

Магістерська дисертація на тему:

**Моделі і методи інтелектуального
аналізу даних для оцінювання
кредитоспроможності фізичних осіб**

Студентки групи КА-42м
Яковенко Мар'яни Миколаївни
Керівник роботи
Бідюк Петро Іванович

Мета дослідження

Побудова математичних моделей для аналізу кредитоспроможності фізичних осіб на основі статистичних даних методами інтелектуального аналізу даних.

Об'єкт дослідження

Позичальники банківських кредитів, представлені статистичними даними.

Предмет дослідження

Математичні методи і моделі інтелектуального аналізу даних, методика рейтингового оцінювання.

Актуальність досліджень

Банківська сфера - невід'ємна складова економіки та всього життя. Саме завдяки розвитку банківської системи фізичні особи можуть брати кредити на вигідніших умовах.

Однак разом зі збільшенням попиту на позиковий капітал, гостріше постає питання адекватного оцінювання кредитоспроможності потенційного позичальника.

Саме тому виникла потреба у розробці спеціальних методик, які враховували б особливості характеристик фізичних осіб і дали можливість вирішити дану проблему.

Постановка задачі

1. Виконати огляд сучасних методів моделювання і прогнозування кредитоспроможності позичальників (фізичних осіб).
2. Розробити моделі для аналізу кредитоспроможності фізичних осіб на основі результатів теорії нечіткої логіки, рейтингового підходу та нейронної мережі.
3. Створити проект та програмно реалізувати інформаційно-аналітичну систему для оцінювання кредитоспроможності позичальників(фізичних осіб).
4. Виконати порівняльний аналіз вибраних методів прогнозування кредитоспроможності позичальників (фізичних осіб) і вибрати кращий для практичної реалізації.

Кредитоспроможність та система кредитного скорингу

Кредитоспроможність клієнта (позичальника) - здатність особи повністю і в строк розрахуватися за своїми борговими зобов'язаннями.

Система кредитного скорингу - це певна математична модель, яка дозволяє оцінювати кредитоспроможність потенційного позичальника числовим показником.

Аналіз існуючих підходів до розв'язання задачі прогнозування кредитоспроможності фізичних осіб

При створенні системи кредитного скорингу використовується такі алгоритми:

- ▶ Класифікаційні дерева
- ▶ Байєсівський підхід на основі мереж довіри
- ▶ Моделювання на основі нечіткої логіки
- ▶ Рейтинговий підхід
- ▶ Нейронні мережі

Для розробки програми було обране моделювання на основі нечіткої логіки, яке може працювати як з якісними, так і з кількісними характеристиками, а також рейтинговий підхід та нейронну мережу.

Нечітка логіка

Використовується у різних експертних і керуючих системах, механізм нечітких висновків у своїй основі має базу знань, сформовану фахівцями предметної галузі у вигляді сукупності нечітких предикатних правил вигляду:

П1 : якщо $x \in A1$ то $y \in B1$

П2 : якщо $x \in A2$ то $y \in B2$

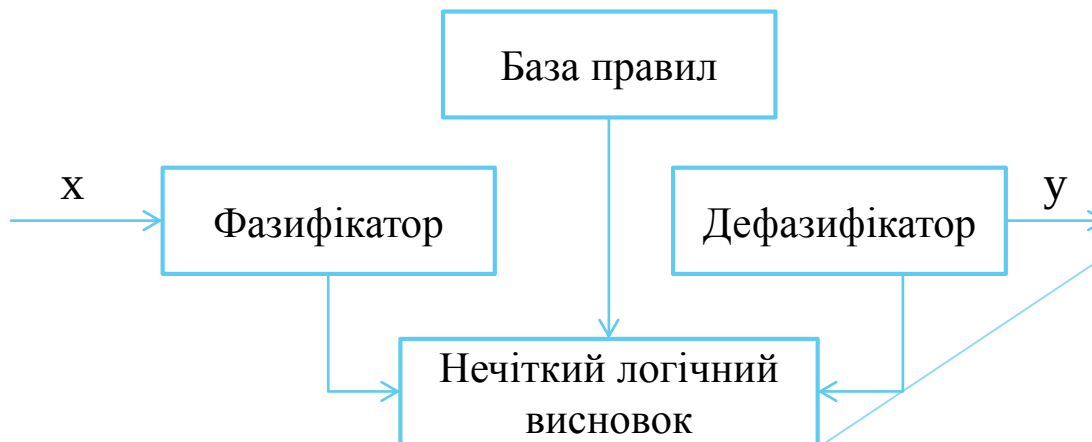
...

Пn : якщо $x \in An$ то $y \in Bn$

де x - вхідна змінна (ім'я для відомих значень даних);

y - змінна висновку (ім'я для значення даних, що буде обчислене);

A та B - функції належності, визначені відповідно на x та y .



Метод Мамдані

Введення нечіткості. Знаходяться ступені істинності для передумов кожного правила: $A_1(x_0)$, $A_2(x_0)$, $B_1(x_0)$, $B_2(x_0)$.

Логічний висновок. Знаходяться рівні “відсікання” для передумов кожного з правил (з використанням операції МІНІМУМ):

$$\alpha_1 = A_1(x_0) \wedge B_1(y_0);$$

$$\alpha_2 = A_2(x_0) \wedge B_2(y_0);$$

де через “ $\wedge$ ” позначена операція логічного мінімуму (min). Потім знаходяться “усічені” функції належності:

$$C'_1 = (\alpha_1 \wedge C_1(z));$$

$$C'_2 = (\alpha_2 \wedge C_2(z)).$$

Композиція. Проводиться об'єднання знайдених усічених функцій з використанням операції МАКСИМУМ (max, позначені далі як “ $\vee$ ”), що призводить до одержання підсумкової нечіткої підмножини для змінної виходу з функцією належності:

$$\mu_{\Sigma}(z) = C(z) = C'_1(z) \vee C'_2(z) = (\alpha_1 \wedge C_1(z)) \vee (\alpha_2 \wedge C_2(z)).$$

Приведення до чіткості. Проводиться для перебування z_0 , наприклад, центроїдним методом.

Рейтинговий підхід

Для визначення узагальненого показника фінансового стану позичальника необхідно розрахувати показники, значення кожного помножити на вагоме значення та знайти суму всіх значень. Після цього визначається клас позичальника за рейтинговою шкалою.

Однією з найпоширеніших моделей даного класу є модель Д. Дюрана, яка дозволяє з достатньою достовірністю визначити ступінь кредитного ризику при отриманні споживчого кредиту.

Він використовував такі фактори: вік, стать, термін проживання, професія, робота в галузі, зайнятість, фінансові показники (наявність банківського рахунку, володіння нерухомістю, наявність полісу страхування життя).

Нейронна мережа на основі TSK

Схема висновку при використанні M правил та N змінних x_j ($i = 1, 2, \dots, M$) має вигляд:

$$\Pi_i: \text{якщо } x_1 \text{ це } A_1^i; x_2 \text{ це } A_2^i; \dots; x_N \text{ це } A_N^i, \text{ то } y_i = p_{i0} + \sum_{j=1}^N p_{ij}x_j$$

A_i^k – значення лінгвістичної змінної x_i , $i = 1, 2, \dots, N$ для правила Π_k з функцією належності:

$$\mu_A^k(x_i) = \frac{1}{1 + \left(\frac{x_i - c_i^k}{\sigma_i^k}\right)^{2b_i^k}}$$

При M правилах агрегований вихідний результат має вигляд:

$$y(x) = \frac{\sum_{k=1}^M w_k y_k(x)}{\sum_{k=1}^M w_k}, y_i(x) = p_{i0} + \sum_{j=1}^N p_{ij}x_j, \text{ де } w_k = \mu_A^k(x) = \prod_{j=1}^N \frac{1}{1 + \left(\frac{x_j - c_j^k}{\sigma_j^k}\right)^{2b_j^k}}$$

Нейронна мережа на основі TSK

Виділяють 5 шарів:

- ▶ I-ий шар. Виконує фазифікацію кожної змінної. Параметричний шар з параметрами $c_j^{(i)}, \sigma_j^{(i)}, b_j^{(i)}$.
- ▶ II-ий шар. Виконує агрегування окремих змінних, визначаючи результуюче значення коефіцієнта належності $w_i = \mu_A^i(x)$.
- ▶ III-ий шар. Розраховуються значення $y_i(x) = p_{i0} + \sum_{j=1}^N p_{ij}x_j$, а також виконується множення $y_i(x)$ на w_i . Параметри p_{ij} підлягають адаптації.
- ▶ IV-ий шар. Розраховується зважена сума сигналів $y_i(x)$ та сума ваг $\sum_{k=1}^M w_k$.
- ▶ V-ий шар. У п'ятому шарі ваги підлягають нормалізації та обчислюється вихідний сигнал:

$$y(x) = \frac{\sum_{k=1}^M w_k y_k(x)}{\sum_{k=1}^M w_k}$$

Критерії якості рішення

Для аналізу якості моделей і встановлення найкращої моделі для розв'язання певної задачі в данній роботі використовуються наступні критерії для оцінювання адекватності моделей:

- 1) загальна точність моделі;
- 2) помилки I-го та II-го роду.

	Прогноз моделі: Повернення кредиту (0)	Прогноз моделі: Дефолт (1)
Фактично: Повернення кредиту(0)	Вірно класифіковані (TP)	Помилки II-го роду (FN)
Фактично: Дефолт (1)	Помилки I-го роду (FP)	Вірно класифіковані (TN)

Програмний продукт

Створено програмний продукт для розв'язання задачі аналізу кредитоспроможності фізичних осіб.

Для реалізації обрано мову програмування C#, тому що C# є простою та багатовункціональною мовою, яка має велику кількість бібліотек, що працюють з максимальною швидкодією.

Реалізований програмний продукт заснований на алгоритмі формування логічного висновку Мамдані, рейтинговій моделі Дюрана для прогнозування кредитоспроможності фізичних осіб та нейронній мережі на основі TSK.

Результати застосування програми для аналізу кредитоспроможності однієї фізичної особи

Прогнозування кредитоспроможності фізичних осіб

ПІБ

Вік Стать

Термін проживання у даній місцевості

Освіта

Сімейний стан

Власність

Стаж роботи

Місце роботи

Професія

Дохід

Кредит

Термін кредиту

Набраний бал

Чи варто надати кредит

Ймовірність повернення кредиту

Чи варто надати кредит

Результати застосування програми для аналізу кредитоспроможності групи фізичних осіб

Form3

	Вік	Стать	Срок проживання	Освіта	Сімейний стан	Власність	Стаж роботи	Галузь	Професія	Дохід	Кредит	Термін
▶	26	Male	15	HIGH	MARRIED	GR	less_0,5year	OTHER	LOW_RISK	2180	1390	12
	24	Female	7	HIGH	SINGLE	CA	less_0,5year	OTHER	LOW_RISK	3600	1700	6
	22	Female	12	HIGHNF	SINGLE	CA	less_0,5year	OTHER	LOW_RISK	3100	740	9
	22	Male	3	HIGH	MARRIED	FL	less_0,5year	GROMAD	OTHER_P	2077	480	12
	54	Female	11	SEC	MARRIED	CA	more_10_years	OTHER	OTHER_P	2200	2395	12
	35	Male	20	SECSP	MARRIED	FL	5-10_years	GROMAD	LOW_RISK	1460	2882	24
	22	Male	11	SECSP	SINGLE	NO	0,5-1year	GROMAD	HIGH_RISK	300	1300	12
	22	Male	7	SECSP	SINGLE	FL	2-5years	GROMAD	OTHER_P	500	2374	12
	27	Male	6	SECSP	CIVLMARRIAGE	FL	2-5years	GROMAD	HIGH_RISK	1600	2410	18
	30	Male	17	HIGH	MARRIED	CA	0,5-1year	OTHER	OTHER_P	8450	370	12
	32	Female	4	HIGH	CIVLMARRIAGE	NO	0,5-1year	OTHER	OTHER_P	2350	480	9
	49	Male	25	HIGH	MARRIED	NO	more_10_years	OTHER	HIGH_RISK	950	4995	24
	25	Male	14	SEC	SINGLE	NO	2-5years	OTHER	OTHER_P	600	3875	24
	26	Male	14	SEC	SINGLE	NO	1-2years	GROMAD	HIGH_RISK	1130	2495	12

Завантажити дані

Розрахувати

Рейтинговий метод

Загальна точність моделі

0,59150530918176

Помилки I-го роду

0,19237976264834

Помилки II-го роду

0,21549031855090

Зберегти результати(рейтинговий метод)

Метод Мамдані

Загальна точність моделі

0,69456589631480

Помилки I-го роду

0,26483447845096

Помилки II-го роду

0,03997501561524

Зберегти результати(метод Мамдані)

Результати застосування програми для аналізу кредитоспроможності за допомогою нейронної мережі

ТСК

Параметры сети

Количество правил

Количество переменных

Количество итераций

Точность

Начальный шаг

Порог

Обучить

Создать сеть

Проверить

Результаты

Критерий

Общая точность

Ошибки 1го рода

Ошибки 2го рода

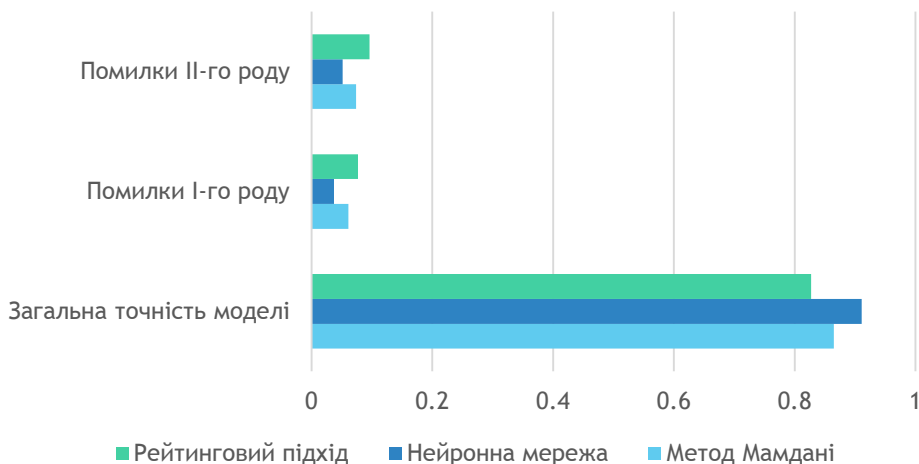
PLATEG	VOZRAST	CREDIT	SROKRABOTI	DOHOD	VOZVRATCREDITA
143,0212	26	1390	0	2180	1
265,4618	24	1700	0	3600	1
98,8836	22	740	0	3100	1
45,0671	22	480	0	2077	1
246,4286	54	2395	34	2200	1
147,2715	35	2882	6	1460	1
102,8929	22	1300	0	300	0

Загрузить данные

Возврат кредита	Значение в сети	Решение по кредиту
1	0,940835388285719	1
1	0,929081201713133	1
1	1,05334167631637	1
1	0,580463184767677	1
1	0,815773013090946	1
1	0,373352349568627	0
0	0,374223653830057	0

Порівняльна таблиця характеристик для різних моделей

Точність моделей



Назва методу	Загальна точність моделі	Помилки I-го роду	Помилки II-го роду
Метод Мамдані	86,51%	6,12%	7,37%
Нейронна мережа	91,10%	3,75%	5,15%
Рейтинговий підхід	82,73%	7,68%	9,59%

Для практичної реалізації можна обрати метод Мадані, оскільки він дає більшу точність, ніж рейтинговий підхід. Також до його переваг відноситься те, що він може працювати у випадку, коли дані про фізичну особу неповні. Нейронна мережа дає найбільшу точність, але вона найскладніша у реалізації.

Новизна отриманих результатів

1. Побудована множина моделей для оцінювання кредитоспроможності фізичних осіб на основі фактичних статистичних даних, які забезпечують отримання високоякісного результату прогнозування повернення кредиту.
2. Розроблено оригінальний програмний продукт на основі побудованих математичних моделей, який відрізняється простотою та зручністю практичного застосування, можливістю розширення функцій та отриманням даних з локальної або глобальної мереж.

Висновки до роботи

1. Виконано огляд сучасних методів моделювання і прогнозування кредитоспроможності позичальників (фізичних осіб).
2. Розроблено моделі для аналізу кредитоспроможності фізичних осіб на основі результатів теорії нечіткої логіки, рейтингового підходу та нейронної мережі, які забезпечують отримання високоякісних оцінок прогнозу.
3. Створено програмний продукт на C#: інформаційно-аналітична система для оцінювання кредитоспроможності позичальників (фізичних осіб).
4. Виконано порівняльний аналіз вибраних методів прогнозування кредитоспроможності позичальників (фізичних осіб) і обрано кращий для практичної реалізації.

Дякую за увагу!