УДК 330.43:519.2

Ю. А. МАРЖИНА, М. С. МАЗОРЧУК, Н. С. БАКУМЕНКО

Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Україна

ВИЗНАЧЕННЯ ПАРАМЕТРІВ МОДЕЛІ КРЕДИТНОГО СКОРИНГУ НА ОСНОВІ АНАЛІЗУ СТАТИСТИЧНИХ ДАНИХ

Проведено аналіз особливостей вирішення завдань кредитного скорингу та обґрунтовано вибір методів для аналізу даних з метою визначення надійності позичальників. Розроблено алгоритмічні моделі та методику класифікації позичальників на основі логістичного регресійного аналізу та методів дерев рішень. Побудовано моделі кредитного скорингу та класифіковано позичальників за статистичними даними вибіркової популяції з використанням програмних засобів SPSS та WEKA. Проведено оцінювання якості отриманих моделей та зроблено висновки щодо використання розробленої методики.

Ключові слова: *кредитний скоринг, логістична регресія, дерево рішень, SPSS, WEKA, статистичний аналіз даних, DataMining.*

Вступ

Кредити стають все популярнішими серед жителів України і що таке кредитна історія знають практично всі, проте термін «кредитний скоринг» відомий одиницям. Про скорингові бали і наслідки масового впровадження систем кредитного скорингу для пересічних українських позичальників мають уявлення лише професіонали.

Кредитний скоринг є системою оцінки позичальника банком або іншим кредитором [1, 2]. Результатом цієї оцінки стає рішення по кредитній заявці: якщо позичальник набирає певну кількість балів, то йому видають кредит.

В основі скорингових систем лежить припущення, що люди зі схожими соціальними показниками поводяться однаково. Апріорно приймаючи такий постулат, можна будувати різні статистичні моделі, які можуть бути дуже корисні в процесі прийняття рішень щодо видачі кредитів. Якщо деяким соціальним характеристикам клієнта (стать, вік, місце проживання, посада, тривалість роботи в одному місці та ін.) присвоїти певні ваги, то кожного нового клієнта можна, на основі його анкети, віднести до групи сильно або слабо відповідних надійному позичальнику. Тобто, клієнту автоматично присвоюється ранг, який вказує ступінь довіри і увагу, яку йому слід надавати з боку банка.

До основних етапів побудови моделі кредитного скорингу можна віднести: визначення характеристик позичальників, збір відомостей про клієнтів, розробку скорингової моделі на основі наявних даних, автоматичне ранжування нових клієнтів за пріоритетними групами за допомогою скорингової моделі. Якщо в якості характеристики взяти здатність клієнта повернути кредитну позику, тоді в результаті ми отримаємо дві групи: клієнти, яким можна видати кредит і клієнти, кредитування яких дуже ризиковано.

Для вирішення цих завдань широко використовуються різні методи і моделі на основі двох основних підходів: статистичних методах аналізу даних, які базуються на розрахунку основних статистичних параметрів за ключовими ознаками і виявлення зв'язків на основі статистичних критеріїв та методах DataMining, де в процесі прийняття рішень використовуються дерева рішень, штучні нейронні мережі, генетичні алгоритми, еволюційне програмування, нечітка логіка [3-5].

Метою даної роботи є розробка ефективної методики для визначення параметрів моделі кредитного скорингу та класифікація позичальників за ключовими ознаками з використанням різних методів аналізу.

Постановка завдання дослідження

По суті, завдання кредитного скорингу являє собою класифікаційну задачу, де виходячи з наявної інформації необхідно отримати функцію, яка найбільш точно розділяє вибірку клієнтів на «поганих» і «хороших» [3-5].

Методи класифікації досить різноманітні і включають в себе:

– статистичні методи, засновані на дискримінантному аналізі (лінійна регресія, логістична регресія);

– різні варіанти лінійного програмування;

© Ю. А. Маржина, М. С. Мазорчук, Н. С. Бакуменко

– прийняття рішень на основі дерев класифікації;

– нейронні мережі;

– генетичні алгоритми та інші.

Серед методів статистичного аналізу для вирішення поставленої задачі найбільш прийнятний є метод логістичної регресії, хоча даний метод часто відносять і до методів DataMining. Логістична регресія використовується для передбачення ймовірності виникнення деякої події за значеннями множини ознак [5, 6]. Наприклад, надійність позичальника це ознака з двома категоріальними значеннями: 1 – надійний та 0 – ненадійний позичальник. За методом логістичної регресії можна оцінити ймовірність належності позичальника до надійних клієнтів чи ненадійних. Надійність позичальника y – це залежна змінна, яка приймає значення 1 чи 0, а множина ознак $x_1, x_2, \ldots, x_n$ - незалежні змінні, які можуть вимірюватися у різних шкалах, але в ході аналізу припускається, що усі дані вимірюються у метричній шкалі.

Ймовірність події для $y = 1$ дорівнює

$$P\{y = 1 | x\} = f(z),$$

$$z = A^T x = a_1 x_1 + \ldots + a_n x_n,$$

де A і x – вектори-стовпці значень незалежних змінних $x_1, x_2, \cdots, x_n$ та параметрів коефіцієнтів регресії чисел $a_1, a_2, \ldots a_n$ відповідно, а функція $f(z)$ – логістична функція, яка дорівнює:

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

Ймовірність події $y = 0$ дорівнює $P\{y = 0 | x\} = 1 - f(z)$.

Для побудови моделі кредитного скорингу на основі логістичної регресії потрібні дані вибіркової сукупності, яку можна визначити як вибірку для навчання.

Після побудови моделі та оцінки помилок можна робити оцінювання нових позичальників. В процесі використання цієї моделі важливо визначитися з результуючою ознакою та перевірити дані незалежних ознак на відповідність нормального закону розподілу – це основна умова використання регресійних методів. Це обмеження може привести до того, що метод логістичної регресії не завжди можна використовувати.

У протилежність методам регресійного аналізу, методи дерев рішень [5, 7], що відносяться до методів DataMining можуть бути використані для вирішення завдань кредитного скорингу практично без обмежень, але при великій кількості незалежних ознак побудова дерева може зайняти певний час.

Дерева рішення створюють ієрархічну структуру класифікаційних правил типу "ЯКЩО ... ТО ..." (if-then), що має вигляд дерева. Для ухвалення рішення, до якого класу віднести деякий об'єкт або ситуацію, потрібно відповісти на питання, що стоять у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд "значення параметра А більше х?". Якщо відповідь позитивна, здійснюється перехід до правого вузла наступного рівня, якщо негативна – то до лівого вузла; потім знову йде питання, пов'язане з відповідним вузлом. Популярність підходу пов'язана з наочністю і зрозумілістю.

Структура дерева містить такі елементи: «листя» і «гілки». На ребрах («гілках») дерева прийняття рішення записані ознаки, від яких залежить результуюча змінна y, в «листі» записані значення залежної змінної, а в інших вузлах – ознаки $x_1, x_2, \ldots, x_n$, за якими розрізняються випадки. Щоб класифікувати нові події, треба спуститися по дереву до листа і видати відповідне значення результуючої ознаки y.

Основна проблема використання дерев рішень – це побудова такого дерева, яке дозволяє з мінімальними помилками класифікувати нові події. Проблема побудови дерева полягає у виборі чергової ознаки, яка відповідає певним значенням результуючої ознаки. Для вибору чергової ознаки використовуються різні алгоритми, такі як:

1. Алгоритм ID3, де вибір ознаки відбувається на підставі приросту інформації, або на підставі коефіцієнту Джині.

2. Алгоритм C4.5 (поліпшена версія ID3), де вибір атрибута відбувається на підставі нормалізованого приросту інформації.

3. Алгоритм CART і його модифікації – IndCART, DB-CART.

4. Автоматичний детектор взаємодії хі-квадрат (CHAID). Виконує багаторівневий поділ при розрахунку класифікації дерев та інші.

На практиці в результаті роботи цих алгоритмів часто виходять занадто деталізовані дерева, які при їх подальшому застосуванні дають багато помилок, що пов'язано з явищем перенавчання. Але методи дерев рішень дають кращі результати, коли незалежні змінні не є дійсними числами, а вхідні дані не відповідають нормальному розподілу.

Таким чином, існуючі методи мають свої переваги та недоліки і не дозволять в повній мірі побудувати якісну модель кредитного скорингу, використовуючи один з підходів.

Математична модель визначення параметрів логістичної регресії

Розглянемо модель розрахунку параметрів моделі логістичної регресії для двох незалежних змінних (двох ознак).

Вихідними даними для розрахунку параметрів регресійної моделі кредитного скорингу є:

$X = \left\{x_1^{(i)}, x_2^{(i)}\right\}$ – незалежні змінні (стать та сукупний дохід);

$Y = \left\{y^{(i)}\right\}, y \in \{0,1\}$ – залежна змінна, яка відповідає надійності позичальника.

Результат розрахунків представлено такими параметрами:

r – коефіцієнт кореляції;

P_0, P_1, P_2 – ймовірності настання деякої події X;

a_0, a_1, a_2 – коефіцієнти логістичної регресії.

Для знаходження параметрів a_0, a_1, a_2 необхідно сформувати вибірку, яка навчає, тобто визначити відповідні

$$\left(x_1^{(1)}, x_2^{(1)}, y^{(1)}\right), \ldots, \left(x_1^{(m)}, x_2^{(m)}, y^{(m)}\right).$$

Для визначення параметрів моделі логістичної регресії використовується метод максимальної правдоподібності, відповідно якому параметри a_0, a_1, a_2 вибираються так, щоб отримати максимальні значення функції правдоподібності.

Для цього необхідно представити p для $y = 1$ як $p(x) = P(y = 1 \mid x)$ та для $y = 0$ як $1 - p(x) = P(y = 0 \mid x)$. Тоді функція правдоподібності незалежних спостережень буде дорівнювати

$$l(a|x) = \prod_{i=1}^{n} \left\{\left[p(x_i)\right]^{y_i} \times \left[1 - p(x_i)\right]^{(1-y_i)}\right\}.$$

Максимум функції правдоподібності відповідає максимуму і логарифму цієї функції, тоді можна знайти логарифм функції правдоподібності:

$$L(a|x) = \ln[l(a \mid x)] =$$

$$= \sum_{i=1}^{n} \{y_i \ln\left[p(x_i)\right] + (1 - y_i) \ln\left[1 - p(x_i)\right]\},$$

де $p(x) = \dfrac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$;

$$\hat{g}(x) = \ln \frac{p(x)}{1 - p(x)} = a_0 + a_1 x_1 + a_2 x_2.$$

Знаходження логістичних змінних здійснюється за наступними формулами за методом послідовного включення ознак [3]:

$$a_0 = \ln\left(\frac{1}{(1 - P_0) - 1}\right), \quad (1)$$

$$a_1 = \ln\left(\frac{1}{(1 - P_1) - 1}\right) - a_0, \quad (2)$$

$$a_2 = \ln\left(\frac{1}{(1 - P_2) - 1}\right) - a_0 - a_1. \quad (3)$$

Алгоритм визначення параметрів логістичної регресії полягає у наступному. На першому кроці знаходиться ймовірність P_0 для залежної змінної та розраховується значення параметру a_0 за формулою (1). Оцінюється якість моделі за критеріями хі-квадрат та Фішера. Далі обчислюється ймовірність P_1 за даними вже однієї незалежної та однієї залежної ознак та знаходиться a_1 за (2). Знову обчислюються критерії. Якщо якість моделі задовольняє умовам, то незалежна ознака залишається в моделі як істотна, якщо ні, то вона виключаються з моделі і здійснюється перехід до наступного кроку. Далі розраховується P_2 та знаходиться a_2 за двома незалежними ознаками (3). Якщо отримано якісну модель, тоді класифікуються нові ознаки за цією моделлю та визначається ймовірність для кожного випадку.

Математична модель побудови дерева рішень

Нехай нам задано безліч прикладів T, де кожен елемент цієї множини описується m атрибутами. Кількість прикладів в безлічі T будемо називати потужністю цієї множини і будемо позначати |T|. Нехай мітка класу приймає наступні значення $C_1, C_2, \ldots, C_k$.

Наше завдання полягатиме в побудові ієрархічної класифікаційної моделі у вигляді дерева з безлічі прикладів T. Процес побудови дерева відбуватиметься зверху вниз. Спочатку створюється корінь дерева, потім нащадки кореня і так надалі. На першому кроці ми маємо порожнє дерево (мається тільки корінь) і вихідна безліч T (асоційоване з коренем). Потрібно розбити вихідну безліч на підмножини. Це можна зробити, вибравши один з атрибутів в якості перевірки. Тоді в результаті розбиття виходять n (по числу значень ознаки) підмножин і, відповідно, створюються n нащадків кореня, кож-

ному з яких поставлено у відповідність своя підмножина, отримана при розбитті безлічі T. Потім ця процедура рекурсивно застосовується до всіх підмножин (нащадкам кореня).

Розглянемо критерій вибору ознаки, за яким має піти розгалуження. Очевидно, що в нашому розпорядженні m (по числу ознак) можливих варіантів, з яких треба вибрати найбільш суттєвий. Деякі алгоритми виключають повторне використання ознаки при побудові дерева, але в даному випадку такі обмеження не накладаються. Будь-який з атрибутів можна використовувати необмежену кількість разів при побудові дерева.

Нехай маємо перевірку X (в якості перевірки може бути обраний будь-який атрибут), яка приймає n значень $A_1, A_2, ..., A_n$. Тоді розбиття T з перевірки X дасть підмножини $T_1, T_2, ..., T_n$, при X рівному відповідно $A_1, A_2, ..., A_n$. Єдина доступна для аналізу інформація це те, яким чином класи розподілено в безлічі T і його підмножинах, отриманих при розбитті за X. Саме цим у даному випадку можна скористатися при визначенні критерію.

Нехай $\mathrm{freq}(C_j, T)$ - кількість прикладів з деякої безлічі T, які стосуються одного й того ж класу C_j. Тоді ймовірність того, що випадково обраний приклад з безлічі T буде належати до класу C_j становитимуть

$$P = \frac{\mathrm{freq}(C_j, T)}{|N|}.$$

Згідно теорії інформації, кількість I, що міститься у повідомленні інформації, залежить від її ймовірності і є логарифмічною функцією

$$I = \log_2\left(\frac{1}{P}\right). \quad (4)$$

Оскільки ми використовуємо логарифм з двійковою основою, то вираз (1.4) дає кількісну оцінку в бітах.

Вираз

$$\mathrm{Info}(T) = -\sum_{j=1}^{k} \frac{\mathrm{freq}(C_j, T)}{|T|} * \log_2\left(\frac{\mathrm{freq}(C_j, T)}{|T|}\right) \quad (5)$$

дає оцінку середньої кількості інформації, необхідної для визначення класу прикладу з безлічі T. У термінології теорії інформації вираз (5) називається ентропією множини T.

Таку ж оцінку, але тільки вже після розбиття множини T по X, дає такий вираз:

$$\mathrm{Info}_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * \mathrm{Info}(T_i). \quad (6)$$

Тоді критерієм для вибору атрибута буде наступна формула:

$$\mathrm{Gain}(X) = \mathrm{Info}(T) - \mathrm{Info}_X(T). \quad (7)$$

Критерій (7) використовується для всіх ознак. Вибирається атрибут, що максимізує даний вираз. Ця ознака буде перевіркою в поточному вузлі дерева, а потім з цієї ознаки проводиться подальша побудова дерева. Тобто у вузлі буде перевірятися значення з цієї ознаки і подальший рух по дереву буде проводитися в залежності від отриманої відповіді.

Такі ж міркування можна застосувати до отриманих підмножин $T_1, T_2, ..., T_n$ і продовжити рекурсивно процес побудови дерева, до тих пір, поки у вузлі не опиняться приклади з одного класу.

Одне важливе зауваження: якщо в процесі роботи алгоритму отримано вузол, асоційований з порожньою безліччю (тобто жоден приклад не потрапив в даний вузол), то він позначається як лист, і як рішення листа вибирається найбільш часто зустрічаємий клас у безпосереднього попередника даного листа.

На наступному кроці потрібно вибрати поріг, з яким повинні порівнюватися всі значення ознаки. Нехай числова ознака має кінцеве число значень. Позначимо їх $\{v_1, v_2, ..., v_i\}$. Попередньо відсортуємо всі значення. Тоді будь-яке значення, що лежить між v_i і v_{i+1}, і ділить всі приклади на дві множини: ті, які лежать зліва від цього значення $\{v_1, v_2, ..., v_i\}$, і ті, що праворуч $\{v_{i+1}, v_{i+2}, ..., v_n\}$. В якості порога можна вибрати середнє між значеннями v_i і v_{i+1}:

$$TH_i = \frac{v_i + v_{i+1}}{2}.$$

Таким чином, ми суттєво спростили завдання знаходження порога, і привели до розгляду всього $n-1$ потенційних порогових значень $TH_1, TH_2, ..., TH_{n-1}$. Формули (4) і (5) послідовно застосовуються до всіх потенційних порогових значень і серед них вибирається те, яке дає максимальне значення за критерієм (7). Далі це значення порівнюється зі значеннями критерію (7), підрахованими для решти ознак. Якщо з'ясується, що серед всіх ознак дана числова ознака має максимальне значення за критерієм (7), то в якості перевірки вибирається саме вона.

Слід зазначити, що всі числові тести є бінарними, тобто ділять вузол дерева на дві гілки.

Отже, ми маємо дерево рішень і хочемо використовувати його для розпізнавання нового об'єкта. Обхід дерева рішень починається з кореня дерева. На кожному внутрішньому вузлі перевіряється значення об'єкта Y за ознакою, яка відповідає перевірці в даному вузлі, і, залежно від отриманої відповіді, знаходиться відповідне розгалуження, і з цієї дузі рухаємо до вузла, що знаходиться на рівень нижче і т.д. Обхід дерева закінчується, як тільки зустрінеться вузол рішення, який і дає назва класу об'єкта Y.

Визначення параметрів моделі кредитного скорингу за допомогою програмних інструментаріїв

Формування моделі логістичної регресії та побудову дерева рішень пропонується зробити з використанням програмних інструментаріїв SPSS та WEKA [8, 9].

Визначення параметрів рівняння логістичної регресії в SPSS здійснюється за допомогою методу послідовного включення змінних. Вхідними змінними були такі ознаки позичальника, як статус позичальника (checking_status), строк кредитування (duration), кредитна історія (credit_history), мета позики (purpose), розмір кредиту (credit_amount), наявність депозитів (savings_status), заплановані витрати (other_payment_plans), наявність забезпечення кредиту (housing), умови праці (foreign_worker) та інші. На рис. 1 наведено фрагмент вибірки позичальників, за якою здійснюється навчання.

У результаті отримано наступну модель логістичної регресії:

$$z = -3{,}891 - 0{,}582 * \text{checking\_status} + 0{,}033 * \text{duration} + 0{,}363 * \text{credit\_history} - 0{,}232 * \text{savings\_status} + 0{,}241 * \text{other\_payment\_plans} + 0{,}256 * \text{housing} + 1{,}354 * \text{foreign\_worker},$$

де значення, на які помножуються кожна з істотних ознак, є коефіцієнтами моделі кредитного скорингу. Саме ці значення і дають нам інформацію щодо надійності позичальника. Наприклад, якщо ми розглянемо першого позичальника, дані за яким можна бачити на рис. 1, то z = -1,867, тобто цей клієнт може повернути позику лише з ймовірністю

$$p = \frac{1}{1+e^{-z}} = 0{,}1338 = 13\% .$$

В таблиці 1 наведено дані щодо якості отриманої моделі, тобто приведено змінні, які було включено до рівняння на останньому кроці та відображено критерії, за якими можна визначити істотність змінних та вплив їх на результат класифікації. У колонці «B» таблиці 1 відображені коефіцієнти регресії, у колонці «Стд. похибка» - значення стандартної похибки змінної, що включається у рівняння, «Вальд» - тест Вальда, що оцінює істотність коефіцієнта при незалежній змінній регресійної моделі, «Значення істотності» - відображає якість моделі при включенні певної ознаки, «Ст. свободи» - ступінь свободи для розрахунку критеріїв, «Exp(B)» показує експоненту в ступені B, що необхідно для знаходження ймовірності настання певного випадку.

В таблиці 2 наведено результати класифікації позичальників в SPSS за методом логістичної регресії. Виходячи зі спостережень, використаних для побудови моделі, 88,7% клієнтів, у яких були борги по кредиту («bad»), класифіковані коректно, і 40,3% «добрих» позичальників («good») також класифіковані правильно. Всього 75,2% спостережень в навчальній вибірці класифіковані вірно.

В таблиці 3 наведено результати класифікації позичальників за методом дерева рішень CHAID. За цим методом вдалося коректно класифікувати 74,3% випадків.

Аналогічно було проведено розрахунки за методом логістичної регресії в WEKA (рис. 2), а саме отримано: коефіцієнти рівняння регресії, час за який були оброблені і розраховані дані, відсоток та кількість коректних та некоректних класифікованих даних, ймовірність помилок при розрахунку. Як бачимо, результати, отримані в WEKA, збігаються з результатами, отриманими в SPSS.

В WEKA реалізовано інші методи дерев рішень, ніж в SPSS. Розглянемо результати класифікації за методом J48.

	checking_status	duration	credit_history	purpose	credit_amo...	savings_status	employment	installment_commitment	personal_status	other_parties	residence
1	1,00	6,00	1,00	1,00	1169,00	5,00	4,00	4,00	2,00	1,00	
2	2,00	48,00	2,00	1,00	5951,00	1,00	2,00	2,00	1,00	1,00	
3	4,00	12,00	1,00	5,00	2096,00	1,00	3,00	2,00	2,00	1,00	
4	1,00	42,00	2,00	6,00	7882,00	1,00	3,00	2,00	2,00	3,00	
5	1,00	24,00	3,00	4,00	4870,00	1,00	2,00	3,00	2,00	1,00	
6	4,00	36,00	2,00	5,00	9055,00	5,00	2,00	2,00	2,00	1,00	
7	4,00	24,00	2,00	6,00	2835,00	3,00	4,00	3,00	2,00	1,00	
8	2,00	36,00	2,00	3,00	6948,00	1,00	2,00	2,00	2,00	1,00	
9	4,00	12,00	2,00	1,00	3059,00	4,00	3,00	2,00	3,00	1,00	

Рис. 1. Статистичні дані для навчання

Таблиця 1

Критерії якості отриманої моделі

Змінні в рівнянні	B	Стд. похибка	Вальд	Ст. свободи	Значення істотності	Exp(B)
checking_status	-0,582	0,068	73,613	1	0,000	0,559
duration	0,033	0,006	27,653	1	0,000	1,034
credit_history	0,363	0,081	20,281	1	0,000	1,438
savings_status	-0,232	0,056	17,101	1	0,000	0,793
other_payment_plans	0,241	0,106	5,154	1	0,023	1,272
housing	0,256	0,097	6,970	1	0,008	1,292
foreign_worker	1,354	0,597	5,139	1	0,023	3,875
Константа	-3,891	1,225	10,092	1	0,001	0,020

Таблиця 2

Результати класифікації за методом логістичної регресії в SPSS

Спостережені	Передбачені		
	good	bad	Відсоток коректних
good	621	79	88,7%
bad	179	121	40,3%
Загальний відсоток			75,2%

Таблиця 3

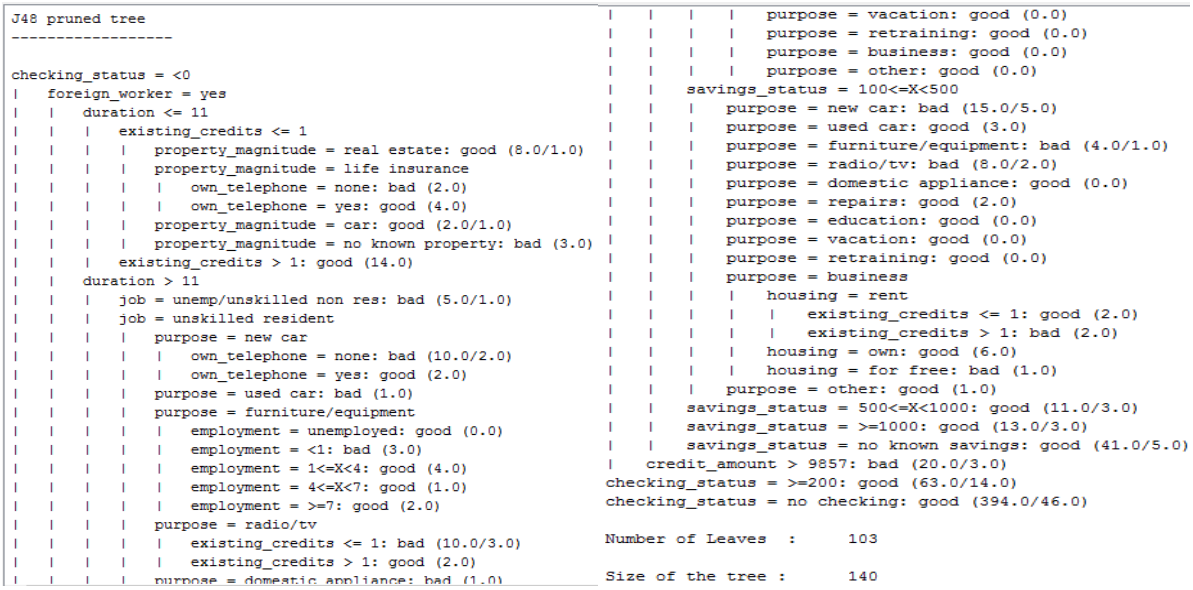
Результати класифікації за методом CHAID в SPSS

Спостережені	Передбачені		
	good	bad	Відсоток коректних
good	593	107	84,7%
bad	150	150	50,0%
Загальний відсоток			74,3%

```
Odds Ratios...
                                                  Class
Variable                                           good
=======================================================
checking_status=<0                               0.4593
checking_status=0<=X<200                         0.6682
checking_status=>=200                            1.2064
checking_status=no checking                      2.5443
duration                                         0.9725
credit_history=no credits/all paid               0.4436
credit_history=all paid                          0.3843
credit_history=existing paid                     0.7971
credit_history=delayed previously                1.0411
credit_history=critical/other existing credit    1.8643
purpose=new car                                  0.5006
purpose=used car                                 2.6497
purpose=furniture/equipment                      1.1047
purpose=radio/tv                                 1.2209
purpose=domestic appliance                       0.8443
purpose=repairs                                  0.6215
purpose=education                                0.4827
purpose=vacation                                      1
purpose=retraining                               3.9251
purpose=business                                 1.0493
purpose=other                                    2.2184
credit_amount                                    0.9999
savings_status=<100                              0.6439
savings_status=100<=X<500                        0.9208
savings_status=500<=X<1000                       0.9379
savings_status=>=1000                            2.4569
savings_status=no known savings                  1.6594
```

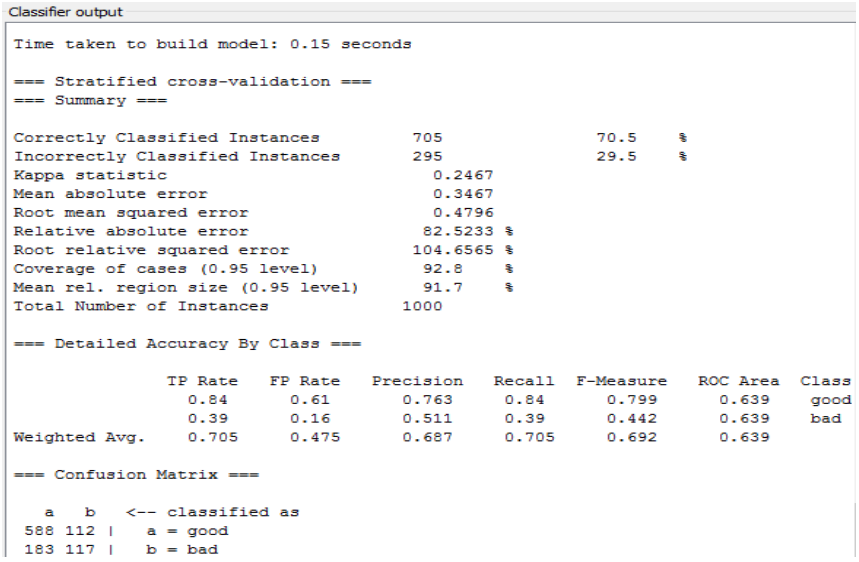
```
Classifier output

Time taken to build model: 0.35 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         752               75.2    %
Incorrectly Classified Instances       248               24.8    %
Kappa statistic                          0.375
Mean absolute error                      0.3098
Root mean squared error                  0.4087
Relative absolute error                 73.727  %
Root relative squared error             89.1751 %
Coverage of cases (0.95 level)          98.6    %
Mean rel. region size (0.95 level)      91.3    %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.864     0.51       0.798     0.864     0.83       0.785    good
                 0.49      0.136      0.607     0.49      0.542      0.785    bad
Weighted Avg.    0.752     0.398      0.741     0.752     0.744      0.785

=== Confusion Matrix ===

   a   b   <-- classified as
 605  95 |   a = good
 153 147 |   b = bad
```

Рис. 2. Побудова моделі логістичної регресії в WEKA



```
J48 pruned tree
------------------

checking_status = <0
|   foreign_worker = yes
|   |   duration <= 11
|   |   |   existing_credits <= 1
|   |   |   |   property_magnitude = real estate: good (8.0/1.0)
|   |   |   |   property_magnitude = life insurance
|   |   |   |   |   own_telephone = none: bad (2.0)
|   |   |   |   |   own_telephone = yes: good (4.0)
|   |   |   |   property_magnitude = car: good (2.0/1.0)
|   |   |   |   property_magnitude = no known property: bad (3.0)
|   |   |   existing_credits > 1: good (14.0)
|   |   duration > 11
|   |   |   job = unemp/unskilled non res: bad (5.0/1.0)
|   |   |   job = unskilled resident
|   |   |   |   purpose = new car
|   |   |   |   |   own_telephone = none: bad (10.0/2.0)
|   |   |   |   |   own_telephone = yes: good (2.0)
|   |   |   |   purpose = used car: bad (1.0)
|   |   |   |   purpose = furniture/equipment
|   |   |   |   |   employment = unemployed: good (0.0)
|   |   |   |   |   employment = <1: bad (3.0)
|   |   |   |   |   employment = 1<=X<4: good (4.0)
|   |   |   |   |   employment = 4<=X<7: good (1.0)
|   |   |   |   |   employment = >=7: good (2.0)
|   |   |   |   purpose = radio/tv
|   |   |   |   |   existing_credits <= 1: bad (10.0/3.0)
|   |   |   |   |   existing_credits > 1: good (2.0)
|   |   |   |   purpose = domestic appliance: bad (1.0)
|   |   |   |   purpose = vacation: good (0.0)
|   |   |   |   purpose = retraining: good (0.0)
|   |   |   |   purpose = business: good (0.0)
|   |   |   |   purpose = other: good (0.0)
|   |   |   savings_status = 100<=X<500
|   |   |   |   purpose = new car: bad (15.0/5.0)
|   |   |   |   purpose = used car: good (3.0)
|   |   |   |   purpose = furniture/equipment: bad (4.0/1.0)
|   |   |   |   purpose = radio/tv: bad (8.0/2.0)
|   |   |   |   purpose = domestic appliance: good (0.0)
|   |   |   |   purpose = repairs: good (2.0)
|   |   |   |   purpose = education: good (0.0)
|   |   |   |   purpose = vacation: good (0.0)
|   |   |   |   purpose = retraining: good (0.0)
|   |   |   |   purpose = business
|   |   |   |   |   housing = rent
|   |   |   |   |   |   existing_credits <= 1: good (2.0)
|   |   |   |   |   |   existing_credits > 1: bad (2.0)
|   |   |   |   |   housing = own: good (6.0)
|   |   |   |   |   housing = for free: bad (1.0)
|   |   |   |   purpose = other: good (1.0)
|   |   |   savings_status = 500<=X<1000: good (11.0/3.0)
|   |   |   savings_status = >=1000: good (13.0/3.0)
|   |   |   savings_status = no known savings: good (41.0/5.0)
|   credit_amount > 9857: bad (20.0/3.0)
checking_status = >=200: good (63.0/14.0)
checking_status = no checking: good (394.0/46.0)

Number of Leaves  :     103

Size of the tree :      140
```

Рис. 3. Побудова дерева рішень в WEKA



```
Classifier output

Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         705               70.5    %
Incorrectly Classified Instances       295               29.5    %
Kappa statistic                          0.2467
Mean absolute error                      0.3467
Root mean squared error                  0.4796
Relative absolute error                 82.5233 %
Root relative squared error            104.6565 %
Coverage of cases (0.95 level)          92.8    %
Mean rel. region size (0.95 level)      91.7    %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.84      0.61       0.763     0.84      0.799      0.639    good
                 0.39      0.16       0.511     0.39      0.442      0.639    bad
Weighted Avg.    0.705     0.475      0.687     0.705     0.692      0.639

=== Confusion Matrix ===

   a   b   <-- classified as
 588 112 |   a = good
 183 117 |   b = bad
```

Рис. 4. Результати побудови дерева рішень J48

Дерево рішень J48 в програмі будується ієрархічно, спочатку будується корінь дерева, а потім листя с певною долею ймовірності, також відображається розмір та кількість рівнів дерева, а також час за який будується модель (рис. 3 - 4). Як видно, алгоритм J48 дав гірші результати, ніж попередньо розглянуті методи – лише 70,5% випадків вдалося коректно класифікувати, при цьому середня похибка досить висока – 0,3467.

Висновки

Виходячи з результатів можна зробити висновки, що логістична регресія краще підходить для аналізу даних кредитного скорінгу, оскільки не тільки дозволяє провести класифікацію з меншими похибками, але і побудувати логістичну модель, тобто визначитися з вагами для кожної ознаки, які є по суті скоринговими балами позичальників. Але слід звернути увагу, що побудована модель добре працює лише з даними числової природи. Усі ознаки, які мали мало категорій, було виключено з аналізу, оскільки виявилися неістотними.

Методи дерев рішень дозволили включити до аналізу усі дані, але побудувати скорингову модель не вдалося. Середня похибка при використанні методів дерев рішень також більше, ніж при використанні логістичної регресії. Тому в процесі прийняття рішень слід звертати особливу увагу на типи вхідних даних та шкали їх вимірювання.

Література

1. Владичин, У. В. Банківське кредитування [Текст] : навч. посіб. / У. В. Владичин ; за ред. С. К. Реверчука. – К. : Атіка, 2008. - 648 с.

2. Вовчак, О. Д. Підвищення ефективності управління кредитними ризиками банків на основі використання сучасних методів оцінки кредитоспроможності позичальників [Текст] / О. Д. Вовчак, В. В. Пірог // Бізнес Інформ. – 2011. – № 2(2). – С. 7-10.

3. Бондаренко, М. Скоринг оценка кредитоспособности заемщика [Текст] / М. Бондаренко // Банковский аудитор. – 2004. – № 11. – С. 5-8.

4. Гребенник, Т. В. Управление качеством кредитного портфеля коммерческого банка в период посткризисного развития [Текст] : дис. ... канд. экон. наук : 08.00.10 ; защищена 21.05.2015 / Гребеник Татьяна Викторовна. – М., 2014. – 214 с.

5. Marque´s, A. I. Literature review on the application of evolutionary computing to credit scoring [Text] / A. I. Marque´s, V. Garcı´a, J. S. Sa´nchez // Journal of the Operational Research Society. – 2013. – № 64. – P. 1384-1399.

6. Логистическая регрессия [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title= Логистическая_регрессия – Загл. с экрана. 30.06.2015.

7. Wilkinson, L. Classification and Regression Trees [Electronic resource] / L. Wilkinson. – Access mode: http://www.cs.uic.edu/~wilkinson/Publications /c&rtrees.pdf. – Загл. с экрана. 30.06.2015.

8. Груздев, А. В. Применение метода деревьев решений для задач банковского скоринга [Текст] / А. В. Груздев // Управление финансовыми рисками. – 2012. – № 2(30). – С. 104-123.

9. Witten, Ian H. Data Mining: Practical Machine Learning Tools and Techniques [Text] / Ian H. Witten, Eibe Frank, Mark A. Hall. – Third edition. – Elsevier Inc., 2011. – 664 p.

Надійшла до редакції 30.06.2015, розглянута на редколегії 11.09.2015

ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ МОДЕЛИ КРЕДИТНОГО СКОРИНГА НА ОСНОВЕ АНАЛИЗА СТАТИСТИЧЕСКИХ ДАННЫХ

Ю. А. Маржина, М. С. Мазорчук, Н. С. Бакуменко

Проведен анализ особенностей решения задач кредитного скоринга и обоснован выбор методов для анализа данных с целью определения надежности заемщиков. Разработаны алгоритмические модели и методика классификации заемщиков на основе логистического регрессионного анализа и методов деревьев решений. Построены модели кредитного скоринга и классифицированы заемщики согласно статистическим данным выборочной популяции с использованием программных средств SPSS и WEKA. Проведена оценка качества полученных моделей и сделаны выводы относительно использования разработанной методики.

Ключевые слова: кредитный скоринг, логистическая регрессия, дерево решений, SPSS, WEKA, статистический анализ данных, DataMining.

DETERMINATION OF PARAMETERS OF MODEL OF CREDIT SCORING ON THE BASIS OF THE ANALYSIS OF STATISTICAL DATA

Y. A. Marzhina, M. S. Mazorchuk, N. S. Bakumenko

Problems of credit scoring analysis are solved and choices of methods to analyze the data to determine the reliability of borrowers are justified. Developed algorithmic models and methods of classification of borrowers based on logistic regression analysis methods and decision trees. Credit scoring models are constructed and classified borrowers sample population statistics using software of SPSS and WEKA. An estimation of the quality of the obtained models and conclusions about the use of the developed methodology are considered.

Keywords: credit scoring, logistic regression, tree of decisions, SPSS, WEKA, statistical analysis of data, DataMining.

Маржина Юлія Анатоліївна – студент кафедри інформатики, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна, e-mail: m-julia-m@mail.ru.

Мазорчук Марія Сергіївна – канд. техн. наук, доцент, доцент кафедри інформатики, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна, e-mail: mazorchuk.mary@gmail.com.

Бакуменко Ніна Станіславівна – канд. техн. наук, доцент кафедри інформатики, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна, e-mail: nina@bigline.net.