

УДК 311.2

МОДЕЛЬ КРЕДИТНОГО СКОРИНГА КАК АЛГОРИТМ ТИПОЛОГИИ НЕЧЕТКИХ СОВОКУПНОСТЕЙ

В.И. Сонникова, К.О. Кулиджоглян

Новосибирский государственный университет
экономики и управления «НИНХ»
E-mail: veronika_sonniko@mail.ru

В работе обсуждаются теоретические и методические вопросы построения моделей кредитного скоринга, приводятся результаты авторского исследования на реальных данных одного из новосибирских банков, построены дискриминантные функции для разных групп клиентов, выполнена статистическая оценка надежности полученных результатов, сделаны выводы и рекомендации.

Ключевые слова: скоринг, обучающая выборка, модель, классификация, логистическая регрессия.

MODEL OF CREDIT SCORING AS ALGORITHM OF TYPOLOGY OF FUZZY AGGREGATIONS

V.I. Sonnikova, K.O. Kulidzhoglyan

Novosibirsk State University of Economics and Management
E-mail: veronika_sonniko@mail.ru

The paper discusses theoretic and methodic issues of credit scoring model development. The results of the authors research using real date of Novosibirsk bank are presented. Discriminant functions for various groups of clients are formed. The statistical evaluation of reliability of obtained results is made. Conclusions are drawn and recommendations are made.

Key words: scoring, learning sample, model, classification, logistic regression.

Введение. По данным Центрального Банка РФ доля просроченной задолженности по кредитам физических лиц на 01.01.2012 г. составляла 5,56 %. По сравнению с данными на 01.01.2005 г. ее объемы увеличились в 35 раз, в то время как объемы кредитования всего в 8 раз. Соответственно, есть необходимость разобраться в причинах и возможных последствиях данного процесса. Так как кредитные организации функционируют в условиях жесткой конкуренции, им приходится бороться за каждого клиента, поэтому необходимо проявлять большую изобретательность в области разработки новых методов кредитования, тем самым привлекая новых клиентов. Существует острая необходимость в наукоемких механизмах, способных в определенном смысле заменить кредитных экспертов и сократить время анализа заявки и долю субъективизма в принятии решений. Можно предположить, что изучаемое явление в скором времени станет основным конкурентным преимуществом кредитной организации, ориентированной на быстрые потребительские кредиты

Постановка задачи, информационная база исследования. Объектом исследования является совокупность клиентов одного из новосибирских банков, занимающегося в том числе выдачей потребительских кредитов.

Информационную базу статистического анализа составили данные по генеральной совокупности клиентов банка – физическим лицам. Объем генеральной совокупности – более 15 000 чел. Анализ проведен по случайной выборке: 1000 наблюдениям (500 благонадежных/500 дефолтных клиентов банка). Задача классификации — в базе данных клиентов банка содержатся их основные параметры, по которым проводится анализ, необходимо предложить методику сегментирования нечеткого множества клиентов с использованием статистических пакетов, а также сделать выводы по полученным результатам, таким образом, чтобы был указан статус клиента: 0 (хороший клиент) или 1 (плохой клиент) [5, 8, 9]. Для решения поставленной задачи применен пакет прикладных программ STATISTICA и база данных по 1000 выданным кредитам (500 хороших/500 плохих), отобранных случайным образом и состоящих из двух частей: обучающая выборка (80 %) – 800 наблюдений и контролирующая (20 %) – 200 наблюдений.

Инструментарий. В специальной литературе принято выделять три основных метода определения степени надежности заемщика: изучение кредитной истории; оценка на основе финансовых показателей платежеспособности; скоринговая оценка.

Оценка надежности клиента на основе первых двух методов трудоемка, не оперативна и в большинстве случаев неэффективна. Кроме того, изучением кредитной истории занимается банковский работник, в результате на принимаемое решение очень сильно воздействует субъективный фактор, а неточность финансовых показателей, возникающая по разным причинам, часто приводит к ошибкам и возникновению рискованных ситуаций. Поэтому существует необходимость в наукоемких механизмах, способных в определенном смысле заменить кредитных экспертов, сократить время анализа заявки и одновременно уменьшить долю субъективизма в принятии решения. В качестве такого механизма может быть использована скоринговая модель.

«Скоринг – это метод оценки благонадежности клиента на основании обработки информации о поведении аналогичных клиентов в прошлом либо экспертных знаний. Скоринговая модель – это математическая модель, предсказывающая, вернет или нет клиент кредит в срок» [13].

«Таким образом, скоринг является классификационной задачей, в которой исходя из имеющейся информации необходимо получить функцию, наиболее точно разделяющую выборку клиентов на платежеспособных и неплатежеспособных» [14].

Проще говоря, скоринговая модель – взвешенная сумма определенных характеристик, в результате чего получается объединенный показатель, и чем он выше, тем выше надежность клиента. Данный показатель сравнивается с неким пороговым значением, или линией раздела, которая является линией безубыточности и рассчитывается из отношения, сколько в среднем нужно клиентов, которые платят в срок, для того, чтобы компенсировать убытки от данного должника.

Сложные скоринговые системы используют три составляющие анкеты:

1) «жесткие» параметры, подтвержденные документами, которые заемщик принес с собой (паспорт, водительские права, справка о доходах с места работы и пр.);

2) «мягкие» параметры (образование, знание языков, количество выездов за рубеж, наличие электронной почты, как корпоративной, так и личной, место работы, наличие недвижимости, автомобиля и пр.);

3) «замаскированный» в разных частях анкеты заемщика небольшой психологический тест, задача которого – определить достоверность информации, представленной клиентом, и его психологический портрет» [10, с. 118].

При построении скоринговой модели возникает две основные проблемы. Первой является определение характеристик, которые нужно включать в построение модели и которые должны быть наиболее тесно связаны с ненадежностью или надежностью клиента. Данные характеристики должны содержать в себе необходимый объем информации, с помощью которого можно будет их классифицировать. С данными характеристиками связана одна из основных проблем скоринга, которая заключается в динамичности развития социальных и экономических процессов, с течением времени люди, обстоятельства, условия, могут измениться. Поэтому скоринговые модели необходимо разрабатывать на выборке из наиболее «свежих» клиентов, периодически проверять качество работы системы и при ухудшении обновлять или создавать новую модель [10].

Второй сложностью является тот факт, что типология выборки производится только на клиентах, которым дали кредит. Информация по состоявшимся кредитам не может быть использована в качестве обучающей выборки, поскольку она не содержит нужных сведений. Это создает методологическую проблему: фактически неизвестно, как бы повели себя клиенты, которым в кредите было отказано, т.е. вполне возможно, что какая-то часть оказалась бы вполне приемлемыми заемщиками. Но даже если бы все из отклоненных соискателей на самом деле дополнили подвыборку только отрицательных претендентов, то и в этом случае с ненулевой вероятностью скоринговые расчеты отличались бы от тех, что получены по фактическим данным. Таким образом, если в скоринговых расчетах опираться только на фактические данные по выданным кредитам (т.е. по состоявшимся заемщикам), то оценки кредитоспособности новых соискателей будут содержать некоторую систематическую ошибку. Смещение результатов скоринга происходит из-за того, что аппликант – это еще не заемщик, и, оставляя в обучающей выборке только состоявшихся заемщиков, мы изначально ее искажаем. То есть новые соискатели кредита принадлежат к другой генеральной совокупности, чем та, из которой была взята обучающая выборка. Таким образом, в данной ситуации возможна так называемая «ошибка исчезающей совокупности» [3, 4, 7].

Результаты. Заемщики в нашем примере измерены пятнадцатью параметрами, каждый из которых имеет градации. Из заявлений – анкет на кредит были получены следующие данные (табл. 1).

Проведем дискриминантный анализ клиентов банка с целью построения скоринговой модели. Он является эффективным способом построения классификации с помощью обучающей выборки. Анализ проводится в модуле Discriminant Analysis пакета STATISTICA. Для построения модели воспользуемся 800 наблюдениями, зависимый фактор которых принимает два значения: возврат кредита и его дефолт.

Таблица 1

Характеристики заемщиков из заявлений на кредит

Характеристика	Тип переменной, шкала измерения	Возможные значения (атрибуты)
Возраст заемщика	Непрерывная, интервальная	От 16 до 100
Доход по основному месту работы	Непрерывная, отношений	От 0 до ∞
Образование	Дискретная, ранговая	1. Начальное, неполное среднее, нет 2. Среднее, в том числе специальное 3. Неполное высшее 4. Высшее 5. Два и более высших образований, ученая степень
Населенный пункт	Дискретная, номинальная	1. Село, ПГТ 2. Областной центр 3. Город
Квартира в собственности	Дискретная, номинальная	1. Нет 2. Есть
Пол	Дискретная, номинальная	1. Мужской 2. Женский
Семейное положение	Дискретная, номинальная	1. Вдовец/вдова 2. Женат/замужем 3. Разведен/разведена 4. Сожительство 5. Холост/не замужем
Количество машин в собственности	Дискретная, интервальная	1. Нет 2. Одна 3. Два и более
Количество иждивенцев	Дискретная, интервальная	1. Нет 2. Один 3. Два и более
Тип организации	Дискретная, номинальная	1. Государственная 2. Негосударственная 3. Нет
Период проживания	Непрерывная, отношений	От 0 до ∞
Стаж работы на текущем месте	Непрерывная, отношений	От 0 до ∞
Должность	Дискретная, номинальная	1. Индивидуальный предприниматель 2. Неруководящая должность 3. Руководящая должность
Тип места жительства	Дискретная, номинальная	1. Свое 2. Коммерческая аренда 3. С родителями 4. Социальное (общежитие) 5. Другое
Сумма кредита	Непрерывная, отношений	От 0 до ∞

Главной целью дискриминантного анализа является выявление признаков, которые оказывают наибольшее влияние на различение зависимой переменной.

Программа предлагает 3 метода отбора значимых переменных. Стандартный метод, когда все переменные будут одновременно включены в модель, пошаговый с включением – на каждом шаге в модель выбирается переменная с наибольшим F -значением, и пошаговый с исключением –

Таблица 2

Матрица классификации наблюдений

Classification Matrix (ДИПЛОМ) Rows: Observed classifications Columns: Predicted classifications			
Group	Percent Correct	надежный p=,50000	невозврат p=,50000
надежный	77,77778	315	90
невозврат	81,77215	72	323
Total	79,75000	387	413

Таблица 3

Матрица классификации после исключения нетипических наблюдений

Classification Matrix (ДИПЛОМ) Rows: Observed classifications Columns: Predicted classifications			
Group	Percent Correct	надежный p=,50000	невозврат p=,50000
надежный	97,13376	305	9
невозврат	98,75389	4	317
Total	97,95276	309	326

в уравнение включаются все выбранные пользователем переменные, которые затем удаляются в зависимости от величины *F*-значения.

Выбираем стандартный метод и для начала включаем все переменные в расчет классификационной функции. В результате переменные, которые не являются значимыми, были постепенно исключены из рассмотрения. Остались только переменные, которые будут участвовать в построении модели. К ним относятся: возраст, доход, наличие квартиры в собственности, семейное положение, тип организации, период проживания по текущему адресу, тип места жительства и сумма кредита.

После того, как информативные переменные отобраны, проверяется корректность обучающих выборок. Для этого используется матрица классификации (табл. 2). Она содержит информацию о количестве и проценте корректно классифицированных наблюдений в каждой группе. Строки матрицы — исходные классы, столбцы — предсказанные классы.

По данным табл. 2 можно сделать вывод, что 80 % объектов было правильно отнесено экспертным способом к выделенным группам. Причем модель лучше угадывает неблагонадежных клиентов, что является плюсом для банка, так как убытки по выданным кредитам неплатежеспособным клиентам намного превышают сумму недополученной прибыли.

Задача получения корректных обучающих выборок состоит в том, чтобы исключить из обучающих выборок те объекты, которые по своим показателям не соответствуют большинству объектов, образующих однородную группу. Для этого определили с помощью метрики Махаланобиса расстояние от всех *n* объектов до центра тяжести каждой группы, определяемых по обучающей выборке. Исключим наблюдения, расстояние которых до центра их группы значительно выше, чем от них до центра другой группы. В табл. 3 представлены полученные данные.

Исключив нетипичные наблюдения, получили модель, которая верно угадывает 98 % наблюдений. Получены классификационные функции:

– для группы благонадежных заемщиков

$$Y = -2,7X_1 + 503,8X_2 + 103,2X_3 + 787,6X_4 + 255,6X_5 + 150,2X_6 - 91224,$$

– для группы неблагонадежных заемщиков

$$Y = -3X_1 + 501,6X_2 + 104,5X_3 + 785,3X_4 + 257,6X_5 + 151,8X_6 - 91261,$$

где *X*₁ – возраст заемщика; *X*₂ – наличие квартиры в собственности; *X*₃ – семейное положение; *X*₄ – тип организации; *X*₅ – период проживания по адресу; *X*₆ – тип места жительства.

Теперь, на основе полученных обучающих выборок можно провести классификацию тех объектов, которые не попали в обучающие выборки. Для этого используется модуль Апостериорные вероятности. Результаты классификации наблюдений, содержащихся в контролирующей выборке (200 объектов) представлены в табл. 4. Таким образом, полученная модель может правильно предсказывать 75 % наблюдений. Следовательно, это ее процент достоверности.

Таблица 4

Матрица классификации объектов контролирующей выборки

	Достоверность, %	Надежные клиенты	Ненадежные клиенты
Надежные клиенты	75	75	25
Ненадежные клиенты	75	25	75
Итого	75	100	100

Оценить группу надежных заемщиков можно также с помощью аппарата множественной логистической регрессии и разработкой статистической карты.

Перед проведением анализа все атрибуты кодируются двоичным кодом. Например, отформатированная переменная «Образование», состоящая из пяти атрибутов, получает следующую кодировку (табл. 5).

Таблица 5

Закодированные двоичным кодом атрибуты признака «Образование»

Два и более высших образования, ученая степень	Высшее	Неполное высшее	Среднее	Начальное, нет
0	0	0	1	0
0	1	0	0	0
1	0	0	0	0
0	0	0	0	1
0	0	0	0	1

После этого параметры будущей модели оцениваются при помощи логистической регрессии [8]. В качестве зависимой переменной выступит показатель того, что клиент является надежным, в качестве независимых – атрибуты, которые будут использованы при построении модели. Результаты представлены в табл. 6. Для удобства интерпретации полученные коэффициенты функции умножены на 1000 с округлением до целых значений.

Проведя определенные преобразования в пакете анализа Excel, получен сводный балл по каждому наблюдению отдельно и построен интервальный ряд распределения по каждому виду кредита (табл. 7).

Используя данные распределения, строятся графики кривых распределений кредитов (рис. 1).

Результаты классификации данных представлены в табл. 8.

Таблица 6

Баллы, полученные на основе модели логистической регрессии

Группа признаков	Признак	Балл
Начальный балл		-1534
Возраст	Менее 34	-959
	35–48	453
	Более 48	1269
Образование	Два и более высших образований, ученая степень	1392
	Высшее	403
	Неполное высшее	-355
	Среднее, в том числе специальное	-356
	Нет, начальное, неполное среднее	-2218
Населенный пункт	Город	-172
	Областной центр	-668
	Село, ПГТ	-495
Квартира в собственности	Есть	-433
	Нет	-1 001
Пол	Мужской	-960
	Женский	-474
Семейное положение	Женат/замужем	834
	Разведен/разведена	-557
	Вдовец/вдова	-1602
	Сожительство	25
	Холост/не замужем	166
Количество машин	Нет	-969
	Один	-644
	Две и более	278
Количество иждивенцев	Нет	-293
	Один	-227
	Два и более	-814
Период проживания	Более 5 лет	323
	От 1 года до 5 лет	-197
	Менее 1 года	-1461
Стаж работы на месте	Менее полугода	-122
	От 6 мес. до 1 года	561
	От 1 года до 3 лет	-562
	Более 3 лет	11
Должность	Индивидуальный предприниматель	-1757
	Неруководящая должность	138
	Руководящая должность	284
Отношение суммы кредита к величине дохода заемщика	0,5–3,2	5304
	3,2–5,9	4495
	5,9–8,6	5185
	8,6–11,3	3542
	11,3–14	-19 660

Таблица 7

Ряд распределения суммарных баллов

Интервал	Количество «хороших» кредитов	Количество «плохих» кредитов	Интервал	Количество «хороших» кредитов	Количество «плохих» кредитов
–3300—–3000	0	1	1500—1800	26	30
–3000—–2700	0	1	1800—2100	29	26
–2700—–2400	0	3	2100—2400	26	10
–2400—–2100	0	4	2400—2700	35	9
–2100—–1800	1	10	2700—3000	31	14
–1800—–1500	1	6	3000—3300	35	8
–1500—–1200	0	16	3300—3600	23	3
–1200—–900	3	14	3600—3900	28	0
–900—–600	1	11	3900—4200	26	0
–600—–300	5	18	4200—4500	15	0
–300—0	9	28	4500—4800	12	1
0—300	6	25	4800—5100	10	0
300—600	12	45	5100—5400	7	0
600—900	20	51	5400—5700	3	0
900—1200	17	27	5700—6000	2	0
1200—1500	22	31	6000—6300	0	0

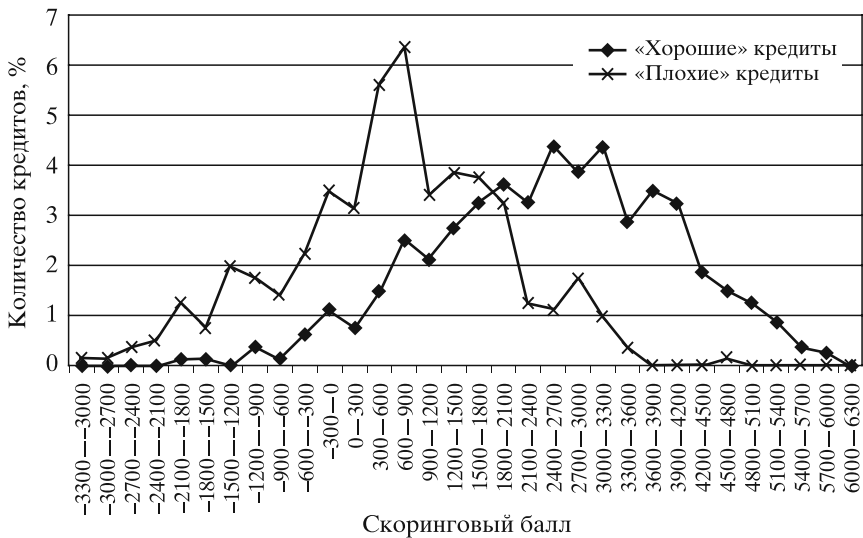


Рис. 1. Кривые распределения для кредитов в относительном выражении

Таблица 8

Матрица классификации наблюдений

Observed	Classification of Cases (ЛОГИСТИЧЕСКАЯ) Odds ratio: 9,6464 Perc. correct: 75,63%			Percent Correct		
	Pred. невозврат	Pred. надежный				
невозврат	302	93	76,45570			
надежный	102	303	74,81481			



Рис. 2. Кумулятивные кривые распределения

Полученная модель верно предугадывает неблагонадежных заемщиков на 76,4 %, а благонадежных на 74,8 %. Общий процент достоверности равен 75,63 %. Для того чтобы определить качество модели, используется статистика Колмогорова–Смирнова. Она позволяет понять, насколько хорошо модель разделяет «хороших» и «плохих» клиентов, показывает, как далеко друг от друга находятся функции распределения вероятности для «плохих» и «хороших» клиентов. Данный показатель определяется как максимальная разница между накопленным количеством «плохих» и «хороших» кредитов. Для этого строятся кумулятивные кривые распределения для кредитов (рис. 2).

Значение статистики для логрегрессионной скоринговой карты К–С максимально в диапазоне от 1500 до 1800 баллов и составляет 52 %, что свидетельствует о среднем качестве построенной модели. Данный интервал можно рассматривать как порог отсека для классификации заемщиков [7, 12].

Выводы. В результате использования дискриминантного анализа и аппарата множественной логистической регрессии были получены скоринговые модели оценки надежности заемщиков. Несмотря на различия данных методов, обе модели показали одинаковый уровень достоверности – 75 %. Таким образом, оба способа анализа являются сильными средствами идентификации нечетких совокупностей: дискриминантный анализ позволяет выделить наиболее значимые переменные, по которым в дальнейшем строится разграничение, логистическая регрессия более удобна для пользователей, так как результат – просуммированные баллы, которые сравниваются с пороговым значением. Проверка показала, что полученные модели соответствуют требованиям, предъявляемым к качеству, и могут быть применены на практике для идентификации нечетких совокупностей.

Литература

1. Бессокирная Г.П. Дискриминантный анализ для отбора информативных переменных // Социология. 2003. № 16. С. 25–35.
2. Бокун Н. Ч., Чернышева Т.М. Методы выборочных обследований: учебник-справочник. Минск, 1997. 416 с.
3. Глинский В.В. Мифическая статистика малого бизнеса. Проблемы статистического изучения турбулентных совокупностей // ЭКО. 2008. № 9. С. 51–61.
4. Глинский В.В. Как измерить малый бизнес? // Вопросы статистики. 2008. № 7. С. 73–75.
5. Глинский В.В. Статистические методы поддержки управленческих решений: монография. Новосибирск: Издательство НГУЭУ, 2008. С. 108–118.
6. Глинский В.В., Макаридина Е.В. О модели жизненного цикла высшего профессионального образования России // Национальные интересы: приоритеты и безопасность. 2011. № 3. С. 12–18.
7. Глинский В.В., Серга Л.К. Нестабильные совокупности: концептуальные основы методологии статистического исследования // Вестник НГУЭУ. 2009. № 2. С. 137–142.
8. Глинский В.В., Серга Л.К. О государственном регулировании малого предпринимательства в России // Национальные интересы: приоритеты и безопасность. 2011. № 19. С. 2–8.
9. Глинский В.В., Гусев Ю.В., Золотаренко С.Г., Серга Л.К. Портфельный анализ в типологии данных: методология и применения в поддержке управленческих решений // Вестник НГУЭУ. 2012. № 1. С. 25–54.
10. Жариков В.В., Жарикова М.В., Евсейчев А.И. Управление кредитными рисками: учеб. пособие. Тамбов: Изд-во Тамб. гос. техн. ун-та, 2009. 244 с.
11. Кулиджоглян К.О. Критерии принадлежности и проблема оценки среднего класса // Бизнес-статистика, финансы и банки: Теоретические и методические аспекты исследования: сб. науч. тр. Новосибирск: НГУЭУ, 2011. 220 с.
12. Серга Л.К. Об одном подходе к определению пороговых значений в решении задач классификации // Вестник НГУЭУ. 2012. №1. С. 54–60.
13. Уланов С.В. Оценка качества и сравнение скоринговых карт // Экономические науки. 2009. № 9 (58). С. 330–335.
14. Черкашенко В.Н. Этот «загадочный» скоринг // Банковское дело. 2006. № 3. С. 42–48.

Bibliography

1. Bessokirnaja G.P. Diskriminantnyj analiz dlja otbora informativnyh peremennyh // Sociologija. 2003. № 16. P. 25–35.
2. Bokun N.Ch., Chernysheva T.M. Metody vyborochnyh obsledovanij: uchebnik-spravochnik. Minsk, 1997. 416 p.
3. Glinskij V.V. Mificheskaja statistika malogo biznesa. Problemy statisticheskogo izuchenija turbulentnyh sovokupnostej // JeKO. 2008. № 9. P. 51–61.
4. Glinskij V.V. Kak izmerit' malyj biznes? // Voprosy statistiki. 2008. № 7. P. 73–75.
5. Glinskij V.V. Statisticheskie metody podderzhki upravlencheskih reshenij: monografija. Novosibirsk: Izdatel'stvo NGUJeU, 2008. P. 108–118.
6. Glinskij V.V., Makaridina E.V. O modeli zhiznennogo cikla vysshego professional'nogo obrazovaniya Rossii // Nacional'nye interesy: prioritety i bezopasnost'. 2011. № 3. P. 12–18.
7. Glinskij V.V., Serga L.K. Nestabil'nye sovokupnosti: konceptual'nye osnovy metodologii statisticheskogo issledovaniya // Vestnik NGUJeU. 2009. № 2. P. 137–142.
8. Glinskij V.V., Serga L.K. O gosudarstvennom regulirovanii malogo predprinimatel'stva v Rossii // Nacional'nye interesy: prioritety i bezopasnost'. 2011. № 19. P. 2–8.

9. *Glinskij V.V., Gusev Ju.V., Zolotarenko S.G., Serga L.K.* Portfel'nyj analiz v tipologii dannyh: metodologija i primenenija v podderzhke upravlencheskih reshenij // Vestnik NGUJeU. 2012. № 1. P. 25–54.
10. *Zharikov V.V., Zharikova M.V., Evsejchev A.I.* Upravlenie kreditnymi riskami: ucheb. posobie. Tambov: Izdatel'stvo Tamb. gos. tehn. un-ta, 2009. 244 p.
11. *Kulidzhogljjan K.O.* Kriterii prinadlezhnosti i problema ocenki srednego klassa // Biznes-statistika, finansy i banki: Teoreticheskie i metodicheskie aspekty issledovanija: sb. nauch. tr. Novosibirsk: NGUJeU, 2011. 220 p.
12. *Serga L.K.* Ob odnom podhode k opredeleniju porogovyh znachenij v reshenii zadach klassifikacii // Vestnik NGUJeU. 2012. №1. P. 54–60.
13. *Ulanov S.V.* Ocenka kachestva i sravnenie skoringovyh kart // Jekonomicheskie nauki. 2009. № 9 (58). P. 330–335.
14. *Cherkashenko V.N.* Jetot «zagadochnyj» skoring // Bankovskoe delo. 2006. № 3. P. 42–48.