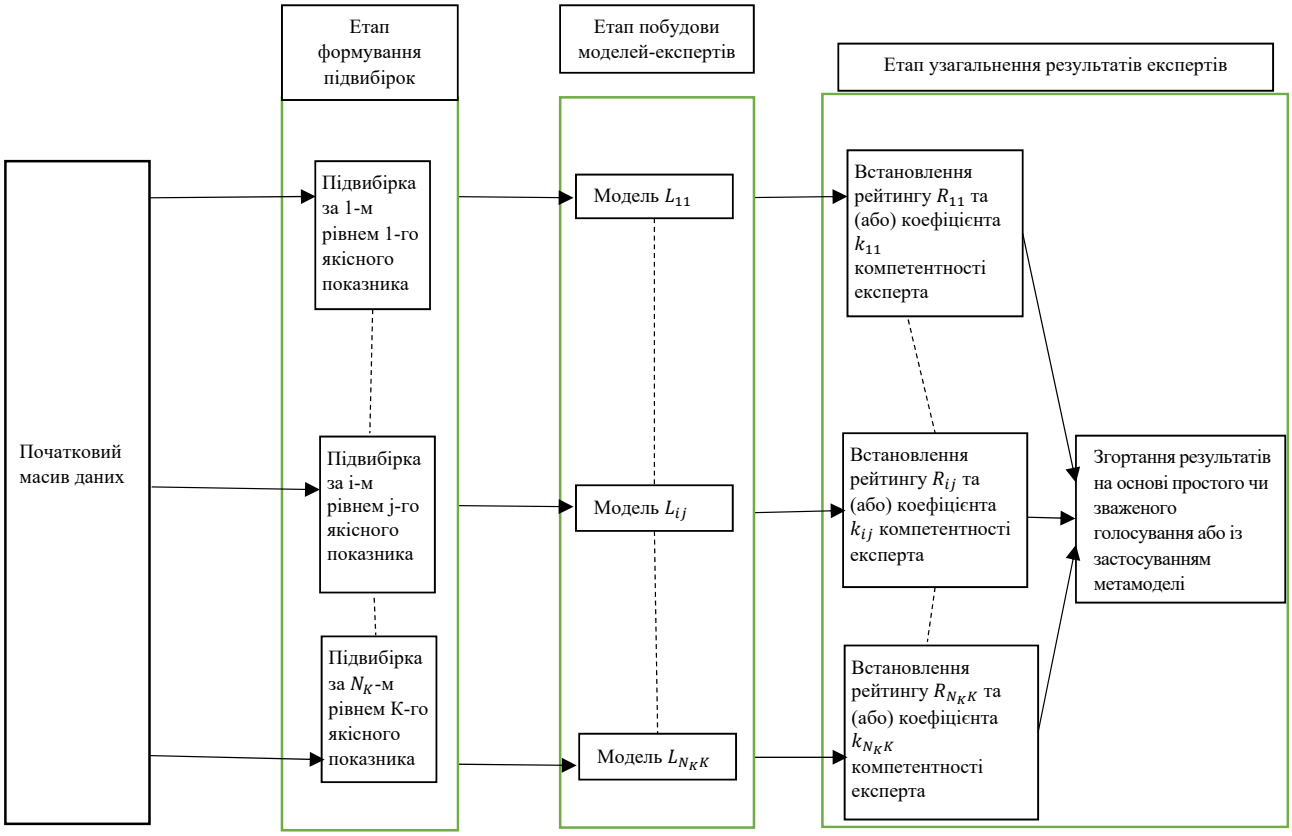

Рисунок 6. Схематичне зображення процесу побудови ансамблю моделей на основі спеціалізації експертів

Узагальнення результатів розрахунків усіх моделей доцільно здійснювати шляхом простого або зваженого голосування. Ваги для моделей відповідатимуть рівню точності класифікації експерта (чим він є вищим, тим більшою є його вага).

Сформована таким чином технологія побудови ансамблю найкращим чином пристосована саме для специфіки задачі оцінювання кредитоспроможності позичальників – фізичних осіб. Вона поєднує в собі й ідею спеціалізації експертів, і логічний спосіб формування масивів даних для навчальних вибірок, і класичний спосіб поєднання результатів розрахунків окремих моделей ансамблю.

Варто зазначити, що за реалізації описаного підходу також не вдається уникнути складностей, пов'язаних із зазначеною вище проблемою недостатньо великого обсягу наявних даних. Крім того, утворені підгрупи мають значні розбіжності в обсягах: найменша містить 145 спостережень, найбільша – 1.547. До того ж, за кожною з них потрібно утворити навчальну та тестову вибірки. Зауважимо, що при визначенні розміру навчальної вибірки недоцільно орієнтуватись на найменший масив даних, оскільки тоді втрачається можливість кращого налаштування інших моделей на більших вибірках. Таким чином виникає проблема неможливості формування навчальних вибірок однакового розміру. А отже, не вдасться повністю коректно застосувати запропоновану схему алгоритму побудови ансамблю моделей.

Враховуючи особливості розподілу позичальників між класами та підгрупами в досліджуваній базі даних, у кожній підгрупі виділялось близько 80% записів для навчальної вибірки та 20% – для тестової. Для вибору експертів за кожною підгрупою розглядались три кращі мережі за показниками точності класифікації на навчальній та тестовій вибірках, а також по узагальненим даним. Результати проведених обчислень наведено в Таблиці 9. Найефективніші моделі за кожною підгрупою тут виділені напівжирним шрифтом.

Таблиця 9. Показники ефективності нейромереж різної архітектури для вибору окремих моделей-експертів при формуванні ансамблю

Архітектура мережі, кількість входів, кількість нейронів проміжного шару	Відсоток правильно класифікованих спостережень у навчальній вибірці, %	Відсоток правильно класифікованих спостережень у тестовій вибірці, %	Узагальнені дані по всьому масиву (без поділу на навчальну та тестову вибірки)		Правильно класифікованих, %
			Клас	Всього	
Підгрупа позичальників із «наявністю однієї чи більше вищих освіт», відібраних за показником «рівень освіти»					
Тришаровий персептрон, 6, 6	64	58	0	439	61.73
			1	297	60.27
Радіально-базисна, 6, 9	59	53	0	439	56.04
			1	297	55.89
Радіально-базисна, 6, 4	57	53	0	439	55.35
			1	297	55.22
Підгрупа позичальників із «наявністю середньої та спеціальної середньої освіти», відібраних за показником «рівень освіти»					
Тришаровий персептрон, 6, 7	62	61	0	692	63.15
			1	747	60.1
Радіально-базисна, 6, 7	62	61	0	692	63.29
			1	747	60.37
Радіально-базисна, 6, 30	65	58	0	692	63
			1	747	59.97
Підгрупа позичальників із «власною справою», відібраних за показником «статус працівника»					
Радіально-базисна, 6, 3	60	56	0	76	60.53
			1	69	55.07
Радіально-базисна, 6, 6	63	58	0	76	64.47
			1	69	56.52
Радіально-базисна, 6, 4	58	58	0	76	60.53
			1	69	55.07
Підгрупа «найманих працівників», відібраних за показником «статус працівника»					
Радіально-базисна, 6, 4	60	58	0	809	58.47
			1	738	60.43
Радіально-базисна, 6, 16	61	59	0	809	58.96
			1	738	60.98
Радіально-базисна, 6, 8	60	59	0	809	58.96
			1	738	60.98
Підгрупа позичальників із «іншим статусом», відібраних за показником «статус працівника»					
Тришаровий персептрон, 6, 9	64	50	0	246	56.5
			1	237	57.8
Радіально-базисна, 6, 7	61	53	0	246	57.8
			1	237	57.32
Радіально-базисна, 6, 3	60	56	0	246	58.65
			1	237	54.88
Підгрупа даних «відсутні», відібраних за показником «наявність утриманців»					
Радіально-базисна 5, 11	60	59	0	491	59.06
			1	489	60.33
Радіально-базисна, 5, 15	63	59	0	491	60.29
			1	489	61.55
Радіально-базисна, 6, 23	63	58	0	491	60.29
			1	489	61.55

http://dx.doi.org/10.21511/nfmte.7.2018.02

Підгрупа даних «одна особа», відібраних за показником «наявність утриманців»					
Тришаровий персептрон, 3, 8	56	58	0	435	57.01
			1	355	57.46
Радіально-базисна, 5, 2	57	55	0	435	56.32
			1	355	56.06
Радіально-базисна, 5, 5	59	58	0	435	57.93
			1	355	58.59
Підгрупа даних «дві та більше особи», відібраних за показником «наявність утриманців»					
Радіально-базисна, 5, 3	49	54	0	205	54.15
			1	200	49
Радіально-базисна, 5, 6	68	61	0	205	68.29
			1	200	60.5
Радіально-базисна, 5, 8	67	60	0	205	66.83
			1	200	59.5

Аналіз даних, наведених у Таблиці 9, підтверджує отримані вище висновки, що більш ефективну класифікацію за представленими даними здійснюють мережі з радіально-базисною архітектурою, і лише в окремих випадках кращий результат демонструє тришаровий персептрон (він виявився ефективнішим для масивів меншої розмірності).

Результати розрахунків усіх моделей ансамблю узагальнювались простим голосуванням. Висновок про точність класифікації, що проведена ансамблем моделей за запропонованою технологією, можна зробити на основі перевірки за усім початковим масивом даних, проводячи порівняння отриманого результату з результатами розрахунків інших комітетів або окремих нейромереж.

За всім масивом даних комітетом було правильно класифіковано 63.5% надійних позичальників та 60% дефолтних. Відповідні показники для ансамблю моделей, сформованих за алгоритмом бустінгу, складали 58.7% правильно класифікованих надійних позичальників та 56% – дефолтних. Як можна бачити, для досліджуваних даних кращі результати демонструє ансамбль, який враховує спеціалізацію експертів.

Метою створення алгоритму бустінгу було підвищення точності класифікації у великих масивах інформації, яких при проведенні дослідження у нас в наявності не було. Також цей алгоритм не враховує особливостей даних, на яких проводяться розрахунки. Оскільки задача класифікації позичальників банку значною мірою пов'язана зі специфікою початкових даних, наприклад, наявністю значної кількості якісних показників, то від коректного врахування особливостей даних залежить і точність класифікації. Отже, при розв'язанні поставленої задачі доцільно застосовувати такий підхід для формування ансамблю, в якому враховується спеціалізація окремих моделей-експертів. Крім того, запропонований підхід може бути застосований для випадків, коли початковий масив даних має невелику розмірність і більшість алгоритмів формування ансамблів не дають задовільних результатів. Такі задачі виникатимуть, зокрема, при дослідженнях нових банківських продуктів, які ще не мають широкого розповсюдження, а, отже, обсяги статистичної інформації за ними недостатні для застосування класичних технологій формування ансамблів.

Оскільки реалізований алгоритм побудови ансамблю було також застосовано для випадку використання в якості експертів *logit*-моделей, то проведемо порівняння ефективності обох цих комітетів. Однак безпосередньо використовувати для порівняння розрахунки, описані в праці [10], неможливо, оскільки побудова *logit*-регресій здійснювалась на всіх даних підгрупи і не містила поділу на навчальну та тестову вибірки. Для більш коректного порівняння ефективності комітетів, утворених з *logit*-моделей та нейромереж, по кожній підгрупі проведено розрахунки параметрів *logit*-регресій на основі лише тієї частини даних, яка була використана в якості навчальних вибірок для нейронних мереж. У результаті отримано вісім логістичних регресій, показники точності яких наведено в Таблиці 10.

Таблиця 10. Показники точності *logit*-регресій

Показник, що використовувався для формування підгрупи даних	Навчальна вибірка		Тестова вибірка		Значення X2
	Відсоток правильно класифікованих значень «0»	Відсоток правильно класифікованих значень «1»	Відсоток правильно класифікованих значень «0»	Відсоток правильно класифікованих значень «1»	
Рівень освіти – «наявність одної чи більше вищих освіт»	60.8	66.3	31	67	47.561
Рівень освіти – «наявність середньої та спеціальної середньої освіти»	59.6	66.8	62	48	98.497
Статус працівника – «власна справа»	60.0	58.2	45	48	16.164
Статус працівника – «найманий працівник»	62.5	67.0	45	56	124.110
Статус працівника – «інший статус»	56.3	62.6	56	65	29.389
Наявність утриманців – «відсутні»	58.8	66.5	52	30	67.646
Наявність утриманців – «одна особа»	58.5	62.0	40	30	37.573
Наявність утриманців – «дві та більше особи»	59.4	69.4	11	78	36.753

Аналіз даних, наведених у Таблиці 10, свідчить, що всі моделі, крім побудованої на даних по статусу працівника «власна справа», є статистично значущими на навчальних вибірках (масив навчальних даних для вказаної моделі є найменшим у дослідженні та складає лише 110 прикладів). Точність класифікації на навчальних вибірках для всіх моделей майже однакова із дещо вищою ефективністю визначення ненадійних позичальників. На тестових вибірках у більшості моделей спостерігається значне зниження точності класифікації. Отримані результати свідчать про недоцільність застосування *logit*-регресій для розв'язання задачі класифікації за відсутності достатньої кількості навчальних даних. Узагальнений результат розрахунків цих моделей за описаним вище алгоритмом побудови ансамблю продемонстрував точність передбачення надійних позичальників на рівні 56.6%, а дефолтних – 61.8%.

Підсумовуючи результати експериментальних розрахунків можна зробити висновок, що комітети моделей, які були утворені на основі всього масиву даних (без поділу на підгрупи) або ж на основі використання *logit*-регресій, недоцільно використовувати для розв'язання поставленої задачі оцінювання кредитних ризиків позичальників-фізичних осіб. Комітет моделей, сформований на повному масиві даних, в окремих випадках демонструє точність оцінювання менше 50%. Так само неприйнятні показники точності класифікації мають окремі моделі *logit*-регресій, які входять у комітет.

Таким чином, за умов неоднорідності в масиві початкових даних доречно використовувати лише два комітети: отриманий за алгоритмом бустінгу та комітет нейромереж із урахуванням спеціалізації моделей-експертів.

Як наголошувалось вище, більш ефективними є комітети, при формуванні яких враховано дві умови – в їх основу покладено моделі, які мають вищі показники ефективності (у даній роботі продемонстровано відповідність зазначеній умові процедури пошуку та відбору кращих моделей при формуванні всіх описаних видів комітетів), а також забезпечується незалежність похибок окремих моделей ансамблю. У праці [14] зазначається, що для перевірки статистичної незалежності розподілу похибок окремих моделей комітету можна скористатись спрощеним підходом аналізу попарних кореляцій похибок всіх пар моделей. З цією метою розглянемо кореляційні матриці похибок моделювання, які утворені двома зазначеними видами комітетів моделей. Для аналізу кореляційної матриці при застосуванні алгоритму бустінгу обрано найкращий варіант ансамблю (навчальна вибірка 500 значень). У цьому випадку кореляційна матриця має вигляд:

http://dx.doi.org/10.21511/nfmte.7.2018.02

$$R_{Boost} = \begin{pmatrix} 1 & 0,88 & 0,64 \\ 0,88 & 1 & 0,65 \\ 0,64 & 0,65 & 1 \end{pmatrix}.$$

Досить тісний зв'язок спостерігається між похибками першого та другого експертів, менш тісний – для інших пар.

Кореляційна матриця похибок для ансамблю на основі спеціалізації експертів відображає всі пари зв'язків восьми експертів та має вигляд:

$$R_{Expert} = \begin{pmatrix} 1 & 0.002 & 0.28 & 0.24 & 0.07 & 0.16 & 0.36 & 0.09 \\ 0.002 & 1 & 0.44 & 0.36 & 0.28 & 0.11 & 0.54 & 0.34 \\ 0.28 & 0.44 & 1 & -0.0003 & 0.0002 & 0.08 & 0.39 & 0.25 \\ 0.24 & 0.36 & -0.0003 & 1 & 0.001 & 0.14 & 0.42 & 0.22 \\ 0.07 & 0.28 & 0.0002 & 0.001 & 1 & 0.08 & 0.21 & 0.15 \\ 0.16 & 0.11 & 0.08 & 0.14 & 0.08 & 1 & 0.0003 & 0.00009 \\ 0.36 & 0.54 & 0.39 & 0.42 & 0.21 & 0.0003 & 1 & -0.0004 \\ 0.09 & 0.34 & 0.25 & 0.22 & 0.15 & 0.00009 & -0.0004 & 1 \end{pmatrix}.$$

Найвище значення коефіцієнту кореляції у матриці для восьми експертів сягає 0.54, що є нижчим, ніж найменше значення кореляції похибок окремих моделей у комітеті, сформованим за алгоритмом бустінгу. Причому, більшість показників кореляції похибок моделей в ансамблі на основі спеціалізації експертів знаходяться в межах від 0 до 0.2. Таким чином, на основі проведеного порівняльного аналізу можна зробити висновок щодо незалежності похибок окремих моделей ансамблю, сформованого за алгоритмом врахування спеціалізації експертів, що свідчить про його вищу ефективність за алгоритм бустінгу [14].

Розглянемо результати класифікації, які отримані двома зазначеними ансамблями моделей. Показники їх ефективності наведені в Таблиці 11.

Таблиця 11. Показники точності класифікації ансамблями моделей, сформованими за різними алгоритмами

Види ансамблів	Відсоток правильно класифікованих значень «0»	Відсоток правильно класифікованих значень «1»
Ансамбль, сформований із урахуванням спеціалізації моделей-експертів	63.5	60.0
Ансамбль, сформований з нейромереж за алгоритмом бустінгу (кращий результат з чотирьох ансамблів)	51.9	70.4

Як бачимо за даними Таблиці 11, запропонована процедура формування ансамблю моделей на основі спеціалізації експертів демонструє більш рівномірну ефективність моделювання в розрізі класів та вищу загальну точність класифікації – 61.8% (тоді як кращий результат класифікації за обома класами на основі алгоритму бустінгу становив 59.4%).

ВИСНОВКИ

Узагальнюючи отримані результати можна зробити наступні висновки. Найпростішим для реалізації при розробці системи скорингової оцінки позичальників банку є тип ансамблю, робота якого полягає в усередненні результатів розрахунків моделей-експертів. Однак проведене експериментальне дослідження показало, що застосування такого типу ансамблю не дало змоги підвищити ефективність класифікації позичальників – точність ансамблю майже співпадає з точністю кращої з його моделей. Перевагою такого виду ансамблю є підвищення стійкості результату його розрахунків.

Зростання точності класифікації понад точність окремої моделі забезпечує реалізація алгоритму бустінгу. Проте підвищення ефективності моделі є недостатньо високим. Крім того, цей алгоритм рекомендується застосовувати для випадку однорідних початкових даних.

Найбільш адаптованим для проведення скорингової оцінки позичальників-фізичних осіб можна вважати запропонований авторами алгоритм побудови ансамблю на основі спеціалізації експертів. Процедура формування навчальних вибірок для експертів ансамблю забезпечує побудову моделей на однорідних масивах даних, згрупованих за значеннями якісних характеристик позичальників. При цьому окремі експерти комітету можуть бути реалізовані на основі різних типів моделей (логістичні регресії, нейромережі, дерева рішень тощо). В алгоритмі враховується компетентність експертів за рахунок введення вагових коефіцієнтів при узагальненні результатів.

Серед переваг авторського підходу є і те, що більшість відомих ансамблевих технологій вимагають значних обсягів та однорідності початкових даних. Запропонований підхід до утворення ансамблю моделей на основі спеціалізації експертів враховує неоднорідність початкових даних та може бути використаний і для невеликої початкової сукупності спостережень. Такі задачі виникатимуть, наприклад, при дослідженнях нових банківських продуктів, які ще не мають широкого поширення, а, отже, обсяги статистичної інформації за ними недостатні для застосування класичних технологій формування ансамблів.

Проведені в дослідженні експериментальні розрахунки підтвердили переваги запропонованого авторами алгоритму побудови ансамблю на основі спеціалізації експертів при розв'язанні задач оцінювання кредитного ризику.

СПИСОК ЛІТЕРАТУРИ

1. Breiman, L. (2001). *Random forests* (33 p.). Berkeley: University of California. Retrieved from https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
2. Flach, P. A. (2012). *Machine learning: the art and science of algorithms that make sense of data* (409 p.). UK: Cambridge University Press.
3. Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences, 55*(1), 119-139. Retrieved from http://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic_generalization.pdf
4. Haykin, S. (1998). *Neural networks - a comprehensive foundation* (842 p.). (2nd ed.). New Jersey: Prentice Hall.
5. Kuznetsov I. A., & Kirieiev, V. S. (2016). Разработка ансамбля алгоритмов классификации с использованием энтропийного показателя качества для решения задачи поведенческого скоринга [Razrabotka ansamblya algoritmov klassifikacii s ispolzovaniem entropijnogo pokazatelya kachestva dlya resheniya zadachi povedencheskogo skoringa]. In *Proceedings of XVIII International conference «Analytics and data management in data-intensive areas»* (pp. 11-42). Retrieved from http://ceur-ws.org/Vol-1752/paper07.pdf
6. Lavrenkov, Y. N. (2014). *Исследование и разработка комбинированых нейросетевых технологий для повышения ефективности безопасной маршрутизации информации в сетях связи* [*Issledovanie i razrabotka kombinirovanyh nejrosetevyh tekhnologij dlya povysheniya effektivnosti bezopasnoj marshrutizacii informacii v setyah svyazi]* (Ph.D. Thesis). Kaluga: MSTU named after N. E. Baumana. Retrieved from https://mpei.ru/Science/Dissertations/dissertations/Dissertations/LavrenkovYN_diss.pdf
7. Matviichuk, A. V. (2011). *Штучний інтелект в економіці: нейронні мережі, нечітка логіка [Shtuchnyi intelekt v ekonomitsi: neironni merezhi, nechitka lohika]* (439 p.). Kyiv: KNEU.
8. Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review, 63*(2), 81-97. http://dx.doi.org/10.1037/h0043158
9. Paklin, N. B., & Oreshkov, V. I. (2013). *Бизнес-аналитика: от данных к знаниям* [*Biznes-analitika: ot dannykh k znaniyam]* (704 p.). Saint Petersburg: Piter.

http://dx.doi.org/10.21511/nfmte.7.2018.02

10. Savina, S. S., & Ben, V. P. (2015). Об'єднання моделей logit-регресій як комітету експертів для оцінки кредитоспроможності позичальника [Obiednannia modelei logit-rehresii yak komitetu ekspertiv dlia otsinky kredytospromozhnosti pozychalnyka]. *Neuro-fuzzy modeling technigues in economics*, *4*, 154-188. Retrieved from http://nfmte.com/article-4-8.html
11. Savina, S. S., & Ben, V. P. (2016). Вибір архітектури нейромережі для розв'язання задачі класифікації надійності позичальників-фізичних осіб [Vybir arkhitektury neiromerezhi dlia rozviazannia zadachi klasyfikatsii nadiinosti pozychalnykiv-fizychnykh osib]. *Neuro-fuzzy modeling technigues in economics*, *5*, 123-151. Retrieved from http://nfmte.com/article-5-6.html
12. Schapire, R. E. (1999). Theoretical views of boosting and applications. In *proceedings of 10th International conference* «*Algorithmic learning theory*» *(Tokyo, Japan)*. Retrieved from http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG445/literatur/schapire_99a.pdf
13. Schapire, R. E. (1990). The strength of weak learnability. *Machine learning, 5*(2), 197-227. Retrieved from http://rob.schapire.net/papers/strengthofweak.pdf
14. Terekhov, S. A. (2007). Гениальные комитеты умных машин [Genialniye komitety umnykh mashyn]. In *IX All-Russian scientific and technical conference* «*Neuroinformatics-2007*»: *lectures on neuroinformatics* (pp. 11-42).
15. Tsarehorodtsev, V. G. (2004). Оптимизация экспертов boosting-коллектива по их кривым обучения [Optimizacija ekspertov boosting-kollektiva po ih krivym obuchenija]. In *Proceedings of XIII All-Russian Seminar* «*Neuroinformatics and its applications*» (pp. 152-157).