УДК 519.86+330.46
DOI 10.33111/nfmte.2019.094

ДОСЛІДЖЕННЯ СПОСОБІВ ТРАНСФОРМАЦІЇ ДАНИХ В КОНТЕКСТІ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ МОДЕЛЕЙ КРЕДИТНОГО СКОРИНГУ

Ю. В. Клебан

Магістр з економічної кібернетики,
викладач кафедри економіко-математичного моделювання
та інформаційних технологій
Національний університет «Острозька академія»

вул. Семінарська, 2, м. Острог, Рівненська обл., 35800, Україна
yuriy.kleban@oa.edu.ua

У статті проведено дослідження з пошуку найефективнішого підходу до попередньої обробки характеристичних ознак позичальників з метою підвищення точності передбачення дефолтів за кредитними зобов'язаннями. Проаналізовано три основних способи подання даних на входи моделей кредитного скорингу: застосування початкових пояснюючих змінних без трансформації, переведення категоріальних характеристик у набір фіктивних змінних, біннінг показників із розрахунком вагомості ознаки (WOE) для кожної категорії.

Для отримання висновків щодо систематичного впливу цих підходів було проведено по 10 повторюваних ітерацій з побудови нейромережевих моделей персептронного типу за кожним із цих трьох способів підготовки вхідних факторів. Кожна скорингова модель оцінювалась за широким набором показників інтегральної та точкової ефективності.

Результати проведених експериментів засвідчили практично за всіма критеріями перевагу запропонованого автором методологічного підходу до попередньої обробки даних шляхом розбиття кількісних змінних на категорії із забезпеченням тренду їх показників вагомості ознаки та дотриманням обмежень на обсяг спостережень у кожній групі.

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

ИССЛЕДОВАНИЕ СПОСОБОВ ТРАНСФОРМАЦИИ ДАННЫХ В КОНТЕКСТЕ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ МОДЕЛЕЙ КРЕДИТНОГО СКОРИНГА

Ю. В. Клебан

Магистр по экономической кибернетике,
преподаватель кафедры экономико-математического
моделирования и информационных технологий
Национальный университет «Острожская академия»

ул. Семинарская, 2, г. Острог, Ровенская обл., 35800, Украина
yuriy.kleban@oa.edu.ua

В статье проведено исследование, посвященное поиску наиболее эффективного подхода к предварительной обработке характеристических признаков заемщиков с целью повышения точности предсказания дефолтов по кредитным обязательствам. Проанализированы три основных способа представления данных на входы моделей кредитного скоринга: применение исходных объясняющих переменных без трансформации, преобразование категориальных характеристик в набор фиктивных переменных, биннинг показателей с расчетом весомости признака (WOE) для каждой категории.
Для получения выводов относительно систематического влияния данных подходов было проведено 10 повторяющихся итераций с построением нейросетевых моделей персептронного типа на основе каждого из этих трех способов подготовки входных факторов. Все скоринговые модели оценивались по широкому набору показателей интегральной и точечной эффективности.
Результаты проведенных экспериментов продемонстрировали практически по всем критериям преимущество предложенного автором методологического подхода к предварительной обработке данных путем разбиения количественных переменных на категории с обеспечением тренда их показателей весомости признака и соблюдением ограничений по объему наблюдений в каждой группе.

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

STUDYING THE METHODS OF DATA TRANSFORMATION IN THE CONTEXT OF INCREASING THE EFFECTIVENESS OF CREDIT SCORING MODELS

Yuriy Kleban

Master's Degree in Economic Cybernetics
Lecturer of Department of Economic and Mathematical Modeling and Information Technology
National University of Ostroh Academy

2 Seminarska Str., Ostroh, Rivne region, 35800, Ukraine
yuriy.kleban@oa.edu.ua

The article highlights a study on the search for the most effective approach to pre-processing the characteristics of borrowers in order to improve the accuracy of predicting defaults on credit obligations.

Three main ways of providing data to the inputs of credit scoring models are analyzed: the use of the initial explanatory variables without transformation, the conversion of categorical characteristics into a set of dummy variables, binning the indicators with the calculation of the weights of evidence (WOE) for each category.
To obtain conclusions about the systematic impact of these approaches, 10 repeated iterations were carried out with the construction of perceptron-type neural network models based on each of these three methods of preparing input factors. All scoring models were evaluated by a wide range of indicators of integrated and point efficiency.
The results of the experiments showed by almost all criteria the advantage of the methodological approach proposed by the author for preliminary data processing by dividing quantitative variables into categories, ensuring the trend in values of their weights of evidence and observing restrictions on the volume of observations in each group.

Keywords: *scoring model, neural network, creditworthiness, binning, weight of evidence (WOE), informational value (IV), Gini coefficient.*

JEL: C45, C51, C52, C53

I. Вступ

Складна геополітична та економічна ситуація в Україні з 2014 року призвела до ліквідації та реорганізації більше половини комерційних банків [1]. За даними Національного банку України частка непрацюючих кредитів на балансах банків протягом останніх років стабільно перевищує 50 %. Щоправда, вона поступово знижується, але лише завдяки статистичному ефектові від зростання загального портфеля [2, с. 5]. За часткою непрацюючих кредитів українські банки лідирують у світі: на 1 квітня 2019 року вона становила 51,7 % кредитного портфеля платоспроможних банків [2, с. 17].

Кредитний ризик є найбільш значущим з усіх банківських ризиків, що визначається Національним банком України в нормативному документі [3] як «наявний або потенційний ризик для надходжень і капіталу, який виникає через неспроможність сторони, що взяла на себе зобов'язання, виконати умови будь-якої фінансової угоди із банком або в інший спосіб виконати взяті на себе зобов'язання». Його значущість зумовлена тим, що кредитні операції є основним напрямком діяльності банків. Це підтверджується статистичними даними [4], за якими частка доходів від кредитування в останні роки стабільно перевищує 2/3 загального доходу по банківській системі. Таким чином, кредитний ризик, у

результаті його реалізації, може нести за собою найбільші збитки і найнегативніші наслідки порівняно з іншими ризиками банківської діяльності.

З метою недопущення зростання проблемної кредиторської заборгованості та подальшого виникнення негараздів у самої кредитної установи виникає потреба в проведенні процедур оцінки надійності позичальника на різних етапах кредитного циклу. Суттєвого підвищення ефективності процесу оцінки кредитних ризиків, збільшення швидкості та якості обслуговування клієнтів, а також зменшення вартості операційних послуг можна досягти за рахунок впровадження механізму скорингової оцінки в системі підтримки прийняття рішень з управління кредитною діяльністю комерційного банку.

У контексті інструментарію управління кредитним ризиком скоринг на основі інтегрального показника (скорингового балу) дозволив здійснити впорядкування та розбиття на класи позичальників за рівнем кредитоспроможності на основі множини їх характеристик, інформації про кредит та інших факторів, що можуть здійснювати вплив на спроможність виконання зобов'язань за кредитним договором. Швидкому поширенню скорингу сприяли масовість, ідентичність споживчого кредитування та обмежений час на оцінку ризиків. У результаті скоринг став найпоширенішим інструментом оцінки кредитного ризику позичальників і залишається таким по сьогоднішній день [5, с. 38–55].

Адекватне оцінювання кредитоспроможності клієнтів банківських установ передбачає пошук механізмів підвищення ефективності скорингових моделей. Одним із головних напрямів у вирішенні цього завдання стає забезпечення адекватного врахування найрізноманітнішої інформації щодо позичальника, що визначатиме ризик його дефолту за зобов'язаннями. Оскільки рівень кредитоспроможності позичальника обумовлюється як кількісними факторами, так і якісними, то важливо передбачити ефективну обробку в тому числі таких характеристик, як форма власності, пов'язані особи 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 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» може бути представлена сімома 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 пропущеними 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 зростатиме кількість фіктивних бінарних змінних. До того ж, переважна більшість їх значень дорівнюватиме нулю. Тому подібне розширення бази даних і кількості факторів суттєво ускладнюватиме процес побудови скорингових моделей і знижуватиме їх ефективність.

Метою цієї статті є удосконалення підходу до трансформації даних за рахунок категоризації факторів моделей кредитного скорингу, призначенням якого є підвищення точності ідентифікації неплатоспроможності позичальників.

Проведене у роботі дослідження розширює наявний апарат методів попередньої обробки даних і формування категоріальних змінних для задач моделювання бінарних показників.

ІІ. Проблеми трансформації даних у скорингових моделях

З метою категоризації показників (розбиття множини значень кількісних змінних на інтервали та переведення якісних змінних у числову форму) було вирішено скористатись загальноприйнятим у скорингу підходом, що ґрунтується на розрахунку показника вагомості ознаки *WOE* (*Weight Of Evidence*), який для кожної підгрупи (категорії) позичальників визначає узагальнену кількісну оцінку їх кредитної поведінки. Така оцінка базується на обчисленні часток поганих угод (за якими був оголошений дефолт) і хороших (закритих згідно умов договору) за кожною підгрупою показника відносно загальної кількості поганих і хороших угод, відповідно, з подальшим розрахунком *WOE* за формулою:

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

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

У спеціалізованій літературі [6, 7] 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 30 тис. грн вказує на значно менший кредитний ризик позичальника порівняно з тим, хто зазначив у кредитній заявці зарплату 6 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 штучного заниження рівня офіційної зарплатні в комерційних організаціях з метою зменшення податкових відрахувань (при високому рівні доходу, що особа отримує на руки). Таким чином, категорія позичальників із зарплатою у 6 тис. грн може виявитись навіть на-

дійнішою, ніж позичальники із заявленими надвисокими доходами. І специфіку поведінки кожної з таких підгруп дозволить виявити саме розрахунок показників *WOE*. Натомість, оперування моделі з початковими значеннями 30 тис. грн та 6 тис. грн вказувало б на п'ятикратну перевагу першого позичальника з відповідним нарахуванням скорингового балу, що далеко не завжди відповідає логіці економічних процесів.

Також варто зазначити, що *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 і кількісних показників різної розмірності до нормалізованих числових значень, придатних для побудови скорингових моделей будь-якого типу.

Згідно з дослідженням Дж. Германа [8] переваги застосування показника *WOE* при побудові скорингових моделей полягають, насамперед, у можливості:

1) ефективної обробки моделлю пропущених значень змінних (оскільки часто у базах даних кредитних організацій різні позичальники характеризуються різними наборами показників, то без цієї властивості доводиться або відкидати спостереження, або видаляти пояснюючі змінні, що суттєво звужує застосовність моделі);

2) виключення впливу екстремальних викидів на якість моделі, що підвищує її стійкість і робастність;

3) приведення всіх вхідних змінних до однієї розмірності (для певних типів економіко-математичних моделей це є суттєвим, оскільки дозволяє нівелювати надмірний вплив окремих показників 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*) [6, 9, 10]:

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

Чим вищою є інформаційна значимість предиктора, тим сильнішою є залежність від нього вихідної змінної. Коефіцієнти *IV*, отримані в результаті розрахунку (2), за [6, 9] 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.

Проте існують й інші погляди на інтерпретацію рівня інформаційної значимості. Наприклад, за Н. Сіддікі [7] категоризовані змінні, для яких *IV* перевищує 0,5, мають бути перевірені на завищення прогностичної сили (overpredicting). Їх слід або виключити з процесу моделювання, або використовувати з обережністю.

Проблемною ділянкою побудови моделі кредитного скорингу є створення ефективної процедури розбиття множини значень кожної з кількісних характеристик 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*) [6].

Розробка ефективного алгоритму біннінгу зводиться до розв'язання задач визначення оптимального числа категорій та їх діапазонів для кожної з кількісних пояснюючих змінних. Загальноприйнятим правилом при розв'язанні цих задач є те, що кожна категорія має об'єднувати множину сусідніх значень 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 [6] описано процес побудови скорингової моделі, починаючи з поділу даних на тестову та навчальну вибірки та закінчуючи оцінкою параметрів моделі. Також детально описано процедуру біннінгу кількісних змінних із розрахунком

показників вагомості ознаки *WOE* та інформаційної значимості *IV*. 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 суттєвого перепаду значень *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 [8], 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.

У роботі [7] Н. Сіддікі пропонує такі базові рекомендації щодо проведення біннінгу:

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

У роботі Н.Б. Палкіна та В.В. Афанасьєфа [11] досліджено проблему 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.

Питання оптимізації процедури біннінгу розкрито не лише у наукових працях, але й висвітлюється у дослідницьких оглядах і патентах аналітичних компаній і вендорів спеціалізованого програмного забезпечення. Зокрема, компанія FICO, яка є «законодавцем» в області конструювання скорингових карт, надає ряд рекомендацій до проведення категоризації кількісної змінної [12]:

- кожна категорія має містити достатньо елементів, аби нівелювати вплив екстремальних значень 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.

Наприклад, у статистичному пакеті STATISTICA 13 реалізований широкий перелік методів і алгоритмів категоризації кількісних змінних. За замовчуванням для біннінгу застосовується метод C&RT (англ. *Classification and Regression Trees* — дерева класифікації та регресії). Якщо кількість категорій не перевищує 20, алгоритм шукатиме усі можливі комбінації груп для максимізації *IV*. Коли кількість груп перевищує 20, STATISTICA застосовує CHAID алгоритм (англ. *Chi-squared Automatic Interaction Detection* — автоматичний детектор взаємодії хі-квадрат). Для поділу кількісних змінних на категорії застосовуються різні типи перетворень, що максимізують зв'язок предикторів із залежною змінною [13], зокрема:

1) монотонне — *WOE* значення усіх сусідніх категорій (інтервалів) будуть або зростати (позитивний монотонний зв'язок інтервалів предиктора та *WOE*), або спадати (негативний монотонний зв'язок);

2) квадратичне — функція зв'язку між кодованими інтервалами значень і *WOE* має U-подібну або перевернуту U-подібну форму;

3) кубічне — функція зв'язку між кодованими інтервалами значень і *WOE* має вигляд кубічної параболи;

4) користувацьке перетворення — задана за замовчуванням схема біннінгу з використанням C&RT або 10 груп приблизно однакового розміру;

5) кодування без обмежень (no restrictions) — пошук методом повного перебору або CHAID алгоритмом.

Попри те, що у проаналізованих вище публікаціях алгоритм біннінгу був достатньо детально описаний і прокоментований, принципи розбиття кількісних змінних на категорії у цих роботах є надто відмінними між собою. Різняться рекомендації щодо розмірів і кількості категорій, правил їх об'єднання, доцільності застосування показника інформаційної значимості тощо.

Отже, проведений аналіз існуючих підходів щодо попередньої обробки даних при побудові скорингових моделей, призначених для вирішення задач бінарної класифікації, дозволив виявити

значну кількість протиріч у численних рекомендаціях і реалізованих процедурах категоризації пояснюючих змінних. Це обумовлює доцільність проведення подальших досліджень з удосконалення методик трансформації даних для побудови математичних моделей кредитного скорингу.

III. Алгоритм формування категорій числових факторів моделі

Вирішення завдання розробки ефективної процедури біннінгу зводиться до таких етапів: сформулювати гіпотези щодо оптимальної категоризації 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 за різних варіантів біннінгу вхідних даних; систематизувати результати досліджень з обґрунтуванням відповідних висновків і рекомендацій.

Здійснити дослідження впливу процедури біннінгу на якість класифікатора можна в рамках методологічного підходу до проведення категоризації кількісних змінних у процесі побудови скорингових моделей [14], зміст якого полягає у виконанні таких дій:

1) збір інформаційної бази, формування навчальної та тестової вибірок;

2) розбиття множин 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 *AUROC*;

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

У процесі аналізу спеціалізованої літератури з питань біннінгу та проведення численних експериментів на реальних даних автором сформульовано низку вимог [14], 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 закриті згідно умов договору, так і дефолтні кредити;
- одне значення показника не може бути поділене між різними категоріями (усі записи з однаковим значенням змінної зводяться до одної категорії);
- 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).

Доцільність встановлення мінімального розміру категорії обумовлюється необхідністю нівелювання окремих випадкових викидів чи помилок у даних і врахування систематичних впливів у змінах показника на результати розрахунку кредитного ризику. Проте прописувати це окремою вимогою до алгоритму сенсу не було, адже навіть без чіткого обмеження мінімального розміру всі категорії будуть охоплювати досить широкий діапазон значень відповідного показника через необхідність забезпечення представництва обох класів кредитів і дотримання тренду змін *WOE*. Тож, встановлювати мінімальний розмір категорії чи ні, залишається на розсуд окремого аналітика або програміста.

В алгоритмі поелементного формування категорій, розробленому нами у рамках методологічного підходу проведення категоризації кількісних змінних із дотриманням сформульованих вище вимог, було вирішено ввести обмеження на мінімальний розмір категорії (зрештою це обмеження за потреби можна встановити на нульовому рівні).

Процес утворення категорій доречно розпочати з поступового об’єднання впорядкованих значень показника, доки їх кількість не перевищить заданий мінімальний розмір групи (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 мінімального або максимального його значення. 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 *WOE* не завжди вдається правильно визначити. Адже після першої категорії мінімального розміру можуть йти кілька категорій із поступовим зменшенням *WOE*, але загальний тренд виявиться зростаю-

чим. 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 показника. 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 в них представників обох класів і неможливості поділу одного значення показника між різними категоріями):

- розмір категорії має бути не меншим встановленого мінімального обмеження;
- розширення діапазонів крайніх 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 цих категорій, щоб вирівняти загальний тренд. У такому випадку алгоритм сам змінить напрям тренду на протилежний (зростаючий на спадний чи навпаки).

IV. Опис даних

Для побудови моделей взято набір даних із відкритого доступу [15] для навчання студентів курсу «Data Mining» у Політехнічному університеті Каталонії. Вибірка містить інформацію про кредитоспроможність фізичних осіб і їх характеристики. База даних складається з 4446 спостережень і 14 факторів, які мають як числовий, так і категоріальний характер, опис яких наведено у табл. 1.

Таблиця 1

ЗМІСТ БАЗИ ДАНИХ ДЛЯ ДОСЛІДЖЕННЯ КРЕДИТОСПРОМОЖНОСТІ ФІЗИЧНИХ ОСІБ

№ з/п	Назва показника	Тип	Короткий опис
1	*Seniority*	Числовий	Стаж, років
2	*Home*	Категоріальний	Інформація про житло: ignore, other, owner, parents, priv, rent
3	*Time*	Числовий	Період кредитування, місяців
4	*Age*	Числовий	Вік особи, років
5	*Marital*	Категоріальний	Сімейний стан: divorced, married, separated, single, widow
6	*Records*	Категоріальний	Наявність або відсутніть записів про попередні кредитні договори: no_rec, yes_rec
7	*Job*	Категоріальний	Забезпеченість та тип працевлаштування особи: fixed, freelance, others, partime
8	*Expenses*	Числовий	Витрати на обслуговування кредиту, гр.од.
9	*Income*	Числовий	Щомісячний дохід, гр.од.
10	*Assets*	Числовий	Активи, гр.од.
11	*Debt*	Числовий	Борг, гр.од.
12	*Amount*	Числовий	Сума кредиту, гр.од.
13	*Price*	Числовий	Вартість товару, на який виданий кредит, гр.од.
14	*Status*	Категоріальний	Статус повернення: bad, good

Важливим для трансформації даних є розуміння типу фактора, оскільки від цього залежить принцип перетворення елементів, якими він представлений. Для числових змінних кількісна величина елемента має значення, для категоріальних — не має. Наприклад, якщо

в базі даних значення «no_rec» та «yes_rec» змінної *Records* замінені на 1 та 2, то це не вказує на перевагу «yes_rec» над «no_rec».

Результат виконання умов кредитної угоди *Status* є бінарним показником із можливими варіантами події «bad» (у числовому виді замінюється на «1» — невиконання умов договору) і «good» («0» — умови договору виконані). Передбаченню підлягає саме подія «bad» (тоді результат розрахунку моделі в інтервалі [0; 1] означатиме ймовірність дефолту за кредитом). Як видно з табл. 2, частка невиконаних договорів у вибірці значно менша, ніж виконаних.

Таблиця 2

СТРУКТУРА ВИБІРКИ ЗА ОЗНАКОЮ ВИКОНАННЯ УМОВ КРЕДИТНОГО ДОГОВОРУ

	Подія «bad»	**Подія «good»**	**Разом**
Значення бінарної події	1	0	
Кількість подій	1249	3197	4446
Частка подій	0,281	0,719	1

Аналіз взаємозв'язку між категоріальними змінними та вихідним показником можна здійснити на основі крос-таблиць, формування яких на прикладі співставлення конкретних значень характеристики працевлаштування особи (*Job*) та статусу кредитного договору (*Status*) наведено у табл. 3.

Таблиця 3

КРОС-ТАБЛИЦЯ ПОКАЗНИКІВ «ПРАЦЕВЛАШТУВАННЯ» (*JOB*) ТА «СТАТУС КРЕДИТНОЇ УГОДИ» (*STATUS*)

Status / ***Job***		**fixed**	**freelance**	**others**	**partime**	**Разом**
bad	Кількість	580	331	68	270	1249
	Частка	0,464	0,265	0,054	0,216	1
good	Кількість	2223	690	103	181	3197
	Частка	0,695	0,216	0,032	0,057	1
Разом	Кількість	2803	1021	171	451	4446
	Частка	0,630	0,230	0,038	0,101	1

З табл. 3 видно, що частка неповернення кредитів особами, які мають неповну зайнятість, складає 21,6 % серед усіх проблемних

позик. При цьому загальна частка неповернених позик особами з неповною зайнятістю $\frac{270}{451}\times 100\% \approx 60\%$. Тобто такі позичальники частіше не повертають кредити, ніж виконують свої зобов'язання. Аналогічний аналіз проводиться і для інших змінних.

V. Трансформація пояснюючих змінних для застосування у скорингових моделях

З метою визначення найефективнішого способу попередньої обробки пояснюючих змінних для їх використання в моделях кредитного скорингу вибробуємо кілька різних підходів до перетворення даних і проведемо порівняльний аналіз точності побудованих на їх основі моделей. Для отримання висновків щодо систематичного впливу цих підходів проведемо по 10 повторюваних ітерацій з побудови моделей одного типу за трьома різними способами підготовки вхідних факторів: на основі даних без трансформації (тип 1), категоріальні характеристики трансформуються у набір фіктивних змінних (тип 2), всі показники категоризуються з розрахунком відповідних *WOE* (тип 3). При цьому біннінг змінних проводиться лише на тренувальній вибірці (тренувальна та тестова вибірки для кожної моделі на кожній ітерації формуються випадковим чином у пропорції 70/30 зі збереженням початкового співвідношення класів результуючого показника).

Отже, перш за все проводиться звичайне перетворення словесних описів категорій на відповідні числові порядкові значення. Так, наприклад, набору значень «divorced», «married», «separated», «single», «widow» змінної «Сімейний стан» (*Marital*) відоповідатимуть числа 1, 2, 3, 4 та 5, у цьому порядку. Приклад перетворення значень низки категоріальних показників до числових відповідників для модельного набору даних наведено в табл. 4.

Для моделі типу 2 категоріальні показники замінюються на фіктивні змінні. При цьому для кожного такого показника утворюється по одній бінарній змінній на кожну категорію, з яких він складається (де бінарній змінній, відповідній поточній категорії, присвоюється значення «1», а усім іншим — «0»). Приклад такого перетворення наведено у табл. 5.

Таблиця 4

ПРИКЛАД ПЕРЕТВОРЕННЯ КАТЕГОРІАЛЬНИХ ПОКАЗНИКІВ ДО ЧИСЛОВИХ ВІДПОВІДНИКІВ

№	***Status***	***Home***	***Marital***	***Records***	***Job***
1	good	rent	married	no_rec	freelance
2	good	rent	widow	no_rec	fixed
3	bad	owner	married	yes_rec	freelance
4	good	rent	single	no_rec	fixed
5	good	rent	single	no_rec	fixed
			⇩		
№	***Status***	***Home***	***Marital***	***Records***	***Job***
1	0	6	2	0	2
2	0	6	5	0	1
3	1	3	2	1	2
4	0	6	4	0	1
5	0	6	4	0	1

Таблиця 5

ПЕРЕТВОРЕННЯ КАТЕГОРІАЛЬНИХ ПОКАЗНИКІВ НА ФІКТИВНІ ЗМІННІ ДЛЯ МОДЕЛІ ТИПУ 2

№	***Marital***		***Marital***		***M_1***	***M_2***	***M_3***	***M_4***	***M_5***
1	married		2		0	1	0	0	0
2	widow	⇨	5	⇨	0	0	0	0	1
3	married		2		0	1	0	0	0
4	single		4		0	0	0	1	0
5	single		4		0	0	0	1	0

Такий підхід збільшує кількість незалежних змінних, що у свою чергу на великих об’ємах даних призводить до зростання часу на побудову, навчання та тестування моделей.

Уникнути таких ускладнень вдається із застосуванням моделей типу 3, в яких кожна змінна (як якісна, так і кількісна) поділяється на категорії з відповідними ним кількісними значеннями *WOE*.

Оскільки якісні змінні вже представлені у категоріях, то для них одразу можна здійснити розрахунок *WOE* та *IV* за формулами (1) та (2), відповідно. Оберемо для демонстрації процесу розрахунку цих показників фактор *Marital* (табл. 6 і рис. 1).

Таблиця 6

ПРИКЛАД РОЗРАХУНКУ *WOE* ТА *IV*
ДЛЯ КАТЕГОРІЙ ЗМІННОЇ «СІМЕЙНИЙ СТАН» (*MARITAL*)

Назва категорії	Номер катег.	Кількість «good»	Кількість «bad»	G_i	B_i	B_i-G_i	*WOE*	*IV*
divorced	1	18	10	0,0081	0,0112	0,0031	0,3198	0,0010
married	2	1664	587	0,7502	0,6559	-0,0944	-0,1344	0,0127
separated	3	50	43	0,0225	0,0480	0,0255	0,7567	0,0193
single	4	451	240	0,2033	0,2682	0,0648	0,2767	0,0179
widow	5	35	15	0,0158	0,0168	0,0010	0,0602	0,0001
		2218	895					0,0510

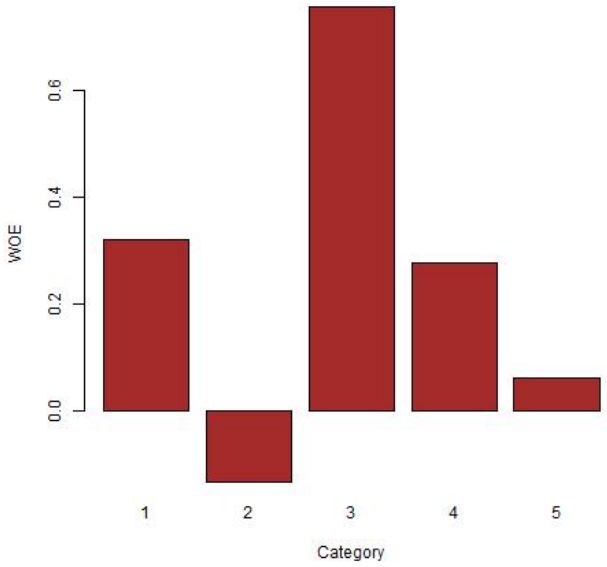

Рис. 1. Гістограма значень груп *WOE* для категоріальної змінної «Сімейний стан» (*Marital*)

Як вказано вище, *WOE* розраховується через співвідношення часток ненадійних «bad» і надійних «good» клієнтів банку відповідної групи у загальній кількості клієнтів того ж класу. Наприклад, для категорії «divorced» значення *WOE* становить 0,3198, що вказує на більшу частку ненадійних клієнтів у цій групі відносно середнього рівня надійності всіх клієнтів банку. Відповідно, на основі рис. 1 і табл. 6 можна попередньо оцінити ризиковість видачі кредитів для категорій осіб із різним сімейним станом. Так, наприклад, особи, що проживають окремо але фактично є одруженими («separated»), є найбільш ризиковими позичальниками. Також офіційно розведені особи («divorced») та одинокі («single») мають досить високу частку несплати кредитів. При цьому одружені особи («married») є досить надійною категорією позичальників банку.

Наступним етапом перетворення якісної змінної для її використання в моделі є заміна значення показника на обчислений *WOE* (табл. 7).

Таблиця 7

ПЕРЕТВОРЕННЯ КАТЕГОРІАЛЬНИХ ЗМІННИХ У ЧИСЛОВУ ФОРМУ ДЛЯ МОДЕЛІ ТИПУ 3

№	***Marital***		***Marital***		***Marital_WOE***
1	married		2		-0,1344
2	widow	⇨	5	⇨	0,0602
3	married		2		-0,1344
4	single		4		0,2767
5	single		4		0,2767

Для числових факторів процес категоризації має ряд особливостей. З метою біннінгу кількісних змінних застосуємо власний алгоритм формування категорій зі збереженням тренду [14].

Проілюструємо процедуру біннінгу за цим алгоритмом на прикладі змінної «Вік» (*Age*). Перш за все було сформовано тренувальну вибірку та задано мінімальний розмір категорії на рівні 5 % (156 спостережень). Всього у вибірці 50 унікальних значень показника *Age* (від 18 до 68 років). На початку процедури кожному такому значенню відповідатиме окрема категорія (табл. 8).

Таблиця 8

ПОЧАТКОВЕ РОЗБИТТЯ НА КАТЕГОРІЇ КІЛЬКІСНОЇ ЗМІННОЇ «ВІК» (*AGE*)

Номер категорії	Значення змінної	Кількість «good»	Кількість «bad»	*WOE*	*IV*
1	18	2	6	—	—
2	19	13	7	—	—
3	20	20	14	—	—
…	…	…	…	…	…
49	66	6	0	—	—
50	68	1	0	—	—

Далі проводиться об'єднання сусідніх елементів, що мають нульові значення у стовпці «good» чи «bad». Так, наприклад, рядки 49–50 будуть трансформовані в один (тоді кількість «good» сягне 7, а «bad» — 0). При такому об'єднанні категорія міститиме вже деяку множину значень показника. Відповідно, замість вказівки окремих елементів для позначення категорії задається діапазон — мінімальне та максимальне значення елементів, що увійшли до неї (у даному випадку [66; 68]).

Після цього проводиться об'єднання окремо крайніх «верхніх» і крайніх «нижніх» груп із відповідним розрахунком для них *WOE* та *IV*, доки не буде досягнуто цими категоріями заданого мінімального розміру та зростатиме різниця *WOE* між ними. Фрагмент бази після подібного укрупнення крайніх категорій наведено у табл. 9.

Таблиця 9

КАТЕГОРИЗОВАНА ЗМІННА «ВІК» ПІСЛЯ УКРУПНЕННЯ КРАЙНІХ КАТЕГОРІЙ

Номер категорії	Мінімальне значення	Максимальне значення	Кількість «good»	Кількість «bad»	*WOE*	*IV*
1	18	23	169	113	0,5050	0,0253
2	24	24	73	36	0,2006	0,0015
…	…	…	…	…	…	…
28	50	50	41	14	-0,1670	0,0005
29	51	68	357	79	-0,6008	0,0437

Після формування крайніх категорій продовжується процес укрупнення тих, що залишились, у напрямку до середини всієї

множини значень показника. Проміжний варіант біннінгу змінної «Вік» за авторським алгоритмом [14] наведено у табл. 10.

Таблиця 10

КАТЕГОРИЗОВАНА ЗМІННА «ВІК» НА ПЕРЕДОСТАННІЙ ІТЕРАЦІЇ ТРАНСФОРМАЦІЇ

Номер категорії	Мінімальне значення	Максимальне значення	Кількість «good»	Кількість «bad»	*WOE*	*IV*
1	18	23	169	113	0,5050	0,0253
2	24	25	135	68	0,2218	0,0034
3	26	37	913	373	0,0124	0,0001
4	38	40	181	92	0,2308	0,0049
5	41	44	196	83	0,0483	0,0002
6	45	47	142	51	-0,1165	0,0008
7	48	50	125	36	-0,3373	0,0054
8	51	68	357	79	-0,6008	0,0437

У випадку, якщо нові сформовані категорії не відповідають вже заданому тренду показника *WOE*, відбувається їх об'єднання з попередніми/наступними категоріями за близькістю *WOE*. Наприклад, з табл. 10 видно, що групи 4 та 5 не відповідають спадному тренду. На наступному етапі вони приєднуються до групи 3. Кінцевий результат проведення біннінгу кількісних показників за запропонованим алгоритмом поелементного формування категорій з дотриманням тренду значень *WOE* на прикладі змінної «Вік» (*Age*) при мінімально допустимому розмірі категорій 5 % наведено у табл. 11 і на рис. 2.

Таблиця 11

РЕЗУЛЬТАТ КАТЕГОРИЗАЦІЇ КІЛЬКІСНОЇ ЗМІННОЇ «ВІК» (*AGE*)

Номер категорії	Мінімальне значення	Кількість «good»	Кількість «bad»	*WOE*	*IV*
1	18	169	113	0,5050	0,0253
2	24	135	68	0,2218	0,0034
3	26	1290	548	0,0514	0,0016
4	45	142	51	-0,1165	0,0008
5	48	125	36	-0,3373	0,0054
6	51	357	79	-0,6008	0,0430

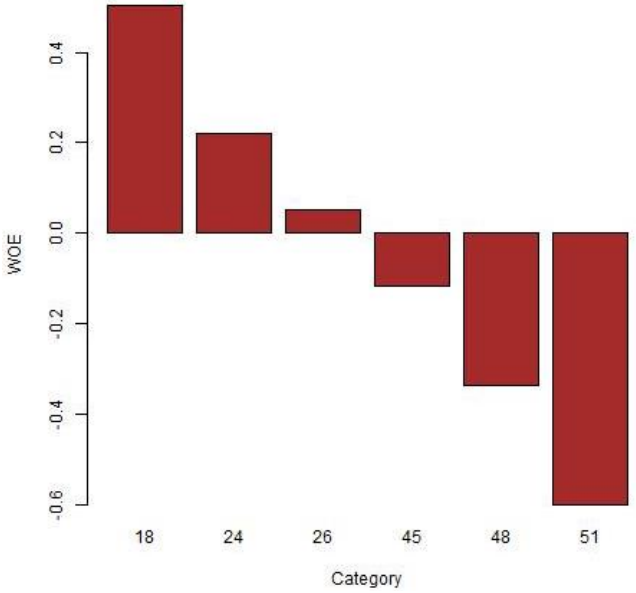

Рис. 2. Гістограма *WOE* категорій кількісної змінної «Вік» (*Age*)

Як видно з табл. 11 і рис. 2, сформовані за запропонованим алгоритмом категорії мають спадний тренд показника *WOE*, який відповідає зростанню частки надійних і зменшенню частки ненадійних клієнтів зі збільшенням віку позичальників.

У другому стовпчику табл. 11 вказується мінімальне значення показника, з якого починається відповідна категорія. Тобто ця категорія покриває діапазон від даного значення до мінімума наступної категорії, що міститься рядком нижче. Для останньої категорії (шостої у табл. 11) верхня границя не задається. Так само й перша категорія не обмежується нижньою границею (не обмежується математично, проте у даному випадку законодавчо).

Принцип застосування в моделях кредитного скорингу категоризованих значень кількісних змінних такий. Наприклад, згідно табл. 11 вік 46 років потрапляє у групу 4, до якої належать усі значення цього показника в діапазоні [45; 48). При побудові скорингової моделі та подальшому оцінюванні кредитного ризику нових клієнтів усі значення *Age* з цього діапазону замінюватимуться на *Age_WOE* = -0,1165, як представлено у табл. 12.

Таблиця 12

ПЕРЕТВОРЕННЯ КІЛЬКІСНОЇ ЗМІННОЇ «ВІК» В КАТЕГОРИЗОВАНУ ФОРМУ ДЛЯ ПОБУДОВИ МОДЕЛІ ТИПУ 3

№	***Age***		***Age_WOE***
1	22		0,5050
2	65	⇨	-0,6008
3	44		0,0514
4	33		0,0514
5	46		-0,1165

Підкреслимо, що перетворення кількісних змінних до категоріального виду здійснюється на основі запропонованого алгоритму поелементного формування категорій, який передбачає можливість встановлення мінімально допустимого їх розміру. При цьому алгоритм дозволяє оптимізувати мінімальний розмір групи за критерієм максимізації інформаційної значимості показників.

У результаті виконання описаних вище процедур з трансформації пояснюючих змінних було отримано три набори даних для побудови відповідних моделей оцінки кредитоспроможності позичальників.

VI. Побудова математичних моделей та оцінювання їх ефективності

Як зазначалось вище, статус виконання кредитного договору *Status* є бінарним показником. Тобто, при побудові скорингової моделі залежна змінна набуває значення «1», якщо відбувся дефолт за зобов'язаннями, та «0» — якщо ця подія не наступає. Результатом розрахунку моделі при її застосуванні в процесі оцінювання кредитного ризику буде вже не бінарне число, а дійсне, визначене на інтервалі [0; 1], що вказує на ймовірність дефолту за зобов'язаннями позичальника.

Основою для побудови математичних моделей у даному дослідженні слугували нейронні мережі персептронного типу з одним прихованим шаром та одним нейроном вихідного шару із логістичною сигмоїдною функцією активації (щоб результатом розрахунку мережі було число в межах від 0 до 1). Кількість нейронів прихованого шару та їх функції активації підбирались експериментально для отримання найбільшої точності класифікації позичальників із тестової вибірки.

Для порівняння ефективності скорингових моделей у даному дослідженні використовувалися такі показники [16–18]: *Accuracy* — загальна точність класифікації за певного рівня розмежування (cut-off); *Balanced Accuracy* — збалансована загальна точність класифікації, що враховує пропорції позитивних і негативних подій у вибірці; *KS* — критерій узгодженості Колмогорова-Смірнова; *Gini* — коефіцієнт Джині та *AUROC* — площа під ROC-кривою (Receiver Operating Characteristic), які є інтегральними характеристиками якості класифікатора.

ROC-крива дозволяє візуально оцінити коректність бінарної класифікації. Вона по осі ординат відображає частку правильно визначених позитивних результатів[1] *TPR* (True Positive Rate або чутливість — *Sensitivity*), а по осі абсцис — частку помилково діагностованих позитивних результатів *FPR* (False Positive Rate) при варіюванні порога відсікання. На рис. 3 зображено ROC-криву, побудовану в процесі тестування моделі типу 1 на ітерації № 1.

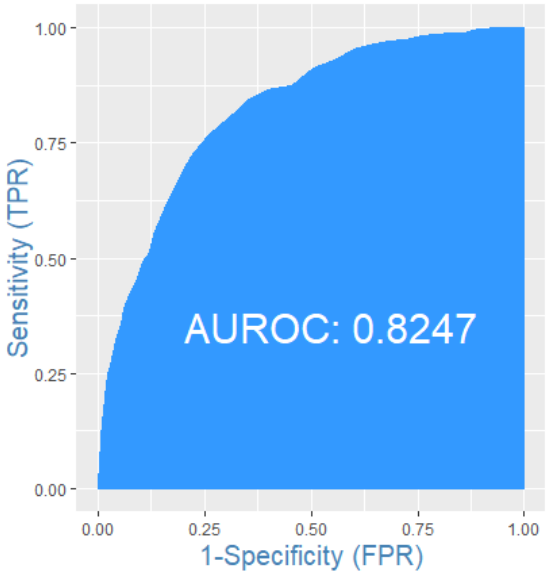

Рис. 3. Графік ROC-кривої моделі типу 1 на першій ітерації дослідження

[1] Оскільки задачею дослідження є моделювання ймовірності дефолту позичальника, то в даному контексті позитивним результатом буде клас «bad» результуючої змінної *Status*.

При порівнянні ефективності моделей лінія поділу класів (cut-off) була встановлена на рівні 0,5, хоча оптимальні лінії cut-off різних моделей мали відмінності, проте незначні (0,5±0,05). У процесі дослідження проведено 10 ітерацій побудови нейронних мереж за трьома підходами до попереденьої обробки даних: NN — на основі даних без трансформації (тип 1), NN_0X — із переведенням категоріальних характеристик у набір фіктивних змінних (тип 2), NN_WOE — на основі категоризованих показників із розрахунком відповідних *WOE* (тип 3). Результати тестування ефективності всіх побудованих моделей наведені у табл. 13.

Таблиця 13

ПОКАЗНИКИ ЕФЕКТИВНОСТІ ПОБУДОВАНИХ НЕЙРОННИХ МЕРЕЖ ЗА ТРЬОМА СПОСОБАМИ ПОПЕРЕДНЬОЇ ОБРОБКИ ДАНИХ

№ експ.	Модель	*AUROC*	*KS*	*Gini*	*Sensitivity*	*Specificity*	*Accuracy*	*Balanced Accuracy*
1	NN	0,8247	0,5170	0,6494	0,4209	0,9265	0,7922	0,6737
	NN_WOE	***0,8490***	***0,5530***	***0,6980***	***0,4548***	***0,9316***	***0,8050***	***0,6932***
	NN_0X	0,8429	0,5439	0,6858	0,4350	0,9275	0,7967	0,6813
2	NN	0,8213	0,5165	0,6426	0,4182	**0,9262**	0,7794	0,6722
	NN_WOE	***0,8424***	***0,5516***	***0,6848***	***0,4519***	0,9241	***0,7877***	***0,6880***
	NN_0X	0,8297	0,5179	0,6594	0,4468	***0,9262***	***0,7877***	0,6865
3	NN	0,8177	0,4901	0,6354	***0,4357***	0,9223	***0,7832***	***0,6790***
	NN_WOE	***0,8294***	***0,5127***	***0,6588***	0,4121	0,9170	0,7727	0,6645
	NN_0X	0,8253	0,5122	0,6506	0,4304	***0,9233***	0,7824	0,6769
4	NN	0,8253	0,5086	0,6506	0,4354	0,9308	0,7899	0,6831
	NN_WOE	***0,8451***	***0,5553***	***0,6902***	***0,4749***	0,9308	***0,8012***	***0,7029***
	NN_0X	0,8329	0,5210	0,6658	0,4591	***0,9361***	0,8005	0,6976
5	NN	0,8087	0,4776	0,6174	0,4154	0,9205	0,7682	0,6680
	NN_WOE	0,8124	***0,4945***	0,6248	0,4229	0,9098	0,7629	0,6663
	NN_0X	***0,8142***	0,4810	***0,6284***	***0,4254***	***0,9313***	***0,7787***	***0,6783***
6	NN	0,8234	0,5322	0,6468	0,4219	0,9168	0,7742	0,6693
	NN_WOE	***0,8352***	***0,5374***	***0,6704***	***0,4583***	***0,9210***	***0,7877***	***0,6897***
	NN_0X	0,8283	0,5300	0,6566	0,4245	0,9189	0,7764	0,6717

Закінч. табл. 13

7	NN	0,8139	0,4865	0,6278	0,4380	0,9151	0,7794	0,6765
	NN_WOE	***0,8239***	0,5087	***0,6478***	***0,4776***	0,9130	0,7892	***0,6953***
	NN_0X	0,8237	***0,5156***	0,6474	0,4644	***0,9203***	***0,7907***	0,6924
8	NN	0,8247	0,5109	0,6494	0,4454	***0,9266***	0,7944	0,6860
	NN_WOE	***0,8429***	***0,5430***	***0,6858***	***0,4836***	0,9255	***0,8042***	***0,7046***
	NN_0X	0,8344	0,5295	0,6688	0,4617	0,9255	0,7982	0,6936
9	NN	0,8079	0,4759	0,6158	0,4251	0,9124	0,7757	0,6688
	NN_WOE	***0,8211***	***0,5069***	***0,6422***	0,4251	0,9135	0,7764	0,6693
	NN_0X	0,8179	0,4810	0,6358	***0,4385***	***0,9155***	***0,7817***	***0,6770***
10	NN	0,8127	0,4737	0,6254	0,4262	***0,9374***	0,7997	0,6818
	NN_WOE	0,8220	0,4900	0,6440	0,4206	0,9312	0,7937	0,6759
	NN_0X	***0,8252***	***0,5082***	***0,6504***	***0,4318***	***0,9374***	***0,8012***	***0,6846***

Як видно з табл. 13, модель типу 3 (на основі категоризованих показників із розрахунком відповідних *WOE*) має найгіршу точність ідентифікації надійних клієнтів (найменші значення *Specificity*). При цьому вона ж дозволяє найточніше ідентифікувати подію дефолт (має найвищі значення *Sensitivity*), що і є метою у вирішенні даної задачі. У табл. 14 представлено зведені з табл. 13 показники переваги моделі кожного типу над іншими за низкою критеріїв ефективності.

Таблиця 14

АГРЕГОВАНІ ПОКАЗНИКИ ПЕРЕВАГ МОДЕЛІ КОЖНОГО ТИПУ ЗА РІЗНИМИ КРИТЕРІЯМИ ЕФЕКТИВНОСТІ

Тип моделі	**Модель**	***AUROC***	***R^2***	***KS***	***Gini***	***Sensitivity***	***Specificity***	***Accuracy***	***Balanced Accuracy***
Тип 1	NN	0	0	0	0	1	3	1	1
Тип 2	NN_0X	2	2	2	2	3	7	5	3
Тип 3	**NN_WOE**	**8**	**8**	**8**	**8**	**6**	**1**	**5**	**6**

Для отримання додаткових висновків відносно дієвості аналізованих підходів до попередньої обробки даних проранжуємо побудовані на їх основі моделі за критерієм *AUROC*, що є інтегральним показником ефективності класифікатора (табл. 15).

Таблиця 15

РАНГИ КРАЩИХ МОДЕЛЕЙ ЗА ПОКАЗНИКОМ *AUROC*

№ експ.	*Тип моделі*	*AUROC*	*Ранг*	*Візуалізація*
1	Тип 3	***0,8490***	1	
4	Тип 3	***0,8451***	2	
8	Тип 3	***0,8429***	3	
2	Тип 3	***0,8424***	4	
6	Тип 3	***0,8352***	5	
3	Тип 3	***0,8294***	6	
10	Тип 2	***0,8252***	7	
7	Тип 3	***0,8239***	8	
9	Тип 3	***0,8211***	9	
5	Тип 2	***0,8142***	10	

Вивчаючи табл. 13–15 можна помітити, що найвищі показники ефективності демонструють моделі третього типу. Так, у табл. 15 перших шість місць рейтингу та 8 переваг над іншими у 10 ітераціях займають саме моделі, побудовані на основі категоризованих змінних із розрахунком *WOE*. Це підтверджує гіпотезу про перевагу запропонованого підходу до розбиття кількісних змінних на категорії із забезпеченням дотримання тренду в значеннях їх показників вагомості ознаки над альтернативними способами подання даних на входи скорингових моделей.

VII. Висновки

У статті вирішується завдання пошуку ефективного підходу до попередньої обробки предикторів з метою підвищення точності моделей кредитного скорингу. У результаті проведеного аналізу наукових праць за даною тематикою та огляду методів обробки даних у спеціалізованих програмних пакетах було вирішено зупинитись на дослідженні трьох основних способів подання даних на входи скорингових моделей: застосування початкових пояснюючих змінних без трансформації, переведення категоріальних характеристик у набір фіктивних змінних, проведення біннінгу показників із розрахунком вагомості ознаки для кожної категорії.

Перевірка придатності обраних підходів до обробки даних проводилась на базі нейронних мереж персептронного типу за широким

переліком критеріїв інтегральної та точкової ефективності. Результати проведених експериментів засвідчили перевагу запропонованого підходу до розбиття кількісних змінних на категорії з дотриманням тренду в значеннях їх показників вагомості ознаки над альтернативними способами подання даних на входи скорингових моделей. Висока ефективність побудованих на базі такого підходу нейромереж вказує на доцільність застосування подібних технологій на практиці з метою підвищення точності оцінювання кредитоспроможності позичальників і зменшення кредитних ризиків банків.

Список літератури

1. Інформація про дати прийняття рішень Національним банком про визнання банків неплатоспроможними та про ліквідацію, рішень ФГВФО про запровадження тимчасової адміністрації з 2014 року. *Національний банк України* : веб-сайт. URL: https://bank.gov.ua/supervision/reorganizat-liquidat/reorganiz-history (дата звернення: 23.07.2019).

2. Звіт про фінансову стабільність, червень 2019 р. *Національний банк України* : веб-сайт. URL: https://bank.gov.ua/admin_uploads/article/FSR_2019-R1.pdf?v=4 (дата звернення: 23.07.2019).

3. Методичні вказівки з інспектування банків «Система оцінки ризиків» : Постанова Правління Національного банку України від 15.03.2004 № 104. URL: https://zakon.rada.gov.ua/laws/show/v0104500-04 (дата звернення: 23.07.2019).

4. Доходи та витрати банків України. *Національний банк України* : веб-сайт. URL : https://bank.gov.ua/files/stat/Inc_Exp_Banks_2019-03-01.xlsx (дата звернення: 23.07.2019).

5. Anderson R. The credit scoring toolkit: theory and practice for retail credit risk management. Oxford: Oxford University Press, 2007. 731 p.

6. Сорокин А. С. Построение скоринговых карт с использованием модели логистической регрессии. *Науковедение*. 2014. Вып. 2. С. 1–29. URL : http://naukovedenie.ru/PDF/180EVN214.pdf.

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

8. Jopia H. R Package ‘smbinning’: Optimal Binning for Scoring Modeling. *Revolutions* : website. 2015. URL : https://blog.revolutionanalytics.com/2015/03/r-package-smbinning-optimal-binning-for-scoring-modeling.html.

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

10. Коляда Ю. В., Бондар В. А. Біннінг у нейромережевих скорингових моделях. *Нейро-нечіткі технології моделювання в економіці*. 2016. № 5. С. 60–80.

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

12. Building Powerful, Predictive Scorecards: An overview of Scorecard module for FICO Model Builder. *Fair Isaac Corporation* : website. 2014. 46 p. URL : http://www.fico.com/en/wp-content/secure_upload/Building_Powerful_Predictive_Scorecards_1991WP.pdf (Last accessed : 10.06.2019).

13. TIBCO Statistica 13.5.0. *statsoft.com* : website. URL : http://documentation.statsoft.com/portals/0/formula%20guide/Weight%20of%20Evidence%20Formula%20Guide.pdf (Last accessed : 10.06.2019).

14. Матвійчук А. В., Клебан Ю. В. Біннінг кількісних змінних з формуванням тренду для задач скорингу. *Моделювання та інформаційні системи в економіці*. 2017. Вип. 93. С. 213–229.

15. Gaston S. CreditScoring. *github.com* : website. https://github.com/gastonstat/CreditScoring/blob/master/CreditScoring.csv (Last accessed : 10.02.2019).

16. Kuhn M. Building predictive models in R using the caret package *Journal of Statistical Software*. 2008. Vol. 28, Is. 5. P. 1–26. URL: http://www.jstatsoft.org/article/view/v028i05/v28i05.pdf.

17. Marsaglia G., Tsang W. W., Wang J. Evaluating Kolmogorov's Distribution. *Journal of Statistical Software*. 2003. Vol. 8, Is. 18. P. 1–4. URL : http://www.jstatsoft.org/v08/i18/paper.

18. Powers D. M. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*. 2011. Vol. 2, Is. 1. P. 37–63. URL : https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.

References

1. The National Bank of Ukraine. (2019). *Informatsiia pro daty pryiniattia rishen Natsionalnym bankom pro vyznannia bankiv neplatospromozhnymy ta pro likvidatsiiu, rishen FHVFO pro zaprovadzhennia tymchasovoi administratsii z 2014 roku*. Retrieved from https://bank.gov.ua/supervision/reorganizat-liquidat/reorganiz-history [in Ukrainian].

2. The National Bank of Ukraine. (2019). *Zvit pro finansovu stabilnist, cherven 2019 r*. Retrieved from https://bank.gov.ua/admin_uploads/article/FSR_2019-R1.pdf?v=4 [in Ukrainian].

3. The National Bank of Ukraine. (2004). *Metodychni vkazivky z inspektuvannia bankiv «Systema otsinky ryzykiv» : Postanova Pravlinnia Natsionalnoho banku Ukrainy vid 15.03.2004 № 104*. Retrieved from https://zakon.rada.gov.ua/laws/show/v0104500-04 [in Ukrainian].

4. The National Bank of Ukraine. (2019, March 1). *Dokhody ta vytraty bankiv Ukrainy*. Retrieved from https://bank.gov.ua/files/stat/Inc_Exp_Banks_2019-03-01.xlsx.

5. Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management*. Oxford, UK : Oxford University Press.

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

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

8. Jopia, H. (2015, March 24). *R Package 'smbinning': Optimal Binning for Scoring Modeling*. Retrieved from https://blog.revolutionanalytics.com/2015/03/r-packagesmbinning-optimal-binning-for-scoring-modeling.html.

9. 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].

10. 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].

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

12. 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.

13. TIBCO. (2019). *TIBCO Statistica 13.5.0*. Retrieved from http://documentation.statsoft.com/portals/0/formula%20guide/Weight%20of%20Evidence%20Formula%20Guide.pdf.

14. Matviychuk, A. V., & Kleban, Yu. V. (2017). Binninh kil′kisnykh zminnykh z formuvannyam trendu dlya zadach skorynhu. *Modelyuvannya ta informatsiyni systemy v ekonomitsi (Modeling and information systems in economics), 93*, 213–229 [in Ukrainian].

15. Gaston, S. (2019). *CreditScoring*. Retrieved from https://github.com/gastonstat/CreditScoring/blob/master/CreditScoring.csv.

16. Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26. Retrieved from http://www.jstatsoft.org/article/view/v028i05/v28i05.pdf.

17. Marsaglia, G., Tsang, W.W., & Wang, J. (2003). Evaluating Kolmogorov's Distribution. *Journal of Statistical Software, 8*(18), 1–4. Retrieved from http://www.jstatsoft.org/v08/i18/paper.

18. Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63. Retrieved from https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.

Стаття надійшла до редакції 28.07.2019