КІБЕРНЕТИКА та КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 519.85 DOI:10.34229/2707-451X.22.3.3

В.О. СТОВБА

СТАТИСТИЧНІ ТА ОПТИМІЗАЦІЙНІ МЕТОДИ В КРЕДИТНОМУ СКОРИНГУ

Зроблено огляд основних математичних методів, які використовуються в задачах кредитного скорингу. В основу огляду покладено матеріал книги «Credit Scoring and Its Applications» 2002 року (L. Thomas, D. Edelman, J. Crook). Розглянуто статистичні та нестатистичні методи, зроблено їх порівняння та описано нюанси їх застосування.

Ключові слова: *кредитний скоринг, статистичні методи, математичне програмування, нейронні мережі, генетичні алгоритми.*

Вступ. Кредитний скоринг (*credit scoring*) – один з найдавніших засобів керування фінансовими ризиками. Початок його використання в США та Великій Британії припадає на 2 половину XX століття. Кредитний скоринг певною мірою заклав основи глибинного аналізу даних (*data mining*), оскільки у цій області в одній з перших використовувались дані про поведінку споживачів. Загалом, найбільш відомі техніки аналізу даних, такі як: сегментація, кластеризація, моделювання схильності – досить успішно використовувались у кредитному скорингу [1].

Кредитний скоринг – одна з областей, де найбільш успішно застосовувались статистичне моделювання та моделювання дослідження операцій у фінансовій галузі та банкінгу [1]. Оскільки скоринг це математична модель, що оперує великими обсягами даних про клієнтів, протягом останніх років широкого вжитку набуло використання методів машинного навчання, зокрема, логістична регресія, метод опорних векторів, метод k середніх (алгоритм Ллойда), k найближчих сусідів тощо.

У роботі зроблено огляд найбільш вживаних математичних методів у області кредитного скорингу, які умовно поділяються на статистичні (розділ 2) та нестатистичні методи: математичне програмування, нейронні мережі, генетичні алгоритми тощо (розділ 3). Наведено порівняння розглянутих методів та описано умови їх застосування до задач кредитного скорингу. Матеріал та структуру статті побудовано з використанням роботи [1].

1. Кредитний скоринг: основні поняття та процеси

Кредитний скоринг це набір моделей прийняття рішень та їх технік, за допомогою яких вирішується питання щодо схвалення або відмови у наданні позики заявнику. Ці техніки визначають хто отримає позику та якого обсягу, а також які операційні стратегії можуть покращити прибутковість позичальників та позикодавців. Зазвичай позикодавцями приймаються рішення щодо двох питань – схвалення чи відмови у наданні кредиту новому клієнту, та політики роботи

з існуючими клієнтами, зокрема, в питанні підвищення їхніх кредитних лімітів. Вирішення другого питання лежить в області поведінкового скорингу (*behavioral scoring*) [1].

Для прийняття рішення щодо схвалення чи відмови у наданні кредиту необхідно проаналізувати наявні дані про заявника та його кредитну історію. В переважній більшості технік кредитного скорингу для віднесення певного клієнта до категорії «хороший» (кредитна заявка схвалена) чи «поганий» (кредитна заявка відхилена) використовуються вибірки заявників з певними даними про них. Для оцінки (скорингу) платоспроможності заявника дані необхідно представити у чисельному вигляді. Для цього використовуються скорингові картки (*scorecards*), які містять набір характеристик заявника. Залежно від відповіді заявника кожна характеристика оцінюється певною кількістю балів. Результат такої скорингової операції це певна числова оцінка платоспроможності заявника, яка й використовується для прийняття рішення щодо надання кредиту.

Існує ціла низка методів роботи з цією оцінкою. В одному з найпростіших варіантів встановлюється певний прохідний бал, якщо оцінка вища ніж цей бал, заявник може бути рекомендованим для надання позики. В іншому випадку, якщо оцінка потрапляє у певний встановлений числовий проміжок («сіру зону»), заявник піддається більш ретельному аналізу, можливо, з залученням додаткових даних. Також часто беруть до уваги пріоритет отриманих даних: наприклад, заявник, оцінка якого вища за прохідний бал, але який у минулому був оголошений банкрутом, переходить у групу заявників, запити яких розглядаються окремо з залученням кредитних аналітиків.

Формулювання задачі, що лежить в основі кредитного скорингу, призвела до появи широкого кола методів, які успішно застосовувались для її розв’язання. На початкових етапах розвитку кредитного скорингу єдиними методами, які використовувались, були класифікаційні методи та статистична дискримінація. Пізніше цей набір методів поповнився багатьма іншими статистичними та нестатистичними методами, які виявились найефективнішими. Задачі кредитного скорингу вдалось сформулювати як оптимізаційні задачі, що дозволило застосовувати цілий клас нових методів. Новий підхід у розв’язанні задач класифікації за допомогою нейронних мереж вдалось успішно застосувати для задач кредитного скорингу.

В наступному розділі розглядаються найбільш вживані статистичні методи у кредитному скорингу.

2. Статистичні методи для побудови кредитних скорингових карток

Статистичні методи володіють переліком важливих властивостей, які дозволяють успішно застосовувати їх до задач кредитного скорингу. Зокрема, ці методи дозволяють оцінювати дискримінантні можливості скорингової картки та відносну важливість різних характеристик, які її утворюють. Це дає можливість вилучити неважливі характеристики та забезпечити оптимальний набір характеристик у картці. Також з’являється можливість з’ясувати які зміни необхідно внести в питання, які задають клієнтам, для побудови більш якісних оцінок. У цьому розділі розглядають найбільш вживані в області кредитного скорингу статистичні методи.

Дискримінантний аналіз. Існує три основні підходи до задачі кредитного скорингу, в яких лінійні дискримінантні функції використовуються як класифікатори: теорія прийняття рішень, поділ на дві групи та лінійна регресія. В першому підході здійснюється пошук правила, яке мінімізує очікувані затрати у прийнятті рішення щодо схвалення кредитної заявки. Другий підхід полягає у побудові функції, яка найкращим чином розділяє заявників на «хороших» та «поганих» у контексті надання кредиту. Третій – ґрунтується на побудові рівняння лінійної регресії, мета якого – знаходження найкращої оцінки правдоподібності заявника бути «хорошим».

Теорія прийняття рішень. Нехай $X = \left(X_1, ..., X_p \right)$ – множина з p характеристик заявника на кредит. Кожна характеристика X_i має скінчену кількість можливих значень (атрибутів), з яких

заявник обирає одне. Таким чином вектор $x=\left(x_1,...,x_p\right)$ – фактичний результат опитування, який описує конкретного заявника. Позначимо A множину всіх можливих векторів x, тобто результатів опитування довільного заявника. Задача кредитного скорингу полягає у знаходженні такого правила, яке розділяє множину A на дві підмножини A_G та A_B. Підмножина A_G містить заявників, що класифікуються як «хороші» та отримують кредит, підмножина A_B містить заявників, класифікованих як «погані», при цьому очікувані витрати мінімальні. Для побудови функції сумарних витрат розглядаються помилки класифікації двох типів. В першому випадку «хороший» заявник класифікується як «поганий», при цьому втрачається потенційний прибуток від такого заявника величини L. В другому випадку «поганий» заявник класифікується як «хороший» і внаслідок неспроможності заявника погасити борг втрачається позика величини D. Тоді функція сумарних витрат має такий вигляд:

$$L\sum_{x\in A_B} p\left(x\,|\,G\right)p_G + D\sum_{x\in A_G} p\left(x\,|\,B\right)p_B, \tag{1}$$

де p_G та p_B – частки заявників, які є «хорошими» та «поганими» відповідно, $p\left(x\,|\,G\right)$ та $p\left(x\,|\,B\right)$ – умовні ймовірності того, що заявник має вектор атрибутів x за умови, що він є «хорошим» або «поганим» відповідно. Правило поділу є таким:

$$A_G=\left\{x\,|\,Dp\left(x\,|\,B\right)p_B\le Lp\left(x\,|\,G\right)p_G\right\}=\left\{x\,|\,\frac{D}{L}\le\frac{p\left(x\,|\,G\right)p_G}{p\left(x\,|\,B\right)p_B}\right\}, \tag{2}$$

тобто заявнику з вектором атрибутів x надається кредит, якщо втрати при хибній класифікації його як «хорошого» будуть не більшими, ніж втрати при хибній класифікації заявника як «поганого». Якщо величини L та D є невідомими, мінімізація сумарних витрат замінюється мінімізацією ймовірності допустити помилку одного з типів, зберігаючи ймовірність допустити помилку іншого типу сталою. В такому випадку виникає задача умовної мінімізації, яка може бути розв'язана за допомогою методу множників Лагранжа.

Варто відзначити, що у тому випадку, коли характеристики заявника є не дискретними, а неперервними випадковими величинами, мають місце аналогічні міркування з використанням функцій щільності випадкових величин та операції взяття інтегралу.

Поділ на дві групи. Нехай $Y=\omega_1X_1+...+\omega_pX_p$ – довільна лінійна комбінація характеристик $X=\left(X_1,...,X_p\right)$. Необхідно знайти таку комбінацію характеристик, які найкращим чином розділяють множину заявників на дві групи, що інтерпретується як схвалення або відхилення заявки про кредитування. Роберт Фішер запропонував використовувати таку міру розділення множин:

$$M=\omega^T\frac{m_G-m_B}{\sqrt{\omega^TS\omega}}, \tag{3}$$

за умови, що вибіркова дисперсія двох груп однакова. Тут m_G та m_G – вибіркові середні груп «хороших» та «поганих» заявників відповідно, S – вибіркова дисперсія цих груп. Використання необхідних та достатніх умов екстремуму дозволяє отримати значення ваг

$$\omega^T\propto S^{-1}\left(m_G-m_B\right)^T, \tag{4}$$

що показує незалежність отриманого результату від форми розподілу випадкових величин $X_1,...,X_p$. Геометрично ваги ω^T утворюють гіперплощину $\omega\cdot x=c$, яка розділяє дві стандартизовані групи «хороших» та «поганих» заявників, причому точка прохідного балу знаходиться на перетині цієї гіперплощини та відрізку, що з'єднує середні значення двох груп.

Лінійна регресія. В цьому підході задача кредитного скорингу полягає у знаходженні лінійної комбінації характеристик

$$\omega_0 + \omega_1 X_1 + ... + \omega_p X_p = \omega^* \cdot \left(X^* \right)^T, \tag{5}$$

яка найкращим чином описує ймовірність несплати. Тут $\omega^* = \left(\omega_0, \omega_1, ..., \omega_p \right)$ та $X^* = \left(1, X_1, ..., X_p \right)$. Тобто якщо p_i – це ймовірність того, що заявник i не сплатив борг, необхідно знайти вектор ω^* такий, що

$$p_i = \omega_0 + x_{i1}\omega_1 + ... + x_{ip}\omega_p \text{ для всіх } i. \tag{6}$$

Для цього мінімізується величина

$$\sum_{i=1}^{n_G} \left(1 - \sum_{j=0}^{p} \omega_j x_{ij} \right) + \sum_{i=n_G+1}^{n_G+n_B} \left(\sum_{j=0}^{p} \omega_j x_{ij} \right)^2, \tag{7}$$

де n_G та n_B – кількість «хороших» та «поганих» заявників відповідно, причому $n_G + n_B = n$. Для простоти викладення матеріалу припускається, що для перших n_G заявників у вибірці $p_i = 1$, для решти – $p_i = 0$. Використання необхідних та достатніх умов екстремуму дозволяє отримати розв'язок $S\omega^T = c\left(m_G - m_B \right)^T$.

Загалом дискримінантний підхід – це один найбільш розповсюджений та установлений метод для розв'язання задачі кредитного скорингу. Один з найперших застосувань цього підходу був аналіз позик на автомобілі. До більш пізніх робіт можна віднести публікації [2, 3].

Логістична регресія. Одним з недоліків лінійної регресії є те, що права частина регресійних рівнянь приймає значення від $-\infty$ до $+\infty$, а ліва частина інтерпретується як ймовірність, тому має належати відрізку $[0,1]$. Для подолання цього недоліку ліву частину можна представити як функцію від p_i, областю значень якої буде інтервал $(-\infty, +\infty)$. Такий підхід має назву логістична регресія, а відповідне рівняння має такий вигляд:

$$\log\left(\frac{p_i}{1 - p_i} \right) = \omega_0 + \omega_1 x_1 + ... + \omega_p x_p = \omega \cdot x^T. \tag{8}$$

Нехай μ_G та μ_B – середні значення серед «хороших» та «поганих» заявників відповідно, а Σ – їх коваріаційна матриця. Тоді відповідна функція щільності характеристик X_i, які мають багатовимірний нормальний розподіл, є такою:

$$f\left(x \mid G \right) = \left(2\pi \right)^{-\frac{p}{2}} \left(\det \Sigma \right)^{-\frac{1}{2}} \exp\left(\frac{-\left(x - \mu_G \right) \Sigma^{-1} \left(x - \mu_G \right)^T}{2} \right). \tag{9}$$

Якщо p_G та p_B – частки «хороших» та «поганих» заявників відповідно, рівняння логістичної регресії має такий вигляд:

$$\log\left(\frac{p_i}{1 - p_i} \right) = \log\left(\frac{p_G f\left(x \mid G \right)}{p_B f\left(x \mid B \right)} \right) = x \cdot \Sigma^{-1} 2\left(\mu_B - \mu_G \right)^T + \left(\mu_G \cdot \Sigma^{-1} \mu_G^T + \mu_B \cdot \Sigma^{-1} \mu_B^T \right) + \log\left(\frac{p_G}{p_B} \right). \tag{10}$$

Підхід для визначення коефіцієнтів за допомогою мінімізації суми квадратів відхилень не застосовується до логістичної регресії, натомість можна використати максимізацію функції правдоподібності.

Підхід логістичної регресії – статистично кращий інструмент порівняно з лінійною регресією для розв'язання задачі бінарної класифікації. Це призвело до інтенсивного застосування цього підходу в області кредитного скорингу [4–6].

Класифікаційні дерева. Якісно іншим підходом до розв'язання задач кредитного скорингу є класифікаційні дерева або алгоритми рекурсивного розділення. Мета таких алгоритмів – це розділення множини заявників на певну кількість підмножин з подальшою їх класифікацією як «хороших» або «поганих» підмножин. Зазвичай класифікація здійснюється так: якщо переважна кількість заявників у підмножині є «хорошою», її класифікують як «хорошу», інакше як «погану». Ще один спосіб класифікації це мінімізація втрат хибної класифікації.

Для роботи таких алгоритмів необхідно визначити три правила. 1. Правило розділення визначає спосіб розділення множини на дві підмножини. 2. Правило зупинки встановлює умову зупинки процесу розділення множин. 3. Потрібне для класифікації термінальних множин на «хороші» та «погані».

Розділення множин здійснюється так: відбувається перебір можливих варіантів розділення з подальшою перевіркою якості кожного варіанту. Для цього вводиться певна міра якості такого розділення. Наприклад, для неперервної характеристики X_i множина розділяється на підмножини $\{x_i < s\}$ та $\{x_i \geq s\}$ для всіх значень s, після чого визначається для якого значення s міра найкраща. Якщо характеристика X_i категорійна змінна, відбувається перебір всіх варіантів розділення множини з перевіркою значень міри на кожному розбитті. Найбільш вживана міра це статистика Колмогорова – Смірнова, однак використовуються також базовий індекс змішаності та індекс Джині, індекс ентропії та максимізація напівсуми квадратів.

Статистика Колмогорова – Смірнова визначається для неперервних характеристик X_i з кумулятивними функціями розподілу $F(s|G)$ та $F(s|B)$ для «хороших» та «поганих» заявників відповідно, і формулюється як задача: знайти величину s таку, що мінімізує

$$LF(s|G)\,p_G + D\big(1-F(s|B)\big)\,p_B\,. \tag{11}$$

Тут L та D – втрати за хибну класифікацію першого та другого типів відповідно. Якщо $Lp_G = Dp_B$, то задача співпадає з задачею знаходження відстані Колмогорова – Смірнова між двома розподілами, тобто мінімізувати $F(s|G)-F(s|B)$ або максимізувати $F(s|B)-F(s|G)$. Якщо дві підмножини, що утворюються у результаті розділення, розглянути як ліву l та праву r множини, задачу можна сформулювати як максимізацію різниці між $p(l|B)$ та $p(l|G)$ – ймовірністю «поганого» заявника потрапити в ліву групу та ймовірністю «хорошого» заявника потрапити у праву групу. Тоді статистику Колмогорова – Смірнова можна використовувати і для дискретних, і для неперервних характеристик X_i у такому вигляді: знайти таке розділення на ліву та праву групи, яке максимізує

$$KS = \big|p(l|B) - p(l|G)\big| = \left|\frac{p(B|l)}{p(B)} - \frac{p(G|l)}{p(G)}\right| \cdot p(l)\,. \tag{12}$$

Базовий індекс змішаності та індекс Джині. Існує ціла низка індексів, які дозволяють визначити рівень змішаності кожного вузла дерева. Вузол чистий (незмішаний), якщо належить лише одному класу. Якщо вузол розділяється на лівий вузол l та правий вузол r з частотою входження у групу лівих вузлів $p(l)$ та частотою входження у групу правих вузлів $p(r)$, оцінити рівень змішаності після такого розділення можна за допомогою величини

$$I = i(v) - p(l)i(l) - p(r)i(r)\,. \tag{13}$$

Чим більша величина I тим більше змішане розділення, тобто нові вершини більш чисті. Отже величину I необхідно максимізувати, що еквівалентно мінімізації виразу

$$p(l)i(l)+p(r)i(r).$$

Якщо відійти від лінійності у пропорціях ймовірності змішаності, можна розглянути квадратичний індекс Джині, в якому більш чистим вузлам відповідають більші вагові коефіцієнти. Визначивши $i(v)=p(G|v)\,p(B|v)$, маємо:

$$G=p(G|v)\,p(B|v)-p(l)\,p(G|l)\,p(B|l)-p(r)\,p(G|r)\,p(B|r). \tag{14}$$

Індекс ентропії. Ще один нелінійний індекс це індекс ентропії, що пов'язаний з кількістю інформації у поділі «хороших» та «поганих» заявників у вузлі:

$$i(v)=-p(G|v)\ln\big(p(G|v)\big)-p(B|v)\ln\big(p(B|v)\big). \tag{15}$$

Максимізація напівсуми квадратів. Ця міра випливає з статистичного χ^2-тесту і визначає чи співпадають частки «хороших» заявників у множині, що розділяється, та двох підмножинах, що при цьому утворились. Чим більша статистика χ^2 тим менш схожі частки, тобто більш значна різниця між ними. Якщо $n(l)$ та $n(r)$ – кількості лівих та правих вузлів, задача полягає у максимізації величини

$$Chi=n(l)n(r)-\frac{\big(p(G|l)-p(G|r)\big)^2}{n(l)+n(r)}. \tag{16}$$

Метод *k* найближчих сусідів. Один найдавніший непараметричний метод класифікації це метод *k* найближчих сусідів. Ідея методу полягає у тому, щоб множину вже класифікованих заявників, кожному з яких відповідає вектор їх атрибутів, розмістити в просторі цих атрибутів. У цьому просторі вводиться метрика для визначення відстані між заявниками. Віднесення нового заявника до певного класу відбувається залежно від властивостей *k* найближчих до нього сусідів. Величина *k* – параметр і регулюється окремо.

Для роботи методу необхідно визначити параметр *k*, частку «хороших» заявників серед k сусідів для віднесення нового заявника до класу «хороших» та метрику простору. Величина параметра *k* суттєво залежить від розміру вибірки заявників та в деяких випадках результат при різних *k* значно відрізняється, тому зазвичай підбирається емпіричним шляхом. Як частка «хороших» заявників серед сусідів для класифікації зазвичай береться більшість таких заявників або обирається таким чином, щоб мінімізувати сумарні втрати за хибну класифікацію. Одна найбільш вдала глобальна метрика для використання в задачах кредитного скорингу це метрика, що є комбінацією евклідової метрики та відстані у напрямку w, який найкращим чином розділяє «хороших» та «поганих» заявників. Вона має такий вигляд:

$$d(x_1,x_2)=\Big\{(x_1-x_2)^T\left(I+Dw\cdot w^T\right)(x_1-x_2)\Big\}^{\frac{1}{2}}, \tag{17}$$

де I – одинична матриця, D – середні втрати у випадку банкрутства заявника.

Варто відмітити, що метод k найближчих сусідів має декілька важливих переваг порівняно з іншими методами, що застосовуються у кредитному скорингу. Зокрема, це можливість динамічно додавати й класифікувати нових заявників, поповнюючи тренувальний набір. Також це можливість змінювати метрику, оскільки правильно підібрана метрика суттєво впливає на якість кінцевого результату.

Мультигрупова дискримінація. Іноді в кредитному скорингу виникає потреба класифікувати заявників на більше ніж 2 групи. Наприклад, необхідно виділити заявників, яким кредитор готовий надати позику, небажаних заявників через банкрутство в минулому та небажаних заявників через

недостатній прибуток для кредитора. В такому випадку застосовують методи мультигрупової дискримінації, до переліку яких входить більшість із вищенаведених методів з відповідними модифікаціями. Розглянемо модифікацію підходу прийняття рішень у дискримінантному аналізі.

Нехай $c(i,j)$ – збитки за віднесення заявника j до групи i, а p_j – частка групи j у вибірці. Позначимо як $p(x|j)$ ймовірність того, що заявники з групи j мають вектор атрибутів x. Тоді ймовірність того, що заявник з вектором атрибутів x належить до групи j, рівна

$$p(j|x)=\frac{p_j p(x|j)}{\sum_i p_i p(x|i)}. \tag{18}$$

Для мінімізації сумарних витрат заявник з вектором атрибутів призначається до групи i якщо

$$\sum_j c(i,j)\,p_j p(x|j) < \sum_j c(k,j)\,p_j p(x|j),\ \forall k,\ k\neq i. \tag{19}$$

3. Нестатистичні методи для побудови кредитних скорингових карток

На початкових етапах розвитку кредитного скорингу здебільшого використовувались статистичні методи розв'язання поставлених задач. Однак період інтенсивного розвитку кредитного скорингу співпав з періодом активних досліджень в області методів дослідження операцій, математичного програмування, штучного інтелекту, а згодом і машинного навчання. Задача класифікації, що лежить у основі кредитного скорингу, була сформульована в багатьох різних формах, що дозволяло застосовувати для її розв'язання широкий набір нестатистичних підходів та методів. У цьому розділі наведено короткий опис таких методів.

Лінійне програмування. Вперше підхід лінійного програмування для розв'язання задач класифікації був використаний у роботі [7], де метою була побудова гіперплощини, що розділяє дві групи. Якщо групи не є лінійно роздільними, проводиться мінімізація суми модулів нев'язок або мінімізація максимуму нев'язки.

Нагадаємо, що задача кредитного скорингу полягає у розділенні множини A – набору всіх комбінацій значень p змінних $X=(X_1,...,X_p)$ – на дві множини A_G та A_B «хороших» та «поганих» заявників відповідно. Нехай n – кількість заявників у вибірці, а n_G та n_B – кількість «хороших» та «поганих» заявників відповідно, причому $n=n_G+n_B$. Не обмежуючи загальності припустимо, що «хорошими» заявниками є n_G перших заявників у вибірці. Нехай i-й заявник має вектор атрибутів $(x_{i1},...,x_{ip})$. Необхідно визначити вектор ваг $(\omega_1,...,\omega_p)$ так, щоб зважена сума $\omega_1 X_1+...+\omega_p X_p$ перевищувала прохідний бал c для «хороших» заявників та не перевищувала для «поганих». Для отримання наближеного розв'язку задачі вводиться вектор нев'язок $a=(a_1,...,a_n)$, тоді умови класифікації формулюються так: якщо заявник i «хороший» має виконуватись нерівність $\omega_1 x_{i1}+...+\omega_p x_{ip}\geq c-a_i$, інакше виконується нерівність $\omega_1 x_{i1}+...+\omega_p x_{ip}\leq c-a_i$. Тоді для пошуку вектора ваг $(\omega_1,...,\omega_p)$, що мінімізує суму модулів відхилень, необхідно розв'язати таку задачу лінійного програмування:

$$a_1+a_2+...+a_{n_G+n_B}\to\min \tag{20}$$

за умов

$$\omega_1 x_{i1}+...+\omega_p x_{ip}\geq c-a_i,\quad i=\overline{1,n_G}, \tag{21}$$

$$\omega_1 x_{i1}+...+\omega_p x_{ip}\leq c-a_i,\quad i=\overline{1+n_G,n_G+n_B}, \tag{22}$$

$$a_i\geq 0,\quad i=\overline{1,n_G+n_B}. \tag{23}$$

Якщо ж мінімізується максимум відхилень, задача має такий вигляд:

$$a \to \min \tag{24}$$

за умов

$$\omega_1 x_{i1} + ... + \omega_p x_{ip} \geq c - a, \quad i = \overline{1, n_G}, \tag{25}$$

$$\omega_1 x_{i1} + ... + \omega_p x_{ip} \leq c - a, \quad i = \overline{1 + n_G, n_G + n_B}, \tag{26}$$

$$a_i \geq 0. \tag{27}$$

Використання лінійного програмування у кредитному скорингу має як переваги, так і недоліки. Зокрема, цей підхід дозволяє встановити бажане відхилення на етапі розробки скорингових карток. Взаємозв'язки між певними змінними легко встановлюються за допомогою лінійних обмежень, які додаються до системи обмежень задачі. Однак обмеження задачі завжди мають форму рівностей або нестрогих нерівностей, виключаючи використання строгих нерівностей. Це може призводити до отримання тривіальних розв'язків, коли всі елементи вектора ваг та прохідний бал рівні нулю. Для уникнення таких ситуацій прохідний бал встановлюється ненульовим, однак у такому разі задачу необхідно розв'язати двічі: для додатного та від'ємного значень прохідного балу.

Ще один важливий недолік застосування лінійного програмування це відсутність можливості оцінити статистичну значимість параметрів, що оцінюються. Один із шляхів подолання цієї проблеми – це спосіб оцінки параметрів за допомогою технік статистичного бутстрепу (*bootstrap*) та методу складного ножа (*jackknife*). Також одна з переваг регресійного підходу над лінійним програмуванням це можливість почергово вводити у рівняння регресори, починаючи з найбільш значимого, що дозволяє будувати моделі з заданим числом регресорів, які будуть найбільш ефективними в дискримінації. В роботі [8] показано як можна використати підхід складного ножа до лінійного програмування, щоб отримати фіксоване число найбільш ефективних характеристик. Насправді, цей підхід також вимагає багаторазового розв'язання задачі лінійного програмування.

Цілочислове програмування. Ще один підхід розв'язання задачі кредитного скорингу це мінімізація числа неправильних класифікацій або сумарних втрат при неправильній класифікації, якщо величина D (втрати через класифікацію «поганого» заявника як «хорошого») значно відрізняється від величини L (втрати через класифікацію «хорошого» заявника як «поганого»). Така модель теж лінійна, причому певні змінні мають бути цілочисловими, отже маємо задачу цілочислового лінійного програмування:

$$L\left(d_1 + ... + d_{n_G}\right) + D\left(d_{n_{G+1}} + ... + d_{n_{G+B}}\right) \to \min \tag{28}$$

за умов

$$\omega_1 x_{i1} + ... + \omega_p x_{ip} \geq c - Md_i, \quad i = \overline{1, n_G}, \tag{29}$$

$$\omega_1 x_{i1} + ... + \omega_p x_{ip} \leq c - Md_i, \quad i = \overline{1 + n_G, n_G + n_B}, \tag{30}$$

$$0 \leq d_i \leq 1, \quad d_i \text{ – цілочислові.} \tag{31}$$

Змінна d_i приймає значення один, якщо заявника i класифіковано неправильно, та нуль в іншому випадку. Для уникнення отримання тривіальних розв'язків, аналогічно до задачі лінійного програмування необхідно ввести умови нормалізації:

$$\sum_{j=1}^{p}\left(s_j^+ + s_j^-\right) = 1,\ s_j^+ \geq 0,\ s_j^- \leq 1, \quad s_j^+,\ s_j^- \text{ – цілочислові},\ j = \overline{1, p}, \tag{32}$$

$$-1 + 2s_j \leq \omega_j \leq 1 - 2s_j, \quad j = \overline{1, p}. \tag{33}$$

Такі умови аналогічні до умов

$$\sum_{j=1}^{p}\omega_j = 1 \text{ та } c = +1 \text{ або } -1.$$

Описана модель виявилась кращою класифікаційною моделлю, ніж модель лінійного програмування. Зокрема, вона дозволяє усунути тривіальні розв'язки та забезпечити інваріантність ваг, якщо дані є зміщеними – це забезпечується за допомогою нормалізації (32), (33). Підхід цілочислового програмування дозволяє вирішити проблему створення оптимальної скорингової картки, використовуючи лише m характеристик. Для цього необхідно додати обмеження

$$\sum_{j=1}^{p}\left(s_j^+ + s_j^-\right) = m,$$

завдяки якому лише m ваг будуть додатними та лише m характеристик будуть ненульовими.

Однак, підхід цілочислового програмування має два суттєвих недоліки. Перший – розв'язання задачі цілочислового програмування потребує значно більше часу, ніж лінійного програмування, тому обсяг вибірки може бути відносно невеликим. Це призводить до малого поширення використання цього підходу в комерційних застосуваннях кредитного скорингу. Другий – полягає у тому, що зазвичай наявно багато оптимальних розв'язків з однаковим числом неправильних класифікацій, але суттєво різною якістю на вихідних вибірках.

Нейронні мережі. Один з найпоширеніших інструментів у сучасному штучному інтелекті це нейронні мережі, які дозволяють розв'язувати широкий спектр різних задач. Найпростіша одношарова нейронна мережа складається з вхідного шару нейронів (набору входів, через які до мережі передаються дані), сумуючої функції, яка підсумовує значення входів з певними вагами, та активаційної функції, яка забезпечує вихід мережі. В термінах кредитного скорингу, через вхідні нейрони до мережі передаються дані про заявника, а на виході мережі отримується результат класифікації заявника як «хорошого» або «поганого». Алгебраїчно модель такої мережі, що має назву перцептрон Розентблата, можна виразити так:

$$u_k = \omega_{k0}x_0 + \omega_{k1}x_1 + ... + \omega_{kp}x_p = \sum_{q=0}^{p}\omega_{kq}x_q\,, \tag{34}$$

$$y_k = F\left(u_k\right). \tag{35}$$

Тут $\left(x_1,...,x_p\right)$ – набір вхідних даних (характеристик), x_0 – відхилення, $\left(\omega_{k0},...,\omega_{kp}\right)$ – ваги, F – функція активації, y_k – вихід мережі, k – індекс нейрону в наступному шарі мережі (для одношарової мережі $k=1$).

Перцептрон дозволяє класифікувати дані тільки у тому випадку, коли вони лінійно роздільні. Для класифікації нелінійно роздільних даних використовуються багатошарові перцептрони з нелінійними функціями активації. Така мережа складається з вхідного та вихідного шарів, а також набору прихованих шарів. Кожен нейрон першого прихованого шару має набір ваг, які разом із значеннями вхідних нейронів використовуються у сумуючій функції. Значення сумуючої функції передається до функції активації, значення якої слугує входом для наступного шару мережі. Таким чином, набір вхідних даних, що надходить до мережі, проходить через низку нейронів, що мають різні ваги та активаційні функції, формуючи набір вихідних даних. Алгебраїчно модель багатошарової нейронної мережі можна представити так:

$$y_k = F_1\left(\sum_{q=0}^{p}\omega_{kq}x_q\right), \tag{36}$$

де індекс 1 вказує, що це перший шар після вхідного. Величини y_k – вихідні значення першого прихованого шару. Оскільки вихідні значення одного шару є вхідними значеннями для наступного, вихід мережі можна записати так:

$$z_v = F_2\left(\sum_{k=1}^{r} K_{vk} y_k\right) = F_2\left(\sum_{k=1}^{r} K_{vk}\left(F_1\left(\sum_{q=0}^{p} \omega_{kq} x_q\right)\right)\right), \tag{37}$$

де z_v – вихід нейрона v вихідного шару, F_2 – функція активації вихідного шару, K_{vk} – матриця ваг між прихованим та вихідним шарами.

Для використання мережі як класифікатора необхідно обчислити ваги всієї мережі. Один з найбільш поширених алгоритмів для цього процесу це метод оберненого розповсюдження помилки (*back-propagation algorithm*), який є методом градієнтного спуску. Він полягає у послідовному коректуванні ваг мережі шляхом надання їй правильно класифікованих прикладів, при цьому мінімізується певна функція похибок.

Для досягнення максимальної ефективності роботи нейронної мережі необхідно правильно визначити архітектуру мережі, а саме кількість прихованих шарів та кількість нейронів у цих шарах, та функцію похибок. Число прихованих шарів зазвичай обирають рівним 2. Перший шар дозволяє отримати значення вище або нижче прохідного балу в опуклій області вхідних змінних, другий – комбінувати ці опуклі області, що може дати неопуклі або повністю розділені області. Число нейронів у прихованих шарах зазвичай підбирається з використанням евристичних процедур.

Використання нейронних мереж дозволяє проводити класифікацію для довільної кількості класів шляхом регулювання кількості вихідних нейронів мережі. Відомо, що багатошаровий перцептрон, натренований з використанням методу оберненого розповсюдження помилки та відповідною функцією похибок скінченою кількістю незалежних прикладів та однаково розподіленими вхідними даними, асимптотично збігається до апроксимації апостеріорних ймовірностей належності до класів. Отже заявник g належить до групи C_g, якщо вихідне значення вихідного шару, на якому ця група була натренована – $F_g(x)$, є більшим, ніж вихідне значення вихідного шару, на якому була натренована довільна інша група – $F_h(x)$, тобто якщо $F_g(x) > F_h(x)$, $g \neq h$. Функцію похибок можна визначити таким чином. Ймовірності отримання вихідних значень з вектором вхідних даних $x(t)$ є y_g, розподіл яких – $P(o(t)\,|\,x(t)) = \prod_{g=1}^{Z}\left(y_g^t\right)^{o_g^t}$. Звідси отримуємо критерій відносної ентропії: $E_2 = -\sum_t \sum_{g=1}^{Z} O_g^t \ln y_g^t$. Оскільки значення y_v інтерпретуються як ймовірності, для забезпечення умов $0 \le y_{vg} \le 1$ та $\sum_{g=1}^{Z} y_{vg} = 1$ використовується активаційна функція softmax:

$$y_g = \frac{e^{u_g}}{\sum_{g=1}^{Z} e^{u_g}}.$$

Нейронні мережі знайшли своє застосування у багатьох прикладних областях, у тому числі в фінансовому та кредитному секторах для прогнозування банкрутства, виявлення махінацій з кредитними картками, аналізу застав, ціноутворення опціонів тощо [9–11].

Градієнтний бустинг. Багато ефективних методів класифікації належать до області машинного навчання. Одним з найбільш широковживаних методів це градієнтний бустинг (*gradient boosting*). Його мета полягає у створенні ансамблю «слабких» класифікаторів для підвищення їх ефективності. Як класифікатори зазвичай обирають дерева прийняття рішень, тому утворений ансамбль має назву градієнтний бустинг над деревами прийняття рішень (*gradient boosted decision tree*).

Нехай $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ – тренувальна вибірка, де $\mathbf{x}_i$ – вектор характеристик *i*-го заявника, $y_i \in \{0,1\}$ – результат класифікації цього заявника. Для класифікації заявників будується функція $F(\mathbf{x}_i)$

така, що мінімізує функцію втрат $L(y_i, F(\mathbf{x}_i))$:

$$F^* = \arg\min_F \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)). \tag{38}$$

Функція $F(\mathbf{x})$ – адитивна та будується ітераційно:

$$F(\mathbf{x}) = \sum_{t=1}^{T} f_t(\mathbf{x}), \tag{39}$$

де кожна функція f – дерево прийняття рішень, T – кількість ітерацій. На ітерації *t* функція f_t реалізує подальшу мінімізацію сукупних втрат ансамблю $\{f_j\}_{j=1}^{t-1}$.

Згодом після відкриття градієнтного бустингу була запропонована його модифікація, що використовує ідею бутстреп-агрегації Бреймана. У новому методі доля навчальних прикладів для кожного дерева знижується, внаслідок чого навчання значно прискорюється, а також більш високе зміщення замінюється низькою дисперсією. Такий метод має назву стохастичний градієнтний бустинг (*stochastic gradient boosting*).

Генетичні алгоритми. Ще один досить ефективний інструмент для розв'язання задач кредитного скорингу це генетичні алгоритми. Їх мета полягає у систематичному пошуку в множині можливих розв'язків задачі таким чином, що якість розв'язків прямо пропорційна ймовірності залишитись у цій множині. Одна з можливих моделей, яка може бути використана для класифікації заявників, така:

$$f(x_i) = a_1 x_{i1}^{b_1} + a_2 x_{i2}^{b_2} + ... + a_p x_{ip}^{b_p} + c, \tag{40}$$

де $a_1, a_2, ..., a_p, b_1, b_2, ..., b_p, c$ – параметри, які необхідно знайти, $x_{i1}, ..., x_{ip}$ – характеристики заявника i. Залежно від значення функції $f(x_i)$ (додатного чи від'ємного) заявник класифікується як «хороший» або «поганий» відповідний. Для кожного параметра встановлюються верхні та нижні границі.

Робота генетичного алгоритму здійснюється у декілька етапів. Спочатку визначається множина вихідних розв'язків задачі (вихідна популяція) випадковим чином за допомогою верхніх та нижніх границь для параметрів. На другому етапі для кожного розв'язку з цієї множини обчислюється його нормалізована продуктивність p_j з використанням функції придатності (*fitness function*). В задачах кредитного скорингу така функція, наприклад, може обчислювати кількість правильно класифікованих заявників. Використовуючи величини p_j як ймовірності бути обраним, формується проміжна вибірка розв'язків (проміжна популяція) обсягу n_p. На третьому етапі виконується

генерація нових розв'язків на основі наявних у проміжній популяція з використанням функцій схрещування (*crossover*) та мутації (*mutation*). Розв'язки зазвичай кодуються за допомогою нуля та одиниці, або набору цифр від нуля до дев'яти. Кожен розв'язок це послідовний набір груп цифр; перша цифра – індикатор і вказує на те володіє чи не володіє заявник певною характеристикою. Решта цифр групи – значення характеристики або її інтервал. Наприклад, розв'язок 11018 означає, що заявник має домашній телефон (перша цифра рівна 1) та не проживає за вказаною адресою від 1 до 8 років (третя цифра рівна 0, дві останні – 1 та 8). Функція мутації дозволяє отримати новий розв'язок шляхом взаємозаміни фіксованих частин двох обраних розв'язків, наприклад, перших n цифр. Функція мутації змінює певні частини розв'язку, наприклад, індикатори з 0 на 1. Обраний розв'язок разом з результатами роботи функцій схрещування та мутації формують нову популяцію. Етапи 2 і 3 повторюються фіксовану кількість разів.

Генетичні алгоритми та їх розширення генетичне програмування – одні з відносно недавніх інструментів, що знайшли своє застосування у кредитному скорингу [12, 13] та прогнозуванні банкрутств [14, 15].

Експертні системи набули широкого поширення наприкінці XX століття. Це набір процесів, мета яких – це емуляція роботи експерта у певній області. Зазвичай експертна система складається з бази знань (правил) та набору фактів, з яких за допомогою генератора висновків отримуються рекомендації до дій. Правила зазвичай мають форму «якщо-тоді», наприклад, «якщо річні виплати перевищують 50 % річних надходжень, тоді кредит не буде погашено». Для побудови бази знань проводиться робота з фахівцем в області, роботу якого система має імітувати. Також для отримання таких правил використовуються нейронні мережі, які після тренування за вхідними даними (характеристиками заявника) видають певне рішення. Однак такі рішення не інтерпретовані, тому одне із завдань експерта їх обґрунтування та пояснення.

Основні переваги експертних систем – це просте розширення бази знань шляхом додавання нових правил та інтерпретація рішень, що видаються системами. Наприклад, побудована в роботі [16] експертна система CLUES для прийняття рішень щодо страхування іпотечної позики видала рішення, практично всі з яких були схвалені страховиками. Система містить близько 1000 правил, які оцінювали кожного заявника на предмет кожного з трьох типів аналізу, які зазвичай проводяться страховиками: аналіз платоспроможності позичальника, аналіз його спроможності до погашення позики та розгляд оціночного висновку. Застосування експертних систем у кредитному скорингу та їх порівняння з іншими методами висвітлено в [17, 18].

4. Порівняння методів

Всі вищерозглянуті методи якісно різні, тому визначити найкращий з них не вдається. Ефективність роботи кожного з методів залежить від низки факторів, які визначають доцільність їх застосування у кожному конкретному випадку. Зокрема, регресійний підхід дозволяє оцінити значимість кожного фактора за допомогою статистичних тестів, а кореляційний аналіз – значимість впливу різних факторів, що в результаті дозволяє послідовно побудувати стійку та надійну модель. Підхід лінійного та цілочислового програмування дає можливість легко додавати до моделі нові обмеження, встановлені позикодавцями, а також розв'язувати задачу з великою кількістю характеристик заявників. Перевага нейронних мереж та класифікаційних дерев полягає у тому, що вони автоматично виявляють та опрацьовують взаємодії між характеристиками, що дозволяє виділяти різні групи заявників та створювати окремі скорингові картки для них. Метод найближчих сусідів та генетичні алгоритми дають можливість будувати скорингові картки, які можна динамічно оновлювати, додаючи нові характеристики та видаляючи старі, які більше не впливають на результат.

Важливий фактор, що впливає на вибір того чи іншого алгоритму для розв'язання задач кредитного скорингу – пріоритети позичальника. Комерційна складова іноді спонукає знизити точність, але підвищити простоту системи. Однак у випадку, коли мова йде про великий обсяг множини

заявників або їх характеристик, доцільним рішенням це покращення обчислювальних можливостей позикодавця та вибір оптимізаційних методів для розв'язання задач кредитного скорингу.

Висновки. Розглянуто основні статистичні та нестатистичні методи та підходи до розв'язання задачі кредитного скорингу, що являє собою задачу бінарної або мультигрупової класифікації. Описано переваги та недоліки розглянутих методів, а також умови їх застосування.

Список літератури

1. Lyn C. Thomas, David B. Edelman, Jonathan N. Crook. Credit Scoring and its Applications. SIAM Monographs on Mathematical Modeling and Computation. Philadelphia, 2002. 243 p. https://doi.org/10.1137/1.9780898718317
2. Sarlija N., Bensic M., Bohacek Z. Multinomial Model in Consumer Credit Scoring. 10th International Conference on Operational Research. Trogir: Croatia. 2004.
3. Abdou H., Pointon J. Credit scoring and decision-making in Egyptian public sector banks. *International Journal of Managerial Finance.* 2009. Vol. 5. N 4. P. 391–406. https://doi.org/10.1108/17439130910987549
4. Abdou H., Pointon J., El Masry A. Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*. 2008. Vol. 35. N 3. P. 1275–1292. https://doi.org/10.1016/j.eswa.2007.08.030
5. Baesens B., Gestel T.V., Viaene S., Stepanova M., Suykens J., Vanthienen J. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*. 2003. Vol. 54. N 6. P. 627–635. https://doi.org/10.1057/palgrave.jors.2601545
6. Юринець З.В., Юринець Р.В., Кунанець Н.Е., Мищишин І.Р. Регресійна модель оцінювання платоспроможності клієнта та банківських ризиків у процесі кредитування. *Соціально-економічні проблеми сучасного періоду України*. 2019. **138** (4). С. 69–73. https://doi.org/10.36818/2071-4653-2019-4-11
7. Mangasarian O.L. Linear and nonlinear separation of patterns by linear programming. *Oper. Res.* 1965. Vol. 13. P. 444–452. https://doi.org/10.1287/opre.13.3.444
8. Nath R., Jones T.W. A variable selection criterion in the linear programming approaches to discriminant analysis. *Decision Sci.* 1988. Vol. 19. P. 554–563. https://doi.org/10.1111/j.1540-5915.1988.tb00286.x
9. Gately E. Neural Networks for Financial Forecasting: Top Techniques for Designing and Applying the Latest Trading Systems. New York: John Wiley & Sons, Inc. 1996.
10. Великоіваненко Г.І., Савіна С.С., Колечко Д.В., Бень В.П. Побудова ансамблів моделей кредитного скорингу. *Журн. Нейро-нечіткі технології моделювання в економіці*. 2018. Т. 7. С. 34–77. https://www.doi.org/10.33111/nfmte.2018.034/
11. Zekic-Susac M., Sarlija N., Bensic M. Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Networks, and Decision Tree Models. 26th International Conference on Information Technology Interfaces. Croatia. 2004.
12. Huang J., Tzeng G., Ong C. Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation.* 2006. Vol. 174. N 2. P. 1039–1053. https://doi.org/10.1016/j.amc.2005.05.027
13. Huang C., Chen M., Wang C. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications.* 2007. Vol. 33. N 4. P. 847–856. https://doi.org/10.1016/j.eswa.2006.07.007
14. Etemadi H., Rostamy A., Dehkordi H. A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Systems with Applications.* 2009. Vol. 36. N 2/2. P. 3199–3207. https://doi.org/10.1016/j.eswa.2008.01.012
15. McKee T., Lensberg T. Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research*. 2002. Vol. 138. N 2. P. 436–451. https://doi.org/10.1016/S0377-2217(01)00130-8
16. Talebzadeh H., Mandutianu S., Winner C. Countrywide loan underwriting expert system. In: Proceedings of the Sixth Innovative Applications of Artificial Intelligence Conference. AAAI Press, Menlo Park, CA. 1994. https://doi.org/10.1609/aimag.v16i1.1123
17. Ben-David A., Frank E. Accuracy of machine learning models versus "hand crafted" expert systems – a credit scoring case study. *Expert Systems with Applications.* 2009. Vol. 36. N 3/1. P. 5264–5271. https://doi.org/10.1016/j.eswa.2008.06.071
18. Kumra R., Stein R., Assersohn I. Assessing a knowledge-based approach to commercial loan underwriting. *Expert Systems with Applications.* 2006. Vol. 30. N 3. P. 507–518. https://doi.org/10.1016/j.eswa.2005.10.013

Одержано 21.10.2022

Стовба Віктор Олександрович,
доктор філософії, молодший науковий співробітник
Інституту кібернетики імені В.М. Глушкова НАН України, Київ.
https://orcid.org/0000-0003-3023-5815
vik.stovba@gmail.com

MSC 91B82, 90C05, 90C10, 92B20

Viktor Stovba

Statistical and Optimization Methods in Credit Scoring

V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv
Correspondence: vik.stovba@gmail.com

Introduction. The role of credit scoring in the work of financial institutions is difficult to overestimate. Accurate and efficient scorecards allow lenders to assess risks correctly and monitor their investments. Such cards should be based on reliable statistical data about previous and current customers using statistical analysis methods.

Over the years of its development, the toolkit of credit scoring has also been supplemented with non-statistical methods based on the use of optimization procedures, decision trees, intelligent databases and knowledge bases, building network models, etc. Given the wide range of available methods, there is a need for their classification and application analysis.

The purpose of the article is to provide a brief description of all relevant statistical and non-statistical methods that allow solving credit scoring tasks in modern formulations. To reveal the features of using the methods described and conduct their comparison.

Results. Statistical methods allow to investigate the significance of all the factors included in the model, as well as to obtain a set of statistical estimates that help to assess the quality of the model. Thus, these methods allow to build an optimal and reliable model. Non-statistical methods allow you to add arbitrary restrictions to the model, automatically detect and process interactions between characteristics, and solve problems with a large number of applicants and their characteristics, which is facilitated due to the development of computational methods.

Conclusions. Modern mathematical methods allow to solve credit scoring tasks effectively, among which one of the main ones is the binary and multigroup classification. The choice of the optimal method depends on the type of system (static or dynamic), the creditor's computing capabilities and the importance of the results interpretation.

Keywords: credit scoring, statistical methods, mathematical programming, neural networks, genetic algorithms.