УДК 303.732.4:519.254

ВДОСКОНАЛЕНІ МЕТОДИ РОЗРАХУНКУ СТАТИСТИКИ КОЛМОГОРОВА-СМИРНОВА, ВАГИ КАТЕГОРІЇ ЗМІННОЇ ТА ЗНАЧЕННЯ ІНФОРМАЦІЇ У КРЕДИТНОМУ РЕЙТИНГУ

О.М. СОЛОШЕНКО

Розроблено вдосконалені методи обчислення показників статистики Колмогорова-Смирнова, ваги категорії змінної та значення інформації без явного розбиття оригінальної вибірки на дві підмножини, з виведенням відповідних формул для аналізу предикативної (прогностичної) сили категоріальних змінних у задачах кредитного рейтингу та інших областях практичного застосування методів бінарної класифікації. Здійснено узагальнення класичних формул статистики Колмогорова-Смирнова, ваги категорії змінної та показника значення інформації шляхом перетворення агрегатних виразів для дискретних розподілів та кумулятивних функцій розподілу з застосуванням скалярного добутку векторів та операторів проектування, а також оператора умовної перестановки. Запропоновано вдосконалені формули обчислення статистики Колмогорова-Смирнова, ваги категорії змінної та індексу значення інформації, що узагальнено описуються в термінах дискретного безумовного розподілу вхідної змінної та умовного розподілу бінарної цільової змінної.

ВСТУП

Практично у всіх системах та підходах побудови кредитних рейтингових моделей — скорингових моделей (скорингових карт) у задачах ризик-менеджменту щодо моделювання кредитних ризиків, для внутрішньої задачі аналізу прогностичної (предикативної) сили вхідних характеристик з метою оцінювання доцільності їх включення у модель логістичної регресії, використовуються значення *WoE* (Weight of Evidence — вага категорії змінної) для категорій категоріальної або дискретизованої змінної та показник IV (Information Value — значення інформації) на основі відстані Кульбака-Лейблера [1]. Одним з індикаторів оцінки якості ймовірнісних (або рейтингових) прогнозів на тестовій (валідаційній) вибірці з бінарною цільовою змінною є статистика KS (Kolmogorov-Smirnov statistic — статистика Колмогорова-Смирнова) [2], яка дозволяє оцінювати нерівність функцій розподілу для двох взаємовиключних класів. Статистику KS також застосовно на етапі аналізу характеристик на навчальній вибірці як альтернативу показнику IV. У такому разі, ця статистика буде в точності відповідати показнику якості прогнозів однофакторної моделі у термінах класичних показників якості прогнозів бінарного класифікатора. Вона відображає роздільну здатність класифікатора відносно двох підмножин, що відповідають двом значенням цільової змінної, тобто якість ранжування елементів всієї множини відносно цільової змінної [2]. Цей факт пояснюється тим, що у випадку використання класичних моделей зважування факторів типу логістичної регресії, що зберігають монотонність виходу моделі відносно єдиного входу, або у випадку використання дерев рішень, що збігаються з вхідною категоріальною змінною при використанні єдиного вхідного категоріального параметру, зберіга-

ється ранжування елементів вибірки [2]. Статистику KS застосовують як для дискретних (категоріальних), так і для неперервних розподілів.

Класичні формули обчислення *WoE* та IV оперують розбиттям на два окремі умовні розподіли категорій певної змінної на власне виділених окремо класах одиничних та нульових значень бінарної цільової змінної [1], а обчислення статистики Колмогорова-Смирнова передбачає побудову емпіричних функцій розподілу безпосередньо розглядаючи всю відому неагреговану множину елементів [2].

Актуальність дослідження полягає у практичній цінності наведення відповідних формул у термінах та поняттях безумовного дискретного розподілу (total distribution) змінної, що аналізується, та у термінах умовних ймовірностей нульових значень цільової змінної (bad rate) за кожною з категорій змінної, що аналізується, оскільки два наведені розподіли найбільш ілюстративні для відображення таблиць та графіків аналізу характеристик (зокрема, групи ризику фінального рейтингового балу) [2]. Також актуальність розроблення формул розрахунку ключових показників предикативності категоріальних змінних у кредитному скорингу (рейтингу) саме за допомогою різноманітних агрегатних показників обумовлена новітніми технологіями розробки баз даних. Вони не відповідають реляційній моделі та призначені для роботи з надзвичайно великими масивами даних [3]. Ще одним аспектом актуальності пропонованих методів є забезпечення можливості точної кількісної оцінки ключових індикаторів, використовуючи лише класичні таблицю та графік аналізу характеристик [2]. Також з використанням альтернативних формул можлива організація додаткової перевірки коректності розрахунку даних статистичних показників, відновлення індикаторів властивостей оригінальної вибірки. Альтернативні формули відображають важливу інтерпретацію числових значень ваги категорії змінної.

ПОСТАНОВКА ЗАДАЧІ

Об’єктами дослідження є класичні формули KS, *WoE* та IV.

Предметом дослідження є методи перетворення агрегатних виразів для дискретних розподілів та кумулятивних функцій з застосуванням скалярного добутку та операторів проектування, а також оператора умовної перестановки.

Мета роботи — наведення вдосконалених методів обчислення показників статистики Колмогорова-Смирнова, ваги категорії змінної та значення інформації без явного розбиття оригінальної вибірки на дві підмножини, з виведенням відповідних формул для аналізу предикативної (прогностичної) сили категоріальних змінних у задачах кредитного скорингу та інших областях практичного застосування методів бінарної класифікації. Тобто необхідно розробити вдосконалені методи розрахунку ключових показників предикативної сили довільної категоріальної змінної у кредитному скорингу за відомих вхідних ймовірностях безумовного дискретного розподілу категоріальної змінної та умовних ймовірностях частоти нульових значень цільової змінної, тобто за узгодження вхідних векторів ймовірностей розподілу вхідної змінної та умовних ймовірностей цільової змінної.

КЛАСИЧНІ МЕТОДИ ОЦІНЮВАННЯ KS, WOE ТА IV

Класичні формули для обчислення показників IV та *WoE* за відомих категорій та відомих значеннях цільової змінної кожного елемента множини вибірки для певної дискретної або дискретизованої вхідної змінної (категоріальної), щоб оцінити предикативну (прогностичну) силу вхідної характеристики, мають такий вигляд [1, 4]:

$$WoE_i = \ln\left(\frac{g_i}{b_i}\right),$$

$$IV = \sum_{i=1}^{c}(g_i - b_i)\ln\left(\frac{g_i}{b_i}\right) = \sum_{i=1}^{c}(g_i - b_i)\mathrm{WoE}_i\,.$$

Категоріальний показник g_i — це відносна кількість елементів з одиничним («good») бінарним цільовим результатом у сегменті категорії до загальної кількості елементів з одиничним цільовим результатом всіх категорій:

$$g_i = \frac{G_i}{\sum_{i=1}^{c} G_i} = \frac{G_i}{G}\,.$$

Тобто оперуємо розподілом елементів з одиничним цільовим результатом за дискретними або дискретизованими значеннями змінної (категоріями), тому має місце рівність:

$$\sum_{i=1}^{c} g_i = 1\,.$$

Аналогічно, категоріальний показник b_i — це відносна кількість елементів з нульовим («bad») цільовим результатом у сегменті категорії до загальної кількості елементів з нульовим цільовим результатом всіх категорій:

$$b_i = \frac{B_i}{\sum_{i=1}^{c} B_i} = \frac{B_i}{B}\,.$$

Тобто також оперуємо розподілом елементів з нульовим цільовим результатами за дискретними або дискретизованими значеннями змінної (категоріями), тому має місце рівність:

$$\sum_{i=1}^{c} b_i = 1\,.$$

Взаємозв'язок значення інформації з відстанню Кульбака-Лейблера у теорії інформації [5] описується рівністю значення інформації сумі двох несиметричних відстаней Кульбака-Лейблера відносно кожного з розподілів [1, 5]:

$$IV = \sum_{i=1}^{c} g_i \ln\left(\frac{g_i}{b_i}\right) + \sum_{i=1}^{c} b_i \ln\left(\frac{b_i}{g_i}\right) = D_{KL}(\vec{g},\vec{b}) + D_{KL}(\vec{b},\vec{g})\,.$$

Класична формула обчислення статистики Колмогорова-Смирнова має такий вигляд [2]:

$$KS = \max_{x \in X} | F_B(x) - F_G(x) | .$$

Основною модифікацією статистики Колмогорова-Смирнова, що використовується на практиці, є показник рівня статистичної значимості (*p*-value) для розподілу Колмогорова, що пов'язаний з поняттям броунівського мосту [6]. Значення рівня статистичної значимості (*p*-value) записується з використанням функції розподілу таким чином:

$$PV = 1 - F(KS) = 1 - \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 KS^2} =$$

$$= 1 - \left(1 + 2 \sum_{j=1}^{\infty} (-1)^j e^{-2j^2 KS^2} \right),$$

$$PV = -2 \sum_{j=1}^{\infty} (-1)^j e^{-2j^2 KS^2} .$$

Перевагами використання статистики Колмогорова-Смирнова та відповідного значення статистичної значимості є двостороння обмеженість (на відміну від значення інформації), наочність (оскільки статистика Колмогорова–Смирнова — це максимальна абсолютна різниця функцій розподілу на спільній області визначення), зв'язок з поняттям броунівського мосту [6]. Основним недоліком статистики Колмогорова-Смирнова є відображення різниці між розподілами за допомогою максимуму, а не інтегрального показника, прикладом якого може слугувати індекс Джині [4].

ВДОСКОНАЛЕНІ МЕТОДИ ОЦІНЮВАННЯ KS, *WOE* ТА IV

У наведеній задачі мають місце вхідні вже агреговані дані без наведення оригінальної множини вибірки — матриця $M = \left(\vec{t} \quad \vec{p} \right)$ розмірності $c \times 2$, перший стовпець якої відповідає безумовному розподілу категорій вхідної змінної (total distribution), а другий — умовним ймовірностям частот елементів з нульовими значеннями бінарної цільової змінної (bad rate).

Має місце така рівність:

$$\sum_{i=1}^{c} t_i = 1.$$

Відповідну ймовірнісну вхідну матрицю зручно представити у вигляді графіку аналізу вхідної характеристики відносно бінарної цільової змінної, де гістограмі відповідає безумовний розподіл категорій вхідної характеристики, а ламаній лінії — відсоток елементів з нульовим цільовим результатом (умовна ймовірність). Наведемо приклад розподілу клієнтів банку за інтервалами віку клієнта та відсоток випадків некредитоспроможності для кожної вікової категорії (рисунок).

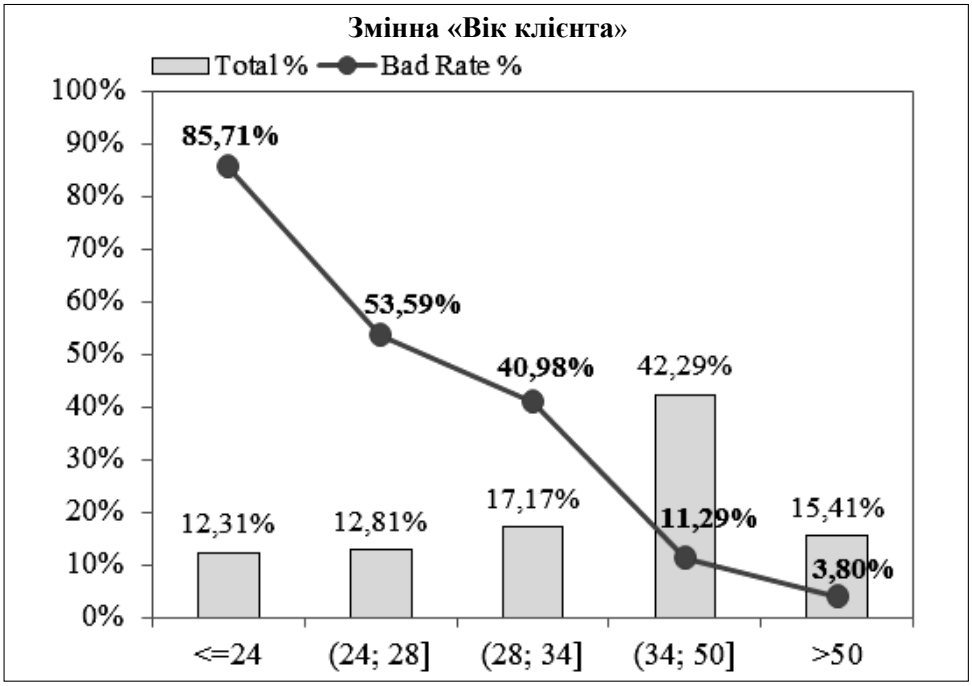

Рисунок. Графік аналізу характеристики «вік клієнта» з наведенням розподілу вибірки за категоріями та графік відсотку некредитоспроможних випадків для категорій

Щодо аналізу характеристик насамперед перепишемо формулу *WoE*, використовуючи «bad rate» на рівні категорії та середньозважений «bad rate»:

$$WoE_i = \ln\left(\frac{g_i}{b_i}\right) = \ln\left(\frac{\frac{G_i}{G}}{\frac{B_i}{B}}\right) = \ln\left(\frac{B}{G}\right) - \ln\left(\frac{B_i}{G_i}\right),$$

$$WoE_i = \ln\left(\frac{\frac{B}{B+G}}{1-\frac{B}{B+G}}\right) - \ln\left(\frac{\frac{B_i}{B_i+G_i}}{1-\frac{B_i}{B_i+G_i}}\right), \quad WoE_i = \ln\left(\frac{p_{w.\text{avg}}}{1-p_{w.\text{avg}}}\right) - \ln\left(\frac{p_i}{1-p_i}\right),$$

$$WoE_i = \ln\left(\frac{\frac{\sum_{j=1}^{c} p_j t_j}{\sum_{k=1}^{c} t_k}}{1-\frac{\sum_{j=1}^{c} p_j t_j}{\sum_{k=1}^{c} t_k}}\right) - \ln\left(\frac{p_i}{1-p_i}\right),$$

$$WoE_i = \ln\left(\frac{(\vec{p},\vec{t})}{1-(\vec{p},\vec{t})}\right) - \ln\left(\frac{p_i}{1-p_i}\right). \quad (1)$$

Враховано рівність:

$$\sum_{i=1}^{c} t_i = 1 .$$

Середньозважений за допомогою скалярного добутку «bad rate» відповідає загальному «bad rate» на всій множині вибірки $\frac{B}{B+G}$.

Наведемо вдосконалену формулу IV:

$$IV = \sum_{i=1}^{c}(g_i - b_i)WoE_i = \sum_{i=1}^{c}\left(\frac{(1-p_i)t_i}{\sum_{l=1}^{c}(1-p_l)t_l} - \frac{p_i t_i}{\sum_{r=1}^{c} p_r t_r}\right)WoE_i .$$

Ще раз врахуємо рівність:

$$\sum_{i=1}^{c} t_i = 1 .$$

Остаточна формула IV:

$$IV = \sum_{i=1}^{c}\left(\frac{(1-p_i)t_i}{1-(\vec{p},\vec{t})} - \frac{p_i t_i}{(\vec{p},\vec{t})}\right)\left(\ln\left(\frac{(\vec{p},\vec{t})}{1-(\vec{p},\vec{t})}\right) - \ln\left(\frac{p_i}{1-p_i}\right)\right). \quad (2).$$

Для прикладу з рисунку округлене значення IV буде дорівнювати 1,97.

Остаточно суть методів обчислення Weight of Evidence (1) та Information Value (2) полягає у використанні скалярного добутку вектору розподілу категорій змінної та відповідного вектору умовних ймовірностей, що відображає ймовірності набуття нульового значення для цільової змінної, а також у використанні інших перетворень від відповідних векторів агрегатних даних. Скалярний добуток відповідає середньому значенню ймовірності набуття нульового значення цільової бінарної змінної на всій вибірці.

Позначимо псевдопроектор з простору R^n на підпростір меншої розмірності R^m, що відповідає m першим координатам, як $P_{n,m}$. Суть псевдопроектора відображається такою формулою:

$$P_{n,m}: \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \\ \dots \\ x_n \end{pmatrix} \to \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}.$$

Цей псевдооператор проектування можна зобразити у вигляді матричного оператора (прямокутної матриці):

$$P_{n,m}:\begin{pmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 \end{pmatrix} \in \mathrm{Mat}(m \times n).$$

Властивість елементів матриці:

$$p_{i,j} = \begin{cases} 1, \ (i = j) \wedge (i \le m); \\ 0, \quad \text{else}. \end{cases}$$

Введемо позначення оператора проектування перших m координат на підпростір розмірності m з довільного простору не меншої розмірності, ніж його образ, тобто довільної (без фіксації) розмірності області визначення n, де $n \ge m$, як P_m.

Основна властивість будь-якого проектора, що збігається з його означенням описується такою формулою (запишемо для P_m) [7]:

$$P_m^2 = P_m.$$

Детальніше основну властивість проектора запропонованого типу можна описати таким чином:

$$\forall n \ge m \quad \forall \vec{x} \in R^n : P_m(P_m(\vec{x})) = P_{m,m}(P_{n,m}(\vec{x})) = P_{n,m}(\vec{x}) = P_m(\vec{x}) \in R^m.$$

Суть означення оператора проектування полягає у властивості ідемпотентності — точній рівності «проекції від проекції» власне значенню проекції [7, 8].

Також лінійний оператор проектування P можна означити як такий, що задається квадратною матрицею $n \times n$, тобто, коли розмірність образу збігається з розмірністю області визначення, а порядок (набір) координат, що проектуються, може бути довільним, при цьому проектування може відбуватись за допомогою лінійних комбінацій координат [7, 8]. Тоді основна властивість $(P^2 = P)$, що збігається з означенням оператора проектування, можлива, наприклад, завдяки таким умовам на елементи матриці такого оператора [7, 8]:

$$\begin{cases} i = j : p_{i,j} \in \{1; 0\}, \\ i \ne j : p_{i,j} = 0. \end{cases}$$

Також для означення лінійного оператора проектування за допомогою квадратної матриці можливе використання довільної ідемпотентної матриці [8].

Надалі будемо використовувати лише P_m — вищеозначений оператор проектування перших m координат на підпростір розмірності m з простору довільної нефіксованої розмірності $n \ge m$. Також формули нижче будуть справедливими у випадку використання замість оператора P_m також звичайних квадратних діагональних матриць P_c^m розмірністю $c \times c$, що мають діагональ з першими m елементами рівними одиниці, а іншими діагональ-

ними елементами рівними нулю, при цьому виконується властивість (означення) проектора: $(P_c^m)^2 = P_c^m$.

Введемо оператор ранжування (перестановки) одного вектора як перестановку його координат, що відповідає сортуванню другого вектора по спаданню координат $R(\vec{x},\vec{y}): R^n \times R^n \to R^n$.

Суть оператора сортування першого вектора відносно другого вектора по спаданню координат можна представити через функцію рангу $r(i,\vec{x})$, яку визначено на натуральних числах (але не більше розмірності власне вектора), що повертає початковий номер позиції координати ще не відсортованого вектора для заданої як аргумент координати вже відсортованого по спаданню вектора:

$$R(\vec{x},\vec{y}): \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \to \begin{pmatrix} x_{r(1,\vec{y})} \\ x_{r(2,\vec{y})} \\ \dots \\ x_{r(n,\vec{y})} \end{pmatrix}.$$

Суть запропонованого методу обчислення статистики Колмогорова-Смирнова полягає у використанні скалярного добутку, спеціального оператора сортування та операторів проектування й описується таким чином:

$$KS = \max_{i=1,..c} \left| \frac{(P_i(R(\vec{t},\vec{p})), P_i(R(\vec{p},\vec{p})))}{(\vec{t},\vec{p})} - \frac{(P_i(R(\vec{t},\vec{p})), P_i(R(\vec{e}-\vec{p},\vec{p})))}{(\vec{t},\vec{e}-\vec{p})} \right|. \quad (3).$$

Тут $\vec{e}$ є одиничним вектором розмірності c.

Введемо позначення для композиції операторів перестановки та проектування:

$$P_{\vec{p},i}(\bullet) = P_i(R(\bullet,\vec{p})). \quad (4).$$

Враховуючи позначення (4), формулу (3) можна переписати таким чином:

$$KS = \max_{i=1,..c} \left| \frac{(P_{\vec{p},i}(\vec{t}), P_{\vec{p},i}(\vec{p}))}{(\vec{t},\vec{p})} - \frac{(P_{\vec{p},i}(\vec{t}), P_{\vec{p},i}(\vec{e}-\vec{p}))}{(\vec{t},\vec{e}-\vec{p})} \right|. \quad (5).$$

Для прикладу з рисунку округлене значення KS для дискретизованої змінної «вік клієнта» буде дорівнювати 56,60%.

ВИСНОВКИ

Запропоновано альтернативні методи обчислення та формули розрахунку статистики KS, *WoE* та IV виходячи з відомого розподілу категорій та відомих умовних ймовірностей нульових значень цільової змінної, що дозволяють аналізувати характеристики та предикативну силу на навчальній та довільній вибірках, маючи лише відповідний графік агрегованих відносних значень. Предикативна сила змінної згідно з класичними методами скорингу [1, 2, 4] дорівнює якості прогнозів однофакторної моделі для категоріальної змінної, оскільки використання монотонних функцій типу логістичного пе-

ретворення або оптимального на навчальній вибірці дерева рішень, що точно відповідатиме власне категоріальній змінній, не змінює порядок категорій відносно умовного розподілу цільової змінної. При цьому присвоєний скоринговий бал відносно довільної шкали у випадку використання логістичної регресії буде монотонною функцією від умовного розподілу цільової змінної — долі нульових значень у категорії.

Основною відмінністю та практичною цінністю запропонованих формул відносно класичних є оперування лише агрегатними величинами без використання розбиття на дві окремі підмножини.

Ключовими особливостями запропонованих методів є використання скалярного добутку векторів з метою зважування величин, умовних перестановок та операторів проектування.

Перевагами запропонованих методів розрахунку статистичних показників прогностичної сили категоріальних змінних є:

- відсутність необхідності розбиття початкової вибірки на дві підмножини, що відповідають двом значенням бінарної змінної;
- використання безумовного дискретного розподілу категоріальної змінної та умовного розподілу бінарної цільової змінної, що відповідає класичним таблиці та графіку аналізу характеристик у кредитному скорингу [2];
- можливість швидкого розрахунку статистичних показників лише на основі наявних агрегованих даних класичного аналізу категоріальної змінної та можливість організації додаткової перевірки розрахунків згідно з класичними формулами;
- математична наочність запропонованих формул в поняттях скалярного добутку, операторів проектування та умовних перестановок;
- можливість точного відновлення значень, що описують детальні властивості оригінальної вибірки (наприклад, *WoE*), за агрегатними ймовірнісними показниками аналізу характеристик, які явно не використовуються в класичних формулах розрахунку значень KS, *WoE* та IV, але мають місце в запропонованих альтернативних формулах;
- зручність використання запропонованих методів у термінах агрегатів нереляційних систем керування базами даних [3], що дозволяє проводити аналіз характеристик та підрахунок показників, які розглядаються, одночасно, а не послідовно, із забезпеченням високої швидкодії на надзвичайно великих масивах даних (Big Data) [3].

Ще одним важливим висновком для ризик-менеджменту [1, 2, 4] є більш наочна інтерпретація показника *WoE* за допомогою запропонованої формули (1) як ступеню відхилення долі нульових значень цільової змінної по окремій категорії вхідної змінної відносно загальної (середньозваженої) долі нульових значень цільової змінної на всій вибірці. Згідно з інтерпретацією, негативне значення *WoE* означає перевищення відносно середнього значення на всій вибірці, а позитивне значення *WoE* означає, що значення долі нульових значень цільової змінної по даній категорії (bad rate) нижче, ніж на всій вибірці, нульове — точна рівність долі по категорії долі на всій вибірці.

Перспективи подальших досліджень містять вдосконалення методів обчислення інших показників предикативності (прогностичної сили) змінних, використовуючи лише агреговані показники значень ймовірностей (умовних та безумовних), а також застосування математичної методології кредитного скорингу поза межами управління ризиками.

ЛІТЕРАТУРА

1. *Siddiqi Naeem*. Credit risk scorecards: developing and implementing intelligent credit scoring. — Hoboken: John Wiley & Sons, Inc., 2006. — 196 p.
2. *Мэйз Элизабет*. Руководство по кредитному скорингу. — Минск: Гревцов Паблишер, 2008. — 464 с.
3. *Фаулер Мартин, Садаладж Дж. Прамодкумар.* NoSQL: новая методология разработки нереляционных баз данных. — Минск: ООО «И.Д. Вильямс», 2013. — 192 с.
4. *Thomas C. Lyn, Edelman B. David, Crook N. Jonathan.* Credit Scoring and its Applications. — Philadelphia: Society for Industrial and Applied Mathematics, 2002. — 248 p.
5. *Kullback Solomon.* Information Theory and Statistics. — Hoboken, NJ: John Wiley & Sons, 1959. — 395 p.
6. *Булинский А.В., Ширяев А.Н.* Теория случайных процессов. — М.: Физматлит, 2005. — 408 с.
7. *Треногин В.А.* Функциональный анализ. — М.: Наука, 1980. — 495 с.
8. *Мальцев А.И.* Основы линейной алгебры. — М.: Наука, 1975. — 400 с.

Надійшла 22.09.2014