doi:https://doi.org/10.15407/emodel.41.02.111

УДК 004.942

П.І. Бідюк, д-р техн. наук, **В.Г. Гуськова**, аспірантка
Інститут прикладного системного аналізу
Національного технічного університету України
«Київський політехнічний ін-т ім. Ігоря Сікорського»
(Україна, 03056, Київ, пр-т Перемоги, 37, кім. 307,
тел. +380972134708, e-mail: pbidyuke_00@ukr.net)

Аналіз кредитоспроможності за допомогою методів інтелектуального аналізу даних

Запропоновано спосіб мінімізації ризику платоспроможності позичальника для банківської системи та інших фінансових компаній, які надають кредити своїм клієнтам. Проведено оцінку кредитоспроможності клієнтів з використанням логістичної регресії, методів на основі нечіткої логіки, нейронної мережі із зворотнім поширенням помилки та дерев рішень. Надано результати оцінки кредитоспроможності позичальників та проаналізовано оцінку стану клієнтів.

К л ю ч о в і с л о в а: ризики, нечітка логіка, логістична регресія, нейронна мережа, дерева рішень, аналіз стану.

Оцінка платоспроможності клієнта є важливою проблемою для економічного, фінансового та банківського секторів, що надають кредити клієнтам. Фінансові системи потерпають від неплатоспроможності клієнтів, що призводить до погіршення фінансових результатів діяльності банків, скорочення обсягів депозитів і кредитів та спричиняє значне падіння фінансової стійкості банківських установ. Відсутність сучасних методик оцінки клієнтів і позичальників та їх фінансової спроможності сприяє зростанню заборгованостей та неповерненню кредитів.

Під час збору статистичних даних завчасно невідомі фактори, які спричиняють будь-який вплив на результат. Через це фінансові установи збирають всю інформацію, на основі якої створюється модель та визначається вплив факторів на результат. На цьому етапі здійснюється аналіз отриманих даних і виявляються суттєві фактори, застосовуються методи логістичної регресії для дослідження залежних та незалежних змінних. Залежна змінна має два значення (в кредитному скорингу: дефолт—відсутність дефолту) і біноміальний розподіл.

При дослідженні застосовано логістичну регресію та нечітки правила, засновані на принципах навчання Л.Х. Ванга та Ж.М. Менделя, для розв'язання задач регресійного аналізу (FRBS (Fuzzy rule-based systems)). Ці принципи були широко застосовані в інженерних та наукових областях в галузі біоінформатики, отриманні даних, задачах управління аналізу фінансових даних, робототехніки і розпізнавання образів [1—3].

Сучасні системи оцінки, засновані на математичних моделях, дозволяють вирішувати проблеми, базуючись на можливостях клієнта щодо його платоспроможності. Проведено оцінку кредитоспроможності з використанням таких підходів, як дерева рішень, моделі лінійного регресійного аналізу, логіт-, пробіт-моделі, байєсовські мережі, системи, основані на нечіткіх правилах FRBS та абстрактна штучна нейронна мережа Backpropagation. Дослідження зосереджено на принципах, методах, процедурах та засобах системного підходу до аналізу та управління фінансовими ризиками [3, 4].

Початковий набір даних складає 9999 позичальників, які мають такі показники (атрибути): AGE — вік клієнта; EDUCATION — освіта клієнта; MARRIAGE — сімейний статус; ADRESS — місце проживання та кількість його змін; LIMIT_BAL — сума кредиту; BILL_AMT1 — лімітний баланс; PAY_AMT1 — рахунки; PAY_AMT2 — попередні платежі; PAYMENT — змінна результату (ймовірність повернення коштів позичальником). Всю вибірку було поділено на навчальну та перевірочну у таких співвідношеннях: 70% — навчальна, 30% — перевірочна; 80% — навчальна, 20% — перевірочна; 90% — навчальна, 10% — перевірочна.

Аналіз на основі логістичної регресії. Логістична регресія — статистична нелінійна регресійна модель, яку застосовують у випадку, коли залежна змінна є категорійною, тобто може набувати тільки двох значень (0 або 1). Нехай ця величина залежить від деякої множини пояснювальних змінних $x = (1, x_1, ..., x_n)^T$. Залежність y від $x_1, ..., x_n$ можна визначити за допомогою введення додаткової змінної y^*:

$$y^* = \theta^T x = \theta_0 + \theta_0 x_1 + ... + \theta_n x_n + \varepsilon.$$

Тоді

$$y^* = \begin{cases} 0, & y \le 0, \\ 1, & y > 0. \end{cases}$$

При побудові логістичної моделі стохастичний доданок ε вважається випадковою величиною з логістичним розподілом ймовірностей. Відповід-

ним конкретним значенням змінних $x = x_1, ..., x_n$ відповідає значення y^* і ймовірність того, що $y = 1$ відповідає функції

$$p(y=1) = p(y \geq 0) =$$
$$= p(\theta^T x + \varepsilon \geq 0) = p(\varepsilon \geq -\theta^T x) = p(\varepsilon \leq \theta^T x).$$

Для максимізації цієї функції можна застосувати, наприклад, метод градієнтного спуску, метод Ньютона чи стохастичний градієнтний спуск [5].

Оцінювання прогностичної здатності моделі будемо виконувати, використовуючи *ROC*-криву, що дає можливість оцінити якість бінарної класифікації. Вона також відома як крива помилок. Кількісну інтерпретацію *ROC* дає показник площі під *ROC*-кривою — *AUC* (Area Under *ROC* Curve). Чим більші значення *AUC*, тим якісніше функціонує класифікатор, при цьому значення 0,5 свідчить про непридатність обраного методу класифікації [6]. Результати побудови логістичної регресії наведено у таблиці.

Системи FRBS, засновані на нечітких правилах, — це добре відомі методи так званих м'яких розрахунків на основі нечітких концепцій для вирішення складних реальних проблем. Вони стали могутнім засобом вирішення різних проблем, таких як невизначеність, неточність та нелінійність. Такі методи зазвичай використовуються для розв'язування задач ідентифікації, класифікації і регресії та поділяються на чотири етапи [3, 7].

Результати оцінки кредитоспроможності позичальника

Тип моделі	Середньо-квадратична похибка (%) при співвідношенні навчальної та перевірочної вибірок (%)		
	70/30	80/20	90/10
Logit model	0,76658	0,76838	0,7612
FRBS			
TRIANGLE	0,6104	0,6105	0,6191
TRAPEZOID	0,23062	0,23055	0,2303
GAUSSIAN	0,22044	0,21703	0,2241
SIGMOID	0,33188	0,33161	0,3297
BELL	0,1969	0,19156	0,2012
BP			
10 HIDDEN LAYERS	0,74196	0,23192	0,7559
3 HIDDEN LAYERS	0,17475	0,18488	0,242
Decision Tree-Regression	0,74426	0,75838	0,7497
Logit model (*AUC*)	0,63068	0,69102	0,6465

Е т а п 1. Рівномірний розподіл вхідних та вихідних просторів початкових даних у нечіткій області. Нечіткі області відповідають інтервалам для кожного лінгвістичного терміну.

Е т а п 2. Створення нечітких правил IF-THEN, які включають дані для навчання, використовуючи базу знань з попереднього етапу. Необхідно обчислити функцію належності для всіх існуючих даних системи адаптації. Для кожного елемента системи визначається лінгвістичний член, який має максимальний ступінь належності для кожної змінної. Цей процес повторюємо для кожного екземпляра системи адаптації при побудові нечітких правил навчальних даних.

Е т а п 3. Ступінь важливості кожного правила визначається підсумовуванням ступенів функцій належності в антицедентній та консеквентній частинах. Використовуються оператори агрегування продукту.

Е т а п 4. Остаточну базу правил отримують за допомогою видалення надлишкових правил. З огляду на наявність ступенів важливості правил, можна видалити надлишкові правила з більш низькими ступенями.

У результаті отримуємо модель Мамдані. На етапі прогнозування виконуються такі чотири кроки: введення нечіткості, перевірка правил, формування висновку і дефазифікація результату.

Введення нечіткості. Визначаємо ступені істинності для передумов кожного правила:

$$A_1(X_0),\ A_2(X_0),\ B_1(X_0),\ B_2(X_0).$$

Логічний висновок. Знаходимо рівні «відсікання» для передумов кожного правила (з використанням операції min):

$$\alpha_1 = A_1(x_0) \Lambda B_1(y_0),$$

$$\alpha_2 = A_2(x_0) \Lambda B_2(y_0),$$

де символом Λ позначена операція логічного мінімуму. Далі знаходимо «усічені» функції належності:

$$C_1' = (\alpha_1 \Lambda C_1(z)),$$

$$C_2' = (\alpha_2 \Lambda C_2(z)).$$

Композиція. Виконуємо об'єднання знайдених усічених функцій з використанням операції max, в результаті чого отримуємо підсумкову нечітку підмножину для змінної на виході з функцією належності:

$$\mu_\Sigma = C(z) = C_1(z)\ v\ C_2(z) = (\alpha_1 \Lambda C_1(z))\ v\ (\alpha_2 \Lambda C_2(z)).$$

Приведення до чіткості. Така операція виконується у тих випадках, коли потрібно перетворити множину нечітких висновків в чітке число для знаходження, наприклад, центроїдним методом [5].

Аргументи моделі. Для реалізації цієї задачі використано програмне середовище для розробки програмного забезпечення RStudio та модель для розв'язання задачі регресії:

WM(data.train, num.labels, type.mf = "GAUSSIAN",
type.tnorm = "PRODUCT", type.implication.func ="ZADEH",
classification = FALSE, range.data = NULL),

де *data.train* — матриця ($m \times n$ нормованих даних для процесу навчання; m — кількість значень; n — кількість змінних; дані повинні бути нормовані між 0 і 1); *num.labels* — матриця ($1 \times n$), елементами якої є кількість мовних термів; n — кількість змінних (значення за замовчуванням — 7); *type.mf* — тип функції належності (значення за замовчуванням — GAUSSIAN). Можливі типи функцій приналежності такі: трикутна функція (TRIANGLE), трапецієвидна функція (TRAPEZOID), гаусова функція (GAUSSIAN), сигмоїдальна функція (SIGMOID), дзвоноподібна функція (BELL); *type.tnorm* — значення, яке надає тип *t*-норми (значення за замовчуванням MIN means standard type (minimum)); *type.implication.func* — значення, що надає тип функції імплікації (значення за замовчуванням — ZADEH); *classification* — булеан, що визначає похибку класифікації; *range.data* — матриця, яка визначає інтервал даних [6].

Алгоритм Backpropagation. Абстрактні штучні нейронні мережі разом із пакетом neuralnet застосовуються в багатьох ситуаціях. Пакет neuralnet побудований для навчання багатошарових персептронів при регресійному аналізі, тобто для наближення функціональних співвідношень між коваріатами. Нейронні мережі використовуються як розширення узагальнених лінійних моделей, а neuralnet є дуже гнучким пакетом серед розробок R.

Реалізовано алгоритм зворотнього поширення і три варіанти повторного перенаправлення, що забезпечує індивідуальний вибір функцій активації [7]. Можна теоретично включити довільне число коваріатів і змінних відповіді, а також прихованих шарів (рис. 1).

Основною ідеєю цього методу є поширення сигналів помилок між виходом та входом мережі у зворотньому напрямку розповсюдження сигналів для звичайного режиму [3]. Алгоритм зворотнього розповсюдження помилок можна застосувати для багатошарового перцептрона. У мережі існує велика кількість входів $x_1, ..., x_n$, виходів *Outputs* та внутрішніх вузлів. Загальна кількість усіх вузлів від одного до N.

Будемо позначати через $w_{i,j}$ ваги, що стоять на ребрі і з'єднують вузли i та j, а через o_i — вихід вузла i. Правильні відповіді мережі t_k, $k \in Outputs$,

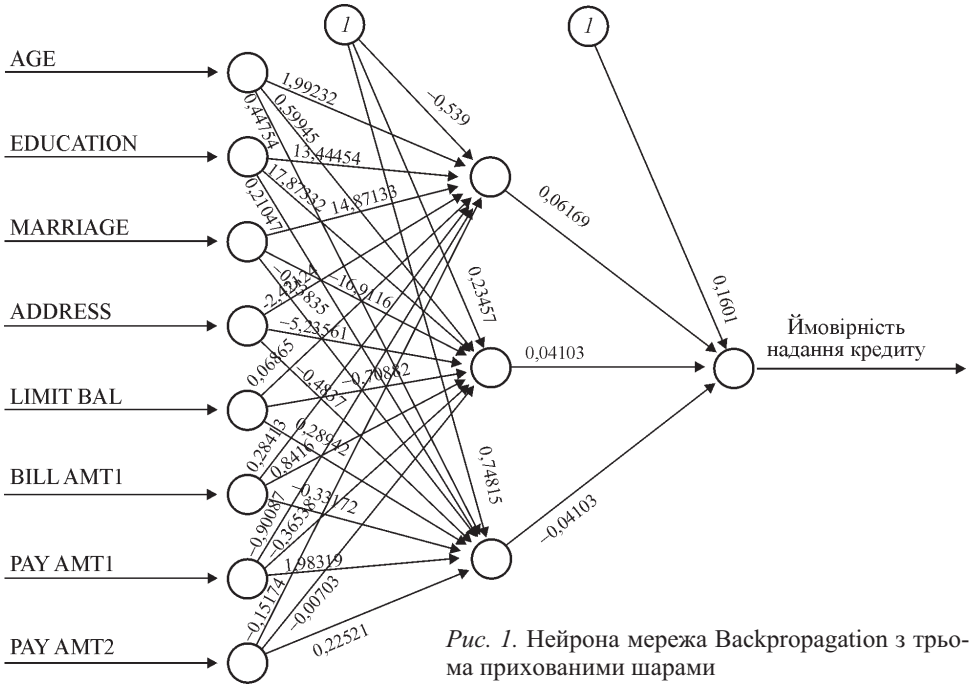

Рис. 1. Нейрона мережа Backpropagation з трьома прихованими шарами

та функція помилки за методом найменших квадратів виглядає так:

$$E(\{w_{i,j}\}) = \frac{1}{2} \sum_{k \in Outputs} (t_k - o_k)^2.$$

Алгоритм Backpropagation:

$$y = (\eta, \alpha, \{x_i^d, t^d\}_{i=1,\, d=1}^{n,\, m}, steps).$$

1. Ініціалізувати $\{w_{i,j}\}_{i,j}$ маленькими випадковими значеннями $\{\Delta w_{ij}\}_{i,j} = 0$.

2. Повторити *number of steps* раз.

2.1. Для всіх d від 1 до m подати $\{x_i^d\}$ на вхід сітки і підрахувати виходи o_i кожного вузла.

2.2. Для всіх $k \in Outputs$ $\delta_k = o_k(1-o_k)(t_k - o_k)$.

2.3. Для кожного рівня I, починаючи з останнього, для кожного вузла j рівня l (див. рис. 1) порахувати δ_i за формулою

$$\delta_i = o_j(1-o_j) \sum_{k \in Chid(j)} \delta_k w_{j,k}.$$

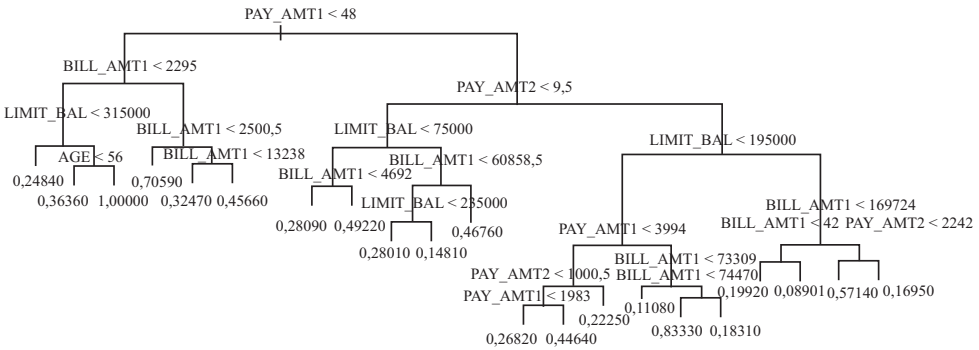

Рис. 2. Дерево рішень за обраними характеристиками

2.4. Для кожного ребра сітки $\{i, j\}$ знайти та видати значення $\Delta w_{i,j}$ за формулою

$$\Delta w_{i,j}(n) = \alpha \Delta w_{i,j}(n-1) + (1-\alpha)\,\eta \delta_j o_i,$$
$$w_{i,j}(n) = w_{i,j}(n-1) + \Delta w_{i,j}(n).$$

Для застосування методу зворотного поширення помилки функція активації нейронів повинна бути диференційованою.

Дерева рішень широко застосовуються в галузі статистичного аналізу даних для прогнозування на основі математичних моделей. До структури дерев рішень можна віднести такі елементи, як «листя» та «гілки». Ребра дерева для ухвалення рішень позначаються атрибутами, які впливають на кінцевий результат. Лист відбиває значення цільової функції, все інше — атрибути для відокремлення випадків. Для класифікації кожного наступного випадку необхідно покроково обходити дерево від листа до листа та відображати відповідне значення. Таким підходом користуються при інтелектуальному аналізі даних, основною метою якого є створення моделі для прогнозування цільових значень на основі змінних на вході [7, 8].

В інтелектуальному аналізі даних дерева рішень можуть бути використані як математичні та обчислювальні методи, що дає змогу описати, класифікувати та узагальнити набір даних, записаний у такому вигляді:

$$(x, Y) = (x_1, x_2, x_3, ..., x_k, Y).$$

Загальна схема побудови дерева ухвалення рішень за тестовими прикладам виглядає так:

1. Обираємо черговий атрибут Q, додаємо його в корінь.
2. Для всіх його значень i:

2.1. Залишаємо з тестових прикладів тільки ті, у яких значення атрибута Q дорівнює i.

2.2. Рекурсивно будуємо дерево в цьому нащадку, як показано на рис. 2.

Скорочення дерева можна здійснювати зверху вниз або знизу догори. Зверху вниз — обрізка починається з кореня, знизу догори — скорочується число листя дерева. Один з найпростіших методів регулювання — зменшення помилки обмеження дерева. Починаючи з листя, кожен вузол замінюється найпопулярнішим класом. Якщо на точність передбачення це не впливає, то заміна зберігається [10].

Результати досліджень. В ході експеременту було обрано п'ять лінгвістичних термів для кожної вхідної змінної: very.small, small, medium, large, very.large. Для досягнення кращого результату в ході побудови моделі були налаштовані такі показники, як функція належності та співвідношення розбиття на навчальну та перевірочну вибірки. Отримані результати оцінювання кредитоспроможності позичальника за допомогою розглянутих методів наведено у таблиці, із якої видно, що найкраще значення оцінки кредитоспроможності, отримане з використанням нейронної мережі Backpropagation та нейронечіткого підходу з дзвоноподібною функцією належності, для співвідношення 70/30 та 80/20 дорівнювало відповідно 0,174 та 0,196. Найгірший результат, а саме 0,76838, отримано із застосуванням логістичної регресії для співвідношення 80/20. Запропонований підхід необхідно застосувати до множини інших вибірок даних, що дасть змогу сформулювати статистично значущий висновок стосовно його потужності.

Висновки

Побудова математичних моделей на основі систем з нечіткими правилами та регресії є одним із перспективних підходів до вирішення складних проблем прийняття управлінських рішень у фінансових установах для оцінки фінансових ризиків та результатів фінансової діяльності. Найкращий результат було отримано при використанні нейронної мережі зворотного розповсюдження помилки Backpropagation. Запропонований підхід свідчить про те, що інтелектуальні методи аналізу даних можна успішно використовувати для вирішення важливих практичних проблем аналізу фінансових ризиків, включаючи оцінку платоспроможності клієнта, та для підвищення стабільності і безпеки функціонування банківської системи. Перспективним напрямком дослідження є поєднання класичних статистичних моделей та моделей, заснованих на інтелектуальному аналізі даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Pearl J.* Probabilistic reasoning in intelligent systems. San Francisco: Morgan Kaufmann Publishers, 1988.
2. *Зайченко Ю.П.* Оценка кредитних банковских рисков с использованием нечеткой логики // Intelligent Information and Engineering Systems, 2008, № 13, с. 190—200.
3. *Зайченко Ю.П.* Оценка кредитных банковских рисков с использованием нечеткой логики // Системні дослідження та інформаційні технології, 2010, № 2, с. 37—54.
4. *Terent'yev A.N., Bidyuk P.I., Korshevnyuk L.A.* Bayesian network as instrument of intelligent data analysis // Journal of Automation and Information Sciences, 2007, Vol. 39, p. 28—38.
5. *Bidyuk P.I., Kuznyetsova N.V., Terentyev O.M.* Decision support system based for analysis of financial data // Research Bulletin of NTUU «KPI». Kyiv: NTUU «KPI», 2011, № 1, p. 48—61.
6. *Verzami J.* Getting Started with RStudio. O'Reilly Media Inc., 2011, 98 p.
7. *Круглов В.В., Дли М.И., Голунов Р.Ю.* Нечеткая логика и искусственные нейронные сети. М.: Физматлит, 2000, 224 с.
8. *Загірська І.О., Бідюк П.І.* Методика побудови сценарного аналізу із використанням байєсівських методів // Електротехнічні та комп'ютерні системи. Інформаційні системи та технології, 2012, № 8 (84), с. 137—142.
9. *Кузнєцова Н.В.* Інформаційні технології аналізу фінансових ризиків за допомогою байєсівських мереж. Дис. ... канд. техн. наук, Київ, 2011, 300 с.
10. *Bidyuk P.I., Bondarenko V.V.* On One Model of Financial Data // Journal of automation and information sciences, 2011, No. 43, p. 76—81.

Отримано 04.03.19

REFERENCES

1. Pearl, J. (1988), Probabilistic reasoning in intelligent systems, Morgan Kaufmann Publishers, San Francisco, USA.
2. Zaychenko, Yu.P. (2008), "Evaluating credit banking risks using fuzzy logic", *Intelligent Information and Engineering Systems*, no. 13, pp. 190-200.
3. Zaychenko, Yu.P. (2010), "Assessing credit risk using fuzzy logic", *Sistemnyie issledovaniya i informatsionnye tekhnologii*, no. 2, pp. 37-54.
4. Terentyev, A.N., Bidyuk, P.I. and Korshevnyuk, L.A. (2007), "Bayesian network as instrument of intelligent data analysis", *Journal of Automation and Information Sciences*, Vol. 39, pp. 28-38.
5. Bidyuk, P.I., Kuznyetsova, N.V. and Terentyev, O.M. (2011), "Decision support system based for analysis of financial data", *Research Bulletin of NTUU «KPI»*, no. 1, pp. 48-61.
6. Verzami, J. (2011), Getting Started with RStudio, O'Reilly Media Inc.
7. Kruglov, V.V., Dli, M.I. and Golunov, R.Yu. (2000), *Nechetkaya logika i iskusstvennye neyronnye seti* [Fuzzy logic and artificial neural networks], Fizmatlit, Moscow, Russia.
8. Zagirska, I.O. and Bidyuk, P.I. (2012), "Method for constructing scenario analysis using Bayesian methods", *Electrical and computer systems. Information systems and technologies*, no. 8 (84), pp. 137-142.
9. Kuznetsova, N.V. (2011), "Information technologies for financial risk analysis using Bayesian networks", Abstract of Cand. Sci. (Tech.) dissertation, Kiev, Ukraine.
10. Bidyuk, P.I. and Bondarenko, V.V. (2011), "On One Model of Financial Data", *Journal of automation and information sciences*, no. 43, pp. 76-81.

Received 04.03.19

П.И. Бидюк, В.Г. Гуськова

АНАЛИЗ КРЕДИТОСПОСОБНОСТИ С ПОМОЩЬЮ
МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Предложен способ минимизации риска платежеспособности заемщиков банковской системы и других финансовых компаний, которые предоставляют кредиты своим клиентам. Проведена оценка кредитоспособности клиентов с использованием логистической регрессии, методов на основе нечеткой логики, нейронной сети с обратным распространением ошибки и деревьев решений. Представлены результаты оценки кредитоспособности заемщиков и дан анализ оценки состояния клиентов.

К л ю ч е в ы е с л о в а: риски, нечеткая логика, логистическая регрессия, нейронная сеть, деревья решений, анализ состояний.

P.I. Bidyuk, V.H. Huskova

ANALYSIS OF SOLVENCY USING DATA MINING METHODS

The focus of the paper is on approach of minimization of solvency risk. The risks for borrowers of the banking system and other financial companies that provide loans to their customers have been investigated. Client creditworthiness was assessed using logistic regression methods based on fuzzy logic, neural network with back propagation of error and decision trees. The results of the assessment of the creditworthiness of borrowers are presented, and the analysis of the assessment of the condition of clients is carried out.

K e y w o r d s: risks, fuzzy logic, logistic regression, neural network, decision trees, state analysis.

БІДЮК Петро Іванович, д-р техн. наук, професор, професор кафедри математичних методів системного аналізу Ін-та прикладного системного аналізу Національного технічного університету України «Київський політехнічний ін-т ім. Ігоря Сікорського». В 1972 р. закінчив Київський політехнічний ін-т. Область наукових досліджень — статистично-ймовірнісне моделювання і прогнозування, інтелектуальний аналіз даних, нейронні мережі та мережі Байєса, інформаційні технології, системний аналіз.

ГУСЬКОВА Віра Геннадіївна, аспірантка кафедри математичних методів системного аналізу Ін-та прикладного системного аналізу Національного технічного університету України «Київський політехнічний ін-т ім. Ігоря Сікорського». В 2016 р. закінчила Київський політехнічний ін-т ім. Ігоря Сікорського. Область наукових досліджень — системний підхід, інтелектуальний аналіз даних, математичне моделювання, інформаційні технології, системний аналіз.