ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, СИСТЕМНИЙ АНАЛІЗ ТА КЕРУВАННЯ

УДК 519.226, 330.322
DOI: 10.20535/1810-0546.2016.2.67487

С.А. Бакун, П.І. Бідюк
Національний технічний університет України "КПІ", Київ, Україна

МЕТОДИКА ПОБУДОВИ СКОРИНГОВИХ КАРТ ІЗ ВИКОРИСТАННЯМ ПЛАТФОРМИ SAS

Background. Development of effective methods for evaluating solvency of individuals and risk of banks in providing consumer loans.
Objective. Determining of the mechanisms for implementation of scoring models in the form of scoring cards. Analysis of the possibility of using scoring cards as a tool for credit risk management.
Methods. Construction of scoring cards and preliminary analysis of input data using specialized component of the SAS Enterprise Miner.
Results. The main stages of scoring cards development were considered. The scoring card was constructed that is based on actual statistical data on granting of the consumer loans. The research also presents comparative analysis of the scoring cards with other statistical methods of subjects classification.
Conclusions. It was established in this study that the scoring cards have better forecasting ability than other statistical methods such as decision trees, neural networks and logistic regression. The format of development the forecasting models in the form of scoring cards is the easiest for interpreting. However, application of this method requires considerable investments as well as continuous updating and renewal of credit histories for borrowers.
Keywords: risk management; data mining; credit scoring; scoring card; logistic regression; classification quality.

Вступ

У період стабільного розвитку в Україні спостерігалося бурхливе зростання ринку іпотечного та побутового кредитування, що спричинило розвиток досі невідомого для нашої країни напряму прикладного математичного моделювання – кредитного скорингу. В умовах економічної та політичної кризи і нестабільності коректне розв'язання задач виявлення платоспроможних клієнтів стає для фінансових установ завданням прямого виживання на ринку кредитних послуг. Незважаючи на те, що споживче кредитування є одним із найбільш прибуткових видів банківської діяльності, воно водночас найбільш ризиковане, оскільки багато банків стикаються з проблемою неповернення виданих фізичним особам кредитів [1, 2]. У зв'язку з цим розробка та застосування нових більш досконалих методик оцінювання кредитоспроможності осіб і ризику банків при наданні кредитів в умовах сучасних кризових явищ у фінансовій сфері є критично важливими.

Ефективним заходом, що дає змогу оптимально розв'язувати задачі оцінювання кредитоспроможності осіб, є кредитний скоринг, який являє собою математичну або статистичну модель, за допомогою якої банк визначає, ймовірність того, що потенційний позичальник поверне кредит у встановлений термін. Метою кредитного скорингу є оптимізація прийняття рішень із надання споживчих кредитів, що й визначає актуальність дійсного дослідження [3, 4].

Класичним видом скорингової моделі є скорингова карта. Перші скорингові карти з'явилися в 40-ві роки XX ст. у США. Скорингові бали тоді виставлялися на основі експертних оцінок, а за основу брали не більше десяти характеристик. Нині для побудови скорингових карт використовуються більш точні математичні методи і спеціальний аналітичний інструментарій, що дає можливість працювати із сотнями характеристик [5]. Звичайно, за наявності достатньої кількості даних статистичний скоринг кращий, оскільки, на відміну від експертного, відображає ті закономірності, які реально спостерігаються в портфелі.

На сьогодні у сфері бізнес-аналітики найбільш відомими й авторитетними вендорами є компанії SAS, SPSS, KXEN, FICO, PROGNOZ й інші. Але всесвітньо визнаним лідером у сфері інтелектуального аналізу даних та спеціалізованих рішень для розв'язання задач скорингу є компанія SAS Institute, продукти якої використовуються в 50 % банківських установ США та покривають 30 % світового ринку рішень для бізнес-аналітики. В Україні відповідні рішення використовуються в ТОП-30 найбільших банках.

Постановка задачі

Метою роботи є побудова скорингової карти для оцінювання кредитних ризиків банківських установ за допомогою системи SAS Enterprise Miner і виконання порівняльного аналі-

зу скорингової карти з іншими статистичними та математичними методами, що дають змогу оцінити кредитоспроможність позичальників, а саме з логістичною регресією, деревами рішень та нейронними мережами. Для побудови скорингової моделі необхідно підготувати навчальну вибірку даних, відібрати найбільш значущі характеристики клієнтів, які потрібно включити в модель, розробити саму модель та оцінити її якість. Для порівняння скорингових моделей і вибору найкращої необхідно застосувати такі критерії оцінки якості, як *ROC*-крива та індекс *GINI*.

Процес розробки скорингової карти

Побудова моделі кредитного скорингу починається зі збору достатньої кількості репрезентативних даних про позичальників банку, для яких вже завершився строк кредитування, тобто наявна інформація про виконання або невиконання ними зобов'язання по кредиту. Якість вихідних даних для побудови статистичної моделі визначає її точність прогнозування і успіх розробки скорингової системи в цілому. Для розробки скорингових карт потрібні надійні та чисті дані з мінімальною допустимою кількістю "хороших" і "поганих" записів. Обсяг необхідних даних може бути різним, але в цілому він повинен відповідати вимогам статистичної значущості та випадковості [6]. Для розробки скорингової карти на практиці, за рекомендаціями фахівців банківського скорингу, зазвичай використовують не менше 2000 "поганих" і 2000 "хороших" записів про клієнтів, які можуть бути випадковим чином відібрані із загальної популяції клієнтів відповідного банку або бюро кредитних історій. Окрім цього, в спеціальних методах скорингування можуть знадобитися додатково 2000 відхилених заявок, за якими необхідно провести аналіз причин відхилення. Вихідні дані для побудови моделі можуть містити внутрішні дані анкет позичальників банку, а також зовнішні дані кредитних історій.

Дані про певний тип клієнтів необхідно виключити з вихідної інформаційної бази. Це можуть бути нетипові клієнти – шахраї, співробітники банку, VIP-клієнти, померлі, неповнолітні, кредити за вкраденими картками тощо. Також з бази повинні бути виключені кредити з аномально великими сумами кредитів, нестандартними умовами погашення, нетиповими цілями позики. Додатковим критерієм відбору даних може бути вид кредитування, для якого буде розроблятися скорингова карта [6].

Більшість фінансових даних характеризуються відсутністю деяких значень або, навпаки, наявністю значень, які недоцільні для тієї чи іншої характеристики. Це можуть бути поля, значення яких не були зафіксовані, які більше не використовуються, які були недоступні або не були заповнені заявниками, а також неправильно введені дані, викиди або значення, що різко виділяються. Є декілька способів позбутися таких значень, наприклад:

1) виключити всі дані з пропущеними значеннями, оскільки аналіз ведеться по всіх змінних. Для випадку фінансових даних такий спосіб у більшості випадків дає дуже мало даних для подальшого аналізу;

2) виключити з моделі характеристики або записи, для яких частка пропущених значень істотна (наприклад, перевищує 20 %);

3) включити до скорингової карти характеристику-ідентифікатор наявності пропуску по атрибуту клієнта;

4) замінити пропущені значення, використовуючи спеціалізовані методи статистики по заповненню даних (синтетичний розподіл) або прогнозуванню (зазвичай дерева рішень або регресійні методи).

Вибір залежної (цільової) змінної визначається метою побудови скорингової моделі. Цілі можуть бути загальними, наприклад скорочення втрат за новими кредитними рахунками, і конкретними, наприклад скорочення числа дефолтів по схвалених заявках протягом 3-х місяців після прийняття позитивного рішення. Залежна змінна може набувати кількісних та якісних значень. Найбільш часто використовується категоріальний вид вимірювання залежної змінної з двома категоріями: "хороший" та "поганий" клієнт. Зазвичай до категорії "поганий" відносять клієнтів, що мають прострочену заборгованість 90 і більше днів [7].

Як незалежні змінні при побудові скорингової моделі можуть використовуватися дані з кредитної заявки: соціально-демографічні дані про позичальника (стать, сімейний стан, вік, посада, наявність дітей, дохід позичальника тощо), інформація про запитуваний кредит (строк погашення кредиту за договором, сума кредиту, розмір початкового внеску, мета кредиту тощо). Одним із основних джерел даних для формування незалежних змінних є дані Бюро кредитних історій на момент подачі заявки позичальником: рейтинг позичальника, детальна інформація про наявні кредити в інших банках, детальна інформація про прострочені

або повністю погашені минулі кредити, наявність інших банківських продуктів і послуг у позичальника тощо. Також для формування незалежних змінних може використовуватися внутрішня кредитна історія позичальника: поточний баланс рахунку, заборгованість на поточний момент, кількість рахунків, число попередніх кредитів у банку, найбільше значення суми заборгованості за попередніми кредитними рахунками [6].

Таким чином, незалежні або скорингові вхідні змінні можуть бути подані в різних шкалах вимірювання залежно від можливості об'єктивних вимірів відібраних характеристик. На практиці можуть бути побудовані скорингові моделі з такими типами незалежних змінних: тільки з кількісними, тільки з категоріальними, з категоріальними і кількісними змінними одночасно. Найчастіше для побудови скорингових карт використовують категоріальні змінні. Категоризація кількісних змінних дає змогу досягти таких основних переваг при побудові скорингової карти: полегшити обробку викидів та екстремальних значень кількісних змінних, спростити інтерпретацію скорингової карти, відобразити складні нелінійні зв'язки.

В остаточну модель включаються найбільш значущі незалежні змінні, які відповідно до статистичних даних краще за інших дають змогу зробити прогноз. Для оцінки ступеня взаємозв'язку між незалежними змінними і залежною в кредитному скорингу прийнято використовувати показник інформаційного значення, або *IV* (Information Value), що обчислюється за формулою

$$IV = \sum_{i=1}^{k} (d_i^{(1)} - d_i^{(2)}) \cdot \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right),$$

де $d_i^{(1)}$ і $d_i^{(2)}$ – відносні частки "поганих" та "хороших" кредитів в *i*-й категорії, k – кількість категорій незалежної змінної.

Чим вище інформаційне значення змінної, тим більшу вагу вона має з точки зору корисності при побудові моделі. Можна керуватися правилами при відборі змінних для побудови скорингової карти згідно з табл. 1.

Важливим етапом побудови скорингової моделі є перевірка її достовірності й апробація на реальних даних. Про ступінь валідації моделі свідчить здатність її правильно класифікувати об'єкти, здатність моделі відрізняти "хороших" позичальників від "поганих". Модель повинна давати коректні прогнози не тільки на навчальній сукупності даних, але і на практиці при її застосуванні. Одна зі стратегій валідації моделі – формування випадковим чином двох вибірок: навчальної – по якій будується модель, і тестової – вона використовується для перевірки моделі. Перевірку достовірності моделі, як правило, проводять на навчальній і контрольній вибірках у пропорціях приблизно 70–80 і 30–20 % відповідно від вихідних даних для побудови моделі. Хороша модель повинна давати прийнятні результати прогнозування як на навчальній, так і на контрольній вибірці. Схожі показники, отримані на обох вибірках, – ознака того, що на практиці модель буде більш стабільною і даватиме коректні прогнози [8].

***Таблиця* 1.** Оцінка значущості незалежної змінної за значенням *IV*

Значення *IV*	Прогнозна здатність
$<0{,}2$	Не має
0,02–0,1	Низька
0,1–0,3	Середня
0,3–0,5	Висока
$>0{,}5$	Чудова

Найпоширеніша статистична модель для побудови скорингової карти при бінарній залежній змінній – логістична регресія. Логістична регресія, як і більшість інших методів прогнозного моделювання, використовує набір характеристик – регресорів для прогнозування ймовірності того чи іншого результату. Математична модель логістичної регресії виражає залежність логарифму шансу (логіта) від лінійної комбінації незалежних змінних (регресорів). Рівняння логіт-перетворення ймовірності події має такий вигляд:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon,$$

де p_i – імовірність настання дефолту по кредиту *i*-го позичальника, x_j – значення незалежної змінної, β_0 – незалежна константа моделі, β_k – параметри моделі, ε – компонент випадкової похибки.

Рівняння логіт-перетворення відображає лінійну залежність ймовірності настання заборгованості по кредиту залежно від значень незалежних змінних. Константа в моделі відображає природний рівень ризику настання події,

що моделюється, за умови рівності всіх незалежних змінних нулю. Значення коефіцієнтів при незалежних змінних, які відображають ступінь їх впливу на шанс дефолту в логарифмічній шкалі, використовуються для побудови скорингової карти [7].

Заключним етапом розробки скорингової моделі є переведення коефіцієнтів логістичної регресії в скорингові бали. Якщо взяти оцінки коефіцієнтів логістичної регресії і помножити їх на значення незалежних змінних, то вийде підсумковий скоринговий бал у шкалі натуральних логарифмів:

$$\textit{підсумковий бал} = \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k,$$

де x_i – значення регресорів для *i*-го позичальника, $\hat{\beta}_j$ – оцінки коефіцієнтів логістичної регресії.

Для зведення скорингових балів у лінійну шкалу використовують прийом масштабування. Масштабування не змінює прогнозної здатності скорингової карти, а лише переводить скорингові бали в нову шкалу, зручну для використання. Скоринговий бал у лінійній шкалі є відношенням шансів "хороших" позичальників до "поганих". Для масштабування необхідно насамперед задати діапазон числової шкали з мінімумом і максимумом (наприклад, від 0 до 1000). На результат масштабування також впливають два показники: кількість балів, які подвоюють шанси стати "хорошим" позичальником, і значення шкали, в якому досягається задане відношення шансів "хороших" до "поганих". Найбільш часто використовують скорингові карти, в яких кожні 20 балів подвоюють шанси стати "хорошим". Інший варіант – кожні 40 балів подвоюють шанси стати "хорошим" позичальником. Для зведення коефіцієнта логістичної регресії в скоринговий бал у лінійній шкалі застосовують таке перетворення:

$$\textit{бал} = R \cdot \hat{b}_j,$$

де R – множник, A – зміщення. Множник визначають за формулою

$$R = \frac{D}{\ln(2)},$$

де D – кількість балів, що подвоює шанси. Зміщення визначають за формулою

$$A = B - R \cdot \ln(C),$$

де B – значення на шкалі балів, у якій співвідношення шансів становить C:1.

Методи і підходи до оцінювання скорингових моделей

На етапі верифікації використовують сукупність критеріїв, засобів і процедур, що дають можливість оцінити якість побудованої скорингової моделі.

Найбільш типовими оцінками якості моделей у задачах прогнозування є середня квадратична (Mean Squared Error – *MSE*) і середня абсолютна (Mean Absolute Error – *MAE*) похибки:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(d_i - y_i)^2,$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|d_i - y_i|,$$

де N – кількість спостережень, d_i – реальне значення цільової змінної *i*-го спостереження, y_i – прогнозне значення.

Середньоквадратичний функціонал сильніше штрафує за великі відхилення порівняно із середнім абсолютним і тому більш чутливий до викидів. При використанні будь-якого з цих двох функціоналів може бути корисно проаналізувати, які об'єкти роблять найбільший внесок у загальну помилку: не виключено, що на цих об'єктах була допущена помилка при обчисленні ознак або цільової величини.

Середньоквадратична похибка підходить для порівняння двох моделей або для контролю якості під час навчання, але не дає змоги зробити висновки про те, наскільки добре модель розв'язує задачу. Наприклад, $MSE = 10$ є дуже поганим показником, якщо цільова змінна набуває значення від 0 до 1, і дуже хорошим, якщо цільова змінна лежить в інтервалі (10000, 100000). У таких ситуаціях замість середньоквадратичної помилки доцільно використовувати коефіцієнт детермінації, або коефіцієнт R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(d_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2},$$

де $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ – середнє значення цільової змінної.

Коефіцієнт детермінації – це співвідношення дисперсій основної змінної: оцінки, отриманої за моделлю, і фактичних значень. Якщо вона близька до одиниці, то модель добре прогнозує дані, якщо ж вона близька до нуля, то прогнози можна порівняти за якістю з константним прогнозуванням.

Ефективний спосіб оцінки точності моделі, навченої класифікувати вхідні дані на два класи, спирається на вивчення її ROC-кривої (Receiver Operating Characteristic – робоча характеристика приймача). ROC-крива показує залежність частки правильно класифікованих позитивних прикладів від частки неправильно класифікованих негативних прикладів [8]. Перші називаються істинно позитивними, а інші – хибно негативними. При цьому передбачається, що класифікатор має певний параметр, змінюючи який, ми будемо отримувати те чи інше розбиття на два класи. Цей параметр називають порогом відсікання (cut-offvalue). Залежно від нього отримуватимуться різні величини помилок I і II роду.

Розглянемо таблицю спряженості (confusion matrix), яка будується на основі результатів класифікації моделі та фактичної приналежності прикладів до класів (табл. 2), де:

TP (True Positives) – правильно класифіковані позитивні клієнти;

TN (True Negatives) – неправильно класифіковані негативні клієнти;

FN (False Negatives) – позитивні клієнти, класифіковані як негативні (похибка I роду);

FP (False Positives) – негативні клієнти, класифіковані як позитивні (похибка II роду).

***Таблиця* 2.** Таблиця спряженості

Результат класифікації моделі	Фактична приналежність	
	Позитивний	Негативний
Позитивний	TP (істинно позитивні)	FP (хибно позитивні)
Негативний	FN (хибно негативні)	TN (істинно негативні)

Для аналізу якості моделі частіше використовують не абсолютні показники, а відносні – частки (rates), виражені у відсотках.

Частка істинно позитивних прикладів (True Positives Rate):

$$TPR = \frac{TP}{TP + FN} \cdot 100\ \%;$$

Частка хибно позитивних прикладів (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} \cdot 100\ \%.$$

Для побудови ROC-кривої вводять ще два поняття: чутливість і специфічність моделі. Ними визначається об'єктивна цінність будь-якого бінарного класифікатора.

Чутливість (*Sensitivity*) моделі – це частка істинно позитивних випадків:

$$Sensitivity = TPR = \frac{TP}{TP + FN} \cdot 100\ \%.$$

Специфічність (*Specificity*) моделі – це частка істинно негативних випадків, які були правильно ідентифіковані моделлю:

$$Specificity = \frac{TN}{TN + FP} \cdot 100\ \% = 100\ \% - FPR.$$

Модель з високою чутливістю частіше правильно класифікує позитивні приклади, а модель з високою специфічністю, навпаки, частіше виявляє негативні приклади.

ROC-криву отримують у такий спосіб:

– для кожного значення порога відсікання, яке змінюється від 0 до 1 з кроком *dx* (наприклад, 0,01) розраховуються значення чутливості та специфічності;

– будується графік залежності: по осі *y* відкладається чутливість, по осі *x* відкладається 100 % – *Specificity* (сто відсотків мінус специфічність).

Вибір оптимального значення порога відсікання залежить від ціни здійснення похибок I і II роду при класифікації. При зниженні порога відсікання чутливість моделі буде збільшуватися, тобто здатність моделі правильно виявляти тих позичальників, які матимуть прострочення платежу. За оптимальний поріг відсікання можна покласти точку балансу між чутливістю та специфічністю.

Для порівняння різних моделей (при побудові декількох моделей з різними параметрами) можна використовувати площу під ROC-кривою – *AUC* (Area Under Curve). Площа під кривою *AUC* змінюється від 0,5 до 1 (табл. 3).

Слід зазначити, що показник площі під кривою призначений тільки для порівняльного аналізу моделей між собою. Значення площі під кривою не містить ніякої інформації про чутливість і специфічність моделі.

***Таблиця* 3.** Оцінка якості моделі за значенням площі *AUC*

Значення *AUC*	Якість моделі
0,9–1	Відмінна
0,8–0,9	Дуже добра
0,7–0,8	Добра
0,6–0,7	Середня
0,5–0,6	Незадовільна

При аналізі якості моделі за значенням площі під ROC-кривою часто обчислюють індекс *GINI*. Цей показник переводить значення площі під кривою в діапазон від 0 до 1, чим більша його величина, тим вища дискримінуюча здатність моделі. Розраховується індекс *GINI* за формулою

$$GINI = 2AUC - 1.$$

Система SAS Enterprise Miner

SAS Enterprise Miner – це спеціалізований інструмент компанії SAS Institute, вартістю до 100 000 євро в базовій комплектації, розроблений спеціально для розв'язання задач інтелектуального аналізу даних, до яких належать насамперед задачі прогнозного моделювання та виявлення структур даних [9].

Цей інструмент створений для виявлення в масивах даних інформації, необхідної для прийняття рішень. Розроблений спеціально для пошуку та аналізу прихованих закономірностей у даних (Data mining) Enterprise Miner включає в себе ефективні методи статистичного аналізу, відповідну методологію виконання проектів дослідження даних (SEMMA) і зручний графічний інтерфейс користувача. Абревіатура SEMMA створена від слів Sample (відбір даних), Explorer (дослідження відношень у даних), Modify (модифікація даних), Model (моделювання взаємозалежностей), Assess (оцінка отриманих моделей і результатів), що позначають відповідні логічні етапи аналітичного проекту [10].

Вся обробка та аналіз даних виконуються за допомогою вузлів Enterprise Miner (nodes). Для кожного типу задач (відповідно до методології SEMMA) існує низка відповідних вузлів.

Своєю чергою задача кредитного скорингу є частковим піднапрямом задач прогнозного моделювання, для розв'язання яких компанією був спеціально розроблений набір інструментів у вигляді інтегрованих компонентів SAS Credit Scoring. Це спеціальні вузли для розв'язання задач, пов'язаних із розробкою і використанням скорингових моделей.

- Scorecard – автоматично обчислює скорингові карти за результатами лог-регресійної моделі, побудованої на даних навчальної вибірки. Крім того, цей вузол надає низку звітів зі статистичними показниками щодо якості (прогнозуючої здатності) побудованої скорингової карти і дає змогу визначити оптимальний бал відсікання.
- Interactive Grouping – забезпечує автоматичний вибір найбільш значущих вхідних змінних і автоматичне (або інтерактивне) формування груп значень (характеристик) для вхідних змінних із неперервними значеннями. Для автоматичного вибору найбільш значущих вхідних змінних використовується коефіцієнт *GINI* або *Information Value*. Для автоматичного формування груп значень як критерії розбиття діапазону значень на групи використовується коефіцієнт *Weight of Evidence* (*WOE*).
- Reject Interface дає можливість доповнити навчальну вибірку даними по претендентах, яким було відмовлено у видачі кредиту з автоматичною сценарною розміткою прецедентів на позитивні/негативні (розмітка Good/Bad).

Практичний приклад побудови скорингових моделей

Для побудови скорингової карти в системі SAS Enterprise Miner було використано вибірку даних M1 з німецького банку, що надає кредити фізичним особам. Вибірка містить 3000 записів по клієнтах, у яких вже закінчився строк кредитування, та включає в себе інформацію щодо 17 показників (анкетних даних) по кожній особі. Опишемо кожен показник більш детально (табл. 4).

Процес побудови скорингової карти включав у себе такі послідовні етапи: розбиття вихідної вибірки на навчальний набір даних (70 %) і тестовий (30 %), категоризацію неперервних змінних на основі значення критерію *WOE*, відбір найбільш значущих вхідних змінних для побудови моделі за допомогою показника інформаційного значення (*Information Value*) та безпосередньо саму побудову скорингової карти (рис. 1).

Результатом моделювання є побудована скорингова карта (табл. 5), яка складається з показників (атрибутів), діапазонів значень кожного атрибута і скорингового балу по кожному з діапазонів. Для оцінки кредитоспроможності нового претендента досить підсумувати бали по кожному показнику скорингової карти.

***Таблиця* 4.** Опис змінних з вибірки M1

Назва змінної	Опис
TITLE	Стать особи
CHILDREN	Кількість дітей
AGE	Вік
NMBLOAN	Кількість кредитів, наданих цим банком
FINLOAN	Кількість закритих кредитів
INCOME	Дохід особи
EC_CARD	Наявність банківської карти
STATUS	Сімейний стан
LOANS	Кількість відкритих кредитів
REGN	Регіон
CASH	Запитувані грошові кошти
PRODUCT	Тип бізнесу
NAT	Національність особи
CAR	Наявність транспортного засобу
CARDS	Тип банківської карти
RESID	Тип місця проживання (власне/орендоване)
GB (Good/Bad)	"Хороший" чи "поганий" клієнт

***Таблиця* 5.** Фрагмент побудованої скорингової карти

Атрибут	Діапазони значень	Скорингові бали
Вік	Вік < 23	–17
	23 ≤ Вік < 28	–5
	28 ≤ Вік < 31	5
	31 ≤ Вік < 46, пропуски	13
	Вік ≥ 46	26
Наявність транспорту	Відсутній	–4
	Наявний	10
Кількість дітей	0, 23, 4, 6, 8, пропуски	8
	3, 5	6
	1	5
	2	4
Наявність банківської карти	0, пропуски	8
	1	3
Дохід, євро	Дохід < 1000, пропуски	9
	1000 ≤ Дохід < 1900	4
	1900 ≤ Дохід < 2500	5
	2500 ≤ Дохід < 3000	7
	3000 ≤ Дохід	8

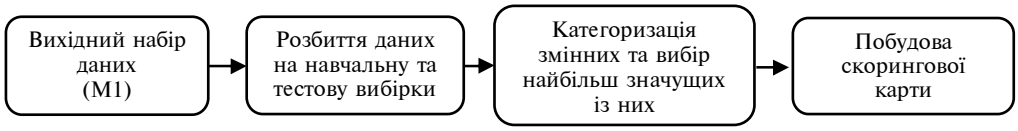

Рис. 1. Етапи побудови скорингової карти в системі SAS Enterprise Miner

Для порівняння скорингової карти з іншими статистичними та математичними методами моделювання, що дають змогу оцінити кредитоспроможність позичальників, було використано три набори даних з різних банків, а саме M1, M2 та M3, які містять у собі інформацію про клієнтів (вік, стать, кількість дітей, заробітну плату, суму кредиту тощо) та інформацію про результат погашення кредиту.

***Таблиця* 6.** Опис наборів даних

Характеристики даних	M1	M2	M3
Кількість спостережень	3000	5837	2002
Кількість показників	17	27	30

Етапи побудови та порівняння скорингових моделей зображені на рис. 2. Спочатку кожна з трьох вихідних вибірок даних була поділена на навчальний та валідаційний набори даних у пропорції 70:30 % відповідно. Після цього були побудовані скорингові моделі за допомогою таких методів: дерева рішень, логістична регресія, нейронні мережі та скорингова карта. У результаті з використанням вузла "Порівняння моделей" були обчисленні різні статистичні параметри оцінки якості моделей та вибрана найкраща модель.

Для вибору найкращої моделі було використано значення площі *AUC* під ROC-кривою та індекс *GINI*. Проаналізувавши значення отриманих статистичних показників якості моделей (табл. 7–9), можна побачити, що скорингова карта дала найкращі результати на всіх трьох вибірках.

***Таблиця* 7.** Порівняльна таблиця характеристик для скорингових моделей, побудованих на даних вибірки M1

Назва методу	Індекс *GINI*	Значення *AUC*	Якість моделі
Дерева рішень	0,403	0,701	Добра
Логістична регресія	0,389	0,695	Середня
Нейронні мережі	0,407	0,704	Добра
Скорингова карта	0,449	0,724	Добра

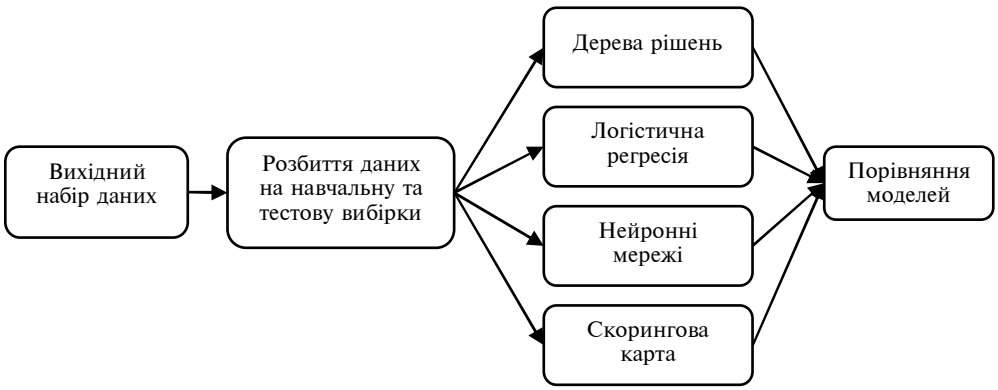

Рис. 2. Етапи побудови й аналізу скорингових моделей у системі SAS EnterpriseMiner

***Таблиця* 8.** Порівняльна таблиця характеристик для скорингових моделей, побудованих на даних вибірки М2

Назва методу	Індекс *GINI*	Значення *AUC*	Якість моделі
Дерева рішень	035	0,675	Середня
Логістична регресія	0,455	0,727	Добра
Нейронні мережі	0,433	0,716	Добра
Скорингова карта	0,529	0,764	Добра

***Таблиця* 9.** Порівняльна таблиця характеристик для скорингових моделей, побудованих на даних вибірки М3

Назва методу	Індекс *GINI*	Значення *AUC*	Якість моделі
Дерева рішень	0,317	0,658	Середня
Логістична регресія	0,632	0,816	Дуже добра
Нейронні мережі	0,639	0,82	Дуже добра
Скорингова карта	0,64	0,82	Дуже добра

Висновки

Розглянуто задачу розробки методики оцінювання кредитоспроможності потенційних клієнтів банків. Для розв'язання цієї задачі вибрано скорингову модель у вигляді скорингової карти, яка будується на результатах оцінювання логістичної регресії. Для визначення ефективності цього методу виконано порівняльний аналіз з іншими математичними та статистичними методами. Критеріями вибору кращої скорингової моделі були такі: значення площі під ROC-кривою та індекс *GINI*. Встановлено, що найкращі результати прогнозування кредитоспроможності надає скорингова карта.

Встановлено, що формат розробки прогнозних моделей у вигляді скорингових карт простіший для інтерпретації порівняно з іншими і привабливий для широкого кола ризик-менеджерів й аналітиків, які не мають глибоких знань щодо сучасних методів статистичного й інтелектуального аналізу даних. Принципи розробки скорингових карт зрозумілі більшості та відповідають нормативним вимогам до забезпечення прозорості методик. Скорингову карту дуже легко діагностувати і контролювати з використанням стандартних форм звітності. Це робить скорингову карту ефективним інструментом управління ризиками.

Отже, використання скорингових карт є досить перспективним та ефективним методом оцінювання кредитоспроможності позичальників. Це дасть можливість банкам знизити ризик неповернення виданих кредитів і відсотків за ними, а отже, позитивно відобразиться на їх фінансових результатах та сумі прибутку. Проте застосування цього методу вимагає значних капіталовкладень та налаштування міжбанківської комунікації з метою неперервного поповнення й оновлення кредитних історій позичальників. Напрямами подальших досліджень є удосконалення розглянутої методики скорингового аналізу та визначення переваг і недоліків впровадження автоматизованих скорингових систем у роботу комерційних банків. Перспективним напрямом досліджень є також використання комбінованих процедур прогнозування кредитоспроможності клієнтів на основі альтернативних, ідеологічно різних, методів аналізу даних, що забезпечують значне підвищення якості оцінок прогнозів.

Список літератури

1. *Bielecki T.R., Rutkowsky M.* Credit Risk: Modeling, Valuation, Hedging. – Berlin: Springer, 2002. – 500 p.
2. *Van Gruening H., Bratanovic S.B.* Analyzing and Managing Banking Risks. – Washington: The World Bank, 2003. – 386 p.
3. *Aven T.* Foundations of Risk Analysis: A Knowledge and Decision-Oriented Perspective. – New York: John Wiley & Sons, Ltd., 2003. – 198 p.
4. *Xuesong G., Zhengwei Z., Shi J.* Corporate credit rating model using support vector domain combined with fuzzy clustering algorithm // Math. Problems Eng. – 2012. – **1**. – P. 1–20.
5. *Лункіна Т.І.* Використання скоринг моделі при управлінні ризиками споживчого кредитування [Електронний ресурс] // Ефективна економіка. – 2015. – № 2. – URL: http://www.economy.nayka.com.ua/?op=1&z=3792
6. *Сорокин А.С.* Построение скоринговых карт с использованием модели логистической регрессии [Электронний ресурс] // Науковедение. – 2014. – Вып. 2 (21). – URL: http://naukovedenie.ru/PDF/180EVN214.pdf
7. *Сиддики Н.* Скоринговые карты для оценки кредитных рисков. Разработка и внедрение интеллектуальных методов кредитного скоринга / Пер. с англ. Е. Ильичева. – М.: Манн, Иванов и Фербер, 2014. – 268 с.
8. *Сорокин А.С.* К вопросу валидации модели логистической регрессии в кредитном скоринге [Электронний ресурс] // Науковедение. – 2014. – Вып. 2 (21). – URL: http://naukovedenie.ru/PDF/ 173EVN214.pdf
9. *Терентьев А.Н., Домрачев В.Н., Костецкий Р.И.* SAS BASE: Основы программирования. – К.: Эдельвейс, 2014. – 304 с.
10. *Anderson B.S., Thompson R.W.* Developing Credit Scorecards Using SAS Credit Scoring for Enterprise Miner 5.3. – Cary: SAS Institute Inc, 2009. – 41 p.

References

1. T.R. Bielecki and M. Rutkowsky, *Credit Risk: Modeling, Valuation, Hedging.* Berlin, Germany: Springer, 2002.
2. H. van Gruening and S.B. Bratanovic, *Analyzing and Managing Banking Risks.* Washington: The World Bank, 2003.
3. T. Aven, *Foundations of Risk Analysis: A Knowledge and Decision-Oriented Perspective.* New York: John Wiley & Sons, Ltd., 2003.
4. G. Xuesong *et al.*, "Corporate credit rating model using support vector domain combined with fuzzy clustering algorithm", *Math. Problems Eng.*, vol. 1, 2012, pp. 1–20.
5. T. Lunkina. (2015). *Using Scoring Models in Consumer Lending Risk Management* [Online]. Available: http://www.economy.nayka.com.ua/?op=1&z=3792 (in Ukrainian).
6. A. Sorokin. (2014). *Building a Scorecard Using a Logistic Regression Model.* [Online]. Available: http://naukovedenie.ru/PDF/180EVN214.pdf (in Russian).
7. N. Siddiki, *Scorecards for Credit Risk Assessment. Development and Implementation of Intelligent Methods of Credit Scoring.* Moscow, Russia, 2014 (in Russian).
8. A. Sorokin. (2014). *On the Question of Validation of a Logistic Regression Model in Credit Scoring.* [Online]. Available: http://naukovedenie.ru/PDF/ 173EVN214.pdf (in Russian).
9. A. Terentyev *et al.*, *SAS BASE: Programming Basics.* Kyiv, Ukraine: Edelveis Publishers, 2015 (in Russian).
10. B.S. Anderson and R.W. Thompson, *Developing Credit Scorecards Using SAS Credit Scoring for Enterprise Miner 5.3.* Cary: SAS Institute Inc, 2009.

С.А. Бакун, П.І. Бідюк

МЕТОДИКА ПОБУДОВИ СКОРИНГОВИХ КАРТ ІЗ ВИКОРИСТАННЯМ ПЛАТФОРМИ SAS

Проблематика. Розробка ефективних методик оцінювання кредитоспроможності осіб і ризику банків при наданні споживчих кредитів.

Мета дослідження. Визначення механізму реалізації скорингової моделі у вигляді скорингової карти. Аналіз можливостей використання методу скорингових карт як інструменту управління кредитним ризиком.

Методика реалізації. Побудова скорингової карти та попередній аналіз вихідних даних за допомогою спеціалізованих компонент системи SAS Enterprise Miner.

Результати дослідження. Розглянуто основні етапи розробки скорингових карт. Побудовано скорингову карту на основі реальних статистичних даних щодо видачі споживчих кредитів. Проведено порівняльний аналіз скорингової карти з іншими статистичними методами класифікації потенційних позичальників кредитів.

Висновки. Встановлено, що скорингові карти мають кращу прогнозну здатність стосовно платоспроможності клієнтів, ніж інші статистичні методи, такі як дерева рішень, нейронні мережі та логістична регресія. Формат розробки прогнозних моделей у вигляді скорингової карти є найбільш простим для інтерпретації. Проте застосування цього методу вимагає значних капіталовкладень і постійного поповнення та оновлення кредитних історій позичальників.

Ключові слова: управління ризиками; інтелектуальний аналіз даних; кредитний скоринг; скорингова карта; логістична регресія; якість класифікації.

С.А. Бакун, П.И. Бидюк

МЕТОДИКА ПОСТРОЕНИЯ СКОРИНГОВЫХ КАРТ С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ SAS

Проблематика. Разработка эффективных методик оценки кредитоспособности лиц и риска банков при предоставлении потребительских кредитов.

Цель исследования. Определение механизма реализации скоринговой модели в виде скоринговой карты. Анализ возможностей использования метода скоринговых карт как инструмента управления кредитным риском.

Методика реализации. Построение скоринговой карты и предварительный анализ исходных данных с помощью специализированных компонент системы SAS Enterprise Miner.

Результаты исследования. Рассмотрены основные этапы разработки скоринговых карт. Построена скоринговая карта на основе реальных статистических данных о выдаче потребительских кредитов. Проведен сравнительный анализ скоринговой карты с другими статистическими методами классификации потенциальных заемщиков кредитов.

Выводы. Установлено, что скоринговые карты имеют лучшую прогнозирующую способность относительно платежеспособности клиентов, чем другие статистические методы, такие как деревья решений, нейронные сети и логистическая регрессия. Формат разработки прогнозных моделей в виде скоринговой карты является наиболее простым для интерпретации. Однако применение этого метода требует значительных капиталовложений, а также постоянного пополнения и обновления кредитных историй заемщиков.

Ключевые слова: управление рисками; интеллектуальный анализ данных; кредитный скоринг; скоринговая карта; логистическая регрессия; качество классификации.

Рекомендована Радою
Навчально-наукового комплексу
“Інститут прикладного системного
аналізу” НТУУ “КПІ”

Надійшла до редакції
12 квітня 2016 року