pISSN 2071-4688
eISSN 2311-8709

Банковская деятельность

СОВРЕМЕННЫЕ ПОДХОДЫ К ПРИМЕНЕНИЮ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ЗАДАЧЕ КРЕДИТНОГО СКОРИНГА

Елена Сергеевна ВОЛКОВА[a,•], Владимир Борисович ГИСИН[b], Владимир Игоревич СОЛОВЬЕВ[c]

[a] кандидат физико-математических наук, доцент Департамента анализа данных, принятия решений и финансовых технологий, Финансовый университет при Правительстве РФ, Москва, Российская Федерация
EVolkova@fa.ru

[b] кандидат физико-математических наук, профессор Департамента анализа данных, принятия решений и финансовых технологий, Финансовый университет при Правительстве РФ, Москва, Российская Федерация
VGisin@fa.ru

[c] доктор экономических наук, профессор, руководитель Департамента анализа данных, принятия решений и финансовых технологий, Финансовый университет при Правительстве РФ, Москва, Российская Федерация
VSoloviev@fa.ru

• Ответственный автор

История статьи:
Получена 04.07.2017
Получена в доработанном виде 09.08.2017
Одобрена 24.08.2017
Доступна онлайн 15.09.2017

УДК 519.226
JEL: C38, C55, D81

Ключевые слова:
кредитный скоринг, машинное обучение, интеллектуальная обработка данных

Аннотация

Предмет. Рост спроса на потребительские кредиты привел к увеличению конкуренции на рынке кредитования. Банки и другие кредитные институты столкнулись с необходимостью обрабатывать большие объемы данных со все возрастающей скоростью. Современные требования к объему обрабатываемых данных и скорости их обработки таковы, что процессы должны быть практически полностью автоматизированы. Эти требования распространяются не только на непосредственную цифровую обработку, но и на процедуры настройки, адаптации и даже построения соответствующих количественных моделей. Традиционно используемые в кредитовании модели, такие как скоринг, стали комбинироваться с новыми вычислительными методами, которые относят к области так называемого машинного обучения или интеллектуального анализа данных. В статье приводится обзор современного состояния исследований в этой области.

Цели. Классификация современных методов кредитного скоринга. Описание моделей сравнения эффективности его различных методов.

Методология. Изучение актуальных научных публикаций по теме статьи, представленных в Google Scholar.

Результаты. Представлена классификация современных методов интеллектуального анализа данных, применяемых в кредитном скоринге.

Выводы. Требующийся в современных условиях уровень эффективности могут обеспечить модели кредитного скоринга, использующие процедуры машинного обучения и гибридные модели, в которых применяются комбинированные методы.

Для цитирования: Волкова Е.С., Гисин В.Б., Соловьев В.И. Современные подходы к применению методов интеллектуального анализа данных в задаче кредитного скоринга // Финансы и кредит. – 2017. – Т. 23, № 34. – С. 2044 – 2060.
https://doi.org/10.24891/fc.23.34.2044

Введение

Кредитный скоринг может быть определен как технология, позволяющая кредитной организации решить вопрос о предоставлении кредита заявителю с учетом его характеристик, таких как возраст, доход, семейное положение и др. Естественно, подобные технологии возникли вместе с появлением торговли и потребностью в кредитовании. Идеи и методы скоринга, соответствующие их современному пониманию, были впервые сформулированы в работе Д. Дюрана [1].

После принятия соглашений Базель II (и особенно Базель III) стало возможным и необходимым применять процедуры

внутреннего рейтинга для оценки общих параметров риска. Это сделало более значительной роль кредитного скоринга и заставило финансовые институты постоянно совершенствовать используемые ими количественные модели.

Достаточно полное представление о работах, посвященных классическим методам кредитного скоринга, дает статья Д. Хенда и В. Хенли [2]. Обзоры более поздних публикаций можно найти в статьях В. Гарсии [3] и С. Лессмана [4] с соавторами. Многочисленные обзоры посвящены отдельным технологиям кредитного скоринга и сравнительному анализу применяемых методов.

Настоящий обзор посвящен прежде всего работам, в которых применяются методы кредитного скоринга, основанные на методологии интеллектуального анализа данных. В последние годы значительно возросло число публикаций, в которых описываются так называемые гибридные методы. В разделе 1 приводится краткий обзор основных методов кредитного скоринга. В разделе 2 дается краткое описание распространенных тестовых наборов данных для сравнения эффективности методов кредитного скоринга. В разделах 3 и 4 описано, каким образом можно сравнивать различные модели и методы кредитного скоринга. В разделе 5 содержится анализ программной реализации алгоритмов машинного обучения.

1. Основные методы машинного обучения в кредитном скоринге

1.1. Линейная регрессия

Линейная регрессия связывает характеристики заемщика, представленные вектором $x \in R^n$ с целевой переменной $y \in \{-1; 1\}$:

$$y = \beta_0 + \langle \beta, x \rangle + \varepsilon,$$

где ε – случайная ошибка с нулевым средним. При решении вопроса об отнесении y к тому или иному классу величина $\beta_0 + \langle \beta, x \rangle$ трактуется как условное математическое ожидание $E(y|x)$. В работе Д. Хенда и М. Келли [5] на основе линейной регрессии построены скоринговые карты. Отметим, что в последние годы линейная регрессия в чистом виде не используется, хотя по-прежнему служит важным инструментом в смешанных моделях.

1.2. Логистическая регрессия

Логистическая регрессия – один из основных инструментов кредитного скоринга. В публикациях логистическая регрессия, как правило, используется для сравнения с другими методами (например, в работах Б. Йапа с соавторами [6], Н. Павлидиса с соавторами [7], З. Хемайса с соавторами [8]) или в комбинации с другими методами, (работы Ф. Лузады с соавторами [9] и З. Ли [10]).

Логистическая регрессия используется в кредитном скоринге для вычисления вероятности $P(y=1|x)$ отказа выдачи заемщику, имеющему характеристики x. Вероятность представляется в виде

$$P(y=1|x) = \frac{1}{1+e^{-(\alpha+\beta^T x)}}.$$

Для оценки коэффициентов α и β_i (координаты вектора β) используется метод максимального правдоподобия. Оценка проводится на обучающем множестве.

1.3. Дискриминантный анализ

Дискриминантный анализ – один из наиболее популярных методов скоринга, и кредитного скоринга в частности. Дискриминантный анализ восходит к работе Р. Фишера [11]. Это был один из первых методов, применявшихся для построения систем кредитного скоринга. Проблемы, связанные с применением дискриминантного анализа в кредитном скоринге, проанализированы в статье Р. Эйзенбейса [12]. В настоящее время дискриминантный анализ продолжает использоваться в кредитном скоринге непосредственно [13]. Дискриминантный анализ часто служит эталоном, с которым сравниваются другие методы, как это делается, например, в статье С. Аккоса [14]. Ряд исследований связан с повышением точности дискриминантного анализа за счет применения новых процедур [15].

1.4. Деревья принятия решений

Метод берет свои истоки в работе Л. Бреймана с соавторами [16]. Представление о современном состоянии дел дает работа В. Лоха [17]. Если говорить о кредитном скоринге, то здесь деревья применяются в основном для классификации [4].

Опишем коротко суть методов, связанных с построением деревьев. Говорят, что переменная X порядковая, если принимаемые ею числовые значения упорядочены существенным для классификации образом. В противном случае переменную называют категориальной. Алгоритм автоматического обнаружения взаимосвязи (Automatic Interaction Detector, AID) – один из первых алгоритмов построения классификационных деревьев – последовательно расщепляет данные в каждом узле. В случае порядковой переменной ветвление происходит по условиям вида $X \le c$, в случае категориальной переменной – по условиям вида $X \in A$. Пусть $S(t)$ – множество номеров данных в обучающей выборке, относящихся к узлу t. Обозначим через $\overline{y_t}$ среднее (по $S(t)$) значение объясняемой переменной Y. Величина $imp(t) = \sum_{i \in S(t)} (y_i - y_t)^2$ служит показателем загрязненности узла t. Алгоритм AID выбирает такое расщепление, которое минимизирует сумму показателей загрязненности по непосредственно следующим узлам. Процесс расщеплений заканчивается, когда снижение загрязненности становится меньше предустановленного порога.

Алгоритмы типа THAID (Theta Automatic Interaction Detector) распространяют описанный метод на категориальные переменные. Здесь в качестве показателя загрязненности используется энтропия или индекс Джини. В более современных алгоритмах CART (Classification And Regression Trees) правила остановки, применяемые в алгоритмах AID и THAID, заменены правилами выращивания и удаления новых ветвей. Применяются также алгоритмы CHAID (Chi-squared Automatic Interaction Detector) и C4.5. В статье С. Финлея [18] приведены сравнительные характеристики различных алгоритмов кредитного скоринга, включая и алгоритмы CART. Отмечено, что алгоритмы CART уступают другим. Однако некоторые новые идеи и усовершенствования моделей, связанных с построением деревьев, позволяют существенно повысить их эффективность (работы Д. Джанга с соавторами [19] и К. Ху с соавторами [20]).

К алгоритмам, связанным с деревьями принятия решений, примыкают алгоритмы извлечения знаний (Rule Extraction, RX), ориентированные на работу с большими данными (И. Хайаши с соавторами [21]).

1.5. Метод опорных векторов

Метод опорных векторов как метод статистической классификации был предложен в работе В. Вапника с соавторами [22]. Суть метода состоит в следующем. Пусть задано обучающее множество $\{(x^{(j)}, y^{(j)})\}_{j=1,2,\ldots,l}$, где $x^{(j)} \in X \subset R^n$ – признаковое описание объекта, $y^{(j)} \in \{-1; 1\}$ – бинарный классификатор. Уравнение вида $\langle w, x \rangle - w_0 = 0, w \in R^n$ задает гиперплоскость с нормальным вектором w, разделяющую в пространстве R^n классы «хороших» $y^{(j)} = 1$ и «плохих» $y^{(j)} = -1$ объектов.

Оптимальная разделяющая гиперплоскость определяется как решение оптимизационной задачи:

$$\|w\| \to \min;$$
$$y^{(j)}(\langle w, x^{(j)} \rangle - w_0) \ge 1, j = 1,2,\ldots,l.$$

В случае, когда разделяющая гиперплоскость существует, величина $\frac{2}{\|w\|}$ – ширина полосы между точками разных классов. Задача нахождения оптимальной разделяющей гиперплоскости может быть решена с использованием теоремы Куна–Таккера. Пусть

$$L(w, w_0, \lambda) = \frac{1}{2}\langle w, w \rangle -$$

$$- \sum_{j=1}^{l} \lambda_j (y^{(j)}(\langle w, x^{(j)} \rangle - w_0) - 1) -$$

соответствующая функция Лагранжа. Объект обучающей выборки $x^{(j)}$ называется опорным

вектором, если $\lambda_j>0$ и $\langle w, x^{(j)}\rangle - w_0 = y^{(j)}$. Вектор w является линейной комбинацией опорных векторов:

$$w=\sum_j \lambda_j y^{(j)} x^{(j)}.$$

Таким образом, для фактического построения вектора w используется сравнительно небольшое число объектов обучающей выборки. Это свойство разреженности отличает метод опорных векторов от классических линейных разделителей типа дискриминанта Фишера.

Если разделяющая плоскость не существует (обучающая выборка линейно не разделима), постановка оптимизационной задачи корректируется: к целевой функции добавляется сумма штрафов за ошибки.

Возможен также переход к нелинейному разделителю с использованием ядра. Под ядром понимается функция $K(x, x')$, $x, x' \in X$ такая, что $K(x, x')=\langle \varphi(x), \varphi(x')\rangle$ для некоторого отображения $\varphi: X \to R^m$. При использовании отображения φ линейный разделитель можно строить в пространстве R^m [23].

Задача квадратичной оптимизации в методе опорных векторов может быть сформулирована в двойственной форме: найти

$$\max_{\lambda}\left(\sum_j \lambda_j + \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j y^{(i)} y^{(j)} K\left(x^{(i)}, x^{(j)}\right)\right)$$

при выполнении условий $0 \le \lambda_j \le C_j$ для всех j и $\sum_j \lambda_j y_j = 0$.

Параметры C_j контролируют относительную ценность показателей. Наиболее употребительны следующие ядерные функции:

$K\left(x^{(i)}, x^{(j)}\right)=\langle x^{(i)}, x^{(j)}\rangle$ – линейная модель;

$K\left(x^{(i)}, x^{(j)}\right)=\left(\langle x^{(i)}, x^{(j)}\rangle+1\right)^d$ – полиномиальная модель степени d;

$K\left(x^{(i)}, x^{(j)}\right)=\exp\left(-\frac{\|x^{(i)}-x^{(j)}\|}{2\sigma^2}\right)$ – гауссова радиальная базисная функция (RBF) с параметром σ.

Для нового объекта предсказание строится по формуле $y = sgn\left(\sum_j \lambda_j y^{(j)} K\left(x^{(i)}, x\right)+b_j\right)$, где $b_j=\sum_j \lambda_j y^{(j)} K\left(x^{(i)}, x^{(j)}\right)$.

Работа В. Шеня с соавторами [24] – одна из первых, в которой метод опорных векторов использовался для решения задачи кредитного скоринга. Система опорных векторов относительно семейства ядер использовалась для кредитного скоринга в работе И. Линга с соавторами [25].

1.6. Байесовские сети

Отправной точкой для применения байесовских сетей в кредитном скоринге послужила, вероятно, работа Н. Фридмана с соавторами [26]. В этом исследовании приведено обобщение так называемого простого (наивного) байесовского метода, в соответствии с которым выбирается решение с наибольшей апостериорной информацией. Применение наивного байесовского метода обосновано в случае, когда атрибуты независимы. Как отмечают авторы, в кредитном скоринге это предположение нереалистично: например, нельзя игнорировать взаимосвязь таких показателей, как возраст, образование, доход. Авторы развили идеи работы Дж. Пирла [27]. В самом общем виде байесовская сеть представляет собой ациклический ориентированный граф. При обучении формируются условные распределения вероятности вида $P(Y|X_1, \ldots, X_k)$, где Y – вершина, а $X_1, \ldots, X_k$ – ее «родители» на графе. Байесовская сеть определяет совместное распределение вершин. Например, наивный Байесовский метод получается, если взять категориальную переменную в качестве корневой вершины, а все атрибуты – в качестве ее «детей». Неформально обучение байесовской сети состоит в ее максимальной адаптации к обучающему набору данных. Оптимизация проводится относительно скоринговой функции. Наиболее употребительными являются байесовская скоринговая функция и функция, основанная на принципе минимальной длины описания (MDL). Эти функции асимптотически приводят к одинаковому результату обучения,

однако на конечных наборах функция MDL зарекомендовала себя лучше.

Пусть $B=(G,\Theta)$ – Байесовская сеть (G – граф, Θ – соответствующее распределение вероятностей), а $D=\{u_1,\ldots,u_n\}$ – обучающий набор (каждое u_i присваивает значения всем вершинам графа). Тогда $MDL(B|D)=\frac{\log N}{2}|B|-LL(B|D)$, где $|B|$ – число параметров сети, а $LL(B|D)=\sum_{i=1}^{N}\log(P_B(u_i))$ измеряет объем информации, необходимой для того, чтобы описать D, основываясь на распределении вероятностей P_B.

Скоринговая функция *MDL* асимптотически корректна.

Укажем несколько работ, в которых Байесовские сети применялись для кредитного скоринга: П. Джиудичи [28], Дж. Гемела [29], А. Антонакис с соавторами [30–31], И. Ву с соавторами [32], Х. Жу с соавторами [33].

1.7. Нейронные сети

Нейронная сеть преобразует набор входных переменных в набор выходных переменных и моделирует как линейные преобразования, так и нелинейные. Преобразования осуществляются с помощью нейронов, представляющих собой упрощенную модель нейронов головного мозга. Нейроны связаны в сеть односторонними каналами передачи информации. Каждый нейрон может быть активирован поступающими входными сигналами, и в активном состоянии выдает выходные сигналы. В нейронной сети имеется слой входных нейронов – это те нейроны, на которые поступают значения входных переменных, слой выходных нейронов – из выходных сигналов этих нейронов формируются выходные переменные, и скрытые слои. Нейронные сети различаются своей структурой, числом скрытых слоев, функциями активации.

В работе Д. Веста [34] проанализированы пять моделей нейронных сетей, используемых в кредитном скоринге: многослойный персептрон (*MLP*); смесь экспертов (*MOE*); сеть радиальных базисных функций (*RBF*); квантование обучающего вектора (*LVQ*); нечеткий адаптивный резонанс (*FAR*). Эффективность применения нейронных сетей перечисленных типов в кредитном скоринге сравнивалась с эффективностью применения классических параметрических методов (линейный дискриминантный анализ и логистическая регрессия), непараметрических методов (*k* ближайших соседей или *k-NN*, ядерной оценке плотности) и классификационных деревьев. Полученные результаты подтвердили, что многослойные персептроны показывают далеко не самую высокую точность, сети типа смеси экспертов и сети радиальных базисных функций показывают в кредитном скоринге вполне удовлетворительный результат. Из классических методов наиболее точным оказался метод логистической регрессии. Сети, основанные на нечетком адаптивном резонансе, оказались наименее точными. Не уступающие другим сетям по распознаванию «плохих» заемщиков, они существенно хуже работают по распознаванию «хороших» заемщиков.

1.8. Генетические алгоритмы

Определенная специфика применения генетических алгоритмов в кредитном скоринге состоит в том, что популяция образована классификационными деревьями. Алгоритмы мутации и скрещивания применяются к деревьям. В остальном структура алгоритмов стандартная. После создания исходной популяции повторяются процессы мутации и скрещивания с последующей оценкой. В качестве оценки берется относительное число ошибок классификации. В работе Ц. Онга с соавторами [35] показано, что на тестовых наборах результаты работы генетических алгоритмов (при 1 000 поколений) оказались в числе лучших.

1.9. Комбинированные методы

К числу гибридных и комбинированных относятся методы, в которых применяются различные техники кредитного скоринга для повышения эффективности. Наиболее употребительны три метода комбинирования (ensemble methods): беггинг (bagging – bootstrap aggregating), бустинг (boosting) и стекинг (stacking).

Беггинг был введен в работе Л. Бреймана [36]. Основная идея метода – построение набора предикторов, которые в совокупности (после определенного агрегирования) дают более совершенный предиктор.

Схематично беггинг применительно к кредитному скорингу выглядит следующим образом. Предполагается, что имеется алгоритм обучения, который по обучающему множеству L строит предиктор $\varphi(x,L)$, выдающий y при заданном x. Основываясь на обучающем множестве L, можно построить набор обучающих множеств $\{L_k\}_{k=1,\ldots,K}$ (как правило, того же объема, что и L). Эти множества состоят из тех же объектов, выбранных случайным образом из L (возможно, с повторениями). Положим K_+ равным числу тех k, для которых $\varphi(x,L_k)$ дает положительный ответ. Агрегированный предиктор $\bar{\varphi}$ выдает положительный ответ, если $K_+ > \frac{1}{2}$.

Применение беггинга оказывается особенно эффективным в тех случаях, когда основной алгоритм обучения неустойчив – сильно зависит от небольших изменений в обучающем множестве.

Основная идея бустинга – сформировать на основе слабого (в смысле точности) алгоритма сильный алгоритм классификации. В процессе формирования сильного алгоритма слабый алгоритм «доучивается» за счет того, что перераспределяются веса примеров из обучающей выборки: в случае верного распознавания вес снижается. В случае неверного – повышается. Представление о бустинге дает следующий пример.

Пусть X – пространство $\{(x^{(j)}, y^{(j)})\}_{j=1,2,\ldots,l}$ – обучающая выборка. Базовый алгоритм запускается в серии раундов $t = 1,\ldots,T$. Обозначим через $D_t(j)$ вес, присвоенный объекту в раунде t (первоначальное распределение весов $D_0(j)$ можно взять равномерным). Задача обучения – в раунде t найти такое отображение $h_t(x)$ со значениями в $\{-1;1\}$, которое минимизирует вероятность ошибки $\varepsilon_t = \sum_{h_t(x^{(j)}) \neq y^{(j)}} D_t(j)$.

Обновление весов происходит следующим образом. Пусть $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.

Тогда $D_{t+1}(j) = \frac{D_t(j)\exp(-\alpha_t y^j h_t(x^{(j)}))}{Z_t}$, где Z_t – нормирующий множитель (так, чтобы $D_{t+1}(j)$ было распределением). Финальный алгоритм распознавания имеет вид

$$H(x) = sign\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right).$$

Этот способ бустинга, основанный на использовании экспоненциальной функции потерь, носит название AdaBoost. Он заставляет алгоритм переобучаться при наличии большого количества шумовых прецедентов. Для минимизации данного эффекта можно использовать логистическую функцию потерь (такой алгоритм называется LogitBoost).

При стекинге происходит комбинирование нескольких алгоритмов с помощью некоторого комбинатора. Как правило, в роли комбинатора выступает логистическая регрессия. Теоретические основы стекинга были заложены в работе Д. Волперта [37].

В кредитном скоринге применение комбинированных методов широко распространено. Характерными примерами могут служить работы С. Аккоса [14], С. Вуковича с соавторами [38], А. Маркеса с соавторами [39].

1.10. Методы, основанные на нечеткой логике

Публикаций, посвященных применению методов нечеткой логики в кредитном скоринге, довольно много. Укажем лишь несколько работ: Ф. Хофман с соавторами [40], Дж. Игнатиус с соавторами [41], А. Лахсасна с соавторами [42], А. Каур с соавторами [43], Р. Малхотра, Д. Малхотра [44].

Однако с учетом огромного числа публикаций по кредитному скорингу их доля сравнительно невелика. Работы, в которых нечеткая логика используется для кредитного скоринга, можно условно разделить на две группы. К первой

относятся те исследования, где элементы нечеткой логики применяются в рамках традиционных методов. Как правило, это работы, связанные с нейронными сетями и методом базисных векторов. Ко второй группе относятся исследования, в которых основной метод заимствован из теории нечетких множеств. Это в первую очередь работы, в основу которых положены системы нечеткого вывода, в частности, системы Мамдани и Такаджи–Сугено. Подробному анализу применения нечеткой логики в кредитном скоринге посвящена вторая часть обзора.

2. Тестовые наборы данных

Широкое распространение среди разработчиков алгоритмов кредитного скоринга получили два набора данных с условными названиями австралийский (Australian scoring data) и немецкий (German Credit Data Set). Австралийский набор содержит в общей сложности данные о 690 заемщиках, из которых 307 состоятельны (выплачивают кредит), а 383 несостоятельны. Описание каждого заемщика включает 14 атрибутов (6 непрерывных и 8 категориальных). Немецкий набор содержит 1 000 записей о заемщиках, из которых 700 состоятельны, 300 несостоятельны. Описание заемщика содержит 20 атрибутов. Оба набора находятся в публичном доступе в UCI Repository of Machine Learning[1].

3. Общее понятие модели кредитного скоринга и сравнение моделей кредитного скоринга

Условимся называть исход хорошим (*Good*), если $y = 0$, и плохим (*Bad*), если $y = 1$. В классической постановке задача прогнозирования состоит в том, чтобы по набору наблюдаемых объектов (x,y) определить $E[Y|x]=E[y=1|x]$. Если бы были известны условные вероятности $P(Bad|x)$ принимать оптимальные решения о предоставлении кредита было бы несложно. Пространство атрибутов обычно слишком велико, чтобы можно было эмпирически оценить вероятности $P(Bad|x)$. Стандартный подход – построение скоринговой функции $s(x)$. Апостериорная вероятность $P(Bad|s)=P(Bad|s(x)=s)$ используется для построения прогнозов, как замена $P(Bad|x)$.

Пусть A – модель кредитного скоринга. Значение скоринговой функции $s_A(x)$ можно рассматривать как реализацию некоторой случайной величины s_A. Обозначим через $f(s_A|y)$ плотность вероятности условного прогноза s_A при данном y, а через $v(s_A)$ вероятность того, что значение скоринговой функции окажется равным s_A.

Основы сравнения моделей кредитного скоринга заложены в работах Р. Клемена с соавторами [45], М. де Гроота с соавторами [46, 47], а также Х. Жу с соавторами [33].

Пусть A и B – скоринговые модели. Говорят, что модель A является достаточной для модели B, если существует функция h, обладающая следующими свойствами:

1) $h(s_B|s_A)\geq 0$ для любых s_A, s_B;

2) $\sum_{S_B} h(s_B, s_A)=1$ при любом s_A;

3) $\sum_{s_A} h(s_B|s_A) f_A(s_A|y)=f_B(s_B|y)$ для любых s_B и y.

Если A является достаточной для B, то B можно считать более неопределенной – функция h придает значениям s_B дополнительную случайность.

Говорят, что модель B не родственна модели A, если y не зависит от s_B при заданном s_A, то есть $P(y|s_A, s_B)=P(y|s_A)$.

Для заданных скоринговых моделей A и B определим комбинированную скоринговую модель C, полагая $s_C=P(Good|s_A, s_B)$. Комбинированная модель является достаточной для моделей A и B. Модель является достаточной для модели C, тогда и только тогда, когда модель B не родственна модели A.

Рассмотрим теперь оценку скоринговой модели с точки зрения полезности.

[1] Statlog (Australian Credit Approval) Data Set. URL: http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval); Statlog (German Credit Data) Data Set. URL: http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

Допустим, что предоставление кредита хорошему заемщику приносит доход в размере 1, а плохому – в размере $-\alpha \le 0$ (убыток). Положим $\pi(s) = P(Good|s)$. Ожидаемый доход R при выдаче кредита заемщику, имеющему скоринг s, составляет:

$$E[R|s] = \pi(s) - \alpha(1 - \pi(s)).$$

Будем также считать, что отказ в кредите приносит доход 0 независимо от типа заемщика. Решение о выдаче кредита принимается, если $E[R|s] \ge 0$. Поскольку функция $\pi(s)$ монотонно возрастает, существует такое значение s^*, что $E[R|s^*] = 0$. Если значение скоринговой функции превосходит s^*, кредит одобряется, если нет – заемщик получает отказ. Таким образом, $E[R] = \sum_{s \ge s^*} E[R|s] v(s)$.

Поскольку s^* зависит от α, математическое ожидание дохода $E[R]$ также зависит от α. Эта величина может использоваться для сравнения моделей кредитного скоринга: скоринговая модель A является достаточной для скоринговой модели B тогда и только тогда, когда $E_A[R] \ge E_B[R]$ для всех α.

4. Оценка качества алгоритмов кредитного скоринга

Одним из способов определения качества модели машинного обучения является разделение выборки на обучающую, которая используется для идентификации параметров алгоритма, и контрольную, для каждого объекта которой проводится сравнение класса, предсказанного алгоритмом, и истинного класса объекта.

При этом наиболее распространенные методы оценки алгоритмов кредитного скоринга основываются на матрице ошибок: все объекты контрольной выборки разбивают на четыре категории в зависимости от комбинации истинного ответа y и ответа $\alpha(x)$, выданного алгоритмом:

	$\alpha(x) = 1$	$\alpha(x) = 0$
$y = 1$	TP	FN
$y = 0$	FP	TN

(TP – сокращение для True Positive, FN – False Negative, и аналогично в двух оставшихся случаях).

Поскольку целью применения алгоритмов классификации в кредитном скоринге является сортировка объектов скоринга на хорошие и плохие, эффективность алгоритмов оценивается путем сопоставления для каждого объекта из контрольного набора данных класса, спрогнозированного алгоритмом, с реальным классом этого объекта.

Задача кредитного скоринга имеет две особенности. Во-первых, классификация плохого кредита как хорошего обходится дороже, чем классификация хорошего кредита как плохого, а во-вторых, в обучающей выборке хороших клиентов всегда больше чем плохих.

В связи с первой особенностью в задаче кредитного скоринга применяются следующие метрики качества алгоритмов:

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ – доля правильно классифицированных кредитов;

$Precision = \frac{TP}{TP + FP}$ – точность, то есть доля правильно классифицированных плохих кредитов среди всех наблюдений, отнесенных алгоритмом к плохим кредитам;

$Recall = \frac{TP}{TP + FN}$ – полнота, то есть оценка способности алгоритма распознавать плохие кредиты;

$Negative\ Predictive\ Value = \frac{TN}{TN + FN}$ – доля правильно классифицированных хороших кредитов среди всех наблюдений, отнесенных алгоритмом к хорошим кредитам;

$Specificity = \frac{TN}{TN + FP}$ – специфичность, то есть оценка способности алгоритма распознавать хорошие кредиты;

$F1\,Score = \frac{2(Precision \cdot Recall)}{Precision + Recall}$ – среднее гармоническое точности и полноты;

$False\ Negative\ Rate = \frac{FN}{TP + FN}$ – доля плохих кредитов, неправильно отнесенных к хорошим;

$False\ Positive\ Rate = \frac{FP}{TN + FP}$ – доля хороших кредитов, неправильно отнесенных к плохим.

Кривая ошибок (Receiver Operating Characteristic, ROC) отображает изменение отношения доли *Recall* верно классифицированных плохих кредитов в их общем количестве к доле *False Positive Rate* хороших кредитов, неправильно отнесенных к плохим, при варьировании порога решающего правила.

ROC-кривая получается следующим образом. Предположим, что результат работы алгоритма $\alpha(x)$ зависит от параметра, например, порогового значения, и алгоритм имеет вид $\alpha(x) = Entier[s(x) > s^*]$.

При $s^* = \infty$ получаем $SEN = 0$ и $FPR = 0$, при $s^* = -\infty$ – $SEN = 1$ и $FPR = 1$. Когда s^* изменяется от $-\infty$ до ∞, точка (FPR, SEN) описывает кривую, которая называется ROC-кривой. Площадь под кривой ошибок AUC служит показателем качества алгоритма. Равенство $AUC = 0{,}5$ означает, что алгоритм относит объекты к категориям наугад. Чем больше AUC, тем качественнее алгоритм. Часто используется также характеристика, называемая индексом Джини (площадь между кривой и диагональю): $Gini = 2\,AUC - 1$.

Важной проблемой при построении скоринговых моделей является тот факт, что доля плохих кредитов в выборке сильно меньше доли хороших (обычно от 2 до 30%). В такой ситуации малую ошибку на обучающей и тестовой выборках может давать модель, которая предлагает всех клиентов признавать хорошими.

Возможными решениями этой проблемы являются введение различных стоимостей ошибок первого и второго рода или модификация обучающей выборки для изменения репрезентативности выборки.

Для изменения репрезентативности выборки используют два основных метода: дублирование миноритарного класса (oversampling) и удаление мажоритарного класса (undersampling). Недостатком первого метода является тот факт, что простое дублирование прецедентов может не влиять никаким образом на одни методы обучения и вести к переобучению других. При удалении объектов, относящихся к мажоритарному классу, возможна потеря важной для классификации информации, что также нежелательно. Наиболее часто для решения проблемы несбалансированности выборки применяется метод синтетического размножения объектов миноритарного класса SMOTE (Synthetic Minority Oversampling Technique). Новый синтезированный объект по методу SMOTE строится следующим образом:

- вычисляется разность $d = x_b - x_a$ между векторами x_a, x_b признаков соседних объектов a, b из миноритарного класса;
- формируется вектор признаков для нового синтезированного объекта $x_{\vec{d}} = x_a + cd$, где $c \sim N(0,1)$.

Существуют разные модификации метода SMOTE, в которых при генерации объектов миноритарного класса используются ближайшие соседи как из миноритарного, так и из мажоритарного класса, и генерируемые объекты располагаются ближе к границе разделения классов или дальше от нее.

Однако на практике метод SMOTE очень часто приводит к переобучению моделей, кроме того, он весьма затратен по используемым вычислительным ресурсам и по времени, к тому же проблема несбалансированности классов в скоринговых выборках, как правило, представляет собой отдельную сложную задачу.

5. Программная реализация алгоритмов машинного обучения в кредитном скоринге

Программные продукты, используемые для автоматизации решения задач интеллектуального анализа данных и машинного обучения, можно разделить на три класса:

– коммерческие статистические пакеты;

– открытые среды;

– облачные решения.

Исторически в банках для решения задач анализа данных, в частности связанных со скорингом, использовались коммерческие программные продукты. Наиболее часто применялся набор продуктов SAS, реже пакеты IBM SPSS и Statistica. Эти три линейки продуктов предоставляют схожие функциональные возможности, включающие в себя средства аналитической подготовки данных, готовые и настраиваемые шаблоны алгоритмов машинного обучения, в том числе моделей линейной и логистической регрессии, деревьев и лесов решений, градиентного бустинга, опорных векторов, нейронных сетей и др. Кроме того, в этих пакетах возможна настройка параметров моделей и использование интерактивных техник оценки качества.

В последние годы банки, не отказываясь полностью от применения коммерческих пакетов типа SAS, стали использовать и открытые среды Python/R/Spark. Преимуществом этих сред является прежде всего возможность использования гораздо большего количества алгоритмов, чем в коммерческих пакетах. Но если, например, в здравоохранении и промышленности с распространением открытых сред Python и R от применения коммерческих пакетов практически отказались в пользу открытых библиотек, то в банках к решениям SAS пока обращаются чаще, чем к Python и R.

Язык R создавался как специальное средство для статистических вычислений, он стал первой открытой средой, которая начала активно использоваться для анализа данных. Наиболее часто используемые библиотеки для машинного обучения в R – это rpart и CARET (алгоритмы классификации и регрессии), randomForest (алгоритм случайных лесов), nnet (нейронные сети), e1071 (одна из первых библиотек машинного обучения в R, содержащая реализацию метода опорных векторов, наивный байесовский классификатор и ряд других методов), kernlab (метод опорных векторов), gbm (градиентный бустинг), ROCR (визуализация метрик качества алгоритмов классификации).

Язык Python стал самым популярным средством для анализа данных после выхода отлично документированной библиотеки scikit-learn, в которой реализовано большое количество алгоритмов машинного обучения. Кроме scikit-learn, популярны также библиотеки TensorFlow и Theano (эти библиотеки также реализуют различные методы анализа данных, но выигрывают у scikit-learn только в количестве реализованных техник работы с нейронными сетями). Для использования аппарата детерминированного, нечеткого и байесовского логического вывода в Python применяется библиотека pyinference. Основное преимущество Python перед R – более высокая скорость выполнения скриптов.

Альтернативным решением для анализа данных в случае, когда быстродействия Python недостаточно, выступает Apache Spark – открытая масштабируемая кластерная вычислительная платформа, ориентированная на вычисления в оперативной памяти. При этом библиотека MLib, которая в Spark реализует возможности машинного обучения, по количеству алгоритмов пока существенно уступает scikit-learn, но активно развивается.

В последние годы появились облачные платформы машинного обучения. Основным преимуществом таких систем является гибкая масштабируемость – выделение и высвобождение вычислительных ресурсов происходит мгновенно в соответствии с решаемыми задачами. Amazon Machine Learning реализует только базовые алгоритмы бинарной и мультиклассовой классификации, а также регрессии. Google Machine Learning Engine предоставляет возможность запуска моделей TensorFlow в облачной среде. Платформа Microsoft Azure Machine Learning Studio предоставляет собой мощное решение, позволяющее с помощью простого графического интерфейса строить модели машинного обучения с использованием множества стандартных алгоритмов классификации, регрессии, кластерного анализа и поиска аномалий, а также встраивать в эти модели собственный код на SQL, Python и R. В ближайшее время ожидается выпуск аналогичного решения от IBM – Watson Machine Learning.

Однако несмотря на все преимущества облачных средств анализа данных, в банках они практически не используются в связи с опасениями по поводу безопасности передачи в облачные хранилища конфиденциальных данных о клиентах.

Список литературы

1. *Durand D.* Risk Elements in Consumer Installment Financing. New York, National Bureau of Economic Research Books, 1941, 163 p.
2. *Hand D.J., Henley W.E.* Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1997, vol. 160, iss. 3, pp. 523–541. doi: 10.1111/j.1467-985X.1997.00078.x
3. *García V., Marqués A.I., Sánchez J.S.* An Insight into the Experimental Design for Credit Risk and Corporate Bankruptcy Prediction Systems. *Journal of Intelligent Information Systems*, 2015, vol. 44, iss. 1, pp. 159–189. URL: https://doi.org/10.1007/s10844-014-0333-4
4. *Lessmann S., Seow H.-V., Baesens B., Thomas L.C.* Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *European Journal of Operational Research*, 2015, vol. 247, no. 1, pp. 124–136. URL: https://www.business-school.ed.ac.uk/waf/crc_archive/2013/42.pdf doi: 10.1016/j.ejor.2015.05.030
5. *Hand D.J., Kelly M.G.* Superscorecards. *IMA Journal of Management Mathematics*, 2002, vol. 13, iss. 4, pp. 273–281.
6. *Yap B.W., Ong S.H., Husain N.H.M.* Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models. *Expert Systems with Applications*, 2011, vol. 38, iss. 10, pp. 13274–13283. doi: 10.1016/j.eswa.2011.04.147
7. *Pavlidis N.G., Tasoulis D.K., Adams N.M., Hand D.J.* Adaptive Consumer Credit Classification. *Journal of the Operational Research Society*, 2012, vol. 63, iss. 12, pp. 1645–1654. doi: 10.1057/jors.2012.15
8. *Khemais Z., Nesrine D., Mohamed M.* Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression. *International Journal of Economics and Finance*, 2016, vol. 8, iss. 4, pp. 39–53. URL: http://dx.doi.org/10.5539/ijef.v8n4p39
9. *Louzada F., Anacleto-Junior O., Candolo C., Mazucheli J.* Poly-bagging Predictors for Classification Modelling for Credit Scoring. *Expert Systems with Applications*, 2011, vol. 38, iss. 10, pp. 12717–12720. URL: https://doi.org/10.1016/j.eswa.2011.04.059
10. *Li Z., Tianb Y., Li K. et al.* Reject Inference in Credit Scoring Using Support Vector Machines. *Expert Systems with Applications*, 2017, vol. 74, pp. 105–114. URL: https://doi.org/10.1016/j.eswa.2017.01.011
11. *Fisher R.A.* The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 1936, vol. 7, iss. 2, pp. 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x
12. *Eisenbeis R.A.* Problems in Applying Discriminant Analysis in Credit Scoring Models. *Journal of Banking & Finance*, 1978, vol. 2, iss. 3, pp. 205–219. doi: 10.1016/0378-4266(78)90012-2
13. *Mylonakis J., Diacogiannis G.* Evaluating the Likelihood of Using Linear Discriminant Analysis as a Commercial Bank Card Owners Credit Scoring Model. *International Business Research*, 2010, vol. 3, no. 2, pp. 9–20.
14. *Akkoç S.* An Empirical Comparison of Conventional Techniques, Neural Networks and the Three Stage Hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) Model for Credit Scoring Analysis: The case of Turkish credit card data. *European Journal of Operational Research*, 2012, vol. 222, iss. 1, pp. 168–178. URL: https://doi.org/10.1016/j.ejor.2012.04.009
15. *Falangis K., Glen J.J.* Heuristics for Feature Selection in Mathematical Programming Discriminant Analysis Models. *Journal of Operational Research Society*, 2010, vol. 61, no. 5, pp. 804–812.

16. *Breiman L., Friedman J., Stone C.J., Olshen R.A.* Classification and Regression Trees. Monterey, CA, Wadsworth & Brooks/Cole Advanced Books & Software, 1984, 368 p.

17. *Loh W.-Y.* Fifty Years of Classification and Regression Trees. *International Statistical Review*, 2014, vol. 82, iss. 3, pp. 329–348. doi: 10.1111/insr.12016

18. *Finlay S.M.* Multiple Classifier Architectures and Their Application to Credit Risk Assessment. *European Journal of Operational Research*, 2011, vol. 210, iss. 2, pp. 368–378. URL: http://dx.doi.org/10.1016/j.ejor.2010.09.029

19. *Zhang D., Zhou X., Leung S.C.H., Zheng J.* Vertical Bagging Decision Trees Model for Credit Scoring. *Expert Systems with Applications*, 2010, vol. 37, iss. 12, pp. 7838–7843. URL: https://doi.org/10.1016/j.eswa.2010.04.054

20. *Hu Q., Che X., Zhang L. et al.* Rank Entropy-Based Decision Trees for Monotonic Classification. *IEEE Transactions on Knowledge and Data Engineering*, 2012, vol. 24, iss. 11, pp. 2052–2064. doi: 10.1109/TKDE.2011.149

21. *Hayashi Y., Tanaka Y., Takagi T. et al.* Recursive-Rule Extraction Algorithm with J48graft and Applications to Generating Credit Scores. *Journal of Artificial Intelligence and Soft Computing Research*, 2016, vol. 6, iss. 1, pp. 35–44. URL: https://doi.org/10.1515/jaiscr-2016-0004

22. *Vapnik V.* Statistical Learning Theory. New York, John Wiley, 1998, 768 p.

23. *Bellotti T., Crook J.* Support Vector Machines for Credit Scoring and Discovery of Significant Features. *Expert Systems with Applications*, 2009, vol. 36, iss. 2-2, pp. 3302–3308. doi: 10.1016/j.eswa.2008.01.005

24. *Chen W., Ma C., Ma L.* Mining the Customer Credit Using Hybrid Support Vector Machine Technique. *Expert Systems with Applications*, 2009, vol. 36, iss. 4, pp. 7611–7616. URL: https://doi.org/10.1016/j.eswa.2008.09.054

25. *Ling Y., Cao Q., Zhang H.* Credit Scoring Using Multi-Kernel Support Vector Machine and Chaos Particle Swarm Optimization. *International Journal of Computational Intelligence and Applications*, 2012, vol. 11, iss. 3, pp. 12500198:1–12500198:13.

26. *Friedman N., Geiger D., Goldszmidt M.* Bayesian Network Classifiers. *Machine Learning*, 1997, vol. 29, iss. 2-3, pp. 131–163. URL: https://doi.org/10.1023/A:1007465528199

27. *Pearl J.* Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988, 552 p.

28. *Giudici P.* Bayesian Data Mining, with Application to Benchmarking and Credit Scoring. *Applied Stochastic Models in Business and Industry*, 2001, vol. 17, iss. 1, pp. 69–81.

29. *Gemela J.* Financial Analysis Using Bayesian Networks. *Applied Stochastic Models in Business and Industry*, 2001, vol. 17, iss. 1, pp. 57–67.

30. *Antonakis A.C., Sfakianakis M.E.* Naïve Bayes as a Means of Constructing Application Scorecards. *Advances in Doctoral Research in Management*, 2008, vol. 2, pp. 47–62.

31. *Antonakis A.C., Sfakianakis M.E.* Assessing Naïve Bayes as a Method for Screening Credit Applicants. *Journal of Applied Statistics*, 2009, vol. 36, iss. 5-6, pp. 537–545.

32. *Wu W.-W.* Improving Classification Accuracy and Causal Knowledge for Better Credit Decisions. *International Journal of Neural Systems*, 2011, vol. 21, iss. 4, pp. 297–309.

33. *Zhu H., Beling P.A., Overstreet G.A.* A Bayesian Framework for the Combination of Classifier Outputs. *Journal of the Operational Research Society*, 2002, vol. 53, iss. 7, pp. 719–727.

34. *West D.* Neural Network Credit Scoring Models. *Computers & Operations Research*, 2000, vol. 27, iss. 11-12, pp. 1131–1152. doi: 10.1016/S0305-0548(99)00149-5

35. *Ong C.-S., Huang J.-J., Tzeng G.-H.* Building Credit Scoring Models Using Genetic Programming. *Expert Systems with Applications*, 2005, vol. 29, iss. 1, pp. 41–47. doi: 10.1016/j.eswa.2005.01.003

36. *Breiman L.* Bagging Predictors. *Machine Learning*, 1996, vol. 24, iss. 2, pp. 123–140. URL: https://doi.org/10.1007/BF00058655

37. *Wolpert D.H.* Stacked Generalization. *Neural Networks*, 1992, vol. 5, no. 2, pp. 241–259.

38. *Vukovic S., Delibašić B., Uzelac A., Suknovic M.* A Case-Based Reasoning Model That Uses Preference Theory Functions for Credit Scoring. *Expert Systems with Applications*, 2012, vol. 39, iss. 9, pp. 8389–8395. doi: 10.1016/j.eswa.2012.01.181

39. *Marqués A.I., García V., Sánchez J.S.* Two-Level Classifier Ensembles for Credit Risk Assessment. *Expert Systems with Applications*, 2012, vol. 39, iss. 12, pp. 10916–10922.

40. *Hoffmann F., Baesens B., Mues C. et al.* Inferring Descriptive and Approximate Fuzzy Rules for Credit Scoring Using Evolutionary Algorithms. *European Journal of Operational Research*, 2007, vol. 177, iss. 1, pp. 540–555. doi: 10.1016/j.ejor.2005.09.044

41. *Ignatius J., Hatami-Marbini A., Rahman A. et al.* A Fuzzy Decision Support System for Credit Scoring. *Neural Computing and Applications*, 2016, vol. 27, no. 1, pp. 1–17. URL: https://doi.org/10.1007/s00521-016-2592-1

42. *Lahsasna A., Ainon R.N., Wah T.Y.* Credit Risk Evaluation Decision Modeling Through Optimized Fuzzy Classifier. Proc. International Symposium on Information Technology, 2008. *IEEE*, 2008, vol. 1, pp. 1–8.

43. *Kaur A. et al.* Fuzzy Rule based Expert System for Evaluating Defaulter Risk in Banking Sector. *Indian Journal of Science and Technology*, 2016, vol. 9, iss. 28, pp. 1–6. doi: 10.17485/ijst/2016/v9i28/98395

44. *Malhotra R., Malhotra D.K.* Differentiating Between Good Credits and Bad Credits Using Neuro-Fuzzy Systems. *European Journal of Operational Research*, 2002, vol. 136, iss. 1, pp. 190–211.

45. *Clemen R.T., Murphy A.H., Winkler R.L.* Screening Probability Forecasts: Contrasts Between Choosing and Combining. *International Journal of Forecasting*, 1995, vol. 11, iss. 1, pp. 133–145.

46. *DeGroot M.H., Fienberg S.E.* The Comparison and Evaluation of Forecasters. *The Statistician: Journal of the Institute of Statisticians*, 1983, vol. 32, no. 1/2, pp. 12–22.

47. *DeGroot M., Eriksson E.A.* Probability Forecasting, Stochastic Dominance, and the Lorenz Curve. *Bayesian Statistics*, 1985, vol. 2, pp. 99–118.

Информация о конфликте интересов

Мы, авторы данной статьи, со всей ответственностью заявляем о частичном и полном отсутствии фактического или потенциального конфликта интересов с какой бы то ни было третьей стороной, который может возникнуть вследствие публикации данной статьи. Настоящее заявление относится к проведению научной работы, сбору и обработке данных, написанию и подготовке статьи, принятию решения о публикации рукописи.

pISSN 2071-4688
eISSN 2311-8709

Banking

DATA MINING TECHNIQUES: MODERN APPROACHES TO APPLICATION IN CREDIT SCORING

Elena S. VOLKOVA[a,•], Vladimir B. GISIN[b], Vladimir I. SOLOV'EV[c]

[a] Financial University under Government of Russian Federation, Moscow, Russian Federation
EVolkova@fa.ru

[b] Financial University under Government of Russian Federation, Moscow, Russian Federation
VGisin@fa.ru

[c] Financial University under Government of Russian Federation, Moscow, Russian Federation
VSoloviev@fa.ru

• Corresponding author

Article history:
Received 4 July 2017
Received in revised form
9 August 2017
Accepted 24 August 2017
Available online
15 September 2017

JEL classification: C38, C55, D81

Keywords: loan scoring, credit score, machine learning, data mining

Abstract

Importance This article examines the current state of research in machine learning and data mining, which computational methods get combined with conventional lending models such as scoring, for instance.

Objectives The article aims to classify the modern methods of credit scoring and describe models for comparing the effectiveness of the various methods of credit scoring.

Methods To perform the tasks, we have studied relevant scientific publications on the article subject presented in Google Scholar.

Results The article presents a classification of modern data mining techniques used in credit scoring.

Conclusions Credit scoring models using machine learning procedures and hybrid models using combined methods can provide the required level of efficiency in the modern environment.

Please cite this article as: Volkova E.S., Gisin V.B., Solov'ev V.I. Data Mining Techniques: Modern Approaches to Application in Credit Scoring. *Finance and Credit*, 2017, vol. 23, iss. 34, pp. 2044–2060.
https://doi.org/10.24891/fc.23.34.2044

References

1. Durand D. Risk Elements in Consumer Installment Financing. New York, National Bureau of Economic Research Books, 1941, 163 p.
2. Hand D.J., Henley W.E. Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1997, vol. 160, iss. 3, pp. 523–541. doi: 10.1111/j.1467-985X.1997.00078.x
3. García V., Marqués A.I., Sánchez J.S. An Insight into the Experimental Design for Credit Risk and Corporate Bankruptcy Prediction Systems. *Journal of Intelligent Information Systems*, 2015, vol. 44, iss. 1, pp. 159–189. URL: https://doi.org/10.1007/s10844-014-0333-4
4. Lessmann S., Seow H.-V., Baesens B., Thomas L.C. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. *European Journal of Operational Research*, 2015, vol. 247, iss. 1, pp. 124–136. URL: https://www.business-school.ed.ac.uk/waf/crc_archive/2013/42.pdf doi: 10.1016/j.ejor.2015.05.030
5. Hand D.J., Kelly M.G. Superscorecards. *IMA Journal of Management Mathematics*, 2002, vol. 13, iss. 4, pp. 273–281.

6. Yap B.W., Ong S.H., Husain N.H.M. Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models. *Expert Systems with Applications*, 2011, vol. 38, iss. 10, pp. 13274–13283. doi: 10.1016/j.eswa.2011.04.147

7. Pavlidis N.G., Tasoulis D.K., Adams N.M., Hand D.J. Adaptive Consumer Credit Classification. *Journal of the Operational Research Society*, 2012, vol. 63, iss. 12, pp. 1645–1654. doi: 10.1057/jors.2012.15

8. Khemais Z., Nesrine D., Mohamed M. Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression. *International Journal of Economics and Finance*, 2016, vol. 8, iss. 4, pp. 39–53. URL: http://dx.doi.org/10.5539/ijef.v8n4p39

9. Louzada F., Anacleto-Junior O., Candolo C., Mazucheli J. Poly-bagging Predictors for Classification Modelling for Credit Scoring. *Expert Systems with Applications*, 2011, vol. 38, iss. 10, pp. 12717–12720. URL: https://doi.org/10.1016/j.eswa.2011.04.059

10. Li Z., Tianb Y., Li K. et al. Reject Inference in Credit Scoring Using Support Vector Machines. *Expert Systems with Applications*, 2017, vol. 74, pp. 105–114. URL: https://doi.org/10.1016/j.eswa.2017.01.011

11. Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 1936, vol. 7, iss. 2, pp. 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

12. Eisenbeis R.A. Problems in Applying Discriminant Analysis in Credit Scoring Models. *Journal of Banking & Finance*, 1978, vol. 2, iss. 3, pp. 205–219. doi: 10.1016/0378-4266(78)90012-2

13. Mylonakis J., Diacogiannis G. Evaluating the Likelihood of Using Linear Discriminant Analysis as a Commercial Bank Card Owners Credit Scoring Model. *International Business Research*, 2010, vol. 3, no. 2, pp. 9–20.

14. Akkoç S. An Empirical Comparison of Conventional Techniques, Neural Networks and the Three Stage Hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) Model for Credit Scoring Analysis: The case of Turkish credit card data. *European Journal of Operational Research*, 2012, vol. 222, iss. 1, pp. 168–178. URL: https://doi.org/10.1016/j.ejor.2012.04.009

15. Falangis K., Glen J.J. Heuristics for Feature Selection in Mathematical Programming Discriminant Analysis Models. *Journal of the Operational Research Society*, 2010, vol. 61, no. 5, pp. 804–812.

16. Breiman L., Friedman J., Stone C.J., Olshen R.A. Classification and Regression Trees. Monterey, CA, Wadsworth & Brooks/Cole Advanced Books & Software, 1984, 368 p.

17. Loh W.-Y. Fifty Years of Classification and Regression Trees. *International Statistical Review*, 2014, vol. 82, iss. 3, pp. 329–348. doi: 10.1111/insr.12016

18. Finlay S.M. Multiple Classifier Architectures and Their Application to Credit Risk Assessment. *European Journal of Operational Research*, 2011, vol. 210, iss. 2, pp. 368–378. URL: http://dx.doi.org/10.1016/j.ejor.2010.09.029

19. Zhang D., Zhou X., Leung S.C.H., Zheng J. Vertical Bagging Decision Trees Model for Credit Scoring. *Expert Systems with Applications*, 2010, vol. 37, iss. 12, pp. 7838–7843. URL: https://doi.org/10.1016/j.eswa.2010.04.054

20. Hu Q., Che X., Zhang L. et al. Rank Entropy-Based Decision Trees for Monotonic Classification. *IEEE Transactions on Knowledge and Data Engineering*, 2012, vol. 24, iss. 11, pp. 2052–2064. doi: 10.1109/TKDE.2011.149

21. Hayashi Y., Tanaka Y., Takagi T. et al. Recursive-Rule Extraction Algorithm with J48graft and Applications to Generating Credit Scores. *Journal of Artificial Intelligence and Soft Computing Research*, 2016, vol. 6, iss. 1, pp. 35–44. URL: https://doi.org/10.1515/jaiscr-2016-0004

22. Vapnik V. Statistical Learning Theory. New York, John Wiley, 1998, 768 p.

23. Bellotti T., Crook J. Support Vector Machines for Credit Scoring and Discovery of Significant Features. *Expert Systems with Applications*, 2009, vol. 36, iss. 2-2, pp. 3302–3308. doi: 10.1016/j.eswa.2008.01.005

24. Chen W., Ma C., Ma L. Mining the Customer Credit Using Hybrid Support Vector Machine Technique. *Expert Systems with Applications*, 2009, vol. 36, iss. 4, pp. 7611–7616. URL: https://doi.org/10.1016/j.eswa.2008.09.054

25. Ling Y., Cao Q., Zhang H. Credit Scoring Using Multi-Kernel Support Vector Machine and Chaos Particle Swarm Optimization. *International Journal of Computational Intelligence and Applications*, 2012, vol. 11, iss. 3, pp. 12500198:1–12500198:13.

26. Friedman N., Geiger D., Goldszmidt M. Bayesian Network Classifiers. *Machine Learning*, 1997, vol. 29, iss. 2-3, pp. 131–163. URL: https://doi.org/10.1023/A:1007465528199

27. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988, 552 p.

28. Giudici P. Bayesian Data Mining, with Application to Benchmarking and Credit Scoring. *Applied Stochastic Models in Business and Industry*, 2001, vol. 17, iss. 1, pp. 69–81.

29. Gemela J. Financial Analysis Using Bayesian Networks. *Applied Stochastic Models in Business and Industry*, 2001, vol. 17, iss. 1, pp. 57–67.

30. Antonakis A.C., Sfakianakis M.E. Naïve Bayes as a Means of Constructing Application Scorecards. *Advances in Doctoral Research in Management*, 2008, vol. 2, pp. 47–62.

31. Antonakis A.C., Sfakianakis M.E. Assessing Naïve Bayes as a Method for Screening Credit Applicants. *Journal of Applied Statistics*, 2009, vol. 36, iss. 5-6, pp. 537–545.

32. Wu W.-W. Improving Classification Accuracy and Causal Knowledge for Better Credit Decisions. *International Journal of Neural Systems*, 2011, vol. 21, iss. 4, pp. 297–309.

33. Zhu H., Beling P.A., Overstreet G.A. A Bayesian Framework for the Combination of Classifier Outputs. *Journal of the Operational Research Society*, 2002, vol. 53, iss. 7, pp. 719–727.

34. West D. Neural Network Credit Scoring Models. *Computers & Operations Research*, 2000, vol. 27, iss. 11-12, pp. 1131–1152. doi: 10.1016/S0305-0548(99)00149-5

35. Ong C.-S., Huang J.-J., Tzeng G.-H. Building Credit Scoring Models Using Genetic Programming. *Expert Systems with Applications*, 2005, vol. 29, iss. 1, pp. 41–47. doi: 10.1016/j.eswa.2005.01.003

36. Breiman L. Bagging Predictors. *Machine Learning*, 1996, vol. 24, iss. 2, pp. 123–140. URL: https://doi.org/10.1007/BF00058655

37. Wolpert D.H. Stacked Generalization. *Neural Networks*, 1992, vol. 5, no. 2, pp. 241–259.

38. Vukovic S., Delibašić B., Uzelac A., Suknovic M. A Case-Based Reasoning Model That Uses Preference Theory Functions for Credit Scoring. *Expert Systems with Applications*, 2012, vol. 39, iss. 9, pp. 8389–8395. doi: 10.1016/j.eswa.2012.01.181

39. Marqués A.I., García V., Sánchez J.S. Two-Level Classifier Ensembles for Credit Risk Assessment. *Expert Systems with Applications*, 2012, vol. 39, iss. 12, pp. 10916–10922.

40. Hoffmann F., Baesens B., Mues C. et al. Inferring Descriptive and Approximate Fuzzy Rules for Credit Scoring Using Evolutionary Algorithms. *European Journal of Operational Research*, 2007, vol. 177, iss. 1, pp. 540–555. doi: 10.1016/j.ejor.2005.09.044

41. Ignatius J., Hatami-Marbini A., Rahman A. et al. A Fuzzy Decision Support System for Credit Scoring. *Neural Computing and Applications*, 2016, vol. 27, no. 1, pp. 1–17. URL: https://doi.org/10.1007/s00521-016-2592-1

42. Lahsasna A., Ainon R.N., Wah T.Y. Credit Risk Evaluation Decision Modeling Through Optimized Fuzzy Classifier. Proc. International Symposium on Information Technology, 2008. *IEEE*, 2008, vol. 1, pp. 1–8.

43. Kaur A. et al. Fuzzy Rule based Expert System for Evaluating Defaulter Risk in Banking Sector. *Indian Journal of Science and Technology*, 2016, vol. 9, iss. 28, pp. 1–6. doi: 10.17485/ijst/2016/v9i28/98395

44. Malhotra R., Malhotra D.K. Differentiating Between Good Credits and Bad Credits Using Neuro-Fuzzy Systems. *European Journal of Operational Research*, 2002, vol. 136, iss. 1, pp. 190–211.

45. Clemen R.T., Murphy A.H., Winkler R.L. Screening Probability Forecasts: Contrasts Between Choosing and Combining. *International Journal of Forecasting*, 1995, vol. 11, iss. 1, pp. 133–145.

46. DeGroot M.H., Fienberg S.E. The Comparison and Evaluation of Forecasters. *The Statistician: Journal of the Institute of Statisticians*, 1983, vol. 32, no. 1/2, pp. 12–22.

47. DeGroot M., Eriksson E.A. Probability Forecasting, Stochastic Dominance, and the Lorenz Curve. *Bayesian Statistics*, 1985, vol. 2, pp. 99–118.

Conflict-of-interest notification

We, the authors of this article, bindingly and explicitly declare of the partial and total lack of actual or potential conflict of interest with any other third party whatsoever, which may arise as a result of the publication of this article. This statement relates to the study, data collection and interpretation, writing and preparation of the article, and the decision to submit the manuscript for publication.