УДК 004.622:519.86

БІННІНГ У НЕЙРОМЕРЕЖЕВИХ СКОРИНГОВИХ МОДЕЛЯХ

Ю. В. Коляда

Кандидат фізико-математичних наук, доцент,
доцент кафедри економіко-математичного моделювання

Державний вищий навчальний заклад «Київський національний економічний університет імені Вадима Гетьмана»
проспект Перемоги, 54/1, м. Київ, 03680, Україна
jukol48@ukr.net

В. А. Бондар

Бакалавр з економічної кібернетики

Державний вищий навчальний заклад «Київський національний економічний університет імені Вадима Гетьмана»
проспект Перемоги, 54/1, м. Київ, 03680, Україна
bondaravn@gmail.com

Стаття присвячена розробці методологічного підходу до категоризації вхідних показників економіко-математичних моделей оцінювання кредитоспроможності позичальників комерційних банків. Основою математичного інструментарію обрано нейронну мережу типу багатошаровий персептрон. Об'єктом дослідження є процес категоризації пояснюючих змінних скорингових моделей. Предметом дослідження є сукупність методів категоризації та способів оцінки їх впливу на точність моделі оцінювання ймовірності невиконання умов договору позичальником. У результаті проведених експериментальних досліджень у рамках запропонованого методологічного підходу було обґрунтовано здійснювати оптимізацію розбиття на категорії вхідних змінних моделі за рахунок максимізації значення коефіцієнта Джині як показника адекватності скорингових моделей. Було отримано висновок, що зниження показника інформаційної значущості не завжди виступає індикатором погіршення якості класифікатора. У статті також був розширений список рекомендацій щодо проведення біннінгу, який може бути використаний для побудови більш точних моделей оцінювання кредитоспроможності позичальників комерційних банків.

Ключові слова: *вагомість ознаки (WOE), інформаційна значущість (IV), біннінг, скорингова модель, нейронна мережа.*

БИННИНГ В НЕЙРОСЕТЕВЫХ СКОРИНГОВЫХ МОДЕЛЯХ

Ю. В. Коляда

Кандидат физико-математических наук, доцент,
доцент кафедры экономико-математического моделирования

Государственное высшее учебное заведение «Киевский национальный экономический университет имени Вадима Гетьмана»
проспект Победы, 54/1, г. Киев, 03680, Украина
jukol48@ukr.net

В. А. Бондарь

Бакалавр по экономической кибернетике

Государственное высшее учебное заведение «Киевский национальный экономический университет имени Вадима Гетьмана»
проспект Победы, 54/1, г. Киев, 03680, Украина
bondaravn@gmail.com

Статья посвящена разработке методологического подхода категоризации входных переменных экономико-математических моделей оценки кредитоспособности заемщиков коммерческих банков. Основой математического инструментария выбрана нейронная сеть типа многослойный персептрон. Объектом исследования является процесс категоризации влияющих факторов скоринговых моделей. Предметом исследования является совокупность методов категоризации и способов оценки их влияния на точность модели расчета вероятности невыполнения условий договора заемщиком. В результате проведенных экспериментальных исследований в рамках предложенного методологического подхода было обосновано проведение оптимизации разбиения на категории входных переменных модели за счет максимизации значения коэффициента Джини как показателя адекватности скоринговых моделей. Было получено заключение, что снижение показателя информационной значимости не всегда выступает индикатором ухудшения точности классификатора. В статье также был расширен список рекомендаций по проведению биннинга, который может быть использован для построения более эффективных моделей оценки кредитоспособности заемщиков коммерческих банков.

Ключевые слова: *весомость признака (WOE), информационная значимость (IV), биннинг, скоринговая модель, нейронная сеть.*

BINNING IN NEURAL NETWORK SCORING MODELS

Yuriy Kolada

PhD (Physical and Mathematical Sciences), Docent,
Associate Professor of Department of Economic and Mathematical Modeling

State Higher Educational Establishment
«Kyiv National Economic University named after Vadym Hetman»
54/1 Peremogy Avenue, Kyiv, 03680, Ukraine
jukol48@ukr.net

Volodymyr Bondar

Bachelor's Degree in Economic Cybernetics,
Master student, Department of Economic and Mathematical Modeling

State Higher Educational Establishment
«Kyiv National Economic University named after Vadym Hetman»
54/1 Peremogy Avenue, Kyiv, 03680, Ukraine
bondaravn@gmail.com

The article is devoted to developing methodological approach of categorizing the input variables of economic and mathematical models of evaluating of creditworthiness of commercial banks' borrowers. The basis of mathematical tools was chosen neural network of type of multilayer perceptron. The object of study is the process of categorizing explanatory variables of scoring models. The subject of study is a set of methods of categorizing and approaches to evaluate their impact on the efficiency of the model of estimation of probability of borrower's default under the contract. As a result of experimental studies within the confines of the proposed methodological approach it was decided to optimize the input variables partitioning into categories by maximizing the Gini coefficient as a measure of the adequacy of scoring models. A conclusion that a drop in the information value does not always act as an indicator of the deterioration of the classifier was obtained. In this article it has been expanded a list of recommendations for binning that may be used to build more accurate models of evaluating the creditworthiness of borrowers of commercial banks.

Keywords: *weight of evidence (WOE), information value (IV), binning, scoring model, neural network.*

JEL Classification: C13, C45, C52, C99, G21

Постановка проблеми

Проведення адекватного оцінювання кредитоспроможності клієнтів банківських організацій є актуальною задачею як для вітчизняних, так і закордонних фінансових структур. Така потреба супроводжується активним пошуком закономірностей підвищення ефективності скорингових моделей, які на основі попередніх спостережень щодо особливостей позичальника прогнозують імовірність його дефолту за зобов'язаннями.

Проблемною ділянкою побудови моделі оцінки кредитоспроможності позичальника є процедура поділу діапазону кожної кількісної пояснюючої змінної на категорії за проявом впливу на надійність угод. Ця процедура дозволяє підвищити робастність моделі (її стійкість до викидів і похибок у даних) та одночасно її адекватність, адже об'єднання дискретних значень змінних у категорії дозволяє виключити негативний вплив екстремальних числових значень (випадковість, неточна розмірність, пропущені дані), замінюючи їх оцінками систематичного впливу категорії на вихідні результати. Процес категоризації вхідних змінних (або перетворення кількісних змінних у категорії) у скорингу ще називається біннінгом (англ. binning) [1].

Застосування у скорингу логістичних регресій або інших нелінійних моделей ускладнює аналіз впливу біннінгу на адекватність класифікації позичальників і, відповідно, потребує спеціального дослідження з розробки алгоритму визначення оптимальної кількості та діапазонів кожної з категорій кількісних вхідних змінних.

Аналіз останніх джерел і публікацій

Загальноприйнятим правилом у літературі є те, що кожна категорія має об'єднувати значення показника з однаковими властивостями 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 значущості змінної та

показника вагомості ознаки. Під час поділу змінних на категорії А. С. Сорокін керується:

- максимізацією показника інформаційної значущості як критерію оптимальності біннінгу;
- об'єднання категорій з близьким значенням вагомості ознаки для посилення тенденції їх зростання або спадання;
- обмеженням максимальної кількості категорій до 50;
- потребою в забезпеченні суттєвої відмінності у різних групах часток ненадійних позичальників та *WOE*;
- необхідністю розбиття значень показників на категорії, які б забезпечували зростаючий або спадаючий тренд *WOE* при переході від однієї категорії до іншої.

Доцільність забезпечення суттєвої різниці *WOE* при переході від однієї категорії до іншої також було доведено у джерелі [2], де Дж. Херманом проаналізовано три способи біннінгу:

- формування меж категорій як відрізків з однаковою довжиною на діапазоні вхідних даних;
- поділ на категорії з однаковою кількістю прикладів;
- посилення різниці значень *WOE* між сусідніми категоріями.

Компанія FICO [3] зауважує, що встановити вичерпну процедуру оптимального біннінгу неможливо, адже це є питанням «мистецтва та науки», але при цьому надає ряд рекомендацій для проведення ефективного розподілу змінної на категорії:

- кожна категорія має вміщувати достатньо значень, аби погасити вплив екстремальних показників та шуму в вибірці;
- кожна категорія має вміщувати лише значення, спільні за мірою впливу на результуючу змінну;
- абсолютні значення інформаційної значущості змінної несуть мінімальне смислове навантаження і мають використовуватися лише для порівняння.

У роботі [4] Н. Сіддікі пропонує класичні рекомендації щодо проведення біннінгу:

- пропущені значення показника мають входити в окрему категорію;
- кожна категорія не може містити менше, ніж 5 % вибірки;
- кількість надійних чи ненадійних угод у категорії не мають дорівнювати 0.

У роботі Н. Б. Палкіна [5] була досліджена проблема оптимального квантування для підвищення точності бінарних класифікаторів (операція укрупнення вже раніше утворених категорій).

У процесі проведення експерименту ним було підвищено точність класифікатора при значній втраті інформаційної значущості змінних. Такий результат обумовлює потребу в перевірці ролі показника інформаційної значущості як критерію класифікації пояснюючих змінних.

А. В. Матвійчук у монографії [6] висвітлив процес категоризації шляхом переведення кількісних змінних у нечіткі множини. Проте представлений інструментарій нечіткої логіки потребує застосування надто малої кількості нечітких множин для кожної змінної та не передбачає встановлення вагомості кожної окремої змінної чи її нечітких множин.

Попри те, що у публікаціях [1—5] алгоритм біннінгу був детально описаний і прокоментований, принципи розбиття кількісних змінних на категорії у цих роботах є надто відмінними між собою. Так, різняться рекомендації щодо: розмірів та кількості категорій, правил їх об'єднання, доцільності застосування показника інформаційної значущості тощо. Тому виникає потреба у перевірці адекватності розроблених методів біннінгу, їх доповненні та розвитку, пошуку нових індикаторів якості проведення цього процесу, перевірки правил біннінгу та перегляду ролі показника інформаційної значущості змінних. Це зумовлює актуальність даного дослідження та доцільність розробки методологічного підходу до категоризації пояснюючих змінних економіко-математичних моделей аналізу кредитоспроможності позичальника з метою підвищення їх точності та адекватності.

Мета і завдання дослідження

Метою дослідження є розробка методологічного підходу щодо проведення ефективної категоризації кількісних змінних у процесі побудови скорингових моделей.

Досягнення поставленої мети потребує вирішення таких завдань: сформулювати гіпотези щодо оптимального поділу діапазону значень кількісних змінних на основі узагальнення світового досвіду з проведення біннінгу; алгоритмізувати процеси поділу значень кількісних змінних на категорії відповідно до висунутих гіпотез, а також побудови скорингових моделей для різних варіантів біннінгу вхідних даних; систематизувати результати експериментальних досліджень із обґрунтуванням відповідних висновків і рекомендацій.

Виклад основного матеріалу

Як уже зазначалось, процедура біннінгу застосовується в процесі розробки скорингових моделей і реалізує трансформацію всього різноманіття значень пояснюючої кількісної змінної до обмеженої кількості категорій, узагальнюючи вплив вхідних даних на результат моделі. Крім цього даний процес дозволяє [2]:

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;
3) привести всі вхідні змінні до однієї розмірності.

У процесі переведення кількісних значень показників у категоріальну форму в сучасному скорингу застосовується показник вагомості ознаки *WOE* (weight of evidence), який розраховується для кожної категорії окремо:

$$WOE_i = \ln\left(\frac{d_i(1)}{d_i(2)}\right), \quad i = \overline{1,k}, \tag{1}$$

де $d_i(1)$ і $d_i(2)$ — відносні частоти ненадійних і надійних угод у вибірці, які відповідають *i*-й категорії, $i = \overline{1,k}$; k — число категорій змінної [1, с. 13].

Показник (1) дозволяє охарактеризувати міру впливу всіх значень змінної з певної категорії на ймовірність дефолту за зобов'язаннями.

Іншим важливим показником, що вказує на величину ефекту від біннінгу для прогнозування бінарної залежної змінної, є інформаційна значущість *IV* (information value) [1, с. 17]:

$$IV = \sum_{i=1}^{k} (d_i(1) - d_i(2)) \cdot WOE_i \,. \tag{2}$$

Причому

$$\sum_{i=e}^{g} (d_i(1) - d_i(2)) \cdot WOE_i \neq (d_j(1) - d_j(2)) \cdot WOE_j \,, \tag{3}$$

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

Вираз (3) вказує, що сума показника *IV* для кількох сусідніх категорій не дорівнює значенню даного показника для нової категорії після об'єднань цих сусідів. Нерівність (3) призводить до того, що пошук оптимальної кількості категорій для змінної не може бути задачею лінійного програмування, а, навпаки, вимагає комбінаторного перебору всіх можливих варіантів. Поставлена задача може бути розв'язана за допомогою порівняння результатів різних варіантів категоризації на основі проведення експерименту.

Оскільки не всі літературні джерела одностайні щодо ролі інформаційної значущості як критерію оптимальності категоризації вхідних даних, у даній роботі доцільно провести оцінювання якості біннінгу одночасно на основі виразу (2) та показника якості моделі в цілому. Існує багато способів оцінки якості моделі, однак, враховуючи бінарну форму вихідної змінної, застосуємо загальновживаний показник адекватності скорингових моделей — коефіцієнт Джині, за яким у тому числі зможемо оцінити ефективність біннінгу.

Значення Джині може бути розраховане як площа під кривою ROC (Receiver Operating Characteristic) на площині з абсцисою «1 — *Specificity*» та ординатою «*Sensitivity*», як представлено на рис. 1.

$$Sensitivity = \frac{TP}{TP + FN}, \tag{4}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{5}$$

де *TP* — істинно позитивні випадки, *TN* — істинно негативні випадки, *FN* — помилка першого роду, *FP* — помилка другого роду [4, с. 49].

Що є позитивним випадком, а що негативним, залежить від конкретного завдання. Коли ми прогнозуємо ймовірність дефолту позичальника, то позитивним результатом буде клас «Позичальник ненадійний», негативним — «Позичальник надійний». Якби ми визначали ймовірність того, що позичальник надійний, то позитивним результатом був би клас «Позичальник надійний», і т. д.

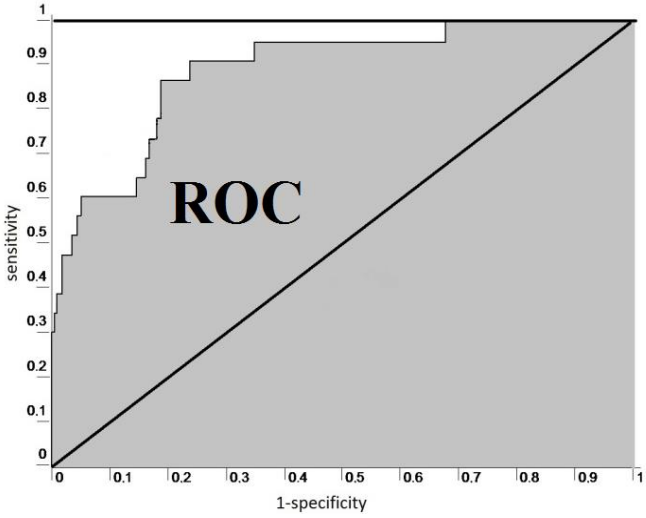

Рис. 1. Типовий вигляд ROC-кривої залежно від чутливості та специфічності

Коефіцієнт Джині розраховується за такою формулою:

$$Gini = \sum_{l=0}^{N-1} \left(Sensitivity(l+1) + Sensitivity(l)\right) \times \\ \times \left(Specificity(l) - Specificity(l+1)\right) - 1, \tag{6}$$

де N — кількість точок для побудови ROC-кривої з кроком дискретизації $\Delta = 1/N$ (у нашому дослідженні було прийнято, що $N = 20$).

За $Gini = 0$ класифікація здійснюється випадковим чином, а при $Gini = 1$ класифікатор працює ідеально (існує такий поріг відсікання, за якого всі позичальники з розрахунком скорингової моделі, вищим за цей поріг, є насправді дефолтними, а всі позичальники, для яких розрахований моделлю рейтинг менше порогу, є надійними). Значення даного коефіцієнта є зручними для порівняння різних моделей між собою: що вище значення *Gini* для моделі, то точнішою вона є [3, с. 43]. Користуючись даною особливістю, результати класифікації на основі скорингових моделей при різних варіаціях біннінгу будуть порівняні шляхом розрахунку даного коефіцієнта.

З метою дослідження впливу біннінгу на якість класифікатора доцільно сформувати методологічний підхід до категоризації вхідних даних 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 з однаковою кількістю прикладів;

3) проведення об'єднання попередньо встановлених категорій різними способами;

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

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

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

7) аналіз отриманих результатів, формулювання висновків про особливості біннінгу;

8) розробка алгоритму автоматизації процесу категоризації вхідних змінних та відповідної побудови скорингових моделей.

У джерелах [1] та [5] скорингова модель представлена логістичною регресією, оскільки це найпоширеніший у скорингу математичний апарат, реалізований у багатьох програмних засобах. Однак у даній роботі перевага була надана багатошаровому персептрону Розенблатта. Даний вибір зумовлений висновками досліджень у роботах [6 — 8] про перевагу штучних нейронних мереж типу персептрон при побудові бінарних класифікаційних моделей, що помітно переважали за точністю інші аналоги, зокрема, моделі на основі логістичної регресії.

Відповідно, п'ятий етап вирішено реалізувати за рахунок побудови багатошарового персептрону, що являє собою нейронну мережу з прямими зв'язками, яка складається з кількох шарів: вхідного, прихованих (одного або кількох) та вихідного. Кожен шар містить певну кількість нейронів, а міжелементні зв'язки можливі між нейронами сусідніх шарів. Приміром, у персептроні з одним прихованим шаром нейрон вхідного шару має зв'язки лише з нейронами прихованого шару, а будь-який нейрон прихованого шару має одночасно зв'язки як і з нейронами входу, так і з нейронами виходу.

Принцип дії кожного нейрону такої мережі був запозичений нейроанатомом Мак-Каллоком і математиком Піттсом з принципів функціонування його біологічного аналога. Штучні нейронні мережі, за аналогією з природними нервовими системами, складаються з нейронів, які поєднуються між собою міжнейронними зв'язками. Міжнейронний зв'язок, який є аналогом синапсів у природних нейронах, здійснює добуток сигналу, що йде до нейрона, на ваговий коефіцієнт, який характеризує силу зв'язку.

Структурно штучний нейрон складається із суматора та функціонального перетворювача. Суматор здійснює додавання зважених сигналів, які надходять по міжнейронних зв'язках від інших нейронів, або зовнішніх вхідних сигналів. Функціональний перетворювач здійснює трансформацію виходу суматора за характеристичною функцією заданого виду [6, с. 44]. Існує багато різних функцій активації. Для побудови персептрону в нашому дослідженні скористаємось логістичною сигмоїдною функцією активації:

$$\psi(s) = \frac{1}{1 + \exp(-rs)}, \qquad (7)$$

де $\psi(s)$ — функція активації; s — значення розрахунку суматора нейрона; r — коефіцієнт стиснення-розтягування функції вздовж осі абсцис.

Таким чином, кожен нейрон мережі перетворює сигнал, який подається на входи, відповідно до внутрішніх конфігурацій. А нейронна мережа, що складається з трьох шарів таких нейронів, може відтворити із заданою точністю будь-яку неперервну функцію з багатьма змінними, якщо кількість нейронів у прихованому шарі є достатньою [6, с. 53].

Процедура побудови персептрону, як математичної основи скорингової моделі, може бути описана таким чином. Перший шар утворюється з такої кількості нейронів, що відповідає кількості пояснюючих змінних моделі. Сигнали з вхідного шару надходять до прихованого шару і після обробки передаються до вихідного. Згідно рекомендацій [6] та [7], кількість нейронів у прихованому шарі має встановлюватись емпірично з метою уникнення явища «перенавчання» нейронної мережі. Оскільки завданням цього дослідження є не конструювання найадекватнішої нейронної мережі, а встановлення принципів ефективного біннінгу, то для проведення моделювання вирішено було обмежитись побудовою персептрону найпростішої структури — до єдиного

прихованого шару якого входить 2 нейрони. У результаті проведення розрахунків на нейроні вихідного шару отримується число в інтервалі $[0,1]$, що вказує на ймовірність дефолту за кредитом, інформацію щодо якого та його позичальника було подано на входи нейромережі. Налаштування параметрів персептрону здійснюється із застосуванням алгоритму зворотного поширення помилки.

Оскільки метод зворотного поширення помилки полягає у знаходженні найближчого локального мінімуму на гіперплощині помилок, то в результаті його застосування для однакових навчальних вибірок даних будуть знаходитись різні параметри персептрону, адже локальних мінімумів на гіперплощині помилок є безліч. Тому для збільшення адекватності моделі в дослідженні було вирішено для кожної нейромережі та обраної процедури категоризації процес налаштування повторювати десять разів (так вдається відібрати модель з найменшою похибкою).

Інформаційною базою для досліджень слугувала вибірка з 7807 кредитних угод. З них 2033 записів увійшли до навчальної вибірки, а до тестової — 5774, які були використані для оцінки ефективності моделі на даних, на яких модель не навчалась. Записи про кожну кредитну угоду містять такі характеристики позичальника на момент її укладання, як його стать і вік, посада, сімейний стан, освіта, заробітна плата, кількість відкритих угод, загальна сума, на яку укладено попередні кредитні угоди, загальна кількість попередніх угод тощо. Результуюча бінарна змінна за кожною угодою набуває значення 1 за умови настання дефолту за кредитним зобов'язанням та 0, якщо кредит був закритий згідно умов договору.

Процедура біннінгу проводиться для вхідних змінних моделі на навчальній вибірці, а потім межі кожної категорії використовуються для заміни поточних значень змінної як у навчальній, так і тестовій вибірці на відповідні їм WOE_i. Процедура поділу діапазону значень змінної на k категорій може проходити таким чином: кожна категорія, за винятком першої та останньої, вміщує всі елементи змінної у межах $(x_{i\cdot\Delta-\Delta}, x_{i\cdot\Delta}]$ для $i=\overline{2,k-1}$ з кроком $\Delta=1/k$, де $x_{i\cdot\Delta}$ — це i-ий квантиль або квантиль рівня $i\cdot\Delta$ (при $\Delta=1/100$ квантиль називається перцентилем, при $\Delta=1/10$ — децилем, $\Delta=1/4$ — квартилем, $\Delta=1/2$ — медіаною). При $i=1$ категорія має межі $(-\infty, x_{\Delta}]$, а при $i=k$ категорія визначається в ме-

жах $\left(x_{(k-1)\Delta},+\infty\right)$. Даний спосіб дозволяє формувати k категорій з практично однаковою кількістю елементів у кожній, залежно від типу вибірки. Зазначимо, що тут дійсні границі аналізованої вибірки змінені на $\pm\infty$ не стільки для розбиття повної множини її елементів на категорії, скільки для послідуючого переведення значень тестової вибірки у категорії (адже вони можуть виходити за межі навчальної вибірки). Варто зауважити, що якщо в базі даних у групи позичальників пропущені значення аналізованої змінної, то під них створюється окрема $k+1$ категорія.

Для наочного прикладу категоризації виконаємо описану вище процедуру для $k = 10$ на діапазоні значень пояснюючої змінної, що відповідає загальній сумі попередніх кредитних угод позичальника. Після об'єднання пропущених значень в окрему категорію NULL та отримання меж 10 категорій для записів, що залишились, проводиться підрахунок суми значень результуючої бінарної змінної для прикладів із кожної категорії. Відношення кількості дефолтних кредитів у i-ій категорії до загальної кількості дефолтів у навчальній вибірці називається відносною частотою ненадійних угод $d_i(1)$ згідно формули (1). Аналогічним чином знаходиться частка надійних угод $d_i(2)$, $i=\overline{1,k+1}$, як відношення кількості погашених кредитів у i-ій категорії до загальної кількості кредитів у навчальній вибірці, за якими позичальниками було виконано зобов'язання. Однак існують такі випадки (особливо за великих значень k), коли $d_i(1)$ або $d_i(2)$ i-ї категорії дорівнюють нулю. У такому разі доцільно об'єднати дану категорію з одною із сусідніх.

За співвідношенням (1) підраховується $k + 1$ окремих значень WOE_i, $i=\overline{1,k+1}$, та відбувається заміна кожного значення змінної відповідними WOE_i. Окрім цього для груп $i=\overline{1,k+1}$ розраховується окремий показник IV_i як доданок виразу (2). Отриманий результат розрахунків для показника подано у табл. 1.

Зазначимо, що якщо категорії характеризуються близькими значеннями *WOE*, це свідчить про те, що в них розподіл дефолтних і погашених кредитів практично однаковий і, відповідно, жодна з цих категорій окремо не несе додаткового смислового навантаження з огляду на класифікаційну здатність моделі. Тому доцільним є об'єднання таких категорій з метою посилення різниці

значень вагомості ознаки (1) між сусідніми категоріями. Варто зауважити, що 11 категорія не є сусідньою до жодної іншої. Тому далі не беремо її до уваги під час виконання процедури об'єднання сусідніх категорій.

Таблиця 1

ПОКАЗНИКИ WOE ТА IV НА ОСНОВІ КВАНТИЛІВ

Категорія №	Права межа категорії, грн.	Кількість значень у категорії	Кількість надійних	Кількість ненадійних	$d(1)$	$d(2)$	WOE_i	IV_i
1	2667,9	148	53	95	0,057	0,086	–0,417	0,012
2	5106,7	147	62	85	0,067	0,077	–0,149	0,002
3	8348,1	146	54	92	0,058	0,084	–0,366	0,009
4	12 521,2	149	62	87	0,067	0,079	–0,172	0,002
5	18 072,0	147	69	78	0,074	0,071	0,044	0,001
6	26 768,8	148	75	73	0,080	0,066	0,194	0,003
7	38 554,0	149	86	63	0,092	0,057	0,478	0,017
8	56 422,8	147	96	51	0,103	0,046	0,799	0,045
9	95 929,0	147	107	40	0,115	0,036	1,151	0,090
10	1 303 000,0	147	115	32	0,123	0,029	1,446	0,136
11	NULL	558	153	405	0,164	0,368	–0,807	0,164
Сума		2033	932	1101	1	1		0,481

Звівши значення WOE_i, $i=\overline{1,10}$, з табл. 1 на рис. 2, можна помітити, що зі збільшенням номеру категорії відбувається зростання значення вагомості ознаки.

У літературі [1, 2] зустрічаються рекомендації щодо реалізації процедури об'єднання сусідніх категорій з близькими значеннями *WOE* за рахунок посилення лінійної залежності вагомості ознаки (1) від номера категорії *i*. З рис. 2 помітно, що перші 5 категорій мають недостатньо виражену тенденцію. За допомогою об'єднання перших трьох, а також 4 і 5 категорій, було отримано біннінг змінної з $k = 7$. При цьому коефіцієнт детермінації апроксимуючої прямої збільшено з 0,916 для 10 категорій з табл. 1 до 0,996 для об'єднаних 7 категорій, що свідчить про посилення лі-

нійної залежності *WOE* від номера категорії *i*. Результат об'єднання зображено на рис. 3.

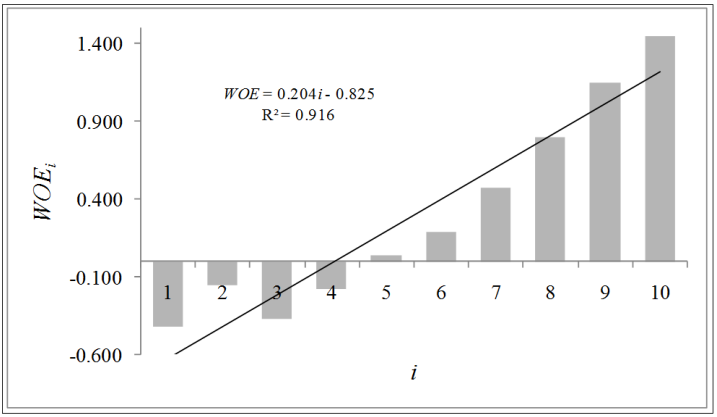

Рис. 2. Значення WOE_i, $i = \overline{1,10}$

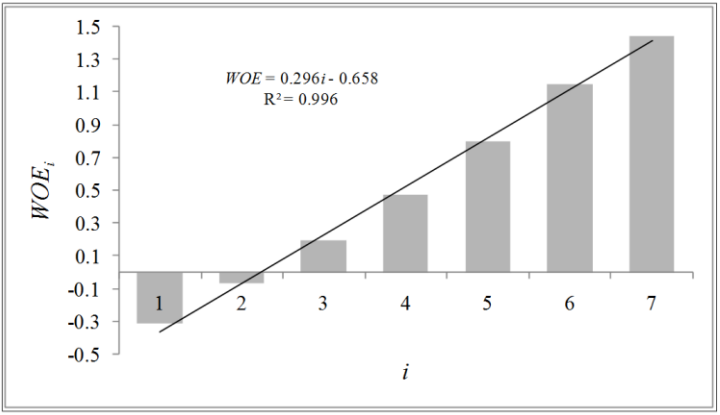

Рис. 3. Значення WOE_i, $i = \overline{1,7}$, після проведення об'єднання категорій з близькими значеннями вагомості ознаки

Показники ефективності скорингових моделей за різних варіантів біннінгу (в тому числі й після об'єднання категорій) можна побачити в табл. 2, де укрупнення категорій при зменшенні показника інформаційної значущості привело до зростання коефіцієнта Джині (6).

Таблиця 2

ЗНАЧЕННЯ КОЕФІЦІЄНТУ ДЖИНІ ТА ІНФОРМАЦІЙНОЇ ЗНАЧУЩОСТІ ЗА РІЗНИХ ВАРІАНТІВ РЕАЛІЗАЦІЇ БІННІНГУ

Кількість категорій, k+1*	8	11	51	101	151	201	301	Без біннінгу
Коефіцієнт Джині, *Gini*	0,729	0,723	0,725	0,727	0,730	0,734	0,718	0,702
Інформаційна значущість, *IV*	0,477	0,481	0,595	0,702	0,776	0,896	1,03	—

*під «+1» мається на увазі категорія зі значеннями NULL

Однак, може так бути, що кожна з перших п’яти категорій на рис. 2 характеризується надто усередненими значеннями елементів, з яких вона складається. Тому і між собою ці категорії можуть мати близькі характеристики. Але при цьому в кожній категорії можуть існувати свої закономірності розподілу дефолтних і надійних позичальників. Тому доцільно перевірити адекватність моделі і при розбитті повної множини значень змінної на більшу кількість груп. Відповідно, у ході проведення експерименту аналогічним чином знайдено межі категорій при k = 50, 100, 150, 200, 300. Для всіх варіантів категоризації були побудовані скорингові моделі та було проведено оцінку їх точності на тестовій вибірці за допомогою коефіцієнту Джині (6). Результати розрахунку коефіцієнта Джині та показника інформаційної значущості при всіх відібраних k представлено також у табл. 2.

З табл. 2 видно, що в результаті об’єднання категорій до восьми було одержано інформаційну значущість змінної IV = 0,477, що є найменшим значенням з-поміж інших варіантів категоризації. Попри зменшення інформаційної значущості, дане укрупнення категорій призвело до зростання точності класифікатора. Також значення інформаційної значущості IV за різних варіантів розбиття на категорії (при всіх відібраних варіантах кількості груп k) графічно зведено на рис. 4.

З рис. 4 видно, що збільшення кількості категорій позитивно впливає на загальний показник IV для всієї вибірки. Досягнення максимального показника IV може бути і за умови найбільшого

можливого значення *k*, при якому кількість надійних і ненадійних угод у кожній категорії не дорівнює нулю. Однак зі збільшенням кількості категорій, деякі з них стають замалими для відображення якоїсь значущої інформації про позичальників, віднесених до таких категорій. Тому інформаційну значущість *IV* можна розглядати як проміжний показник ефективності біннінгу, тоді як основним критерієм процесу попередньої обробки даних і побудови скорингової моделі є якість класифікатора.

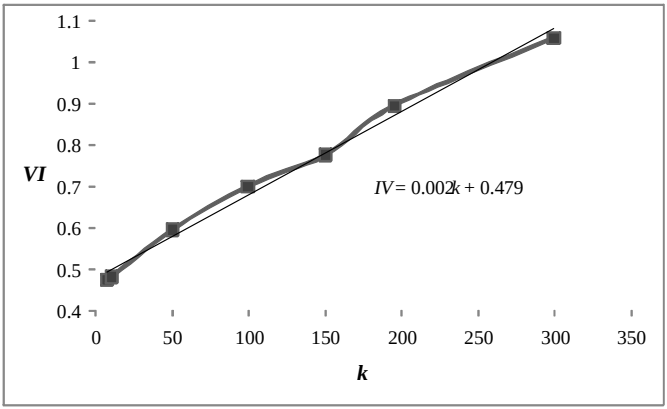

Рис. 4. Залежність величини показника IV
від кількості категорій змінної

У результаті проведених експериментів з побудови скорингових моделей із різними варіантами реалізації біннінгу було отримано ряд висновків, які уточнюють або заперечують ті чи інші рекомендації відомих учених чи дослідницьких груп в області скорингу:

1) показник інформаційної значущості не впливає на точність моделі, а, отже, не повинен слугувати критерієм ефективності біннінгу (найвища точність класифікатора була досягнута за $k = 200$ (*Gini* = 0,734 та *IV* = 0,896), тоді як найвище значення інформаційної значущості (*Gini* = 0,718 та *IV* = 1,03) відповідало варіанту поділу $k = 300$);

2) кожна категорія може містити менше 5 % вибірки, оскільки за $k = 200$ було досягнута найвища точність моделі (коли в середньому кожна категорія містила близько 10 елементів, або 1 % від вибірки);

3) доцільно здійснювати об'єднання сусідніх категорій з близькими значеннями WOE_i для посилення лінійної залежності вагомості ознаки (1) від номеру категорії, адже таким чином вдається підвищити точність класифікації моделі (збільшити коефіцієнт Джині), незважаючи на зменшення інформаційної значущості *IV*.

З урахуванням даних висновків сформовано алгоритм побудови скорингової моделі з проведенням ефективної категоризації кількісної змінної, який полягає у виконанні таких кроків:

1) збір інформації та формування начальної та тестової вибірок;

2) за допомогою квантилів проводиться поділ пояснюючої змінної на k категорій (починаючи з невеликої кількості, наприклад, $k=10$, поступово збільшуючи кількість категорій залежно від особливостей вибірки), а за наявності пропущених елементів — їх об'єднання в окрему $k+1$ категорію;

3) розрахунок значень WOE для кожної категорії;

4) побудова лінії тренду для k категорій і знаходження пари сусідніх категорій, які є найбільш віддаленими від тренду (для яких вираз $\left(\left|WOE_g - ag - b\right| + \left|WOE_{g+1} - a(g+1) - b\right|\right)$ набуває найбільшого значення, де g — номер першої з двох сусідніх категорій, a, b — параметри лінії тренду $ai + b$);

5) об'єднання пари категорій, найбільш віддалених від апроксимуючого тренду, та обчислення значення WOE для новоутвореної категорії;

6) заміна поточних значень обраної пояснюючої змінної на відповідні WOE в навчальній і тестовій вибірках (за умови, що всі інші вхідні змінні моделі попередньо категоризовані та замінені значеннями WOE, залишаючись незмінними упродовж виконання алгоритму);

7) побудова скорингової моделі та налаштування її параметрів на начальній вибірці;

8) оцінювання на основі побудованої моделі ризику дефолту за кредитами тестової вибірки та розрахунок коефіцієнту Джині (6);

9) якщо значення коефіцієнта Джині від поточного об'єднання категорій вище за максимально досягнуте при k, то дане значення зберігається як максимальне для k та здійснюється перехід до кроку 4, якщо ні — то перехід на крок 10 (для початкової розбивки на k категорій зберігається поточне значення Джині як максимальне при k з переходом до кроку 4);

10) скасування поточного об'єднання категорій;

11) збільшення *k* (число збільшень *k* обирається індивідуально, залежно від типу вибірки) та повернення на крок 2; якщо *k* досяг максимального встановленого значення, то здійснюється вибір такого варіанту категоризації та параметрів скорингової моделі, яким відповідає найвище значення з усіх максимальних Джині по *k*.

Висновки

У статті проведено дослідження особливостей процесу поділу кількісних змінних скорингових моделей на категорії. При цьому акцентовано увагу на знаходженні оптимального числа категорій для кожної змінної та побудови класифікаційної моделі, в основу якої покладено інструментарій нейронних мереж персептронного типу.

Було встановлено, що задача знаходження оптимальної кількості категорій не належить до класу задач лінійного програмування. Проведені експерименти засвідчили, що показник інформаційної значущості *IV* не є головним критерієм ефективного поділу змінної на категорії. У статті запропоновано використання коефіцієнту Джині як основного показника оцінки повноти ефекту категоризації.

У рамках запропонованого методологічного підходу до категоризації вхідних даних було проведене експериментальне дослідження залежності точності скорингової моделі від способу категоризації, що надало можливість сформулювати розширений перелік рекомендацій для процесу біннінгу:

1) кількість надійних і ненадійних угод у категорії не мають дорівнювати 0;

2) розмір і кількість категорій для змінної залежать від її особливостей і не можуть бути чітко встановлені без проведення окремого експериментального дослідження;

3) сусідні категорії з близькими значенням *WOE* мають об'єднуватись;

4) всі пропущені у базі даних елементи аналізованої змінної доцільно об'єднати в окрему категорію;

5) зниження показника інформаційної значущості після об'єднання категорій з близькими значеннями *WOE* не виступає індикатором погіршення якості класифікатора;

6) оцінювання якості біннінгу доцільно здійснювати через показники точності скорингової моделі (наприклад, коефіцієнт Джині).

Таким чином, у даній роботі був розширений список рекомендацій проведення ефективного біннінгу, як важливого етапу обробки вхідних даних економіко-математичних моделей оцінювання кредитоспроможності позичальників комерційних банків.

Література

1. *Сорокин А. С.* Построение скоринговых карт с использованием модели логистической регрессии / А. С. Сорокин // Науковедение.— 2014. — № 2. — С. 1—29.

2. *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.

3. 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.

4. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring / N. Siddiqi. — New Jersey : John Wiley & Sons, 2006. — 196 p.

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

6. *Матвійчук А. В.* Штучний інтелект в економіці: нейронні мережі, нечітка логіка : монографія / А. В. Матвійчук. — К. : КНЕУ, 2011. — 439 с.

7. *Dreiseitl S.* Logistic regression and artificial neural network classification models: a methodology review / S. Dreiseitl, L. Ohno-Machado // Journal of Biomedical Informatics. — 2002. — No. 35. — P. 352—359.

8. *Zekic-Susac M.* Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models / M. Zekic-Susac, N. Sarlija, M. Bensic // 26th International Conference on Information Technology Interfaces — ITI 2004 / Cavtat(Croatia) : University of Zagreb. — 2004. — P. 265—270.

References

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

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

3. 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.

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

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

6. Matviychuk, A. V. (2011). *Shtuchnyi intelekt v ekonomitsi: neironni merezhi, nechitka lohika.* Kyiv: KNEU [in Ukrainian].

7. Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics, 35*, 352—359.

8. Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004, June 7—10). Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models. *Proceedings of the 26th International Conference on Information Technology Interfaces (ITI 2004) (Cavtat, Croatia : University of Zagreb)*, 265—270.

Стаття надійшла до редакції 17.06.2016