ПРИМЕНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ДЛЯ ОЦЕНКИ ЗАЕМЩИКА БАНКА

УДК 519.24

ЛЕПА Евгений Владимирович
к.т.н., доцент кафедры информационных технологий Херсонского национального технического университета.
Научные интересы: системы принятия и поддержки решений,
технологии интеллектуального анализа данных, моделирование
e-mail: e.lepa@mail.ru

ВВЕДЕНИЕ

Кредитование является важнейшим источником финансовых средств, как для предприятий (организаций), так и населения. В последнее время в банках накопилось много проблемных кредитов, когда клиенты перестают их обслуживать по различным причинам. Одной из таких причин, является неверная оценка заемщика банком с точки зрения его платежеспособности и возможностей обслуживания кредита. Современные средства обработки и анализа данных могут позволить в значительной мере эту проблему решить [1,2].

ЦЕЛИ И ЗАДАЧИ ИССЛЕДОВАНИЯ

За период работы с заемщиками, банк создает информационную базу данных, как о самих заемщиков, так и данных по выданным или отказанным кредитам. На основании этих данных можно построить модель заемщика, а затем использовать ее для классификации нового клиента. Это позволит отнести клиента к одному из классов, определяющих возможность дать кредит или отказать в его выдаче.

Для повышения достоверности результатов моделирования использованы два разных метода [3,4]. В первом случае модель представляется в виде дерева решений, а во втором - в виде искусственной нейронной сети. Адекватность моделей оценивается, прежде всего, показателями качества классификации, а также опытом и квалификацией специалистов, участвующих в создании и исследовании моделей. Модели используются с разными наборами входных атрибутов, что позволит выбрать наилучшую из них для классификации новых заемщиков банка.

Кроме того предполагается оценить значимость параметров, характеризующих заемщиков банка и исключить из базы те данные, которые мало не влияют на качество модели классификации, но требуют дополнительных затрат на их сбор и хранение [5].

В работе использована аналитическая платформа Deductor [6], в которой для решения задачи классификации реализованы несколько методов.

ИССЛЕДОВАНИЕ И АНАЛИЗ РЕЗУЛЬТАТОВ

В информационной базе банка, которая поддерживается в актуальном состоянии, хранится информация о заемщиках. К такой информации относятся: суммы, сроки и назначение кредитов, личные данные клиента и т.д. (всего 21 параметр). Эти параметры имеют разное значение при оценке заемщика и их, желательно, ранжировать по важности. Кроме того, это позволит исключить в дальнейшем сбор и хранение малозначимой информации.

Для оценки важности параметров выполнен корреляционный анализ, показывающий степень влияния входных параметров, определяющих клиента, на выходной параметр – возможность выдачи кредита (табл.1).

В правом столбце таблицы указаны коэффициенты корреляции. Наиболее значимыми являются шесть параметров, а остальные параметры исключены из рассмотрения. Таким образом, при построении мо-

делей используются только наиболее значимые параметры.

Таблица 1

Оценка значимости входных данных

Целевой атрибут: Давати кредит			
№	Номер	Атрибут	Значимость, %
1	3	Строк кредиту	34,619
2	18	Середньомісячна витрата	22,741
3	1	Сума кредиту	17,438
4	17	Середньомісячний дохід	14,639
5	5	Вік	5,325
6	14	Час роботи підприємства	5,238

Первая модель в виде дерева решений построена с помощью метода C4.5 [7,8], который реализован в аналитической платформе, а результат представлен следующими отображениями.

1. Дерево решений.
2. Правила.
3. Значимость атрибутов (полей базы данных).
4. Прогнозирование с помощью режима **«Что если».**
5. Таблица сопряженности.

Дерево решений приведено на рисунке 1.

Условие	Следствие	Поддержка	Достоверность
ЕСЛИ		142	84
Строк кредиту < 9		45	38
Середньомісячний дохід < 3750	False	2	2
Середньомісячний дохід >= 3750		43	38
Сума кредиту < 6250		13	8
Середньомісячна витрата < 750	True	6	6
Середньомісячна витрата >= 750		7	5
Вік < 22,5	True	2	2
Вік >= 22,5	False	5	5
Сума кредиту >= 6250	True	30	30
Строк кредиту >= 9		97	77
Середньомісячний дохід < 7250		94	76
Сума кредиту < 37500		67	49
Середньомісячний дохід < 4750	False	24	23
Середньомісячний дохід >= 4750		43	26
Середньомісячна витрата < ...		16	14
Час роботи підприємств...	True	13	13
Час роботи підприємств...	False	3	2
Середньомісячна витрата >...	False	27	24
Сума кредиту >= 37500	False	27	27
Середньомісячний дохід >= 7250	True	3	2

Рисунок 1 - Дерево решений

Структура дерева имеет 21 узел и определяется исходными данными, параметрами процесса его построения и самого метода классификации. Для каждого узла указаны значения параметров, характеризующих дерево – **Поддержка** и **Достоверность**.

Поддержка - количество примеров, попавших в узел, от общего количества примеров выборки. Чем выше это значение, тем выше статистическая обоснованность результатов, поскольку классификация в данном узле проводится на большем количестве примеров.

Достоверность - число распознанных примеров от общего числа примеров в данном узле. Чем выше данный показатель, тем достовернее результаты квалификации.

Сформулировано на основании дерева решений 11 правил по каждому из которых выходной параметр данных **Давать кредит** принимает одно из двух логических значений True (Истина – Давать кредит) и False (Ложь – Не давать кредит).

На основании дерева решений можно для каждого выбранного узла дерева отдельно вывести его основные параметры и сформулированные правила. Они отображаются в виде таблицы или словесного описания.

Узел 14; Правило 6

Класс	№	%
False	**23**	**95,80**
True	1	4,17
Поддержка:	24	16,90

ЕСЛИ
Строк кредиту >= 9 **И**
Середньомісячний дохід < 7250 **И**
Сума кредиту < 37500 **И**
Середньомісячний дохід < 4750
ТОГДА
Давати кредит = False

Рисунок 2 – Параметры узла дерева решений

Для выбранного 14 узла действует шестое правило. В дереве решений в этот узел попадает 24 примера из обучающей выборки (Поддержка 24), что составляет 16,9% от общего количества. Правильно классифицировано 23 примера или 95,8% от количества примеров, попавших в этот узел. Неправильно классифицирован только один пример из 24 или 4,17%.

Правило для указанного узла можно сформулировать так (рис.3).

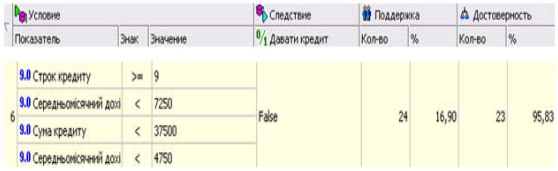

Рисунок 3 – Правила для узла дерева решений

Если срок кредита больше или равен 9 месяцам, среднемесячный доход меньше 7250 грн., сумма кредита меньше 37500 грн., среднемесячный доход меньше 4750 грн., то кредит не может быть выдан, так как выходной параметр исходных данных Давать кредит принимает значение False. Аналогично можно представить правила и для остальных узлов дерева решений.

Для оценки качества классификации в аналитической платформе используется таблица сопряженности [8]

Таблица 2

Оценка качества классификации

	Классифицировано		
Фактически	False	True	Итого
False	89	1	90
True	6	53	59
Итого	95	54	149

Модель правильно классифицирует 89 примеров, которые в таблице данных имеют значение выходного параметра (Давать кредит) False и 53 примера, имеющих это значение равное True. Неверно классифицирован один пример (Давать кредит равно False) и соответственно 6 примеров (Давать кредит равно True). Числа правильной классификации расположены на главной диагонали таблицы сопряженности, а не правильно классифицированные – на побочной диагонали.

Таким образом, модель позволяет правильно классифицировать 142 примера и неправильно 7 примеров из исходной базы данных заемщиков банка, и ошибка классификации составляет около 5%. Использование такой модели для оценки нового заемщика, может принести банку финансовые потери. Повышение качества модели, используя тот же метод классификации, можно попытаться достигнуть за счет расширения исходной базы клиентов и использования дополнительных атрибутов, характеризующих заемщика. Это потребует дополнительного времени и средств.

Более эффективным способом следует признать построение модели заемщика другим методом, на основе той же информационной базы. В качестве такого метода классификации использована искусственная нейронная сеть [3,4].

Искусственные нейронные сети могут быть разной архитектуры, которая определяется количеством слоев и числом нейронов на каждом слое. На обязательном входном слое буде шесть нейронов по количеству входных параметров, и обязательном выходном слое один нейрон (один выходной параметр). Количество внутренних слоев и число нейронов на них, может изменяться. Это дает возможность получить такую сеть, которая обеспечит самое высокое качество классификации.

Простейшая нейронная сеть с одним внутренним слоем и двумя нейронами на нем, представлена на рисунке 4.

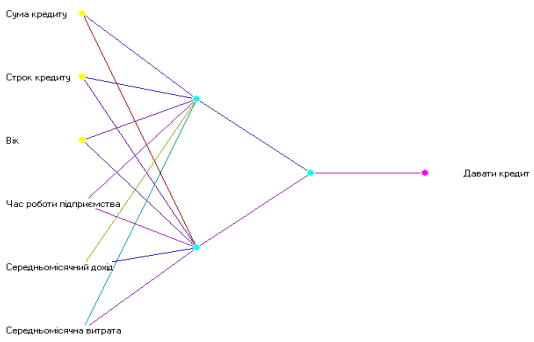

Рисунок 4 - Архитектура нейронной сети

Оценка качества классификации по этой модели проводится на основании таблицы сопряженности (табл.3).

Таблица 3

Оценка качества классификации

	Классифицировано		
Фактически	False	True	Итого
False	90		90
True	4	55	59
Итого	94	55	149

На главной диагонали матрицы приведены количества правильно классифицированных примеров (145), на побочной диагонали – количество неверно классифицированных примеров (4). При сравнении результатов классификации двумя разными методами видно, что вторая модель лучше первой в виде дерева решений.

Для дальнейшего улучшения модели, были построены нейронные сети разной архитектуры и проведена оценка качества классификации. Исследования показали, что наивысшее качество достигнуто при архитектуре нейронной сети, которая содержит один внутренний слой, на которых находится 7 или 8 нейронов. Архитектура такой сети представлена на рисунке.

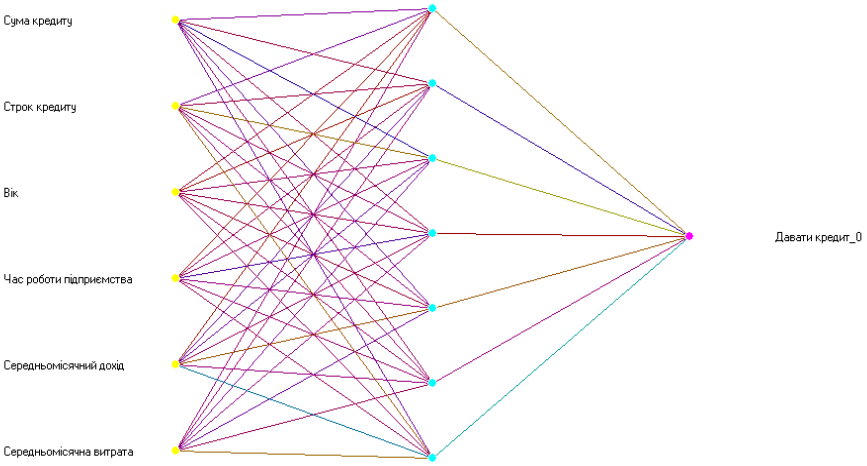

Рисунок 5 - Архитектура оптимальной нейронной сети

Такая нейронная сеть неверно классифицирует только один пример. Дальнейшие попытки повысить качество классификации за счет изменения структуры не дали положительных результатов. Таким образом, рассмотренную выше модель, можно считать наилучшей (оптимальной).

Эта модель может быть использована для оценки новых заемщиков, данных о которых нет в информационной базе банка. Для этого задаются значения шести входных параметров нового клиента, выполняется классификация, в результате которой он относится к одному из двух классов - дать или отказать в выдаче кредита.

Оптимальная модель заемщика была построена для конкретных исходных данных, которые представлены в информационной базе. Поэтому корректной работы этой модели можно ожидать, если параметры новых потенциальных заемщиков будут находиться в тех же пределах. Эти пределы можно определить из статистика, полученной в процессе построения нейронной сети (табл.4).

Таблица 4

Пределы значений входных данных модели

Параметр	Минимальное значение	Максимальное значение
Сумма кредита, грн.	2000	69500
Срок кредита, грн.	6	48
Возраст, лет	19	70
Время работы предприятия, лет	2	45
Среднемесячные доходы, грн.	3500	10500
Среднемесячные расходы, грн.	500	4500

Пусть два новых клиента банка характеризуются определенными значениями входных параметров (табл.5).

Таблица 5

Значения входных параметров новых клиентов

Параметр	Первый клиент	Второй клиент
Сумма кредита, грн.	30000	2000
Срок кредита, грн.	12	6
Возраст, лет	50	20
Время работы предприятия, лет	10	10
Среднемесячные доходы, грн.	8000	5000
Среднемесячные расходы, грн.	3000	3000

Для оценки клиента используется функция **«Что если»**, которая реализована в аналитической платформе Deductor**.** Выходной параметр модели под именем **Выдать кредит** может принимать логическое значение True или False. В первом случае кредит может быть выдан, а во втором будет отказано в выдаче кредита. Результаты моделирования представлены на рисунке 6.

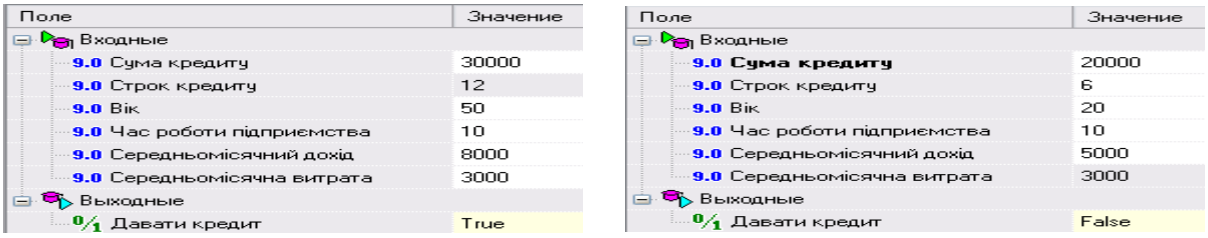

Рисунок 6 –Результаты классификации новых клиентов банка

Так как банк заинтересован в выдаче кредитов, то он может предложить клиенту другие условия кредитования, например, изменить сумму кредита или его срок. Возможно так же предложить клиенту изменить свои среднемесячные доходы и расходы.

Предложенная модель заемщика банка, позволяет определить влияние каждого входного параметров на выходной параметр. На основании таблицы режима «Что если» можно построить диаграммы, определяющие зависимость выходного параметра от каждого входного. В результате будут получены значения входных параметров, при которых возможно получение кредита, т.е. условия кредита, которые удовлетворят и клиента и банк. Глубина анализа и дальнейшие направления исследования будут зависеть от заинтересованности банка в клиенте, а также уменьшения финансовых рисков, связанных с не возвратом выданного кредита.

Качество принятия решения в значительной мере зависит от выполненного прогноза, который в свою очередь зависит от качества модели. Подтверждением этого является выполнение прогноза для новых клиентов с использованием двух разных моделей для одних и тех же значений исходных данных (рис.3). Для первого клиента обе модели дают положительный результат, т.е. кредит может быть выдан. Для второго клиента по первой модели результат будет положительным, а по второй (оптимальной) – отрицательным. Более того исследования показали, что при использовании модели низкого качества кредит может быть выдан уже при среднемесячных доходах свыше 3700 грн. и совсем не зависит от среднемесячных расходов, что является практически невозможным. Таким образом, при использовании такой модели, будет принято решение, которое может привести к финансовым потерям банка.

ВЫВОДЫ

Для оценки заемщика построены модели в виде дерева решений и искусственной нейронной сети с разной архитектурой. Выполнен сравнительный анализ

качества классификации и определена наилучшая модель в виде нейронной сети с одним внутренним слоем с 7 или 8 нейронами. Модель использована для оценки новых клиентов банка и подбора условий кредитования, приемлемых для обеих сторон. Для углубления степени анализа, можно решить многомерную задачу классификации [9,10], когда клиенты распределяются по нескольким классам, для каждого из которых будут свои условия кредитования.

REFERENCES

1. Palkin N.B., Oreshkov V.I. Biznes-analitika: ot dannyh k znaniyam. - SPb : Piter, 2010.-352s.
2. Lepa E.V. Segmentaciya abonentskoj bazy telekomunnikacionnoj kompanii//Problemi informacijnih tekhnologij. – 2014. – №2 (016). – s.119-122.
3. Taha, Hehmdi, A. Vvedenie v issledovanie operacij.: Per. s angl.- M.: Izdatel'skij dom "Vil'yams", 2001. - 912 s.
4. Kompaniya BaseGroup Labs. Deductor. Rukovodstvo analitika. Versiya 5.2.-M.: BaseGroup Labs, 2010. – 122 s.
5. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.A., Holod I.I. Metody i modeli analiza dannyh: OLAP i DATA MINING. – SPb.: BHV-Peterburg, 2004. – 336 s.
6. Dyuk V. A., Samojlenko A. P. Data Mining. Uchebnyj kurs. – SPb.: Piter, 2002. – 368 s.
7. Dyuk V.A. Obrabotka dannyh na PK v primerah. — SPb: Piter, 1997. – 352 s.
8. Базовые навыки работы в Deductor Studio 5.1. Практикум. - BaseGroup Labs, 2007-2008. – 43 с.
9. Lepa E.V., Miheev E.K., Krinitsin V.V. Sistemi pidtrimki priynyattya rishen: Navchalniy posibnik, Chastina 1 – Herson: HEPI, 2006. – 236 s.
10. Lepa E.V., Miheev E.K., Krinitsin V.V. Sistemi pidtrimki priynyattya rishen: Navchalniy posibnik, Chastina 2 – Herson: HEPI, 2006. – 224 s

Рецензент: *д.т.н., проф. Рудакова Г. В.*
Херсонський національний технічний університет