УДК 519.86

А. В. Матвійчук, д.е.н., проф.
професор кафедри економіко-математичного моделювання
ДВНЗ «Київський національний економічний університет імені Вадима Гетьмана»

Andriy Matviychuk, Doctor of Economics, Professor
Professor of Department of Economic and Mathematical Modeling,
Kyiv National Economic University named after Vadym Hetman

Ю. В. Клебан,
завідувач відділу ІТ наукової бібліотеки
Національного університету «Острозька академія»

Yurii Kleban,
Head of IT department at Science Library
National University of Ostroh Academy

БІННІНГ КІЛЬКІСНИХ ЗМІННИХ З ФОРМУВАННЯМ ТРЕНДУ ДЛЯ ЗАДАЧ СКОРИНГУ

BINNING OF QUANTITATIVE VARIABLES WITH FORMATION OF TREND FOR SCORING ISSUES

АНОТАЦІЯ. У статті запропоновано методологічний підхід та алгоритм розбиття кількісних змінних на категорії із забезпеченням дотримання тренду в значеннях їх показників вагомості ознаки (WOE). У процесі аналізу спеціалізованої літератури з питань біннінгу та проведення численних експериментів на реальних даних було сформульовано низку вимог, яким має задовольняти алгоритм категоризації змінних. Реалізований засобами мови програмування R алгоритм дозволяє швидко провести категоризацію факторів, забезпечуючи одночасно тренд WOE та дотримання обмежень щодо обсягу спостережень у кожній групі. Даний підхід показав високу ефективність роботи у тому числі на великих масивах даних.

КЛЮЧОВІ СЛОВА: скорингова модель, біннінг, категоризація кількісної змінної, вагомість ознаки (WOE), інформаційна значущість (IV).

АННОТАЦИЯ. В статье предложен методологический подход и алгоритм разбиения количественных переменных на категории с обеспечением тренда в значениях их показателей весомости признака (WOE). В процессе анализа специализированной литературы по вопросам биннинга и проведения многочисленных экспериментов на реальных данных был сформулирован ряд требований, которым должен удовлетворять алгоритм категоризации переменных. Реализованный средствами языка программирования R алгоритм позволяет быстро провести категоризацию факторов, обеспечивая одновременно тренд WOE и соблюдение ограничений по объему наблюдений в каждой группе. Данный подход продемонстрировал высокую эффективность работы в том числе на больших массивах данных.

КЛЮЧЕВЫЕ СЛОВА: скоринговая модель, биннинг, категоризация количественной переменной, весомость признака (WOE), информационная значимость (IV).

ANNOTATION. The article proposes a methodological approach and an algorithm for dividing of quantitative variables into categories, taking into account the trend in values of their weights of evidence (WOE). In the process of analyzing the specialized literature on binning issues and carrying out numerous experiments on real data, a number of requirements to algorithm of categorization of variables were

formulated. The algorithm implemented by means of the programming language R enables rapid binning of factors, simultaneously ensuring the trend of the WOE and observing the limitations on the volume of observations in each group. This approach demonstrated high efficiency of work including on big data.

KEYWORDS: scoring model, binning, categorization of a quantitative variable, weight of evidence (WOE), information value (IV).

Вступ. Адекватне оцінювання кредитоспроможності клієнтів банківських установ передбачає пошук механізмів підвищення ефективності скорингових моделей. Одним із головних напрямів у вирішенні цього завдання стає забезпечення врахування найрізноманітнішої інформації щодо позичальника, що визначатиме ризик його дефолту за зобов'язаннями. Оскільки рівень кредитоспроможності позичальника обумовлюється як кількісними факторами, так і якісними, про що вже наголошувалось вище, то важливо передбачити ефективну обробку в тому числі таких характеристик, як форма власності, пов'язані особи тощо для юридичних осіб, або ж стать, рівень освіти, посада та ін. для фізичних осіб, щоб на їх основі можна було будувати різноманітні скорингові моделі. Для отримання можливості врахування у статистичних скорингових моделях якісних показників виникає потреба в їх перетворенні у кількісну форму.

Одним з варіантів вирішення цього завдання є застосування підходу, коли певним характеристикам якісної змінної надаються номерні позначення: 0, 1, 2… Наприклад, такий показник, як «Освіта», може приймати значення: «незакінчена середня» – 1; «середня» – 2; «середньотехнічна» – 3, «вища» – 4 тощо. Однак за такого кодування може виникати некоректне впорядкування категорій, адже автоматично встановлюється, що позичальники з середньотехнічною освітою отримують нижчу кількісну оцінку, ніж з вищою освітою. Якщо подібним чином здійснити нумерацію регіонів країни, то такий підхід взагалі стає беззмістовним – кількісне значення одного регіону в кінці алфавіту може у десятки разів перевищувати номер регіону на початку списку, що взагалі нічого не означає з огляду на оцінку ризиковості кредитної поведінки його мешканців. Тож подібний підхід до переведення якісних змінних у числа є некоректним.

Інший підхід до кодування якісних характеристик базується на застосуванні фіктивних змінних, який полягає у позначенні належності досліджуваного об'єкта до певної категорії якісного показника бінарною маскою. Тобто, наприклад, характеристика «Освіта» буде представлена усіма бінарними фіктивними змінними, кожна з яких вказуватиме належність позичальника до відповідної категорії. Зокрема, для позичальника з вищою освітою

фіктивна змінна, що відповідає категорії «вища», матиме значення «1», а у шести інших категорій – «0». Звісно, для кодування семи класів якісного показника буде достатньо і шести фіктивних змінних, проте практично для всіх характеристик (як якісних, так і кількісних) додається ще клас з невідомими значеннями «NULL» (коли в базі даних відсутня інформація за цим показником щодо певного позичальника).

При цьому важливо розуміти, що із збільшенням кількості якісних характеристик за такого підходу випереджаючими темпами зростатиме кількість фіктивних бінарних змінних. До того ж, переважна більшість їх значень дорівнюватиме нулю. Тож, подібне розширення бази даних та кількості факторів суттєво ускладнюватиме процес побудови скорингових моделей та знижуватиме їх ефективність.

Метою статті є вивчення підходів до категоризації факторів для задач кредитного скорингу, а також розробка алгоритму формування категорій зі збереженням тренду.

Для переведення якісних змінних у числову форму було вирішено скористатись загальноприйнятим у скорингу підходом, що ґрунтується на розрахунку показника *WOE* (*Weight Of Evidence*), який для кожної підгрупи позичальників визначає узагальнену кількісну оцінку їх кредитної поведінки. Така оцінка базується на обчисленні часток надійних і ненадійних угод за кожною підгрупою (категорією) показника відносно загальної кількості надійних і ненадійних угод, відповідно, із подальшим розрахунком *WOE* за формулою:

$$WOE_i = ln\left(\frac{B_i}{G_i}\right), \quad i = \overline{1,k}, \tag{1}$$

де B_i – відношення кількості ненадійних позичальників у i-й категорії до загального числа ненадійних позичальників у вибірці; G_i – частка надійних угод за i-ю категорією відносно їх загальної кількості; k – кількість підгруп (категорій) змінної.

У спеціалізованій літературі [1, 2] рекомендується *WOE* розраховувати не тільки для якісних показників, але й для кількісних, попередньо здійснивши розбиття усієї множини значень відповідного показника на інтервали. І вже для кожного такого i-го інтервалу розраховується власне WOE_i.

У принципі, такий підхід має логічне підґрунтя. Адже не можна однозначно стверджувати, що, скажімо, заробітна плата у

20 тис. грн. вказує на значно менший кредитний ризик позичальника порівняно з тим, хто зазначив у кредитній заявці зарплату 4 тис. грн. По-перше, для отримання кредиту в умовах української дійсності зацікавлена особа може отримати практично будь-яку довідку по заробітній платі (хоча зазвичай навіть такої довідки не потрібно), тож високі її показники не гарантують, що вона є дійсно такою. По-друге, поширеною є практика зниження рівня офіційної зарплатні в комерційних організаціях з метою зменшення податкових відрахувань. Таким чином, категорія позичальників із зарплатою у 4 тис. грн. може виявитись навіть надійнішою, ніж позичальники з надвисокими доходами. І специфіку поведінки кожної з таких підгруп дозволить виявити саме розрахунок показників *WOE*. Натомість, оперування моделі з вихідними значеннями у 20 тис. грн. та 4 тис. грн. передбачало б на п'ятикратну перевагу першого позичальника із відповідним нарахуванням скорингового балу, що явно не відповідає логіці економічних процесів.

Також варто зауважити, що близько двох третин кредитних історій, що використовувались у наших попередніх дослідженнях, взагалі не містили інформації щодо рівня заробітної плати. Тож при побудові моделі на абсолютних значеннях цього показника більша частина статистичних спостережень була б виключена з навчальної вибірки. Крім того, якщо для нового позичальника рівень заробітної плати не вказаний, то така модель також не зможе бути застосована в оцінюванні його кредитного ризику.

Проте, *WOE* розраховується як для різних категорій якісного показника чи інтервалів кількісного показника, так і для окремої категорії, відповідної пропущеним даним. Таким чином, застосування *WOE* надає можливість зробити модель універсальною, тобто такою, яку можна використовувати за будь-якого наповнення даних щодо характеристик позичальників. На додаток до цього, при розрахунку *WOE* здійснюється переведення якісних та кількісних показників різної розмірності до нормалізованих числових значень, придатних для побудови скорингових моделей будь-якого типу.

Підсумовуючи зазначене вище підкреслимо переваги застосування показника *WOE* при побудові скорингових моделей [3], що полягають, насамперед, у можливості:

1) включити у модель пропущені значення змінних (оскільки часто у базах даних кредитних організацій різні позичальники характеризуються різними показниками, то без цієї властивості доводиться або відкидати спостереження, або видаляти пояснюючі змінні, що суттєво звужує застосовність моделі);

2) виключити вплив екстремальних викидів на якість моделі (що підвищує її стійкість та робастність);

3) привести всі вхідні змінні до однієї розмірності (для певних типів економіко-математичних моделей це є суттєвим, оскільки дозволяє виключити надмірний вплив окремих змінних на результат розрахунків).

Для оцінювання ефективності розбиття змінної на категорії та визначення загальної прогностичної сили категоризованого фактора (якісної чи кількісної характеристики, переведеної у категорії з розрахунком відповідного *WOE*) застосовується показник інформаційної значимості *IV* (*Information Value*) [4]:

$$IV = \sum_{i=1}^{k} \left(B_i - G_i \right) \cdot WOE_i \, . \qquad (2)$$

Чим вищою є інформаційна значимість предиктора, тим сильнішою є залежність від нього вихідної змінної. Коефіцієнти *IV*, отримані в результаті розрахунку (2), за [4] інтерпретуються таким чином:

- $IV < 0{,}02$ – характеристика не має прогностичної сили;
- $0{,}02 \le IV < 0{,}1$ – слабка прогностична сила;
- $0{,}1 \le IV < 0{,}3$ – середня прогностична сила;
- $0{,}3 \le IV < 0{,}5$ – висока прогностична сила;
- $0{,}5 \le IV$ – відмінна прогностична сила категоризованої змінної.

Автори роботи [5] проблемною ділянкою побудови моделі оцінки кредитоспроможності вважають створення ефективної процедури розбиття кількісної характеристики на категорії, що б забезпечувало підвищення точності класифікації позичальників за рівнем їх надійності. Ця процедура дає можливість посилити робастність моделі (її стійкість до випадкових збурень і похибок у даних) та одночасно збільшити її адекватність, адже об'єднання дискретних значень змінних у категорії дозволяє виключити негативний вплив екстремальних викидів, замінюючи їх оцінками систематичного впливу категорії на результуючий показник. Процес категоризації вхідних змінних (або розбиття кількісних змінних на категорії) у скорингу ще називається біннінгом (англ. binning) [1].

Розробка ефективного алгоритму біннінгу зводиться до розв'язання задач визначення оптимального числа категорій та їх діапазонів для кожної з кількісних вхідних змінних. Загальноприйнятим правилом при розв'язанні цих задач є те, що кожна категорія має об'єднувати значення показника з однаковими вла-

стивостями відносно їх впливу на кредитоспроможність клієнта. Даному питанню була присвячена низка вітчизняних і закордонних публікацій, короткий аналіз яких подається нижче.

У статті А.С. Сорокіна [1] описано процес побудови скорингової моделі, починаючи з поділу даних на тестову та навчальну вибірки, та закінчуючи оцінкою параметрів логістичної регресії. Також детально описується процедура біннінгу кількісних змінних на основі розрахунку інформаційної значимості *IV* та показника вагомості ознаки *WOE*. Під час поділу змінних на категорії Сорокін керується:

- максимізацією показника інформаційної значимості змінної *IV* як критерію оптимальності біннінгу;
- необхідністю розбиття множини значень кількісного показника на категорії, які б забезпечували зростаючий або спадаючий тренд *WOE* при переході від однієї категорії до іншої;
- доцільністю об'єднання категорій з близькими значеннями вагомості ознаки *WOE* для посилення тенденції її зростання або спадання;
- потребою в забезпеченні суттєвої відмінності *WOE* у різних категоріях;
- обмеженням максимальної кількості категорій до 50.

На доцільності забезпечення суттєвої різниці *WOE* при переході від однієї категорії до іншої також було наголошено у джерелі [3], де Дж. Херманом проаналізовано три способи біннінгу:

- встановлення інтервалів категорій однакової довжини на множині можливих значень показника;
- поділ на категорії з однаковою кількістю прикладів;
- посилення різниці значень *WOE* між сусідніми категоріями.

Компанія FICO надає ряд рекомендацій щодо поділу кількісної змінної на категорії [7]:

- кожна категорія має містити достатньо значень, аби нівелювати вплив екстремальних показників і шуму в вибірці;
- кожна категорія має формуватись з елементів, ідентичних за мірою впливу на результуючу змінну;
- абсолютні показники інформаційної значимості *IV* змінної несуть мінімальне змістовне навантаження і мають використовуватись лише для порівняння.

У роботі [2] Н. Сіддікі пропонує такі базові рекомендації щодо проведення біннінгу:

- пропущені значення показника мають входити в окрему категорію;
- кожна категорія не може містити менше 5 % вибірки;

- кількість надійних чи ненадійних угод у категорії не мають дорівнювати 0.

У роботі Н.Б. Палкіна та В.В. Афанасьєфа [6] була досліджена проблема оптимального квантування (укрупнення вже утворених категорій) для підвищення точності бінарних класифікаторів. У процесі проведення експерименту вдавалось збільшити точність класифікації при значній втраті інформаційної значимості змінних. Такий результат ставить під сумнів роль показника інформаційної значимості *IV* як критерію ефективності категоризації пояснюючих змінних.

Попри те, що у проаналізованих вище публікаціях [1-7] алгоритм біннінгу був достатньо детально описаний і прокоментований, принципи розбиття кількісних змінних на категорії у цих роботах є надто відмінними між собою. Так, різняться рекомендації щодо: розмірів і кількості категорій, правил їх об'єднання, доцільності застосування показника інформаційної значимості тощо. Тому виникає потреба у перевірці адекватності розроблених раніше методів біннінгу, їх доповненні та розвитку, пошуку нових індикаторів ефективності поділу змінних на категорії, перевірки правил біннінгу та перегляду ролі показника інформаційної значимості.

Це зумовлює актуальність розробки методологічного підходу та алгоритмів проведення ефективної категоризації кількісних змінних у процесі побудови скорингових моделей.

Вирішення поставленого завдання зводиться до таких етапів: сформулювати гіпотези щодо оптимального поділу діапазону значень кількісних змінних на основі узагальнення світового досвіду з проведення біннінгу; здійснити алгоритмічну та програмну реалізацію процесів поділу кількісних змінних на категорії (відповідно до висунутих гіпотез) та побудови скорингових моделей для різних варіантів біннінгу вхідних даних; систематизувати результати експериментальних досліджень із обґрунтуванням відповідних висновків і рекомендацій.

Як зазначалось вище, створення ефективного алгоритму поділу кількісних змінних на категорії є нетривіальним завданням, оскільки не існує якогось визнаного критерію оптимальності для його розв'язання. Зокрема, компанія FICO, яка є «законодавцем» в області конструювання скорингових карт, наголошує, що встановити вичерпну процедуру оптимального біннінгу неможливо, адже це є питанням «мистецтва та науки» [7]. Проте, якщо спеціаліст у цій галузі може здійснити ефективний біннінг, керуючись своїми досвідом, інтуїцією, знаннями, то подібні навички можна

покласти в основу алгоритмів, підкріпивши їх додатковими перевірками, критеріями тощо.

У роботі [4] вказано, що оцінювання адекватності розбиття змінної на категорії може бути здійснено за показником інформаційної значимості (2). Однак, крім того, що не всі літературні джерела одностайні щодо коректності застосування цього показника у якості критерію оптимальності біннінгу, він несе ряд додаткових ускладнень, позаяк не задовольняє властивості адитивності:

$$\sum_{i=e}^{g}\left(B_i - G_i\right)\cdot WOE_i \neq \left(B_j - G_j\right)\cdot WOE_j, \qquad (3)$$

де j – новоутворена категорія змінної, що складається з усіх категорій між e та g, де $e,g \in [1,k]$, $\left(B_j = \sum_{i=e}^{g} B_i,\ G_j = \sum_{i=e}^{g} G_i \right)$.

Вираз (3) означає, що сума показників *IV* кількох сусідніх категорій не дорівнює значенню даного показника для нової категорії після об'єднання цих сусідів. Нерівність (3) призводить до того, що пошук оптимальної кількості та діапазонів категорій для змінної не може бути задачею лінійного програмування, а вимагає комбінаторного перебору всіх можливих варіантів. Поставлена задача може бути розв'язана лише шляхом порівняння ефективності різних способів категоризації на основі проведення експерименту.

Беручи до уваги зазначені вище обмежені властивості коефіцієнта інформаційної значимості, які не дозволяють скористатись ним як єдиним критерієм оптимальності поділу кількісних змінних на категорії, у даній роботі вирішено здійснювати оцінювання ефективності біннінгу одночасно на основі виразу (2) та узагальненого показника адекватності моделі. Зважаючи на бінарну форму вихідної змінної, з цією метою застосуємо загальновживаний показник ефективності скорингових моделей – коефіцієнт Джині.

Здійснити дослідження впливу біннінгу на якість класифікатора можна в рамках методологічного підходу до проведення категоризації кількісних змінних у процесі побудови скорингових моделей, зміст якого полягає в реалізації таких етапів:

1) визначення інформаційної бази для досліджень, формування навчальної та тестової вибірок;

2) розбиття значень пояснюючих змінних на категорії за різними алгоритмами біннінгу;

3) розрахунок для кожної категорії за всіх варіантів біннінгу показників *WOE* та *IV*;

4) побудова скорингових моделей на навчальній вибірці для різних варіантів категоризації вхідних змінних;

5) оцінка адекватності побудованих скорингових моделей на тестовій вибірці за критерієм Джині;

6) аналіз отриманих результатів, формулювання висновків щодо ефективності алгоритмів біннінгу.

В рамках запропонованого методологічного підходу розробимо низку алгоритмів категоризації кількісних змінних та проведемо порівняльний аналіз їх ефективності з метою вибору найбільш адекватного з них.

У процесі аналізу спеціалізованої літератури з питань біннінгу та проведення численних експериментів на реальних даних автором було сформульовано низку вимог, яким має задовольняти алгоритм поділу кількісних змінних на категорії:

- усі записи показника, за якими відсутня інформація, мають бути об'єднані в окрему категорію з відповідним розрахунком її *WOE* та *IV*;
- у кожній окремій категорії мають бути представлені як виконані згідно умов договору, так і дефолтні кредити;
- одне значення показника не може бути поділене між різними категоріями (якщо кількість записів із таким значенням перевищує встановлений мінімальний розмір категорії, то усі таки записи утворюють єдину категорію);
- з метою забезпечення систематичного впливу вхідного показника на результуючу змінну значення WOE мають бути підпорядковані деякому тренду (тобто, WOE повинні або поступово спадати, або зростати при переході від першої до останньої категорії).

Усі інші специфікації алгоритму (доцільність застосування у біннінгу показника інформаційної значимості IV, обмеження на максимальну кількість категорій чи мінімальний розмір категорії, мінімальну різницю WOE між сусідніми категоріями, доцільність об'єднання сусідніх категорій тощо) можуть варіюватись залежно від особливостей сформованої вибірки даних чи бачення аналітика.

Доцільність встановлення мінімального розміру категорії обумовлюється необхідністю врахування систематичних змін показника та нівелювання впливу окремих випадкових викидів чи помилок у даних на результати розрахунку кредитного ризику.

Проте це прописувати окремою вимогою до алгоритму сенсу не було, адже навіть без чіткого визначення мінімального розміру, всі категорії будуть охоплювати більш-менш тривалий діапазон значень змінної через необхідність вміщення обох класів кредитів та забезпечення тренду змін *WOE*. Тож, встановлювати мінімальний розмір категорії чи ні, залишається на розсуд окремого аналітика або програміста.

В алгоритмі поелементного формування категорій, розробленому нами відповідно до сформульованих вище вимог та методологічного підходу проведення категоризації кількісних змінних, було вирішено ввести обмеження на мінімальний розмір категорії (зрештою, користувач системи, в основі якої покладено цей алгоритм, за бажання може це обмеження встановити на нульовому рівні).

Процес утворення категорій доречно розпочати з поступового об'єднання значень показника, доки їх кількість не перевищить мінімально встановлений розмір категорії (при додатковому аналізі наявності у категорії кредитів з обох класів). Звісно, такий процес немає сенсу розпочинати із середини діапазону значень даного показника – його варто ініціювати від початку або з кінця. І вже поступово розширювати створені категорії та додавати нові, забезпечуючи при цьому дотримання тренду змін *WOE*.

Однак, якщо проводити категоризацію з якогось одного кінця діапазону значень показника, то напрям тренду зміни *WOE* визначити практично неможливо. Адже після першої категорії мінімально встановленої довжини можуть йти кілька категорій із поступовим зменшенням *WOE*, але загальний тренд виявиться зростаючим. І щоб коректно здійснити біннінг такої змінної, алгоритм доведеться постійно повертати до першої категорії, поступово розширюючи її та корегуючи множину значень другої категорії. І так до останнього елементу, рекурсивно повертаючись на початок. Причому в якийсь момент може виявитись, що після тривалого зростання тренд *WOE* таки пішов на спад. І алгоритму доведеться заново здійснювати перерозбивку змінної від першої категорії.

Відповідно, у створеному нами алгоритмі поелементного формування категорій було вирішено розпочати біннінг одночасно з обох кінців діапазону значень змінної. Частіше за все напрям тренду вдається визначити на етапі формування крайніх груп елементів (категорій) за перепадом значень їх *WOE*. Ці групи генеруються з дотриманням двох додаткових умов (у доповнення до встановлених вище щодо наявності у них представників обох

класів та неможливості поділу одного значення показника між різними категоріями):

- розмір групи має перевищувати мінімально допустимий розмір;
- розширення діапазонів категорій (додавання нових елементів вибірки до цих груп) відбувається доти, доки збільшується різниця між їх *WOE*.

Після утворення цих крайніх категорій розпочинається формування нових у напрямку до середини множини значень показника. Якщо новостворені категорії відповідають визначеному тренду, то вони фіксуються і процес біннінгу продовжується далі у напрямі до центру загального діапазону. Якщо ж якась із категорій, що додається до крайньої, йде у розріз із встановленим трендом (наприклад, загальний тренд зміни *WOE* визначений як зростаючий, але друга категорія отримала оцінку WOE нижче за першу), то алгоритм буде поелементно збільшувати крайню категорію та, відповідно, зсувати сусідню, доки вони не відповідатимуть заданому тренду (для вказаного прикладу при розширенні першої категорії в якийсь момент її *WOE* стане нижчим, ніж у другої категорії).

Звісно, може статись, що якась із крайніх категорій отримала значення *WOE*, яке не відповідає загальній тенденції. Це буде виявлено з розширенням крайніх категорій, щоб вирівняти загальний тренд. У такому випадку алгоритм сам змінить напрям тренду на протилежний (зростаючий на спадний чи навпаки).

Розглянемо приклад категоризації кількісної змінної «Дохід». Початковий масив даних містить 4455 спостережень. З них 1254 щодо ненадійних та 3201 – надійних позичальників. У цій вибірці крім числових значень містяться також спостереження, де інформацію за доходом позичальника вказано не було (NA).

На початковому етапі всі унікальні значення доходу об'єднуються у групи з розрахунком для них *WOE* та *IV* (якщо кількість елементів із якимось одним значенням доходу є замалою для визначення *WOE* та *IV*, то для такої групи на даному етапі за цими двома показниками будуть поставлені пропуски). Таких унікальних значень у вибірці виявилось 290. Фрагмент сформованої подібним чином бази наведено на рис. 1, де значення «good» та «bad» показують кількості «хороших» та «поганих» позичальників у відповідній групі.

Для подальшої роботи алгоритму виключимо NA із загального списку, оскільки ця група не може бути об'єднана з іншими. Мінімальний розмір категорії встановимо на рівні 1 % від вибірки,

тобто 45 спостережень. З кожною новою ітерацією роботи алгоритму до першої та останньої групи почергово додаються їхні сусіди, доки кількість елементів у цих крайніх групах не перевищить заданий мінімальний розмір категорії та не припиниться збільшення різниці між їхніми *WOE*.

```
      min max good bad         woe           iv
[1,]   NA  NA   14  20 -1.29379976 1.497616e-02
[2,]    6   6   15   2  1.07777820 3.331563e-03
[3,]    8   8   14   2  1.00878533 2.803149e-03
[4,]   16  16   16   4  0.44916954 8.123884e-04
[5,]   17  17   36  11  0.24849885 6.149242e-04
[6,]   19  19   20   3  0.95999517 3.701456e-03
[7,]   20  20   14   1  1.70193251 6.086426e-03
```

Рис. 1. Фрагмент бази унікальних значень для змінної «Дохід»

Джерело: обчислено автором у RStudio

Процес поетапного формування категорій зображено у табл. 1. Наприклад, з таблиці видно, що на другому етапі було сформовано по дві категорії зверху та знизу. Також можна побачити, що нові категорії додаються з урахування тренду у значеннях WOE (у даному випадку спадного). Кількість нерозділених груп у таблиці – це та кількість груп унікальних значень показника, яка залишилась після формування верхніх та нижніх категорій на даному етапі.

Таблиця 1

ЗНАЧЕННЯ *WOE* ДЛЯ СФОРМОВАНИХ НА КОЖНОМУ ЕТАПІ КАТЕГОРІЙ

№ групи	№ етапу								
	0	1	2	3	4	5	6	7	8
1	-	0,790	0,790	0,790	0,790	0,790	0,790	0,790	0,790
2	-	-	0,603	0,603	0,603	0,603	0,603	0,603	0,603
3	-	-	-	0,478	0,488*	0,495*	0,505*	0,518	0,520
4	-	-	-	-	-	-	0,603**	0,603**	-
5	-	-	-	-	-	0,327	0,327	0,327	0,327

Закінчення табл. 1

1	2	3	4	5	6	7	8	9	10
6	-	-	-	-	0,187	0,187	0,187	0,187	0,187
7	-	-	-	0,024	0,024	0,024	0,024	0,024	0,024
8	-	-	-0,119	-0,119	-0,119	-0,119	-0,119	-0,119	-0,119
9	-	-0,474	-0,474	-0,474	-0,474	-0,474	-0,474	-0,474	-0,474
К-ть нерозділених груп	290	251	226	132	76	38	14	0	0

Джерело: розраховано автором

Позначка «*» у табл. 1 вказує на те, що новоутворена на даному етапі група не відповідала тренду, тому була об'єднана з ближньою категорією зі свого краю діапазону значень показника. Категорія з позначкою «**» також не відповідала тренду і, оскільки межувала одночасно з двома вже сформованими категоріями, на завершальному етапі була об'єднана з одною із них за подібністю *WOE*. Графічна ілюстрація розподілу значень *WOE* для всіх категорій на сьомому етапі наведена на рис. 2.

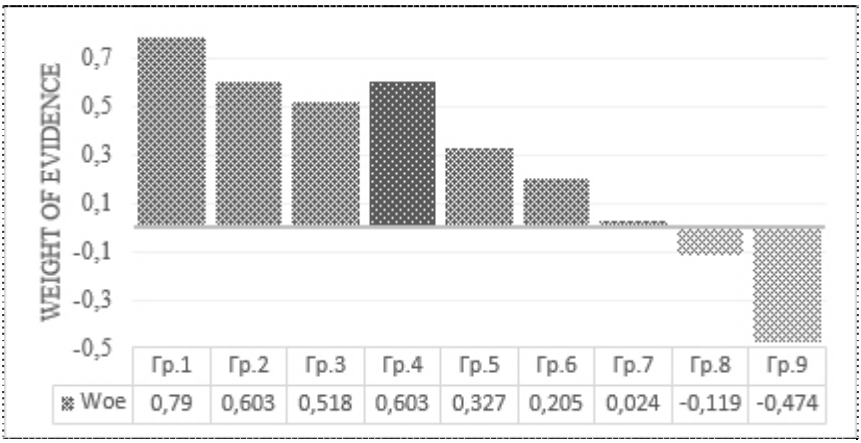

Рис. 2. Гістограма розподілу значень
WOE для всіх категорій до перевірки на відповідність тренду

Джерело: розраховано автором

Після перевірки на дотримання тренду четверта категорія була приєднана до третьої, у зв'язку з чим гістограма розподілу значень WOE у кінцевому результаті набула вигляду, представленого на рис. 3. Вплив четвертої категорії на загальну структуру розподілу «хороших» і «поганих» позичальників з третьої виявився незначним, адже після їх об'єднання показник *WOE* у третій категорії зріс всього на 0,002.

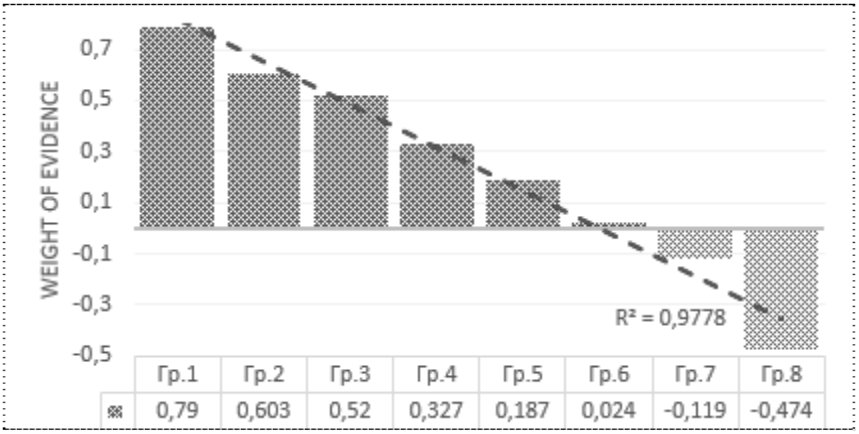

Рис. 3. Гістограма заключного розподілу значень *WOE* для всіх категорій кількісної змінної «Дохід»

Джерело: розраховано автором

Остаточні параметри всіх категорій, включно з групою із відсутніми даними (NA), наведені на рис. 4.

	min	max	good	bad	woe	iv
[1,]	NA	NA	14	20	-1.29379976	1.497616e-02
[2,]	0	16	45	8	0.79009613	6.066770e-03
[3,]	17	20	70	15	0.60332022	5.976758e-03
[4,]	21	152	1520	354	0.52004388	1.001370e-01
[5,]	153	172	92	26	0.32656722	2.614943e-03
[6,]	173	224	117	38	0.18746296	1.171277e-03
[7,]	225	248	34	13	0.02428635	6.189487e-06
[8,]	250	280	34	15	-0.11881450	1.592164e-04
[9,]	283	959	367	231	-0.47418068	3.298347e-02

Рис. 4. Характеристики всіх категорій змінної «Дохід», сформованих із дотриманням тренду в їх *WOE*

Джерело: розраховано автором

При розробці алгоритму біннінгу був досліджений також варіант попередньої обробки даних, коли здійснювалось об'єднання сусідніх груп, утворених із унікальних значень показника, для яких відсутні представники якогось одного класу (нульове значення поля «good» або поля «bad»). Такий підхід зменшує загальну кількість ітерацій, проте значення *WOE* за категоріями формуються менш рівномірно. Наприклад, коефіцієнт детермінації R^2 лінійного тренду для опису *WOE* зображених на рис. 3 категорій, які сформовані без попереднього об'єднання початкових груп із відсутніми представниками одного з класів, становить 0,978, а для варіанту з об'єднанням $R^2 = 0{,}968$.

Зауважимо, що на діапазони та загальну кількість категорій здійснює суттєвий влив обмеження щодо мінімального їх розміру. Гістограми для біннінгу з мінімальним розміром категорії 3 % від загального обсягу вибірки (щонайменше 134 спостереження у категорії для досліджуваного масиву даних) та 5 % (223 спостереження) мають вигляд, представлений на рис. 5 та 6, відповідно (на рисунках не відображено групу з NA).

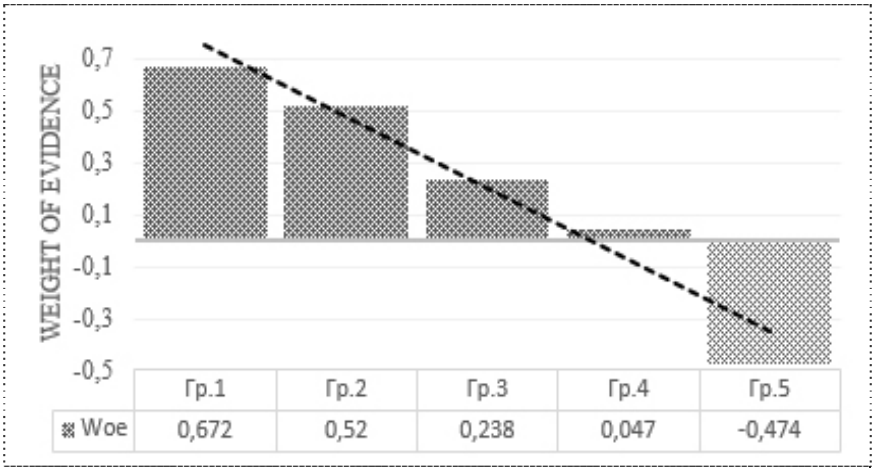

Рис. 5. Заключний вигляд категорій з урахуванням тренду для мінімального розміру групи 3 %

Джерело: розраховано автором

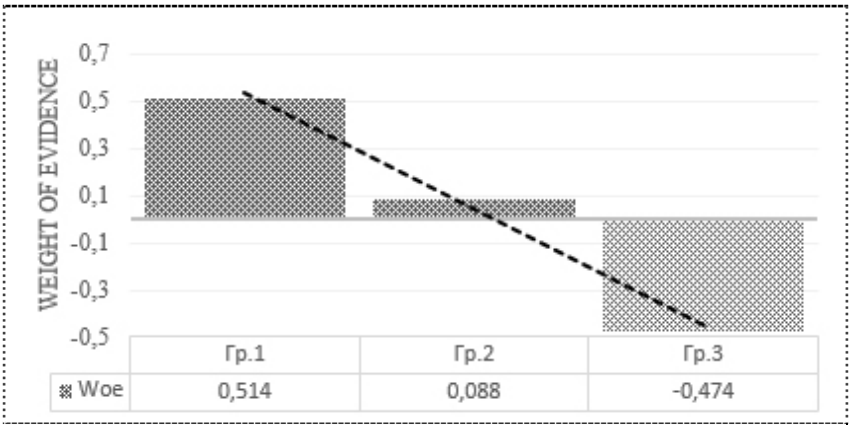

Рис. 6. Заключний вигляд категорій з урахуванням тренду для мінімального розміру групи 5 %

Джерело: розраховано автором

Для визначення оптимального розміру категорій необхідно проводити додаткові експериментальні дослідження із побудовою набору математичних моделей та порівнянням їх показників Джині або інших критеріїв адекватності скорингових моделей.

Висновки. Розроблений методологічний підхід до формування категорій кількісних змінних розширює прикладне застосування біннінгу як для задач кредитного скорингу, так і для інших задач бінарної класифікації. Створений алгоритм стане основою для підвищення точності розроблюваних математичних моделей, їх стійкості до випадкових збурень і похибок у даних, адже об'єднання дискретних значень змінних у категорії дозволяє виключити негативний вплив екстремальних викидів, замінюючи їх оцінками систематичного впливу категорії на результуючий показник.

Література

1. *Сорокин А. С.* Построение скоринговых карт с использованием модели логистической регрессии. [Электронный ресурс] / А.С. Сорокин // Интернет-журнал «Науковедение». – 2014. – Вып. 2. – С. 1–29. – Режим доступу: http://naukovedenie.ru/PDF/180EVN214.pdf.

2. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring / N. Siddiqi. – Hoboken : John Wiley & Sons, 2006. – 196 p.

3. *Herman J.* R Package 'smbinning': Optimal Binning for Scoring Modeling [Електронний ресурс] / J. Herman. — 2015, March 24. — Режим доступу: http://blog.revolutionanalytics.com/2015/03/r-package-smbinning-optimal-binning-for-scoring-modeling.html.

4. *Ковалев М.* Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц / М. Ковалев, В. Корженевская // Вестник Ассоциации белорусских банков. – 2007. – № 46. – С. 16–20.

5. *Коляда Ю. В., Бондар В. А.* Біннінг у нейромережевих скорингових моделях // Нейро-нечіткі технології моделювання в економіці.— 2016.— № 5.— С. 6–80.

6. Building Powerful, Predictive Scorecards: An overview of Scorecard module for FICO Model Builder // Fair Isaac Corporation. – 2014. – March. – 46 p. [Електронний ресурс]. – Режим доступу: http://www.fico.com/en/wp-content/secure_upload/Building_Powerful_Predictive_Scorecards_1991WP.pdf.

7. *Палкин Н.Б.* Оптимальное квантование для повышения качества бинарных классификаторов / Н.Б. Палкин, В.В. Афанасьеф // Штучний інтелект. — 2013. — № 4. — С. 392–399.

References

1. Sorokin, A. S. (2014). Postroyeniye skoringovykh kart s ispolzovaniyem modeli logisticheskoy regressii. *Naukovedeniye (Science of Science), 2*, 1-29 [in Russian].

2. Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. New Jersey, USA: John Wiley and Sons.

3. Herman, J. (2015, March 24). *Optimal Binning for Scoring Modeling*. Retrieved from http://blog.revolutionanalytics.com/2015/03/r-packagesmbinning-optimal-binning-for-scoring-modeling.html.

4. Kovalev, M., & Korzhenevskaya, V. (2007). Metodika postroyeniya bankovskoy skoringovoy modeli dlya otsenki kreditosposobnosti fizicheskikh lits. *Vestnik Assotsiatsii belorusskikh bankov (Bulletin of the Belarusian Banks Association), 46*, 16-20 [in Russian].

5. Kolyada, Y. V., & Bondar, V. A. (2016). Binninh u neyromerezhevykh skorynhovykh modelyakh. *Neyro-nechitki tekhnolohiyi modelyuvannya v ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics), 5*, 60-80 [in Ukrainian].

6. Fair Isaac Corporation. (2014, March). *Building Powerful, Predictive Scorecards: An overview of Scorecard module for FICO Model Builder*. Retrieved from http://www.fico.com/en/wp-content/secure_upload/Building_Powerful_Predictive_Scorecards_1991WP.pdf.

7. Palkin, N.B., & Afanasiev, V. V. (2013). Optimal'noye kvantovaniye dlya povysheniya kachestva binarnykh klassifikatorov. *Shtuchnyy Intelekt (Artificial Intelligence), 4*, 392–399 [in Russian].