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

УДК 519.226, 330.322
DOI: 10.20535/SRIT.2308-8893.2017.1.03

ОЦІНЮВАННЯ КРЕДИТНИХ РИЗИКІВ МЕТОДАМИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

В.Я. ДАНИЛОВ, О.Л. ЖИРОВ, П.І. БІДЮК

Анотація. Проаналізовано кредитні ризики фінансових організацій за допомогою методів інтелектуального аналізу даних. Фактичні статистичні дані, які характеризують позичальників кредитів, використано для побудови математичних моделей у формі рівнянь типу логіт, дерев рішень і байєсівських мереж. Якість побудованих моделей проаналізовано за множиною належних статистичних критеріїв, які забезпечують основу для вибору кращої альтернативної моделі. Із використанням двох вибірок банківських даних виконано ряд обчислювальних експериментів і виявлено кращі моделі у формі рівнянь типу логіт і байєсівські мережі. Передбачається розширити множину методів побудови математичних моделей і реалізувати ідею комбінування оцінок, згенерованих за альтернативними методами. Обґрунтовано доцільність розроблення та реалізацію спеціалізованої системи підтримання прийняття рішень для виконання досліджень у галузі оцінювання та прогнозування фінансових ризиків.

Ключові слова: кредитний ризик, статистичні дані, логіт-модель, байєсівські мережі, параметри якості моделей.

ВСТУП

Для того, щоб ефективно управляти кредитними ризиками, необхідно вміти точно вимірювати їх. Існує досить велика множина методів і моделей оцінювання кредитного ризику. Для створення системи управління кредитним ризиком банки спираються на власний досвід та напрацювання.

Натепер існує багато моделей оцінювання кредитоспроможності на підставі ринкових показників, а саме: моделі Блека–Шоулза–Мертона [1, 2]. Перевагами таких моделей є висока прогнозна спроможність, абсорбувальна інформація про позичальника, доступна всім інвесторам, присутнім на ринку. Недоліками є те, що інформація про позичальника є відповідною лише за умови ефективності ринку, потрібний великий масив даних. Ці методи реалізовані на практиці у вигляді програмного продукту CreditMonitor.

Підходи до розроблення моделей для оцінювання кредитоспроможності на підставі фундаментальних показників ґрунтуються на макроекономічних показниках, фінансових показниках та на даних рейтингових агентств. Особливостями таких підходів є те, що вони враховують циклічність економіки, дають змогу отримувати довгострокову оцінку, виконувати крос-

© *В.Я. Данилов, О.Л. Жиров, П.І. Бідюк*, 2017

аналіз. Їх переваги: доступність інформації, простота розрахунків і прийнятна точність прогнозу. Недоліки: важко визначити періодичність циклів економіки та оцінити ймовірність дефолту конкретного позичальника; не завжди надані дані є достовірними; бухгалтерська звітність показує результати постфактум, тобто недостатній прогноз майбутніх перспектив; переоцінка рейтингу має часовий лаг [1, 3].

Мінімізація кредитного ризику потребує належного управління ними, що являє собою процес виявлення і оцінювання ризиків, а також вибір методів та інструментів для цього. Традиційно кредитний ризик розглядається в розрізі кожного конкретного позичальника. Численні моделі використовують складний математичний апарат для оцінювання кредитного ризику. Ключовим завданням побудови математичної моделі кредитного ризику є оцінювання розподілу збитків усього агрегованого кредитного портфеля. Це так звані системи скорингу [4, 5].

Роботу присвячено аналізу можливості застосування лінійної і нелінійної регресій, а також байєсівського методу аналізу даних; виконанню та аналізу результатів обчислювальних експериментів з оцінювання кредитоспроможності клієнтів за наявними статистичними даними; порівнянню результатів застосованих методів оцінювання кредитного ризику.

ОГЛЯД МОДЕЛЕЙ ДЛЯ ОЦІНЮВАННЯ КРЕДИТНОГО РИЗИКУ

Лінійна та логістична регресія. Традиційними і найбільш поширеними є регресійні методи, насамперед лінійна багатофакторна регресія: $p = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$, де p — імовірність дефолту; w — вагові коефіцієнти; x — характеристики клієнта. Недолік моделі полягає у тому, що ліва частина рівняння містить імовірність, яка набуває значення в інтервалі [0, 1], а змінні в правій частині можуть набувати будь-яких значень від $-\infty$ до $+\infty$. Цей недолік може подолати нелінійна модель у формі логістичної регресії:

$$\ln\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n .$$

Дослідимо імовірність появи події (зі значенням імовірності, близьким до нуля або до одиниці) залежно від значень регресорів $x = \{x_1, x_2, \ldots, x_n\}$. У результаті отримуємо значення логіт-функції в інтервалі [0,1], тобто

$$g(z) = \frac{1}{1+e^{-z}} .$$

Тоді модель матиме вигляд: $\Pr = g(p_0 + p_1 x_1 + p_2 x_2 + \ldots + p_m x_m)$.

Припустімо, що існує вибірка $(x_1^i, x_2^i, \ldots, x_m^i, y^i)$ $(i = 1, \ldots, n)$, де y^i — значення залежної змінної; n — кількість спостережень. Потрібно оцінити коефіцієнти моделі. Для цього скористаємося принципом максимальної правдоподібності, відповідно до якого за оцінки коефіцієнтів беруться ті значення, які дають максимум функції правдоподібності:

$$L(\bar{p}) = \prod_{i=1}^{n} g\left(\overline{px^1}\right)^{y^i} \left[1 - g\left(\overline{px^1}\right)\right]^{1-y^i} .$$

Для зручності позначимо: $\overline{p} = (p_0, p_1, p_2, \ldots, p_m)$, $\overline{x^1} = (x_1^i, x_2^i, \ldots, x_m^i)$; $\overline{px^1} = p_0 + p_1 x_1^i + p_2 x_2^i + \ldots + p_m x_m^i$. Зазвичай використовується логарифм функції правдоподібності, що не змінює суті задачі, але дозволяє позбутись добутку: $l(\overline{p}) = \sum_{i=1}^{n} y^i \ln\left(g\left(\overline{px^1}\right)\right) + (1 - y^i)\ln\left(1 - g\left(\overline{px^1}\right)\right)$.

Сьогодні логістична регресія є лідером скорингових систем. Перевага логістичної регресії полягає ще й у тому, що вона може поділяти клієнтів як на дві групи (0 – поганий, 1 – хороший), так і на кілька груп (1, 2, 3, 4 групи ризику).

Дерева рішень. Дерева рішень — це модель, що будується на логічному ланцюжку правил, які намагаються описати окремі взаємозв'язки між даними щодо очікуваного результату. Структура дерев рішень відкрито показує аргументацію правил і тому дає змогу легко зрозуміти процес прийняття рішення [6].

Критерії якості моделі та оцінок прогнозів. Існує множина критеріїв, які визначають якість побудованої моделі і якість прогнозу. Подамо деякі з них, які використано у цій роботі.

Інформаційний критерій Акайке (AIC). Критерій використовується для порівняння моделей з різною кількістю параметрів, коли потрібно вибрати найкращий набір пояснювальних змінних. Для лінійної моделі множинної регресії значення критерію розраховується за такою формулою:

$$AIC = \ln\left(\frac{\sum_{i=1}^{k} \varepsilon_i^2}{n}\right) + \frac{2k}{n},$$

де n — кількість спостережень; k — кількість параметрів моделі; $\sum_{i=1}^{k} \varepsilon_i^2$ — сума квадратів залишків моделі, отриманих під час оцінювання коефіцієнтів моделі за методом найменших квадратів. Зі збільшенням кількості пояснювальних змінних перший доданок у правій частині зменшується, а другий збільшується. Таким чином, критерій не тільки винагороджує за якість наближення, але і штрафує за використання зайвої кількості параметрів моделі. Серед кількох альтернативних моделей перевага надається тій, значення AIC якої менше.

Інформаційний критерій Шварца (SC). Цей критерій, аналогічно критерію Акайке, дозволяє порівняти моделі з різною кількістю параметрів, коли потрібно вибрати кращу множину пояснювальних змінних. Для лінійної моделі множинної регресії значення критерію визначається за формулою

$$SC = \ln\left(\frac{\sum_{i=1}^{k} \varepsilon_i^2}{n}\right) + \frac{k \ln(n)}{n},$$

де n — кількість спостережень; k — кількість параметрів моделі; $\sum_{i=1}^{k} \varepsilon_i^2$ — сума квадратів залишків моделі, отриманих під час оцінювання коефіцієнтів моделі за методом найменших квадратів. Зі збільшенням кількості пояснювальних змінних перший доданок у правій частині формули зменшується,

а другий — збільшується. Серед кількох альтернативних моделей перевага віддається тій, значення SC якої менше.

Коефіцієнт детермінації розглядають, як правило, як основний показник, що відображає міру якості регресійної моделі, яка описує зв'язок між залежною і незалежними змінними моделі. Коефіцієнт детермінації показує, яка частка варіації пояснювальної змінної y врахована в моделі і зумовлена впливом на неї факторів, включених у модель:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \bar{y})},$$

де y_i — значення спостережуваної змінної; $\bar{y}$ — середнє значення спостережуваних даних; $\hat{y}$ — модельні значення, побудовані за оціненими параметрами. Чим ближче значення R^2 до одиниці, тим вища адекватність моделі. Стандартна похибка регресії (стандартна похибка оцінки) розглядається як міра розсіювання даних спостережень від змодельованих значень. Чим менше значення стандартної похибки регресії, тим якість моделі вища. Сума квадратів величин розбіжності між змодельованими і фактичними значеннями, які пояснюються змінними на періоді ідентифікації, розраховується за формулою

$$\sum_{i=1}^{n}(y_i - \hat{y})^2,$$

де y_i — значення спостережуваної змінної; $\hat{y}$ — модельні значення, обчислені за оціненими параметрами.

Критерій Дарбіна–Уотсона (DW-критерій). Це статистичний критерій, що використовується для знаходження автокореляції залишків першого порядку регресійної моделі; обчислюється за формулою $DW = 2 - 2\rho$, де ρ — коефіцієнт автокореляції першого порядку і

$$\rho = \frac{\frac{1}{N-1}\sum_{k=2}^{N}[e(k) - \bar{e}][e(k-1) - \bar{e}]}{\sigma_e^2};$$

$$\sigma_e^2 = \frac{1}{N-1}\sum_{k=1}^{N}[e(k) - \bar{e}]^2,\ e(k) = y(k) - \overline{y(k)}.$$

Критерій Ханана–Куіна (HQ). Критерій використовується для порівняння моделей за співвідношенням між якістю вибору і кількістю оцінюваних параметрів. Обчислюється тільки для моделей бінарного та множинного вибору за виразом

$$HQ = -2\frac{\ln L}{n} + 2k\frac{\ln(\ln n)}{n},$$

де L — функція правдоподібності; n — кількість спостережень; k — кількість змінних моделі. Вибирається модель з найменшим значенням критерію.

Критерій Макфадена (McFadden Rsquared). Це аналог коефіцієнта детермінації для звичайної регресії:

$$\text{McFaddenRsquared} = 1 - \frac{\ln L}{\text{Re}\,str\ln L},$$

де $\ln L$ — логарифм функції правдоподібності; $\mathrm{Re}str\ln L$ — залишок логарифма функції правдоподібності. Значення міститься в діапазоні (0, 1). Обчислюється тільки якщо модель містить константу. Найкращим вважається значення, що найближче до одиниці.

Середньоквадратична похибка (СКП) використовується для оцінювання адекватності моделі і обчислюється за формулою

$$\text{СКП} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}[y(k) - \hat{y}(k)]^2},$$

де $y(k)$ — значення спостережуваної змінної; $\hat{y}(k)$ — модельні значення, обчислені за оціненими параметрами. Середня абсолютна похибка у відсотках (САПВ) — це середнє абсолютних значень похибок оцінок прогнозу відносно фактичного значення показника:

$$\text{САПВ} = \frac{1}{N}\sum_{k=1}^{N}\frac{|y(k) - \hat{y}(k)|}{|y(k)|}100,$$

де $y(k)$ — значення спостережуваної змінної; $\hat{y}(k)$ — модельні значення, обчислені за оціненими параметрами. Оскільки ця міра характеризує відносну якість прогнозу, то її використовують здебільшого для порівняння точності прогнозів різнорідних об'єктів (процесів) прогнозування. Однак вона завжди корисна для виконання порівняльного аналізу якості прогнозування одного й того ж самого процесу різними методами, оскільки відносна міра є чіткою і зрозумілою для дослідника і практичного користувача [7, 8].

Для аналізу якості моделей і встановлення кращої моделі для розв'язання певної задачі використовують кілька критеріїв для оцінювання адекватності моделей [9]: загальна точність моделі; помилки першого і другого роду; *ROC*-крива та індекс *GINI*. Загальна точність моделі (*CA* — *Common Accuracy*) визначається так: $CA = \dfrac{Correct\,Forecast}{N}$, де $Correct\,Forecast$ — кількість правильно спрогнозованих випадків; N — загальна кількість випадків. Загальна точність моделі є дещо суб'єктивною оцінкою, оскільки вона залежить від частки дефолтів у моделі та від порога відсікання [9]. Для різних значень порога точність моделі також буде набувати різних значень. *ROC-крива* (*Receiver Operation Characteristic* — робоча характеристика приймача) показує залежність кількості правильно класифікованих позитивних прикладів від кількості неправильно класифікованих негативних прикладів. Перші називають істинно позитивними, а другі — негативними множинами. Припускається, що у класифікаторі є певний параметр, варіюючи яким можна отримати певне розбиття на класи. Цей параметр часто називають порогом або точкою відсікання (*cut-off*), залежно від якого будуть отримані різні величини *помилок першого і другого роду* (табл. 1).

Таблиця 1. Помилки першого і другого роду

Повернення/ Неповернення	Прогноз моделі: повернення кредиту (0)	Прогноз моделі: дефолт (1)
Фактично: повернення кредиту (0)	Правильно класифіковані (*TP*)	Помилки другого роду (*FN*)
Фактично: дефолт (1)	Помилки першого роду (*FP*)	Правильно класифіковані (*TN*)

Для аналізу якості моделі найчастіше використовують такі відносні показники (у відсотках):

– частка істинно позитивних прикладів (*True Positives Rate*): $TPR = \dfrac{TP}{TP + FN}$;

– частка хибно позитивних прикладів (*False Positives Rate*): $FPR = \dfrac{FP}{TN + FP}$.

Зазвичай для аналізу якості моделей використовують ще дві характеристики: чутливість та специфічність. Чутливість моделі — це частка істинно позитивних випадків, тобто $Se = TPR = \dfrac{TP}{TP + FN}$.

Специфічність моделі — це частка істинно негативних випадків, які були правильно класифіковані моделлю: $Sp = \dfrac{TN}{TN + FP}$. Очевидно, що $Sp = \dfrac{TN + FP - FP}{TN + FP} = 1 - \dfrac{FP}{TN + FP} = 1 - FPR$.

Модель з високою чутливістю надає істинний результат за наявності позитивних випадків (виявляє позитивні приклади). Навпаки, модель із високою специфічністю найчастіше дає істинний результат за наявності негативних випадків (виявляє негативні приклади). Для побудови графіка *ROC*-кривої по осі Y відкладаються значення чутливості Se, а по осі X — частку хибно позитивних випадків FPR або $1 - Sp$. Графік ідеального класифікатора *ROC*–кривої (рис. 1) проходить через верхній лівий кут, де частка істинно позитивних випадків становить 1 (ідеальна чутливість), а частка хибно позитивних прикладів дорівнює нулю. Тому чим ближче крива наближається до верхнього лівого кута, тим кращою є здатність моделі передбачувати. Діагональна лінія відповідає класифікатору, який не здатний розпізнати ці два класи.

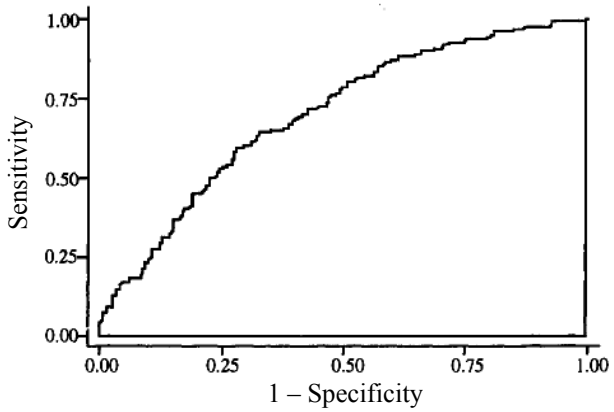

Рис. 1. Графік *ROC*-кривої

Оскільки візуальне порівняння *ROC*-кривих не завжди дає змогу визначити ефективнішу модель, застосовують оцінку площі під кривими. Число-

вий показник площі під кривою *AUC* (*Area Under Curve*) обчислюється, наприклад, за методом трапецій:

$$AUC = \int f(x)dx = \sum_i \left[\frac{X_{i+1} + X_i}{2} \right] (Y_{i+1} - Y_i) .$$

Більш зрозумілим і частіше згадуваним у літературі параметром оцінювання якості моделі є індекс *GINI*, запропонований італійським статистиком К. Джіні, який тісно пов'язаний з числовим показником площі під *ROC*-кривою. Індекс *GINI* — це площа ділянки між діагоналлю і кривою Лоренца, поділена на площу всієї ділянки під діагоналлю. Індекс *GINI* широко використовується для аналізу роздільної здатності системи оцінювання під час управління кредитними ризиками, тобто оцінювання здатності моделі поділяти клієнтів на схильних та несхильних до дефолту. Якщо модель здатна оцінити клієнтів за ймовірністю дефолту, то більшість клієнтів, схильних до дефолту, мають отримати більшу ймовірність дефолту. Відповідно найменша ймовірність дефолту має бути для клієнтів, не схильних до дефолту. Індекс *GINI* проілюстровано графіком, де сукупний відсоток дефолту для клієнтів показано поряд із сукупним відсотком клієнтів, коли вони упорядковані за ймовірністю дефолту (менша ймовірність дефолту — зліва, більша — справа). Цей графік відомий як крива Лоренца (рис. 2) [9].

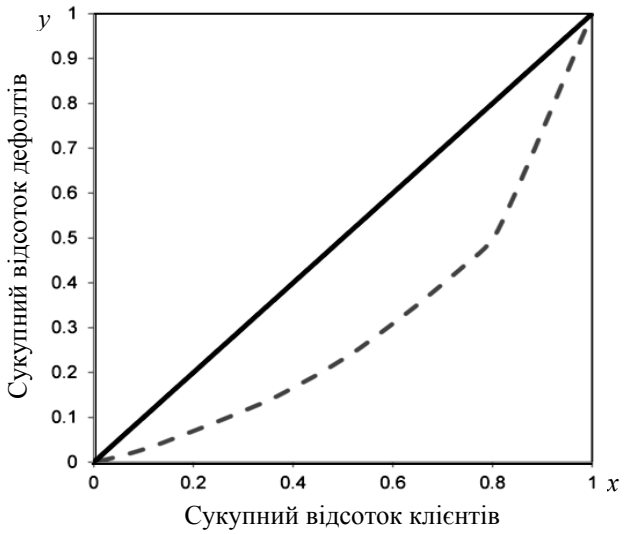

Рис. 2. Крива Лоренца для кредитування

Діагональна лінія — це лінія «випадкової моделі». На осі *X* клієнти впорядковані за заданою ймовірністю дефолту. Очевидно, що 80% дефолтів припадає саме на решту 20% клієнтів (20% клієнтів з найбільшою ймовірністю дефолту). Це означає, що модель виконує прийнятний поділ. Отже, чим більша ділянка між діагоналлю та кривою, тим вища якість моделі. Дефолти серед клієнтів з однаковою ймовірністю дефолту вважаються поділеними випадково серед цих клієнтів. Індекс *GINI* можна визначити через площу фігури, що розміщена 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. Оцінка якості моделі істотно залежить від даних, за якими вона будується. 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 (оцінювання фінансового стану існуючих клієнтів) індекс *GINI* зазвичай набуває значень, вищих за 70%. У літературі наведено шкалу значень індексу *GINI* (табл. 2) [9].

Таблиця 2. Оцінка якості моделі за площею *AUC* та індексом *GINI*

Інтервал *AUC*	**Індекс** *GINI*	**Якість моделі**
0,9 – 1,0	0,8 – 1,0	Відмінна
0,8 – 0,9	0,6 – 0,8	Дуже висока
0,7 – 0,8	0,4 – 0,6	Прийнятна
0,6 – 0,7	0,2 – 0,4	Середня
0,5 – 0,6	0 – 0,2	Незадовільна

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

– мінімальної величини чутливості (специфічності) моделі;

– максимальної сумарної чутливості та специфічності моделі, тобто $cut-off = \max_k (Se_k + Sp_k)$;

– балансу між чутливістю і специфічністю, тобто коли $Sp \approx Se$: $cut-off = \min_k | Se_k - Sp_k |$.

ПРИКЛАД ЗАСТОСУВАННЯ МЕТОДИКИ МОДЕЛЮВАННЯ

Для побудови моделі лінійної регресії, логістичної регресії та дерева рішень використано статистичні дані першої вибірки — All_1 (довжина вибірки даних — 15000 значень).

– x_1 — змінна, яка характеризує стать позичальника (*gender*);

– x_2 — змінна, яка характеризує вік позичальника (*Age*);

– x_3 — змінна, яка характеризує суму кредиту (*Credit_sum*);

– x_4 — змінна, яка характеризує термін кредитування у днях (*Term_of_crediting_in day*);

– x_5 — змінна, яка характеризує сімейний стан позичальника (*Maritial_status*);

– x_6 — змінна, яка характеризує кількість дітей позичальника (*Children*);

– x_7 — змінна, яка характеризує кількість найманих працівників в компанії позичальника (*Number_of_employees_in_company*);

– x_8 — змінна, яка характеризує дохід позичальника (*Income_customer*);

– x_9 — змінна, яка характеризує витрати позичальника (*Costs_customer*);

– y — змінна, яка характеризує результат повернення кредиту (*Result*).

Особливістю цієї вибірки є її висока асиметрія стосовно типів позичальників, тобто вона містить характеристики 750 клієнтів, які не повертають

кредити, і 14250 клієнтів, які повертають кредити. Результат побудови лінійної регресії показано на рис. 3.

Equation: UNTITLED Workfile: ALL_1

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 04/30/14 Time: 22:45
Sample: 1 15000
Included observations: 15000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.965872	0.009212	104.8455	0.0000
X1	0.016636	0.003582	4.644929	0.0000
X2	-0.000539	0.000188	-2.864474	0.0042
X3	7.47E-08	3.17E-08	2.352971	0.0186
X4	-2.23E-09	1.54E-09	-1.442450	0.1492
X5	0.002748	0.001968	1.396726	0.1625
X6	0.003767	0.002594	1.452080	0.1465
X7	-0.005106	0.001223	-4.176210	0.0000
X8	-6.93E-08	9.26E-08	-0.748356	0.4543
X9	3.50E-07	7.02E-07	0.497868	0.6186

R-squared	0.004825	Mean dependent var	0.950000
Adjusted R-squared	0.004227	S.D. dependent var	0.217952
S.E. of regression	0.217491	Akaike info criterion	-0.212651
Sum squared resid	709.0625	Schwarz criterion	-0.207574
Log likelihood	1604.885	F-statistic	8.074477
Durbin-Watson stat	1.854974	Prob(F-statistic)	0.000000

Рис. 3. Результати оцінювання лінійної регресії та її характеристики

Для вибору кращої моделі із множини різних специфікацій надано значення інформаційних критеріїв Акайке (*AIC*) та Шварца (*SC*), логарифмічну функцію правдоподібності. Для визначення міри якості лінійної регресійної моделі вибрано такі статистичні параметри: R^2 — коефіцієнт детермінації; стандартну похибку регресії; статистику Дарбіна–Уотсона; статистику Фішера (*F*-статистику) та відповідну ймовірність. Значення стандартної похибки регресії показує, що лінійна модель дає гірші результати $(SE = 0{,}212)$, ніж нелінійна модель логіт $(SE = 0{,}173)$.

Проаналізовано також іншу вибірку (BASE_ALL) для порівняння зі статистичними даними, які описують такі змінні (потужність вибірки даних становить 7568 значень):

– змінна, яка характеризує вік позичальника (Ageyears);

– змінна, яка характеризує стать позичальника (Sex);

– змінна, яка характеризує освіту позичальника (Education);

– змінна, яка характеризує сімейний стан позичальника (Maritalstatus);

– змінна, яка характеризує суму кредиту (Credit);

– змінна, яка характеризує регіон проживання позичальника (Region);

– змінна, яка характеризує сферу, де працює позичальник (Prev Employer Sector);

– змінна, яка характеризує статус в суспільстві позичальника (Occupation Status);

– змінна, яка характеризує досвід роботи позичальника (Work Experience Month);

– змінна, яка характеризує результат повернення кредиту (Good/Bad).

Для побудови моделі перетворено дані таким чином: стать позичальника (*Sex*): чоловіча (*Male*) — 1; жіноча (Female) — 0; освіта (*Education*): початкова (*Elementary*) — 0, середня (*Meaddle*) — 1, вища (*High*) — 2; дві вищі чи ступінь (*Twohighand/ordegree*) — 3; сімейний стан (*Maritalstatus*): неодружений (*Notmarried*) — 0, одружений (*Registeringmarriage*) — 1, вдова(ець) (*Widow(er)*) — 2, розведений (*Divorced*) — 3, цивільний шлюб (*Civilmarriage*) — 4; сфера праці (*PrevEmployerSector*): невизначена (none) — 0, інші (other) — 1, медицина (*Medicine*) — 2, сільське господарство (*Agriculture*) — 3, послуги (*Services*) — 4, будівництво (*Building*) — 5, освіта (*Education*) — 6, торгівля (*Trade*) — 7, виробництво (*Manufacturing*) — 8, фінанси (*Finance*) — 9, видобуток (*Mining*) — 10; статус у суспільстві (*OccupationStatus*): найманий робітник (*Employee*) — 0, урядовий офіцер (*Governmentofficer*) — 1, пенсіонер (*Pensioner*) — 2, власник/співвласник (*Owner/coowner*) — 3, військовий (*Military*) — 4; результат (*Good/Bad*): повернено кредит (*Good*) — 1, не повернено кредит (*Bad*) — 0.

Узагальнену порівняльну характеристику лінійної моделі та моделі логіт подано в табл. 3:

Таблиця 3. Результати застосування лінійної і нелінійної регресії

Варіант розрахунків	**Якість моделі**			**Якість прогнозу**		
	Коефіцієнт детермінації	Сума квадратів залишків	Статистика Дарбіна–Уотсона	Середня квадратична похибка	Середня абсолютна похибка	Коефіцієнт Тейла
Лінійна регресія						
*All*_1	0,0048	709,062	1,855	0,212	9,463	0,113
*BASE*_743	0,3840	114,416	0,272	0,309	34,056	0,302
Нелінійна регресія						
*All*_1	–	638,054	–	0,173	7,446	0,101
*BASE*_743	–	96,127	–	0,157	22,441	0,189

Отже, нелінійна регресія за всіма показниками якості прогнозу дає кращі результати ніж лінійна регресія. Побудована модель нелінійної регресії є кращою за показниками якості.

ПОБУДОВА ДЕРЕВ РІШЕНЬ ЗА СТАТИСТИЧНИМИ ДАНИМИ

У результаті застосування системи *SPSS* побудовано дерево рішень з відповідними характеристиками для вибірки *All*_1 (рис. 4).

За допомогою методу *CHAID* (*CHi-squared Automatic Interaction Detection*) отримано результат, який свідчить, що рівень доходу (x_8) є кращим предиктором кредитного рейтингу. Для категорії з низьким рівнем доходів кращий предиктор — рівень витрат (x_9), для малих витрат кращий предиктор — сума кредиту (x_3). Оскільки немає ніяких розгалужень, то це термінальний вузол. Клієнти цих вузлів мають досить прийнятний кредитний рейтинг (0,69 або 0,78), незважаючи на суму кредиту. Для великих

значень витрат немає розгалужень; це термінальний вузол. Для категорій середнього та високого доходів кращим предиктором є термін кредитування (x_4). Для середнього терміну кращий предиктор — сума кредиту (x_3), а для більшого терміну — стать позичальника (x_1).

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	y
	Independent Variables	$x_9, x_8, x_6, x_7, x_5, x_1, x_4, x_2, x_3$
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	$x_8, x_9, x_3, x_4, x_1, x_7$
	Number of Nodes	31
	Number of Terminal Nodes	21
	Depth	3

Рис. 4. Зведена таблиця загальних специфікацій дерева рішень

Gain Summery for Nodes

Node	*N*	Percent,%	Mean
30	1091	7,30	1
27	1001	6,70	0,99
13	233	1,60	0,99
11	947	6,30	0,98
7	2042	13,60	0,98
21	1912	12,70	0,97
29	68	0,50	0,97
15	2737	18,20	0,97
23	507	3,40	0,97
28	176	1,20	0,97
25	417	2,80	0,96
6	108	0,70	0,94
20	318	2,10	0,94
24	689	4,60	0,94
22	298	2,00	0,94
9	678	4,50	0,93
26	171	1,10	0,91
12	106	0,70	0,9
5	366	2,40	0,86
19	612	4,10	0,79
18	523	3,50	0,69

Growing Melhod CHAID
Department Variable: *y*

Risk

Estimate	Std Error
0,073	0,006

Growing Melhod CHAID

Рис. 5. Узагальнені результати оцінювання

Для клієнтів з високим рівнем доходів кращим предиктором є термін кредитування (x_4). Залежно від терміну кращими предикторами є дохід позичальника (x_8), кількість найманих робітників в компанії (x_7) або сума кредиту(x_3).Узагальнені результати за деревом рішень (оцінку ризику, його стандартну похибку, тобто міру точності прогнозу) показано на рис. 5.

Отже, за допомогою дерев рішень отримано оцінку ризику неправильного оцінювання клієнта 4,3% та стандартне відхилення 0,001. Узагальнені результати за деревом для другої вибірки (*BASE*_743) подано нижче. Діаграма дерева (рис. 6), що являє собою графічне зображення моделі дерева, показує, що з використанням методу *CHAID* регіон проживання позичальника (x_6) є кращим предиктором кредитного рейтингу. Оскільки глибина дерева одинична, то не відбувається ніякого додаткового розгалуження.

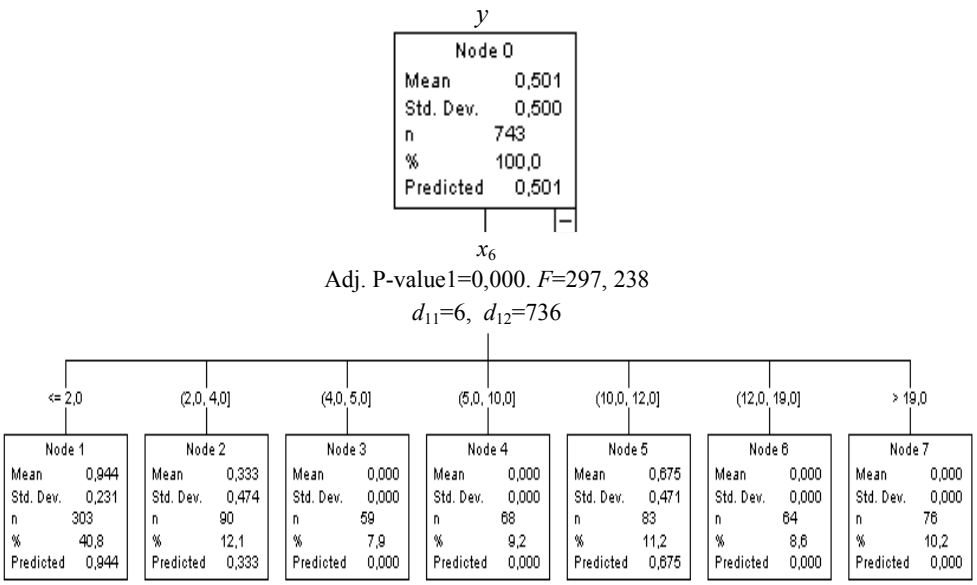

Рис. 6. Діаграма дерева рішень

Оцінку ризику та його стандартну похибку, тобто міру точності прогнозу дерева, подано на рис. 7.

Gain Summary for Nodes

Node	N	Percent, %	Mean
1	303	40,80	0,94
5	83	11,20	0,87
3	90	12,10	0,33
7	76	10,20	0
4	68	9,20	0
6	64	8,6	0
3	59	7,90	0

Growing Method: CHAID Dependent Variable: *y*

Risk

Estimate	Std Error
0,073	0,006

Growing Melhod CHAID

Рис. 7. Узагальнені результаті та ризики

Отже, за допомогою дерев рішень отримано ризик неправильного оцінювання клієнта 7,3%, а стандартне відхилення становить 0,06. Результати ризику неправильного оцінювання та стандартне відхилення для двох вибірок наведено в табл. 4.

Таблиця 4. Оцінка ризику та стандартна похибка дерева рішень для двох множин статистичних даних

Вибірки	Ризик неправильної оцінки	Стандартне відхилення
*ALL*_1	0,043	0,001
*BASE*_743	0,073	0,006

Побудова мережі Байєса. Для побудови першої моделі (вибірка *ALL* — 15000 значень) використано статистичні дані для 15000 виданих кредитів, термін яких закінчився. Вибірку поділено на навчальну (13000 випадків) та перевірну (2000 випадків). Навчальна вибірка завантажується в підсистему побудови моделі. Для побудови моделі необхідно формалізувати дані у зручному для оброблення вигляді, тобто перевести їх у заданий формат, а у випадку неперервних змінних — дискретизувати їх. Для дискретизації використано ієрархічну дискретизацію. На наступному кроці порівнюються характеристики взаємовиключних змінних і вибираються змінні, які будуть використовуватись на етапі побудови мережі. Далі вибирається відповідний алгоритм навчання мережі; у разі потреби використовуються експертні знання і виконується навчання мережі.

Будуючи структуру мережі Байєса в програмі *GeNIe*, слід пам'ятати, що обраний алгоритм впливає на швидкість і якість побудови структури. Фактично найшвидшим є алгоритм *Greedy Thick Thinning*, його і будемо використовувати для аналізу даних. У результаті роботи алгоритму отримуємо тільки одну структуру, яка є цілком логічною і оптимальною за критерієм максимальної правдоподібності (рис. 8).

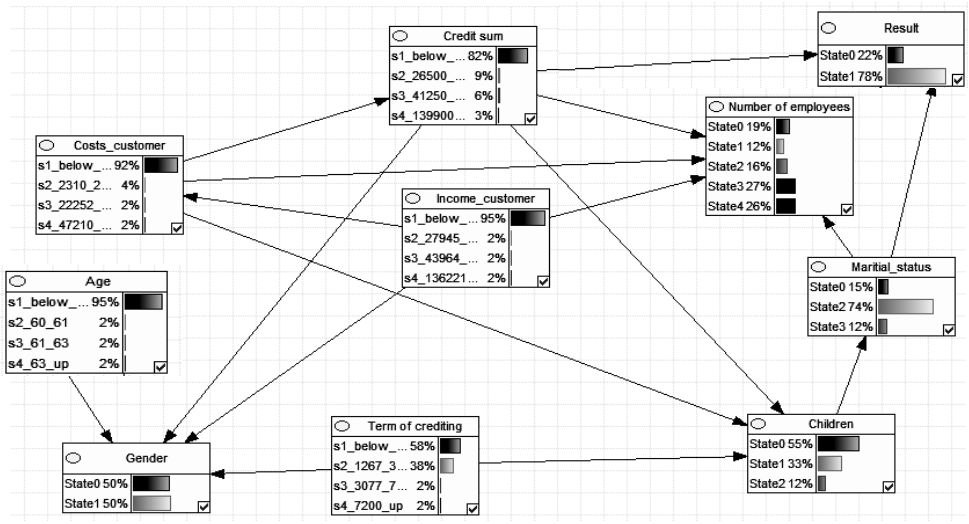

Рис. 8. Структура мережі Байєса у вигляді гістограм вибірки *ALL*

На основі перевірної вибірки перевіряється якість моделі з використанням відомих оцінок: загальної точності, помилок першого і другого роду. Для різних порогів відсікання обчислено помилки першого і другого роду (табл. 5).

Найбільша точність моделі досягається на рівні 0,75 зі встановленням порога 0,3 — буде пропущено 10 дефолтів. Очевидно, що мережа Байєса має схильність до «перестрахування», тобто вона частіше помиляється, усуваючи тих клієнтів, які б повернули кредит. Зрозуміло, що точність моделі та кількість помилок першого і другого роду будуть залежати від порога відсікання, установленого банком. Слід пам'ятати, що, встановлюючи поріг відсікання, варто визначати не лише відсоток відсіяних клієнтів, а і нижню межу ймовірності повернення кредиту, тобто поріг, нижче за який клієнт вважається таким, що не поверне кредиту, або ж нижню межу ймовірності

дефолту, нижче від якої вважається, що клієнту слід видати кредит. Значення ймовірності дефолту 0,1 або 0,2 для клієнта є незначними і статистично малими, а тому поріг відсікання доцільно встановити на рівні 0,25 – 0,3. Зрозуміло, що встановлений поріг відсікання впливає на кількість помилок першого і другого роду. Для мережі Байєса побудовано *ROC*-криву (рис. 9).

Таблиця 5. Загальна точність моделі та помилки першого і другого роду для різних рівнів порога відсікання, отримані для мереж Байєса (*ALL*)

Характеристика: повернення/кредиту	**Прогноз: повернення кредиту (0)**	**Прогноз: дефолт (1)**	**Точність, %**
Cut-off=0,5			
Факт: повернення кредиту (0)	82	50	0,620
Факт: дефолт (1)	12	56	0,82
Загальна точність моделі			**0,69**
Cut-off=0,4			
Факт: повернення кредиту (0)	92	49	0,65
Факт: дефолт (1)	3	56	0,94
Загальна точність моделі			**0,74**
Cut-off=0,3			
Факт: повернення кредиту (0)	101	49	0,67
Факт: дефолт (1)	4	46	0,92
Загальна точність моделі	—	—	**0,735**

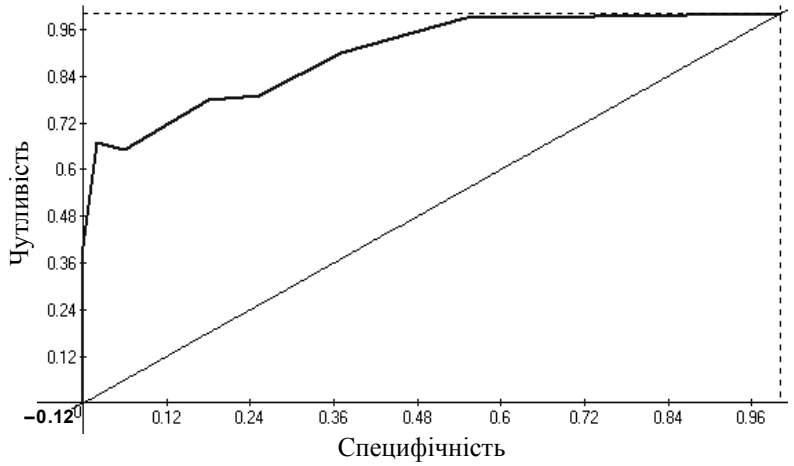

Рис. 9. ROC-крива для мережі Байєса

Для порівняння моделей використаємо індекс *GINI*. Для цього спочатку обчислимо значення площі під кривою: $AUC = 0{,}86$. Відповідно індекс *GINI* становить: $GINI = 2AUC - 1 = 0{,}72$.

АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Таким чином, у праці використано три методи інтелектуального аналізу даних для прогнозування кредитоспроможності фізичних осіб за статистичними даними: лінійну регресію, логістичну регресію та дерева рішень. Для ви-

конання обчислювальних експериментів використано дві вибірки даних: для 15000 та для 7568 клієнтів банку. Вибрано множину статистичних критеріїв для аналізу якості класифікації клієнтів. Кращі результати класифікації клієнтів на дві групи отримано за допомогою дерев рішень (табл. 6), що можна пояснити можливістю досягнення високої якості класифікації на дві групи за допомогою цього методу на конкретних даних.

Таблиця 6. Стандартне відхилення для трьох моделей

Вибірки	Лінійна регресія	Логістична регресія	Дерева рішень
*ALL*_1	0,217	0,176	0,001
*BASE*_743	0,392	0,278	0,006

Результати порівняння якостей моделей на основі бінарної логістичної регресії та мереж Байєса наведено в табл. 7. Так, на основі моделей у формі мереж Байєса і бінарної логістичної регресії отримано дуже високі значення індексу *GINI*.

Таблиця 7. Порівняльна таблиця характеристик для мереж Байєса та бінарної *логістичної регресії*

All				
Назва методу	**Індекс *GINI***	**Значення AUC**	**Точність моделі**	**Якість моделі**
Бінарна логістична регресія	0,74	0,87	0,77	Дуже висока
Мережа Байєса	0,72	0,86	0,74	Дуже висока
BASE				
Назва методу	Індекс *GINI*	Значення *AUC*	Точність моделі	Якість моделі
Бінарна логістична регресія	0,74	0,87	0,79	Дуже висока
Мережа Байєса	0,76	0,88	0,74	Дуже висока

Наведені результати свідчать, що надалі банкам доцільно використовувати відомі скорингові моделі і мережі Байєса, оскільки отримані результати та прогнозні якості мереж Байєса дають можливість усунути недобросовісних позичальників і таким чином зменшити втрати банків від неповернення кредитів. Високі значення параметрів точності моделі також дає логістична регресія. Ці результати ще раз підтверджують доцільність використання логістичної регресії для оцінювання кредитоспроможності позичальника.

Очевидно, що мережа Байєса має схильність до «перестрахування», тобто вона частіше помиляється, усуваючи тих клієнтів, які б повернули кредит. Зрозуміло, що точність моделі та кількість помилок першого і другого роду залежатиме від порога відсікання, який встановлюється банком.

ВИСНОВКИ

Виконано короткий огляд моделей оцінювання кредитоспроможності і проаналізовано їх переваги та недоліки. З наведеного огляду випливає, що існує потреба у створенні нових сучасних комп'ютерних систем для оцінювання ризиків з метою їх мінімізації та ризик-менеджменту. Описано три методи,

які можна використовувати для прогнозування кредитоспроможності фізичних осіб: логістичну регресію, лінійну регресію та дерева рішень. Незважаючи на те, що лінійна регресія використовується нечасто, її також можна використати для попереднього наближеного оцінювання та порівняльного аналізу результатів.

Виконано загальний огляд статистичних критеріїв аналізу якості класифікації клієнтів. Зазначено, що розрахунок критеріїв якості дещо відрізняється для лінійних і нелінійних моделей. Існує можливість вибрати саме ті критерії, які необхідні для аналізу якості класифікації клієнтів. Оцінено якість прогнозу та якість моделі за такими критеріями, як коефіцієнт детермінації, сума квадратів залишків, статистика Дарбіна–Уотсона, середня квадратична похибка та середня абсолютна похибка у відсотках. Нелінійна регресія дає можливість отримати значно кращі якісні показники, ніж лінійна за рядом критеріїв. Установлено, що кращі результати класифікації клієнтів отримано за допомогою дерев рішень (стандартне відхилення —0,1–0,6 %). Показано, що дерева рішень і байєсівські мережі дають змогу отримати прийнятний за якістю результат класифікації.

У подальших дослідженнях для оцінювання кредитоспроможності клієнтів фінансової установи доцільно побудувати спеціалізовану систему підтримки прийняття рішень на основі комбінованого використання методів регресійного та інтелектуального аналізу даних. Для підвищення достовірності результатів доцільно також використати статистичні дані, отримані з альтернативних джерел.

ЛІТЕРАТУРА

1. *Матигорова И.Ю.* Характеристика основных подходов к оценке кредитного риска / И.Ю. Матигорова // Экономическая наука и практика: материалы междунар. науч. конф. (г. Чита, февраль 2012 г.). — Чита: Изд-во «Молодой ученый», 2012. — С. 68–69.
2. *Сиддики Н.* Скоринговые карты для оценки кредитных рисков / Н.Сиддики. — М.: Изд-во «Манн, Иванов и Фербер», 2014. — 268 с.
3. *Liu Y.* The evaluation of classification models for credit scoring / Y. Liu. — Arbeitsbericht 02/2002. — Institut fur Wirtschaftsinformatik, 2002. — 19 p.
4. *Кузминчук Н.В.* Методы оценки кредитного риска в банковской деятельности / Н.В. Кузминчук, О.С. Мандрыка // Бизнесинформ, 2009. — № 1. — С. 113–117.
5. *Bielecki T.R.* Credit Risk: modeling, valuation and hedging / T.R. Bielecki, M. Rutkowski. — Berlin: Springer, 2002. — 500 p.
6. *Hosmer D.W.* Applied Logistic Regression / D.W. Hosmer, S. Lemeshow. — New York: John Wiley & Sons, Inc. 1989. — 400 p.
7. *Бідюк П.І.* Аналіз часових рядів / П.І. Бідюк , В.Д. Романенко, О.Л. Тимощук. — К.: Політехніка, 2013. — 600 с.
8. *Бідюк П.І.* Системний підхід до прогнозування на основі моделей часових рядів / П.І. Бідюк // Системні дослідження та інформаційні технології. — 2003. — № 3. — С. 88–110.
9. *Довгий С.О.* СППР на основі ймовірнісно-статистичних методів / С.О. Довгий, О.М. Трофимчук. — К.: Логос, 2014. — 430 с.

Надійшла 01.11.2016