ПРОБЛЕМИ ПРИЙНЯТТЯ РІШЕНЬ ТА УПРАВЛІННЯ В ЕКОНОМІЧНИХ, ТЕХНІЧНИХ, ЕКОЛОГІЧНИХ І СОЦІАЛЬНИХ СИСТЕМАХ

УДК 004.942
DOI: 10.20535/SRIT.2308-8893.2019.2.04

ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ КРЕДИТІВ МЕТОДАМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

В.Г. ГУСЬКОВА, П.І. БІДЮК

Анотація. Розглянуто актуальне завдання оцінювання кредитоспроможності на основі експертного та скорингово підходів. Виконано аналіз предметної галузі та проаналізовано основні методи математичного моделювання і оцінювання кредитних ризиків, запропоновано математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі альтернативних методів, розроблено математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі дерев рішень, логістичної регресії, мереж Байєса та нечіткої логіки. Установлено, що модель на основі нечіткої логіки для розв'язання задачі визначення ймовірності дефолту для кредитного позичальника є більш точною, на що вказують пораховані точності моделей. Це зумовлено можливістю з використання методу нечіткої логіки з нечітким висновком Мамдані точніше встановлювати причинно-наслідкові зв'язки між характеристиками-факторами задачі та їх вплив на вихідну змінну.

Ключові слова: дерева рішень, логістична регресія, мережа Байєса, нечітка логіка, імовірність дефолту, висновок Мамдані.

ВСТУП

Оцінювання кредитоспроможності — важливе завдання загальної проблеми менеджменту ризиків фінансових організацій, які забезпечують клієнтів кредитами. Коректне вирішення цього завдання забезпечує повне та своєчасне повернення кредитів, зменшення можливих утрат. Є два основні підходи до його розв'язання: експертний та скоринговий.

Експертний підхід передбачає, що фахівці з кредитування індивідуальних позичальників визначають істотні характеристики клієнта банку, які можуть впливати на повернення кредиту, ставлять у відповідність цим характеристикам певні ваги [1–3]. Для кожного клієнта за всіма характеристиками проставляються бали відповідно до вагових коефіцієнтів, установленими експертами, та підраховується сума всіх балів. Для банку завчасно встановлюється певне порогове значення, яке визначає правило: якщо сума всіх балів менша за це значення, то клієнту не слід видавати кредит, а якщо більша, то клієнт може отримати кредит. Банк може встановити ще одне обмеження (верхню межу), яке визначає таке правило: якщо сума балів більша від уста-

новленого значення, то клієнту можна видавати кредит без надмірної обережності. Якщо і верхня, і нижня межі видачі кредиту задані, то значення між ними є компетенцією менеджера банку, який на власний розсуд визначає, чи видавати кредит клієнту, чи ні. Зрозуміла суб'єктивність цього підходу для встановлення як балів, так і порогового та верхнього значень.

Скоринговий підхід ґрунтується на побудові математичної моделі для оцінювання кредитоспроможності позичальників на основі кредитних історій банку та оцінюванні ймовірності дефолту потенційного позичальника з урахуванням його соціально-демографічних характеристик. Маючи статистичні дані «гарних» і «поганих» кредитів за певний період, банк може визначити безпосередньо фактори, які створюють передумови для повернення/неповернення кредиту, і для кожного нового клієнта на основі цих характеристик визначити його можливість повернути кредит [4, 5]. На початковому етапі скоринговий підхід ґрунтується на експертному підході, оскільки необхідно передусім визначити, які саме характеристики щодо клієнта потрібно збирати та як перевірити надані клієнтом дані, і в процесі застосування скорингової моделі змінювати оброблені банком дані про клієнта [6]. Очевидно, що скорингову модель необхідно коригувати у процесі роботи (кожні два-три роки), оскільки ситуація в країні динамічно змінюється, з'являються певні зловмисники, які прораховують фіктивні дані для скорингової моделі з надання кредиту і т. ін. Спільне використання обох підходів дозволило б полегшити процес прийняття рішення щодо видачі кредиту клієнту банком [7].

Більш адекватним є підхід на основі статистики дефолтів за попередні періоди, тобто побудова скорингової моделі. Статистичні методи передбачають визначення ймовірності реалізації певної події на підставі певних вимог:

- об'єкти, для аналізу яких пропонується використати статистичні дані, і об'єкти, на яких збираються статистичні дані, є еквівалентними;
- умови, у яких пропонується використовувати статистичні дані і умови їх збирання є еквівалентними;
- обсяги вибірок статистичних даних є достатніми, методи обробленя — коректними, а джерело інформації надійним.

Такий підхід ставить високі вимоги до статистики дефолтів:

1. Однорідна вибірка (позичальники повинні бути досить схожими).

2. Вибірка має складатися з певної кількості випадків – чим більше дефолтів, тим краще. За експертними оцінками для адекватності моделі обсяг вибірки повинен складати не менше ніж 2000 випадків.

3. Вибірка для побудови моделі повинна нагромаджуватися за доволі обмежений термін. Ця вимога викликана фактом зміни макроекономічного середовища. Позичальник з певними параметрами в одному макроекономічному середовищі без проблем виплачуватиме кредит; а в іншому середовищі виявиться дефолтом. Вважається, що в умовах країн, які розвиваються, модель для оцінювання ймовірності дефолту необхідно змінювати кожні два-три роки.

4. Необхідно нагромаджувати не лише кредитну історію позичальників (дефолт / не дефолт), але й параметри цих позичальників – вік, стать, місце роботи, посада, сім'я тощо. Проблема в тому, що спочатку невідомо, які са-

ме параметри виявляться значущими у моделі. Тому на етапі побудови моделі необхідно нагромаджувати максимальну кількість параметрів за кожним позичальником.

5. Вибірка має включати інформацію про кредити, цикл кредитування яких уже закінчився. Ця вимога необхідна, оскільки обов'язковою є інформація про повернення / неповернення кредиту.

6. Історію потрібно нагромаджувати в межах кредитних продуктів (споживчий кредит, кредит на автомобільний транспорт, іпотечний кредит). Скорингову модель необхідно будувати для кожного кредитного продукту.

Дані про процес кредитування фізичних осіб задовольняють майже всі ці вимоги. Для фізичних осіб параметрами скорингової моделі можуть бути вік, сімейний стан, кількість дітей, освіта, місце проживання, робота, посада, власність, кредитна історія і т. ін.

ПОСТАНОВКА ЗАДАЧІ

Для досягнення поставленої мети у роботі поставлені і вирішуються такі завдання: виконати аналіз предметної галузі і вибрати методи математичного моделювання та оцінювання кредитних ризиків; запропонувати математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі альтернативних методів; розробити математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі дерев рішень, логістичної регресії, мереж Байєса (МБ) та нечіткої логіки; створити проект та програмну реалізацію інформаційної системи підтримання прийняття рішень (СППР) для оцінювання кредитоспроможності позичальників і порівняти показники якості моделей за допомогою ROC-аналізу.

МЕТОДИ МОДЕЛЮВАННЯ І ПРОГНОЗУВАННЯ БАНКІВСЬКИХ КРЕДИТНИХ РИЗИКІВ

Розглянемо особливості побудови моделей у формі логістичної регресії, МБ та нечіткої логіки.

Логістична регресія — це різновид множинної регресії, загальне призначення якої полягає в аналізі зв'язку між декількома незалежними змінними (регресорами або предикторами) і залежною змінною. Бінарна логістична регресія, як випливає з назви, застосовується в разі, коли залежна змінна є бінарною (тобто може набувати тільки двох значень). Інакше кажучи, за допомогою логістичної регресії можна оцінювати ймовірність того, що подія настане для конкретної випробовуваної особи (хворий/здоровий, повернення кредиту/ дефолт і т. ін.). Це досягається застосуванням такого регресійного рівняння (логіт-перетворення) [8, 9]:

$$P = \frac{1}{1+e^{-y}},$$

де P — імовірність того, що відбудеться подія, яка цікавить; e — основа натуральних логарифмів 2,71; y — змінна, що визначається рівнянням лі-

нійної регресії. Залежність, що зв'язує ймовірність події і величину *y*, показано на рис. 1.

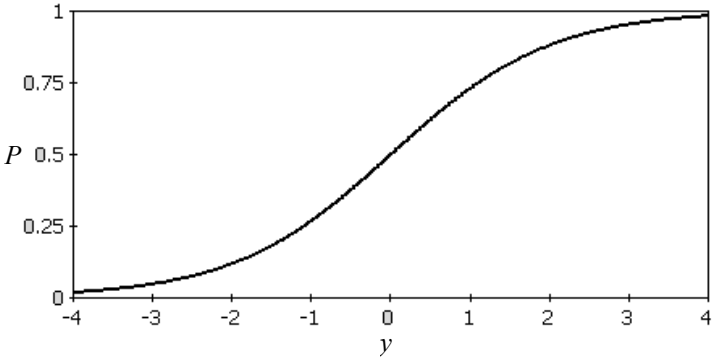

Рис. 1. Логістична функція розподілу

БАЙЄСІВСЬКИЙ ПІДХІД НА ОСНОВІ МЕРЕЖ ДОВІРИ

За допомогою МБ як інструменту аналізу даних вирішують дві математичні задачі: 1) побудову структури МБ і 2) формування імовірнісного висновку. Задача побудови МБ за заданими навчальними даними є NP-складною, тобто завданням нелінійної поліноміальної складності. Кількість усіх можливих нециклічних моделей, які потрібно проаналізувати, обчислюється за рекурентною формулою Робінсона, запропонованою у 1976 р. [10]:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} C_n^i 2^{i\cdot(n-i)} f(n-i),$$

де n — кількість вершин; $f(0) = 1$. Однак на практиці виконати повний перебір моделей можна тільки для мереж не більш ніж із сімома вершинами (вузлами), інакше не вистачить обчислювальних ресурсів.

Завдання формування імовірнісного висновку в МБ є важливим і складним і належить до класу задач прийняття рішень. Однак для реалізації цього методу необхідно звести структуру МБ до вигляду об'єднаного дерева (junction tree), і тільки тоді можна використовувати алгоритм імовірнісного висновку в об'єднаному дереві, який ґрунтується на проходженні λ і π повідомлень по дереву. Очевидно, що існуючі методи побудови МБ і формування висновку потребують трудомістких обчислень. Тому розроблення методів, що дозволяють зменшити обчислювальну складність, є актуальним для моделювання процесів різної природи МБ [10, 11].

Постановка задачі побудови МБ складається з таких кроків:

1) розроблення евристичного методу побудови МБ, що містить два етапи. На першому етапі обчислюється значення інформації між усіма вершинами, на другому — цілеспрямований пошук, який використовує як оцінну функцію оцінки мінімальної довжини, засновану на принципі опису, який застосовується на кожній ітерації алгоритму навчання.

2) розроблення методу ймовірнісного висновку в МБ з використанням двох кроків. На першому кроці обчислюється матриця емпіричних значень спільного розподілу ймовірностей всієї мережі, а на другому кроці — значення ймовірностей всіх можливих станів неінстанційованих вершин [12].

Байєсова мережа являє собою пару (G,B), перша компонента G якої — це напрямлений нециклічний граф, який відповідає випадковим змінним і записується як набір умов незалежності: кожна змінна незалежна від її батьків у G. Друга компонента пари B — це множина параметрів, що визначають мережу; містить параметри $\theta_{X^{(i)}|pa(X^{(i)}} = P(X^{(i)} \mid pa(X^{(i)}))$ для кожного можливого значення $x^{(i)} \in X^{(i)}$ і $pa(X^{(i)}) \in Pa(X^{(i)})$, де $Pa(X^{(i)})$ — набір батьків змінної $X^{(i)} \in G$. Кожна змінна $X^{(i)} \in G$ зображується у вигляді вершини. Якщо розглядають більше одного графу, то для визначення батьків змінної $X^{(i)}$ у графі використовують позначення $Pa^G X^{(i)}$. Повна спільна ймовірність БМ обчислюється за виразом [12, 13]

$$P_B(X^{(1)},...,X^{(N)}) = \prod_{i=1}^{N} P_B(X^{(i)} \mid P_a(X^{(i)})).$$

Гібридна МБ $B = (X, D, P)$ визначається через напрямлений ациклічний граф $G = (X, E)$ і його функції $P_i = \{P(x_i \mid pa_i)\}$, де pa_i — набір батьківських вузлів x_i; X — набір змінних, поділених на дискретні Δ і неперервні Γ змінні, тобто $X = \Gamma \cup \Delta$. Структура графу G обмежена тим, що неперервні змінні не можуть мати дискретних змінних як їх вузли-нащадки. Умовний розподіл неперервних змінних задається лінійною гаусівською моделлю:

$$P(x_i \mid I = i, Z = z) = N(\alpha(i) + \beta(i) \times z, \gamma(i)),\ x_i \in \Gamma,$$

де Z та I — множини відповідно неперервних і дискретних батьків x_i; $N(\mu(\sigma))$ — мультиноміальний нормальний розподіл. Мережа являє собою спільний розподіл усіх його змінних, заданих добутком усіх таблиць умовних імовірностей.

Можна розглядати сформульований висновок як шлях конвертації багатовимірних гаусіанів у МБ. Змінні упорядковуються у певному порядку $X_1,\ldots,X_n$. Потім цей висновок використовується для знаходження умовного розподілу:

$$P(X_i \mid X_1,\ldots,X_{i-1}) = N\left(X_i; \beta_{i0} \sum_{j=1}^{i-1} \beta_{i,j} X_j, \sigma_i^2\right).$$

Створюється ребро з X_j до X_i $(1 \le j < i)$ тоді і тільки тоді, коли $\beta_{ij} \neq 0$. Умовний імовірнісний розподіл (УІР) X_i називають лінійним умовним імовірнісним розподілом, що має вигляд деякого виразу після скорочення усіх нульових значень. Лінійний умовний імовірнісний розподіл для кореневих вузлів є просто одновимірною гаусіаною. Мережа Байєса, у якій всі умовні імовірнісні розподіли є лінійними, називається лінійною гаусіаною (ЛГ). Таким чином, кожна багатовимірна гаусіана може бути подана ЛГ. Обернене твердження також справедливе. Кожна МБ з лінійними УІР являє собою спільний нормальний розподіл.

Умовна лінійна гаусіана (УЛГ) — це МБ, яка містить як неперервні змінні (Γ), так і дискретні змінні (Δ) з такими обмеженнями [14, 15]:

- дискретний вузол не може мати неперервних батьків, таким чином, усі УІР для дискретних вузлів можуть бути подані як у дискретних МБ;
- умовно-імовірнісний розподіл будь-якої неперервної змінної є лінійним УІР, заданим будь-якою комбінацією дискретних батьків. Формально, якщо вузол Y має батьків $\{X_1,\ldots,X_k\}\subseteq\Gamma$ і $D=\{D_1,\ldots,D_l\}\subseteq\Delta$, він визначається як УІР з використанням таких параметрів: для кожного $d\in \mathrm{Dom}(D,\beta_{d0},\ldots,\beta_{dk})$ і σ_d^2:

$$P(Y\mid x,d)=N(Y;\beta_{d,0}+\sum_{i=1}^{k}\beta_{d,i}x_i,\sigma_d^2).$$

Це найбільш поширений вид гібридних моделей. Вони дозволяють мати лише лінійні відношення між неперервними змінними і не дозволяють дискретним змінним мати неперервних батьків. Широке використання таких моделей зумовлено їх математичною зручністю. В УЛГ задано значення дискретних змінних, розподіл неперервних змінних — багатовимірну гаусіану; отже спільний імовірнісний розподіл — це композиція гаусіан, з якими можна працювати, використовуючи аналітичні інструменти.

Якщо всі дискретні змінні задані, то УІР неперервних змінних є лінійними умовними неперервними розподілами. Тоді з наданням будь-яких значень дискретним змінним УЛГ редукується у ЛГ і тому є нормальним розподілом. Звідси випливає, що спільний розподіл, поданий УЛГ, є композицією гаусіан, де кожна композиція компонентів відповідає реалізації дискретних змінних.

Таким чином, використання МБ дає змогу аналізувати причинно-наслідкові зв'язки між окремими змінними (подіями, даними) і формулювати на цій основі обгрунтований імовірнісний прогноз. Використання гібридних МБ дозволяє коректно аналізувати неперервні змінні у моделі (наприклад, вік чи суму кредиту) і досягати вищої точності прогнозу [15].

ЗАСТОСУВАННЯ МЕТОДІВ МОДЕЛЮВАННЯ КРЕДИТНИХ РИЗИКІВ

Прогнозування кредитоспроможності індивідуального позичальника виконується на основі бінарної логістичної регресії. Для побудови моделі використаємо навчальну вибірку даних про клієнта та результати повернення чи неповернення цим клієнтом кредиту (0 — якщо кредит повернено і 1 — якщо ні). Маючи інформацію про 1500 клієнтів, спочатку будуємо модель логістичної регресії на основі даних про 1200 клієнтів, а потім робимо прогноз на основі побудованої моделі для 300 клієнтів, маючи в моделі всі дані про них, окрім результатів повернення чи неповернення кредиту, і порівнюємо результати побудованого прогнозу з реальним станом подій. У цьому випадку як вхідні дані включено такі змінні: суму кредиту, вік позичальника, його дохід, платіж за кредитом, термін кредиту та освіту. Характеристики побудованих логіт і пробіт моделей подано в табл. 1.

Таблиця 1. Статистичні характеристики якості прогнозів

Характеристика	Середньо-квадратична похибка	САПП	Коефіцієнт Тейла
Логіт	0,407083	13,23895	0,493373
Пробіт	0,406719	13,23129	0,493053

Отримані статистичні характеристики прогнозування для обох типів моделей дуже близькі між собою. Тобто тип розподілу не є значущим для побудови моделі.

ЗАСТОСУВАННЯ ДЕРЕВ РІШЕНЬ ДЛЯ ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКІВ

Дерева рішень часто використовують для розв'язання задач класифікації. Наприклад, звернення за кредитом можуть бути класифіковані як такі, що мають високий або низький ризик — дерева рішень дозволяють визначити правила для виконання такої класифікації на основі даних з навчальної вибірки. Порівняно з іншими алгоритмами дерева рішень мають такі переваги, як висока швидкість навчання, прийнятна точність та легко зрозумілі шаблони. На вибірці з 1600 кредитів побудовано дерево рішень, що дозволяє класифікувати нового клієнта як такого, що поверне кредит, або не поверне. Результати побудови дерев рішень зведено для порівняння в табл. 2 та подано у вигляді діаграми (рис. 2).

Таблиця 2. Результати для всіх типів дерев рішень

Тип дерева рішень	Загальна точність: CA	Похибки класифікації		
		I роду	II роду	Сумарна похибка
Chaid	0,843	3,733	11,933	15,666
Exhaustive Chaid	0,825	3,733	13,733	17,466
CRT	0,922	4,4	3,4	7,8
Quest	0,885	5,333	6,1333	11,466

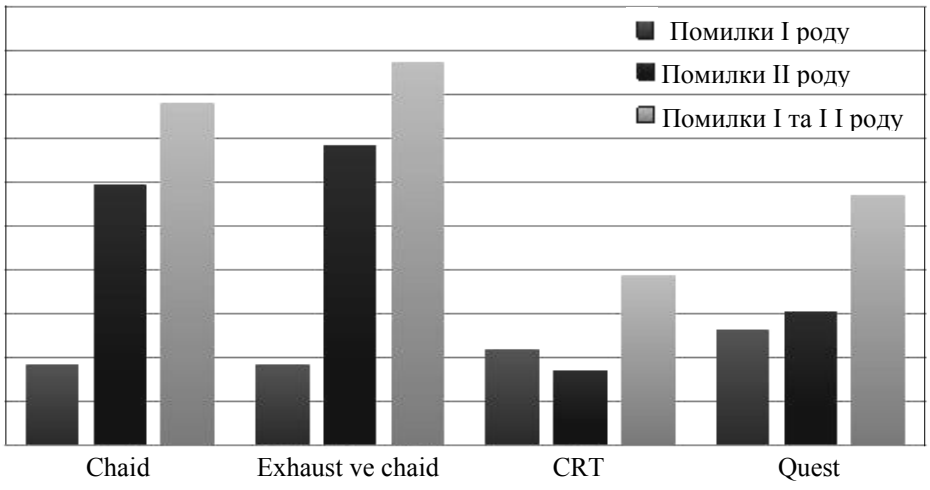

Рис. 2. Порівняльна діаграма похибок алгоритмів дерев рішень

ПРОГНОЗУВАННЯ КРЕДИТОСПРОМОЖНОСТІ ІНДИВІДУАЛЬНОГО ПОЗИЧАЛЬНИКА ЗА ДОПОМОГОЮ БАЙЄСІВСЬКИХ МЕРЕЖ

За рік банком видано 1600 кредитів. Кожний клієнт описується 18 характеристиками. За наявними даними побудовано мережу, яка показує зв'язок між характеристиками клієнта і вершиною — подією повернення кредиту. За наявними навчальними даними визначено ймовірність повернення кредиту новим клієнтом, що звернувся до банку. Отже, треба визначити PD_i — імовірність дефолту потенційного позичальника. Оскільки числові характеристики (сума кредиту, вік, дохід і т.ін.) набувають багато значень, тобто є неперервними, то для розв'язання поставленої задачі потрібно використовувати гібридні МБ. За описаною вище методикою необхідно виконати аналіз проблеми і зробити формалізовану постановку задачі. Розглянемо детальніше схему застосування запропонованої методики.

Крок 1. Для розв'язання задачі оцінювання кредитоспроможності позичальника зібрати статистичні дані про видані кредити, частину з яких повернуто, а частина виявилась дефолтом, тобто отримати позитивні і негативні приклади навчальних даних. На основі вибраних параметрів позичальника та кредиту розробити формальну модель і оцінити ймовірності дефолту позичальника PD_i:

$$PD_i = F(w^j, x_i^j),$$

де w^j — ваги параметрів x_i^j; i — кількість позичальників; j — кількість параметрів кредиту. Модель для оцінювання кредитоспроможності на основі МБ описується таким чином:

$$PD_i = F(v_i^k, G, J) = 1 - PR_i,$$

де v_i^k — змінні, що описують характеристики клієнта і кредиту; J — імовірнісний розподіл змінних v_i^k; G — напрямлений ациклічний граф, вузли якого відповідають випадковим змінним v_i^k модельованого процесу; PR_i — імовірність повернення кредиту.

Крок 2. Дані задачі — це статистичні дані за 1600 виданими кредитами, з яких 450 випадків дефолтів, 1150 — повернутих кредитів.

Крок 3. Взаємовиключними змінними для даної задачі є 18 характеристик, що описують клієнта та кредит.

Кроки 4–5. Оскільки в задачі використовуються неперервні змінні і будується гібридна МБ, то для неперервних змінних використовується дискретизація, тобто область значень неперервної змінної розбивається на проміжки. Кількість проміжків визначається користувачем, ті стовпчики, для яких не вказано кількість інтервалів розбиття, автоматично розбиваються на два проміжки. Доцільно виконувати дискретизацію даних за однаковою шириною класів або за однаковою кількістю точок усередині кластерів. Крім цього, ширина та кількість проміжків можуть бути регламентовані банком, виходячи із соціологічних або демографічних досліджень груп клієнтів [15]. Під час побудови структури МБ слід пам'ятати, що обраний алгоритм впливає на швидкість виконання програми і на саму побудовану

структуру. Найшвидшим є алгоритм Greedy Thick Thinning, його і будемо використовувати для аналізу прикладів. У результаті роботи алгоритму отримуємо тільки одну структуру, яка є логічною і оптимальною. Побудовану за цими даними структуру МБ зображено на рис. 3. Вона наочно демонструє зв'язки між даними.

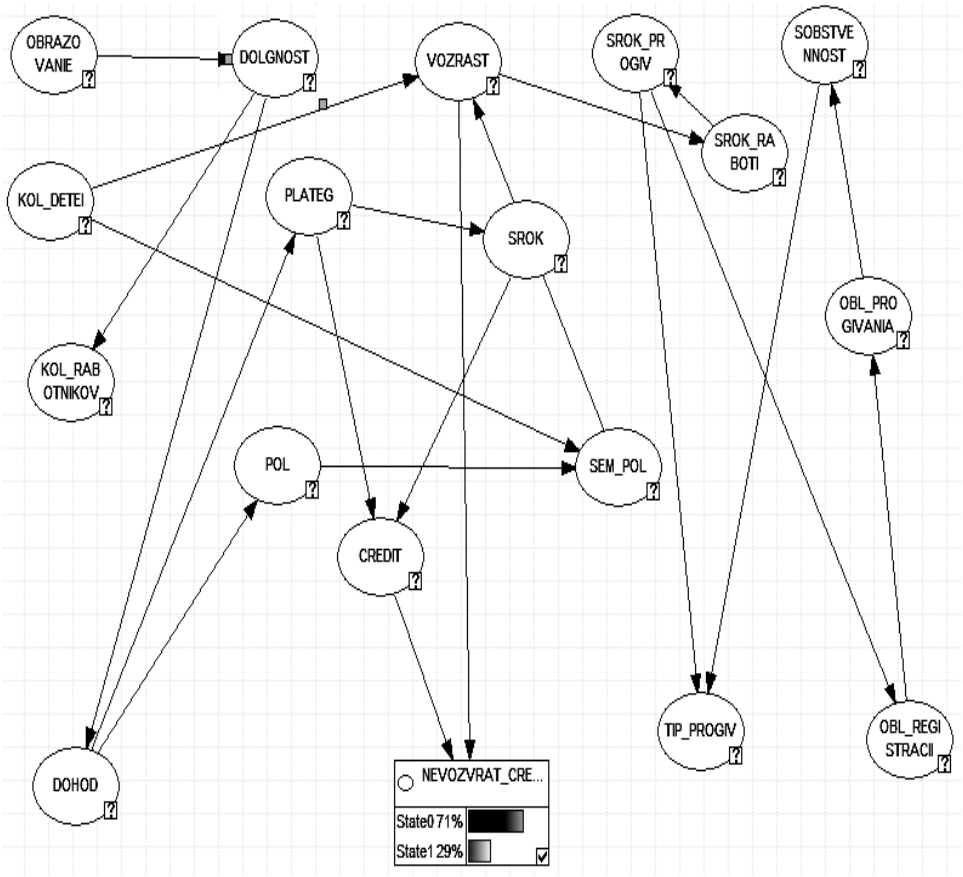

Рис. 3. Структура побудованої МБ

ОЦІНЮВАННЯ КРЕДИТОСПРОМОЖНОСТІ ПОЗИЧАЛЬНИКА З ВИКОРИСТАННЯМ НЕЧІТКОЇ ЛОГІКИ

Для врахування не тільки кількісних, але і якісних характеристик вектор даних можна подати за допомогою лінгвістичних змінних. Лінгвістична змінна визначається як $< b, T, X, G, M >$, де

b — ім'я лінгвістичної змінної;

T — базова терм-множина лінгвістичної змінної, кожний елемент якої (терм) подається як нечітка множина на універсальній множині X;

G — синтаксичне правило, часто у вигляді граматики, що дозволяє оперувати елементами терм-множини T, зокрема, генерувати нові терми (значення);

M — семантичне правило, що задає функцію належності нечітких термів, утворених синтаксичними правилами [16, 17]. Для побудови правила

необхідно побудувати функції належності кожної лінгвістичної змінної, тобто задати вигляд функції та її параметри. Параметри функцій належності визначаються в процесі її побудови. Підхід на основі нечітких нейронних мереж (ННМ) дозволяє автоматично будувати функцію належності, а відповідно і базу знань та реалізувати нечіткий логічний висновок. У загальному вигляді алгоритм має такі етапи.

1. Визначення множини вхідних змінних про позичальника: $\vec{X} = \{X_1, X_2, ..., X_i, ..., X_N\}$.

2. Визначення множини вихідних змінних — можливі групи ризику: $D = \{D_1, D_2, ..., D_i, ..., D_M\}$.

3. Формування базової терм-множини з відповідними функціями належності кожної змінної: $A = \{a_1, a_2, ..., a_N\}$.

4. Формування кінцевої множини нечітких правил, узгоджених щодо змінних, що в них використовуються.

5. Реалізація логічного висновку, тобто визначення істинності для передумов кожного правила та визначення нечітких підмножин для змінних виходу для кожного правила.

6. Композиція нечітких підмножин кожної змінної виходу за всіма правилами.

7. Знаходження чіткого значення для кожної з вихідних лінгвістичних змінних.

У ННМ з логічним висновком Мамдані реалізується логічний висновок, який має такі етапи.

1. **Уведення нечіткості.** Знаходимо ступінь істинності для передумов кожного правила: $A_1(x_0), A_2(x_0), B_1(x_0), B_2(x_0)$.

2. **Логічний висновок.** Знаходимо рівні відсікання для передумов кожного з правил (з використанням операції знаходження мінімуму): $\alpha_1 = A_1(x_0) \cap B_1(y_0)$; $\alpha_2 = A_2(x_0) \cap B_2(y_0)$. Знаходимо рівні відсікання функції належності: $C_1' = (\alpha_1 \cap C_1(z))$; $C_2' = (\alpha_2 \cap C_2(z))$.

3. **Композиція.** Підсумовуємо знайдені відсічені функції належності з використанням операції максимум та отримуємо підсумкову нечітку підмножину для змінної виходу з функцією належності: $\mu_\Sigma = C(z) = C_1'(z) \cup C_2'(z) = (\alpha_1 \cap C_1(z)) \cup (\alpha_2 \cap C_2(z))$.

4. **Зведення до чіткості.** Використовуються центроїдний метод (рис. 4).

Вхідними є такі фактори впливу, як вік, рівень заробітної плати, термін кредитування, а вихідним — імовірність повернення кредиту та відповідний фінансовий клас позичальника. Задачу оцінювання кредитоспроможності можна сформулювати таким чином [17–19]. Кожна кредитна заявка задається вектором $\{x_1, x_2, .., x_i, .., x_n\}$, де x_i — певним чином формалізовані дані з анкети позичальника та параметри кредиту. Далі за заданим вектором треба прийняти рішення про надання кредиту, тобто оцінити ймовірність повернення кредиту. У загальному вигляді нечіткий логічний висновок має такі етапи: визначення множини вхідних змінних; визначення множини вихідних змінних; формування базової терм-множини з відповідними функціями належності кожного терму; формування кінцевої множини нечітких правил, узгоджених щодо використовуваних у них змінних. Знаходження чіткого значення для кожної з вихідних лінгвістичних змінних.

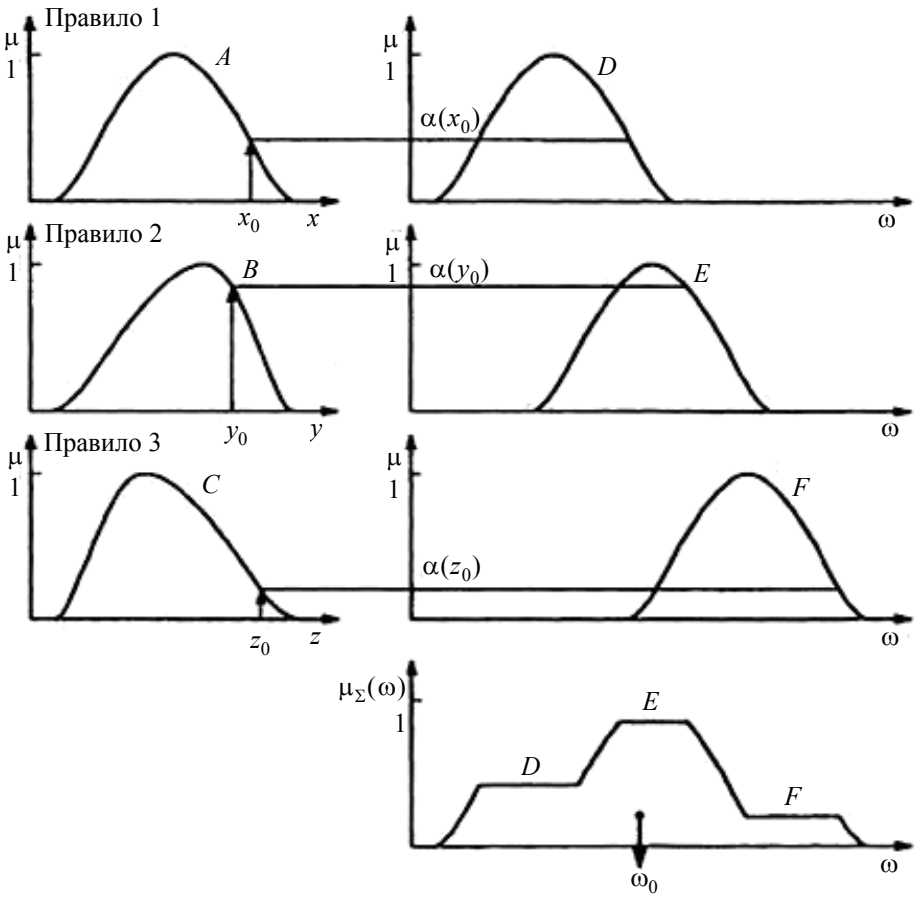

Рис. 4. Ілюстрація логічного висновку Мамдані

Під час виконання аналітичних досліджень визначено перелік факторів, що впливають на формування рівня кредитоспроможності клієнта. Нижче подано перелік вхідних факторів, використаних для розроблення системи, зокрема: 1) вік позичальника (x_1); 2) стаж роботи (x_2); 3) місячна заробітна плата (x_3); 4) наявність майна (x_4); 5) наявність кредитної історії (x_5); 6) сума кредиту (x_6); 7) термін кредитування (x_7).

Для оцінювання кредитоспроможності фактори, що впливають на формування рівня кредитоспроможності клієнта, можуть набувати таких значень: вік = 18–65 (років); стаж роботи = 0–49 (років); заробітна плата = = 0–30000 (грн); наявність майна = 1 — ні; 2 — машина; 3 — квартира; 4 — дім; 5 — два та більше з перелічених значень. Наявність кредитної історії: 1 — є кредити, оплата за якими прострочена; 2 — не було; 3 — є кредити в інших банках і погашаються своєчасно; 4 — є кредити у банку і погашаються своєчасно; 5 — немає відкритих кредитів, попередні сплачувались своєчасно; сума кредиту: 1000–100000 (грн); термін кредитування: від 4 місяців до 5 років.

Як вихідну маємо лінгвістичну змінну «імовірність повернення кредиту», що має такі терми: «дуже низька ймовірність повернення», «низька ймовірність повернення», «середня ймовірність повернення», «висока ймо-

вірність повернення», «дуже висока ймовірність повернення». Приклади функцій належності для деяких змінних подано на рис. 4, 5.

Для запису правил, що використовувались у програмі, уведемо позначення: «дуже низький» — ДН, «низький» — Н, «середній» — Ср, «високий» — В, «дуже високий» — ДВ.

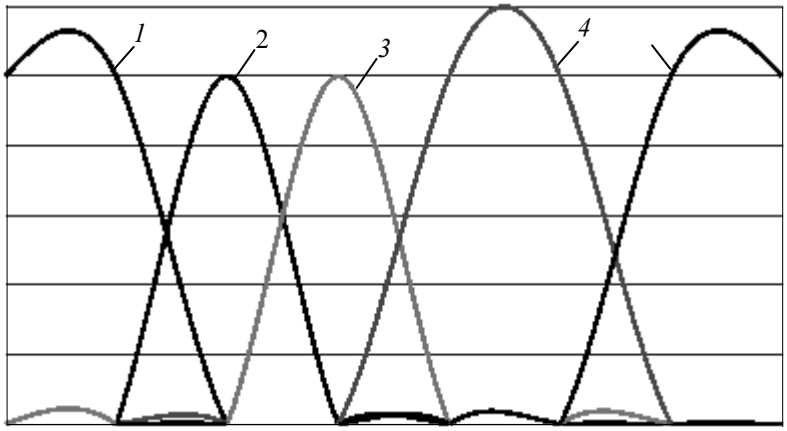

Рис. 4. Функція належності змінної «вік»:: *1* — ДН, *2* — Н, *3* — Ср, *4* — ДВ, *5* — В

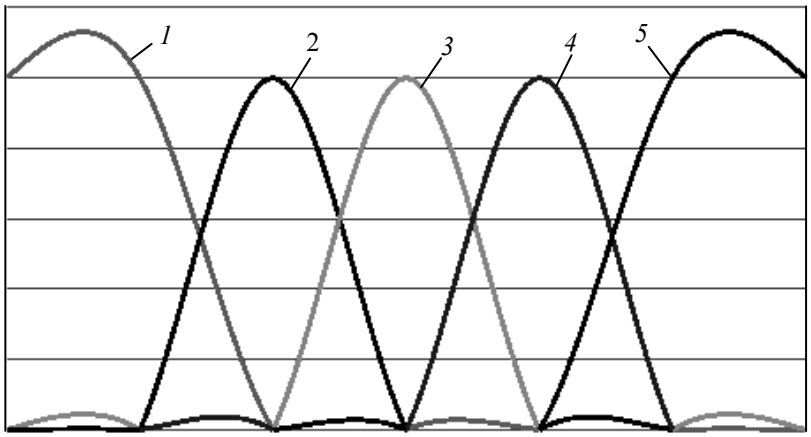

Рис. 5. Функція належності змінної «спроможність погашення»: *1* — ДВ, *2* — В, *3* — Ср, *4* — Н, *5* — ДН

Змінна «спроможність погашення» визначається так:

$$Спроможність\ погашення = \frac{Сума\ кредиту \times Строк\ кредитування}{Заробітна\ плата}.$$

РОЗРОБЛЕНЕ ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

Для розроблення автоматизованої системи підтримання прийняття рішень (СППР) використано Microsoft Visual Studio (мова програмування — C#). Приклад інтерфейсу розробленої системи для оцінювання кредитоспроможності фізичних осіб, вхідними змінними якої є зазначені фактори впливу, а

вихідною змінною — імовірність повернення кредиту, подано на рис. 6. Після заповнення полів з інформацією про позичальника необхідно натиснути кнопку «Обчислити». Результатом роботи команди є відповідна інформація навпроти поля «Імовірність повернення кредиту» та фінансовий клас позичальника.

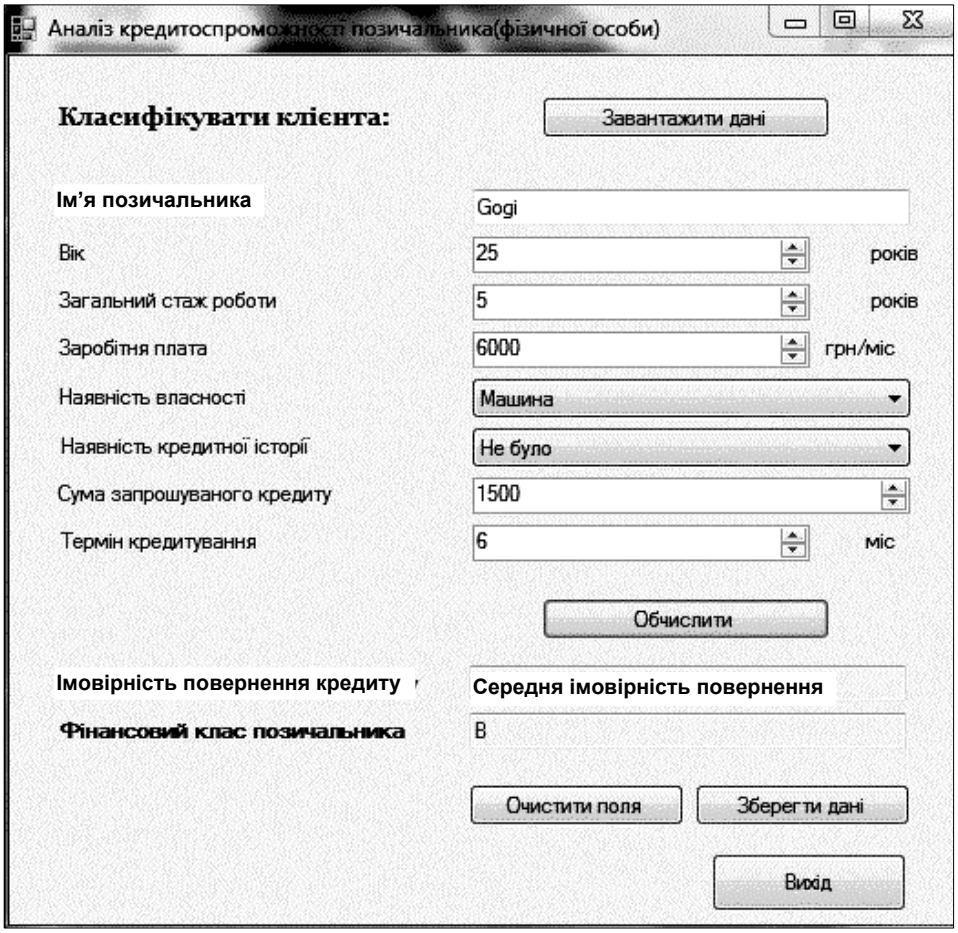

Рис. 6. Приклад роботи програми

АНАЛІЗ РЕЗУЛЬТАТІВ

Для аналізу якості моделей і встановлення найкращої моделі для розв'язання певної задачі використовують декілька критеріїв для оцінювання адекватності моделей: загальну точність моделі; помилки I та II роду; ROC-криву та індекс GINI та загальну точність моделі (CA — Common Accuracy).

Індекс GINI можна визначити через площу фігури, що міститься під ROC-кривою, таким чином: $GINI = 2 \cdot AUC - 1$. Діапазон значень індексу GINI становить $0 \le G \le 1$, а моделі з найвищою роздільною здатністю, тобто моделі, які виконують високоякісне сортування схильних до дефолту клієнтів і клієнтів, не схильних до дефолту, отримають найвищі коефіцієнти. На практиці оцінка якості моделі суттєво залежить від даних, за якими вона бу-

дується. Для застосування на практиці скорингу (оцінки фінансового стану нових клієнтів) індекс GINI на рівні 55% є вже дуже високим, у той час як для скорингу поведінки (оцінки фінансового стану існуючих клієнтів) індекс GINI зазвичай набуває значень вищих за 70%.

Значення точок ROC-кривої використано для знаходження порога відсікання — компромісу між чутливістю та специфічністю моделі. Критеріями вибору порога відсікання можуть бути:

1) вимога мінімальної величини чутливості (специфічності) моделі;

2) вимога максимальної сумарної чутливості та специфічності моделі, тобто

$$cut-off = \max_k (Se_k + Sp_k);$$

3) вимога балансу між чутливістю і специфічністю [6], тобто коли

$$Sp \approx Se: \; cut-off = \min_k | Se_k - Sp_k |.$$

За побудованими моделями виконано аналіз прогнозів останніх 300 значень (з 1600) про кредити, отриманих за методами логістичної регресії, дерев рішень, МБ та нечіткої логіки, порівняно з реальними значеннями. Після проведення ROC-аналізу і встановлення порогів відсікання на певних рівнях, тобто, якщо ймовірність дефолту перевищує вказаний рівень, то клієнт вважається таким, що не поверне кредиту, отримано таблицю результатів для логістичної регресії (табл. 3).

Таблиця 3. Загальна точність моделі та помилки I та II роду для різних рівнів порога відсікання, отримані для логістичної регресії

Характеристика	**Прогноз: Повернення кредиту (0)**	**Прогноз: Дефолт (1)**
cut-off = 0,1		
Факт: Повернення кредиту (0)	57	183
Факт: Дефолт (1)	0	60
Загальна точність моделі = 61,9 %		
cut-off = 0,15		
Факт: Повернення кредиту (0)	86	154
Факт: Дефолт (1)	1	59
Загальна точність моделі = 67 %		
cut-off = 0,2		
Факт: Повернення кредиту (0)	109	131
Факт: Дефолт (1)	3	57
Загальна точність моделі = 70,3 %		
cut-off = 0,25		
Факт: Повернення кредиту (0)	151	89
Факт: Дефолт (1)	6	54
Загальна точність моделі = 76,5 %		
cut-off = 0,3		
Факт: Повернення кредиту (0)	184	56
Факт: Дефолт (1)	11	49
Загальна точність моделі = 79,2 %		

Максимальна точність моделі на рівні 79,2 % досягається за значення порога відсікання 0,3, при цьому модель пропускає 11 дефолтів та відсіює 23,3% добросовісних позичальників. Значення площі під ROC-кривою становить: $AUC = 0{,}775$, а індекс GINI відповідно: $GINI = = 2AUC - 1 = 0{,}549$. На основі навчальної вибірки з 1300 клієнтів також побудовано МБ. Для перевірної вибірки (300 випадків) обчислено ймовірності дефолтів (табл. 4).

Таблиця 4. Загальна точність моделі та помилки I та II роду для різних рівнів порога відсікання, отримані для МБ

Характеристика	**Прогноз: Повернення кредиту (0)**	**Прогноз: Дефолт (1)**
cut-off = 0,1		
Факт: Повернення кредиту (0)	3	237
Факт: Дефолт (1)	0	60
Загальна точність моделі = 50,6 %		
cut-off = 0,15		
Факт: Повернення кредиту (0)	3	237
Факт: Дефолт (1)	0	60
Загальна точність моделі = 50,6 %		
cut-off = 0,2		
Факт: Повернення кредиту (0)	22	218
Факт: Дефолт (1)	0	60
Загальна точність моделі = 54,6 %		
cut-off = 0,25		
Факт: Повернення кредиту (0)	23	217
Факт: Дефолт (1)	0	60
Загальна точність моделі = 54,8 %		
cut-off = 0,3		
Факт: Повернення кредиту (0)	146	94
Факт: Дефолт (1)	3	57
Загальна точність моделі = 77,9 %		

Найбільша точність моделі досягається на рівні 77 % під час встановлення порога 0,3; при цьому буде пропущено три дефолти та відкинуто 39 % добросовісних позичальників. Значення площі під кривою становить: $AUC = 0{,}879$, а індекс GINI відповідно: $GINI = 2AUC - 1 = 0{,}758$. Результати оцінювання кредитоспроможності для різних порогів відсікання під час використання нечіткої логіки подано в табл. 5.

Як видно з отриманих результатів нечітка модель показує вищі значення загальної точності для порогів відсікання 0,2 і 0,25, а також відсіює менше добросовісних позичальників порівняно з даними для МБ. Значення площі під кривою для моделі на основі нечіткої логіки: $AUC = 0{,}8875$, а індекс GINI відповідно: $GINI = 2AUC - 1 = 0{,}775$, що є кращим результатом.

Таблиця 5. Загальна точність моделі та помилки I та II роду для різних рівнів порога відсікання, отримані для нечіткої логіки

Характеристика	Прогноз: Повернення кредиту (0)	Прогноз: Дефолт (1)
cut-off = 0,1		
Факт: Повернення кредиту (0)	42	198
Факт: Дефолт (1)	0	60
Загальна точність моделі = 58,75 %		
cut-off = 0,15		
Факт: Повернення кредиту (0)	88	152
Факт: Дефолт (1)	0	60
Загальна точність моделі = 68,3 %		
cut-off = 0,2		
Факт: Повернення кредиту (0)	121	119
Факт: Дефолт (1)	0	60
Загальна точність моделі = 75,3 %		
cut-off = 0,25		
Факт: Повернення кредиту (0)	161	79
Факт: Дефолт (1)	2	58
Загальна точність моделі = 81,9 %		
cut-off = 0,3		
Факт: Повернення кредиту (0)	173	67
Факт: Дефолт (1)	3	57
Загальна точність моделі = 83,5 %		

Точність моделі та кількість помилок I та II роду залежать від порога відсікання, який буде встановлений банком. Установлення порога відсікання визначає не лише відсоток відсіяних клієнтів, а і нижню межу ймовірності повернення кредиту, тобто поріг, нижче за який клієнт вважається таким, що не поверне кредит. При цьому значення ймовірності дефолту 0,1 або 0,2 для клієнта є статистично незначними, а тому поріг відсікання доцільно встановлювати на рівні 0,25–0,3. Таким чином, підхід за допомогою нечіткої логіки до оцінювання кредитоспроможності позичальника дозволяє підвищити якість моделі, а також її здатність розрізняти надійних та ненадійних клієнтів та зменшити кількість некоректно відсіяних клієнтів. Результати застосування використаних методів оцінювання кредитоспроможності наведено на табл. 6.

Таблиця 6. Порівняльна таблиця характеристик для різних моделей

Назва методу	Індекс GINI	Значення AUC	Точність моделі, %	Якість моделі
Бінарна логістична регресія	0,5491	0,7745	79,2	Прийнятна
Дерева рішень	0,548	0,774	71,25	Прийнятна
Мережа Байєса	0,758	0,879	77,9	Висока
Нечітка логіка	0,775	0,8875	83,5	Висока

Отримані результати обчислювальних експериментів свідчать, що найкращі результати дає нечітка логіка (83,5%), також хороший результат дає модель у формі МБ (77,9%). Високі значення точності моделі отримано за логістичною регресією (79,2%). Ці результати ще раз підтверджують доцільність використання нечіткої логіки, логістичної регресії, дерев рішень та МБ під час оцінювання кредитоспроможності позичальників кредитів.

ВИСНОВКИ

Виконано аналіз якості моделей, побудованих за методами логістичної регресії, дерев рішень, МБ та нечіткою логікою. Установлено, що модель на основі нечіткої логіки є кращою (у розгляданого випадку) для розв'язання задачі визначення ймовірності дефолту для кредитного позичальника, на що вказує порахована точність моделі, яка становить 83,5% для моделі нечіткої логіки та 77,9% для МБ. Це зумовлено тим, що використання методу нечіткої логіки (за нечітким висновком Мамдані) дає змогу точніше встановити причинно-наслідкові зв'язки між характеристиками-факторами задачі, їх вплив на вихідну змінну.

Для аналізу якості результатів використано кілька критеріїв. Для кожної з моделей обчислено помилки I та II роду, загальну точність моделей для заданих порогів відсікання, значення яких безпосередньо впливають на точність моделі та кількість помилок I та II роду. Побудовано ROC-криві і обчислено площі під кривою та індекс GINI.

Таким чином, застосування методів інтелектуального аналізу даних до розв'язання задачі прогнозування кредитоспроможності клієнтів фінансової установи уможливлює отримання високоякісних результатів, придатних для подальшого практичного використання для прийняття рішень.

У подальших дослідженнях доцільно використати більші обсяги статистичних даних і розширити кількість змінних, які характеризують позичальників кредиту. Необхідно також створити інформаційну систему підтримання прийняття рішень комерційного типу на основі комбінованого використання ідеологічно різних типів моделей: статистичного типу та інтелектуального аналізу даних для розв'язання задач такого класу. Це суттєво спростить роботу працівникам банків, які займаються розв'язуванням задач оцінювання та менеджменту ризиків.

ЛІТЕРАТУРА

1. *Крючковский В.В.* Экспертная система оценки кредитоспособности банковских клиентов на основе методов нечеткой логики и сети Байеса / В.В. Крючковский, С.А. Бабичев, А.В. Шарко // Экономика научно-технического прогресса. — 2009. — № 1. — С.197–205.
2. *Allen S.* Financial risk management: A practitioner's guide to managing market and credit risk / S. Allenб. — Hoboken, N.J.: John Wiley & Sons, Inc., 2003. — 567 p.
3. *Hennie van Greuning.* Analyzing and managing banking risk: a framework for assessing corporate governance and financial risk / Hennie van Greuning, Sonja Bratanovic. — 2nd ed.
4. *Костюченко Н.С.* Анализ кредитных рисков / Н.С. Костюченко. — СПб.: ИТД «Скифия», 2010. — 440 с.

5. *Бідюк П.І.* Моделі оцінки ризиків кредитування фізичних осіб / П.І. Бідюк, Є.О. Матрос // Кібернетика та обчислювальна техніка. — 2007. — № 153. — С. 87–95.
6. *Бідюк П.І.* Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / П.І. Бідюк, Н.В. Кузнєцова // Наукові вісті НТУУ «КПІ». — 2010. — № 1. — С. 42–53.
7. *Колпаков В.М.* Теория и практика принятия управленческих решений: учеб. пособие / В.М. Колпаков. — 2-е изд., перераб. и доп. — К.: МАУП, 2004. — 504 с.
8. *Agresti A.* Building and applying logistic regression models. An Introduction to Categorical Data Analysis / A. Agresti. — Hoboken, New Jersey: Wiley, 2007. — 138 p.
9. *Shariff A.* The Comparison Logit and Probit Regression Analyses in Research / A. Shariff, A. Zaharim, K. Sopian. — 2009. — Vol. 27, N 4. — P. 548–553.
10. *Терентьев А.Н.* Сравнение методов интеллектуального анализа данных при оценивании кредитоспособности физических лиц / А.Н. Терентьев, П.И. Бидюк, А.В. Миронова, Н.Ю. Мединн // Проблемы управления и информатики. — К.: ИКИ НАНУ-НКАУ, 2009. — № 5. — С. 141–149.
11. *Терентьев А.Н.* Метод вероятностного вывода в байесовских сетях по обучающим данным / А.Н. Терентьев, П.И. Бидюк // Кибернетика и системный анализ. — 2007. — № 3. — С. 93–99.
12. *Бідюк П.І.* Основні етапи побудови і приклади застосування мереж Байєса / П.І. Бідюк, Н.В. Кузнєцова // Системні дослідження та інформаційні технології. — 2007. — № 4. — С. 26–39.
13. *Heckerman D.* Bayesian Networks for Data Mining / D. Heckerman // Data Mining and Knowledge Discovery. — 1997. — № 1. — P. 79–119.
14. *Кузнєцова Н.В.* Гібридні мережі Байєса: основні особливості і точні методи формування висновку / Н.В. Кузнєцова // Праці Одеського політехн. ун-ту. — Одеса, 2009. — Вип. 1(31). — С. 114–121.
15. *Кузнєцова Н.В.* Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса / Н.В. Кузнєцова, П.І. Бідюк // Наукові вісті НТУУ «КПІ». — 2008. — № 3. — С. 11–24.
16. *Недосекин А.* Методологические основы моделирования финансовой деятельности с использование нечетко множественных описаний / А. Недосекин. — СПб, 2003. — 280 с.
17. *Зайченко Ю.П.* Оценка кредитных банковских рисков с использованием нечеткой логики // Системні дослідження та інформаційні технології. — 2010. — № 2. — С. 37–54.
18. *Шовгун Н.В.* Аналіз кредитоспроможності позичальника за допомогою методів з нечіткою логікою / Н.В. Шовгун // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: зб. наук. праць. — К.: Век+, 2012. — № 55. — С.169–173.
19. *Shovgun Natalia.* Fuzzy neural networks for evaluating the creditworthiness of the borrowers / Natalia Shovgun // Information theories & applications. ITHEA IBS ISC. — 2014. — Vol. 21. — P. 54–257.

Надійшла 28.01.2019