Фот Н.П.
ГОУ ВПО «Оренбургский государственный университет»

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОЦЕНКИ КРЕДИТОСПОСОБНОСТИ

В статье рассматривается возможность применения методов многомерного статистического анализа при оценке кредитоспособности заемщиков коммерческих банков. Также проведен анализ применения ряда рассмотренных методов к известным кредитным историям.

Оценке кредитоспособности заемщиков коммерческих банков сопутствует явное противоречие: с одной стороны кредитный отдел не способен рассматривать каждую заявку индивидуально ввиду большой трудоемкости и длительности процесса, с другой – есть необходимость увеличения количества и объема выдаваемых кредитов. Одним из путей разрешения данного противоречия является скоринг-кредитование, получившее в последнее время большую популярность в крупных кредитных организациях. Скоринг-кредитование позволяет исключить риск влияния субъективного мнения менеджера на принятие решения о выдаче кредита, значительно упрощает процесс кредитования, способствует дальнейшему наращиванию объемов кредитования [4].

Скоринг-модель обладает такими ключевых преимуществами, как:

1) Сокращение сроков принятия решения о предоставлении кредита. Увеличение числа и скорости обработки заявок за счет минимизации документооборота при выдаче кредита частным клиентам;

2) Эффективная оценка и постоянный контроль уровня рисков конкретного заемщика;

3) Снижение влияния субъективных факторов при принятии решения о предоставлении кредита. Обеспечение объективности в оценке заявок кредитными инспекторами во всех филиалах и отделениях банка;

4) Оценка и управление риском портфеля кредитов частным лицам банка в целом, включая его отделения. Учет уровня доходности и риска кредитного портфеля при определении параметров новых кредитов.

В основе математической модели скоринга, предложенной зарубежными представителями банковской деятельности, заложены методы прогнозирования поведения заемщика на основе имеющейся обширной статистики. Однако при использование зарубежных данных в разработке скоринг-модели для отечественных кредитных организаций возникает проблема, связанная с отсутствием достаточного количества достоверных кредитных историй (вследствие кратковременности работы отечественных организаций на рынке банковских услуг).

Методология решения базируется на анализе специфики деятельности банка. При этом учитываются как группы клиентов (отраслевая и региональная принадлежность и др.), так и кредитные продукты банка для физических лиц. Исходя из потребностей банка в развитии бизнеса и имеющихся данных, могут быть построены скоринг-модели, основанные на экспертных знаниях банковского менеджмента, на статистических данных (модели обучения «с учителем» и «без учителя»), на учете макроэкономических данных о социально-экономическом развитии конкретных регионов и отраслей. Наиболее мощными по точности оценки кредитного риска являются модели, использующие комплексный подход, т.е. учет всех данных и экспертных знаний менеджмента банка [4].

При построении скоринг-моделей используют три основных варианта. Первый – это настройка модели под профиль некоего целевого клиента, второй— приобретение уже готовой модели, списанной с другой страны (однако здесь банку придется оттачивать требования к заемщикам на основе собственного опыта использования этой модели) и третий – создание собственной модели, настроенной на индивидуальные особенности банка.

Третий вариант зачастую используют мощные банки, выходящие на рынок с новыми продуктами. Для этого на первом этапе от применения скоринга отказываются в принципе, давая кредиты всем желающим. Банк устраивает клиентам лишь стандартный security check, чтобы отсечь мошенников. Набрав историю по паре тысяч дефолтов, банк может приступать к разработке своей скоринговой модели, причем такой подход выливается в достаточно крупные денежные затраты (так, при средней сумме кредита в 5-10 тыс. долларов цена опыта составит порядка 10-20 млн. долларов) [4].

В данной работе, посвященной разработке универсального алгоритма оценки кредитоспособности заемщиков в кредитных организациях, проведен анализ совместного использования методов многомерного статистического анализа и методов непараметрической статистики на базе известных 1209 кредитных историй одного из банков Оренбургской области, с последующим выявлением факторов влияния и оценки влияния на показатели кредитоспособности.

На первом этапе исследования, посредством методов кластерного анализа, проведена классификация всех имеющихся кредитных историй.

Под классификацией мы подразумеваем разделение рассматриваемой совокупности объектов на однородные в определенном смысле группы. В общем случае понятие однородности объекта определяется правилом задания вычисления величины ρ_{ij}, характеризующей либо расстояние $d(O_i,O_j)$ между объектами O_i и O_j, либо степень близости (сходства) $r(O_i,O_j)$. Если задана функция $d(O_i,O_j)$, то близкие в смысле этой метрики объекты считаются однородными, принадлежащими одному классу (2). .

В качестве метрики между объектами также выбирается обычное Евклидово расстояние, а между группами объектов – расстояние, вычисленное по принципу «ближайшего соседа»:

$$\rho_{\min}(S_l,S_m) = \min_{x_i \in S_l, x_j \in S_m} d(x_i,x_j) \qquad (1)$$

Однако ряд показателей, представляемых в анкетах банков, не имеют количественной оценки (пол, возраст, образование и т.п.), и поэтому использование обычных метрик при вычислении расстояний между объектами, предусмотренных в методах классификации «без учителя» (кластерном анализе), будет неправильным. В этой связи, представляется возможным 2 подхода: исключение качественных показателей из рассмотрения и проведение классификации только по количественным признакам или применение вида расстояния, предложенного Б.А. Лагошей [3].

В первом подходе предлагается рассмотреть отдельно качественные и количественные признаки объектов, а потом на их основе построить выпуклую линейную комбинацию.

Высчитывается коэффициент связности c_{ij}, учитывающий одновременно качественные и количественные признаки, по формуле:

$$c_{ij} = \frac{n-n_1}{n}\bar{c}_{ij} + \frac{n_1}{n}\tilde{c}_{ij}\ , \qquad (2)$$

где n – общее число признаков;

n_1 – число количественных из них;

$\bar{c}_{ij}$; $\tilde{c}_{ij}$ – коэффициенты связности, обусловленные качественными и количественными признаками, при этом больший вес получают те признаки, количество которых больше.

Использование данного метода позволяет всесторонне оценить кредитоспособность заемщиков, учитывая разную физическую природу идентифицируемых показателей, однако, его использование возможно только при разработке соответствующего программного обеспечения, реализации которого нет ни в одном из разработанных статистических модулей.

Во втором подходе – при исключении качественных показателей, как правило, пользуются агломеративными методами кластерного анализа [2].

На базе известных кредитных историй, с целью определения количества классов заемщиков банка, автором была произведена классификация агломеративными методами кластерного анализа заемщиков по 11 социально-экономическим признакам (результаты представлены в таблице 1) и выявлены 4 социально-экономических класса. В таблице 2 представлены средние значения показателей для каждого класса.

В результате анализа данных были выявлены некоторые особенности представителей каждого класса. Первый класс преимущественно составляют женатые мужчины в возрасте 25-30 лет, имеющие высшее или неоконченное высшее образование, с высоким уровнем доходов. В кредит берут суммы,

Таблица 1. Количественное распределение кредитных историй по классам

Номер класса	Количество кредитных дел	Количество должников	Процент невозврата (в процентах)
класс 1	76	8	10,67
класс 2	420	78	18,62
класс 3	378	13	3,45
класс 4	322	19	5,92

Таблица 2. Средние значения показателей каждого класса

Показатель	Класс 1	Класс 2	Класс 3	Класс 4
Сумма кредита, руб.	333167,1	92503,64	101076,1	116063,5
Процентная ставка, %	16,58	19,35	14,72	15,1
Срок кредита, мес.	44,79	33,74	45,27	48,33
Сумма поручительств, руб.	498216,1	138292,7	92645,51	152595,6
Сумма залога, руб.	170680,7	1976,88	19017,46	19230,8

превышающие среднюю величину, под высокий процент.

Холостые мужчины в возрасте 35-45 лет, имеющие среднее образование с небольшим совокупным доходом, берущие в кредит небольшие суммы на небольшой период и под высокую процентную ставку, а также не имеющие поручителей и не предоставляющие имущество под залог, составляют второй класс.

Третий класс в равной степени состоит как из мужчин, так и из женщин со средне-специальным образованием, имеющих семью и не высокий уровень совокупного дохода. В кредит берут небольшую сумму, под низкий процент на длительный срок, имеют одного поручителя.

И четвертый класс – в основном состоит из одиноких женщин в возрасте 45-50 лет, имеющих средне-специальное образование, невысокий уровень дохода. В кредит берут небольшие суммы под низкую процентную ставку, на длительный период.

Учитывая опыт зарубежных банков, достаточно весомыми при решении о выдаче кредита являются такие показатели, как образование, возраст, пол и т.п. (например, в Германии приоритет при получении кредита имеют пенсионеры). Выявление таких признаков для отечественных представителей кредитных организаций было проведено методами непараметрического корреляционного анализа (ранговой корреляции), в частности, с использованием коэффициентов ранговой корреляции Спирмэна и Кендалла [1].

Коэффициент корреляции Спирмэна служит показателем степени тесноты связи и направления связи между ранжировками $x^{(k)}=(x_1^{(k)},x_2^{(k)},...,x_n^{(k)})^T$ и $x^{(j)}=(x_1^{(j)},x_2^{(j)},...,x_n^{(j)})^T$ между признаками k и j:

$$\tau^{(s)}{}_{kj}=1-\frac{6}{n^3-n}\sum_{i=1}^{n}(x_i^{(k)}-x_j^{(j)})^2. \qquad (3)$$

Для совпадающих ранжировок (т.е. при $x_i^{(k)}=x_i^{(j)}$ $i=1,..n$) $\tau^{(S)}{}_{kj}=1$, а для противоположных (т.е. при $x_i^{(k)}=n-x_{i+1}^{(j)}$ $i=1,..n$) – $\tau^{(S)}{}_{kj}=-1$.

Во всех остальных случаях $|\tau^{(S)}{}_{kj}|<1$.

Другой широко используемой характеристикой является ранговый коэффициент корреляции Кендалла $\tau^{(K)}{}_{kj}$ между признаками $x^{(k)}$ и $x^{(j)}$, вычисляемый по формуле:

$$\tau^{(K)}{}_{kj}=1-\frac{4\nu(x^{(k)},x^{(j)})}{n(n-1)}, \qquad (4)$$

где $\nu(x^{(k)},x^{(j)})$ – минимальное число обменов соседних элементов последовательности $x^{(j)}$ ($x^{(k)}$), необходимых для приведения ее к упорядочению $x^{(k)}$ ($x^{(j)}$).

Аналогично, при совпадающих ранжировках (т.е. при $x^{(k)}=x^{(j)}$) $\tau^{(K)}{}_{kj}=1$ (так как $\nu(x^{(k)},x^{(j)})=0$), а при противоположных ранжировках (т.е. при $x_i^{(k)}=n-x_{i+1}^{(j)}$) $\tau^{(K)}{}_{kj}=-1$ (так как $\nu(x^{(k)},x^{(j)})=\frac{1}{2}n(n-1)$), во всех остальных случаях – $|r^{(K)}{}_{kj}|<1$.

В программном продукте «Statistica 6.0» были вычислены значения ранговых коэффициентов Спирмена и Кендала между показателем «статус заемщика» (данному показателю присваивалось значение «1», если заемщик имеет задолженность и «0», если ее нет) и показателями, являющиеся неколичественными (такие как, пол, образование и т.п.). Необходимо отметить, что данный программный продукт позволяет сразу проверить значимость полученных коэффициентов – дополнительно вычисляется вероятность принятия гипотезы о значимости коэффициентов, при этом значимые показатели выделяются на экране красным.

Причиной для отказа в выдаче кредита заемщику – мужчине, отнесенного после предварительной классификации в первый класс, может являться взятие кредита на короткий срок и под высокую процентную ставку.

При отнесении потенциального заемщика во второй класс банк принимает на себя меньший риск при выдаче кредита людям среднего и пожилого возраста, имеющим семью.

При оценке кредитоспособности заемщика, отнесенного к третьему классу, менеджер должен отдавать предпочтение людям, имеющим высшее или неоконченное высшее образование, берущим кредит на длительный период и имеющим два и более поручителей.

В четвертом классе также, как и для первого класса, в невыигрышном положении оказываются заемщики, берущие кредит под высокую процентную ставку на короткий срок.

Однако даже совместное применение методов кластерного анализа и ранговой корреляции позволяют только выделить количество классов и сделать выводы о социальном статусе заемщиков, а граница между полученными классами носит формальный характер. В связи с этим, на следующем этапе исследования были использованы методы дискриминантного анализа, что позволило определить численное значение параметра, выполняющего роль границы между выделенными классами.

Основной принцип классификации в дискриминантном анализе заключается в следующем: объект следует отнести к тому классу (т.е. к той генеральной совокупности), в рамках которого он выглядит более правдоподобным [1].

Ключевым является понятие процедуры классификации (решающее правило дискриминантной функции) δ(x). Данная функция может принимать только целые положительные числа от 1 до k, причем те значения x, при которых она принимает значение, равное j, относятся к классу j, т.е. $S_j = \{x : \delta(x) = j\}$. S_j – это n-мерные области в пространстве P(x) возможных значений анализируемого признака x, причем δ(x) строится таким образом, чтобы $S_1 + ... + S_k$ заполняли все пространство P(x) и чтобы они попарно не пересекались.

Процедура классификации называется оптимальной (байесовской), если она сопровождается минимальными потерями среди всех других процедур классификации. Процедура классификации, при которой потери будут минимальными, определяется следующим образом:

$$S_j^{опт} = \left\{ x : \sum_{i=1}^{k} \pi_i f_i(x) c(j/i) = \min_{1 \le l \le k} \sum_{\substack{i=1 \\ i \ne l}}^{k} \pi_i f_i(x) c(l/i) \right\} \quad (5)$$

Т.е. объект x_v (v=1…N) будет отнесен к классу j тогда, когда средние удельные потери от его отнесения в этот класс окажутся минимальными по сравнению с потерями при отнесении этого объекта в любой другой класс.

В случае равных потерь правило классификации приобретает более простой вид. Наблюдение x_v будет отнесено к классу j тогда, когда

$$\pi_j f_j(x_v) = \max_{1 \le l \le k} \pi_l f_l(x_v), \quad (6)$$

т.е. максимизируется взвешенная правдоподобность этого наблюдения в рамках класса, где в качестве весов выступают априорные вероятности π_j.

В вышеприведенных отношениях задаются лишь теоретически оптимальные классификации. Для их реализации необходимо знание априорных вероятностей $\pi_1,...,\pi_k$ и знание распределения в каждом классе, т.е. $f_1(x),...,f_k(x)$.

На практике решение этой задачи осуществляется при замене теоретических характеристик соответствующими оценками, построенными на базе обучающих выборок.

Априорные вероятности π_j оценивается следующим образом:

Таблица 3. Показатели, выявленные посредством методов ранговой корреляции

№ класса	Значимые показатели
1	пол, процентная ставка, срок кредита в месяцах
2	семейное положение, возраст
3	образование, срок кредита, число поручителей
4	пол, процентная ставка, срок кредита в месяцах

$$\pi_j = \frac{n_j}{n_{общ}}, \; n_{общ} = n_1 + ... + n_k \qquad (7)$$

Правило классификации в случае k=2 заключается в следующем: наблюдение x_v относится к классу $j_0 \Leftrightarrow \frac{f_{j_0}(x_v)}{f_j(x_v)} \geq \frac{\pi_j}{\pi_{j_0}} \; \forall j = \overline{1..k}$

или $\ln\frac{f_{j_0}(x_v)}{f_j(x_v)} \geq \ln\frac{\pi_j}{\pi_{j_0}} \; \forall j = \overline{1..k}$

В случае нормальных плотностей $f_j(x)$, т.е. $f_j(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|^n}} \exp\left(-\frac{1}{2}(x-a_j)^T \Sigma^{-1}(x-a_j)\right)$, соотношение эквивалентно следующему:

$$\left(x_v - \frac{1}{2}(\hat{a}_{j_0} - \hat{a}_j)\right)^T \hat{\Sigma}^{-1}(\hat{a}_{j_0} - \hat{a}_j) \geq \ln\frac{\hat{\pi}_j}{\hat{\pi}_{j_0}} \qquad (8)$$

Данное соотношение задает вид дискриминантной функции в задаче различения нормальных классов при постоянных значениях потерь от неправильной классификации. Если обучающие выборки одинакового объема, то соотношение можно представить в следующем виде:

$$\left(x_v - \frac{1}{2}(\hat{a}_{j_0} - \hat{a}_j)\right)^T \hat{\Sigma}^{-1}(\hat{a}_{j_0} - \hat{a}_j) \geq 0 \qquad (9)$$

При выполнении этого соотношения объект x_v следует отнести к классу j_0.

Если обозначить $\hat{\Sigma}^{-1}(\hat{a}_{j_0} - \hat{a}_j) \equiv B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, то получают соотношение

$$x_v B \geq \frac{1}{2}\left(\hat{a}_{j_0}^T B + \hat{a}_j^T B\right) \Rightarrow b_1 x_{1v} +$$

$$+ b_2 x_{2v} \geq \frac{1}{2}\left[\left(b_1 \hat{a}_{1j_0} + b_2 \hat{a}_{2j_0}\right) + \left(b_1 \hat{a}_{1j} + b_2 \hat{a}_{2j}\right)\right], \qquad (10)$$

где b_1, b_2 – коэффициенты дискриминантной функции.

В двухмерном случае дискриминантная линейная функция – это прямая, по одну сторону от которой лежат объекты, принадлежащие одному классу, по другую – объекты, принадлежащие другому классу. Эта прямая задается уравнением

$$f = C, \qquad (11)$$

где $f = b_1 x + b_2 x$, $C = \frac{1}{2}(C_1 + C_2)$, $C_1 = f(\hat{a}_{j_0})$, $C_2 = f(\hat{a}_j)$.

Оценка ковариационной матрицы определяется следующим образом:

$$\hat{\Sigma} = \frac{1}{N-k}\left[(n_1 - 1)\hat{\Sigma}_1 + \right.$$

$$\left. + (n_2 - 1)\hat{\Sigma}_2 + \cdots + (n_k - 1)\hat{\Sigma}_k\right], \qquad (12)$$

где $N = n_1 + n_2 + \ldots + n_k$;

k – число классов или обучающих выборок;

n_j – объем j-й обучающей выборки;

Σ_j – оценка ковариационной матрицы, найденной по j-й обучающей выборке.

Для первого класса, выделенного методами кластерного анализа, были сформированы обучающие выборки (все заемщики разделены на 2 класса – должники и вовремя уплатившие заемщики). С помощью «Statistica 6.0» для него получены следующие значения дискриминантных функций:

$$\begin{aligned} f_1^1(sb, sk, sp, sz) &= -3{,}183 + 0{,}849sb - \\ &- 0{,}055sk + 1{,}443sp + 0{,}218sz \\ f_2^1(sb, sk, sp, sz) &= -3{,}968 - 1{,}011sb + \\ &+ 0{,}649sk + 1{,}603sp - 1{,}061sz \end{aligned} \qquad (13)$$

$$b_v^1 = \begin{pmatrix} 4{,}302 \\ -31{,}262 \\ 54{,}08 \\ 25{,}799 \end{pmatrix} b_z^1 = \begin{pmatrix} 5.695 \\ -30{,}063 \\ 51{,}765 \\ 24{,}209 \end{pmatrix} \qquad (14)$$

$$b_v^1 x_k + b_z^1 x_k \geq 111{,}25$$

где sb – совокупный баланс; sk – сумма кредита; sp – сумма поручительств;

sz – сумма залога; b_v, b_z – коэффициенты дискриминантной функции

Полученные результаты применения дискриминантного анализа позволили определить количественную границу первого класса. То есть, при классификации нового заемщика его стоит отнести к первому выделенному классу, если выполняется следующее неравенство:

$$b_v^1 x_k + b_z^1 x_k \geq 111{,}25 \qquad (15)$$

Здесь необходимо добавить, что на предварительном этапе применения дискриминантного анализа было проверено соблюдение условия о законе распределения всех рассмотренных значений показателей. Установлено, что все количественные показатели, используемые для анализа, имеют нормальный закон распределения.

Аналогичным образом определяются значения для всех классов.

Для каждого класса должно выполнятся условие:

– для второго: $b_v^l x_k + b_z^l x_k \geq 0{,}424$ (16)

– для третьего: $b_v^l x_k + b_z^l x_k \geq 3{,}076$ (17)

– для четвертого: $b_v^l x_k + b_z^l x_k \geq 4{,}027$ (18)

Общие выводы

1. Совместное применение методов многомерного статистического анализа при оценке кредитоспособности потенциального заемщика позволяет подбирать наиболее приемлемые условия кредита как для заемщика, так и для банка, уменьшая тем самым общий кредитный риск банка.

2. Предложенные методы могут быть реализованы при разработке соответствующего программного обеспечения индивидуально для каждой кредитной организации, в соответствии со спецификой его деятельности и возможностью выбора наиболее важных с этой точки зрения показателей, удобным интерфейсом.

Список использованной литературы:

1. Айвазян, С.А. Прикладная статистика и основы эконометрики: учебник для вузов / С. А. Айвазян, В. С. Мхитарян. – М.: ЮНИТИ, 1998. – 1022 с.
2. Дубров, А.М. Многомерные статистические методы / А.М. Дубров, В. С. Мхитарян, Л. И. Трошин – М.: Финансы и статистика, 2000. – 152с.
3. Лагоша, Б.А. Оптимальное управление в экономике : учеб. пособие для вузов / Б.А. Лагоша. - М.: Финансы и статистика, 2003. - 192 с.
4. Скоринг– оценка заемщиков – физических лиц. [Электронный ресурс]. – Режим доступа: http://www.consumerlending.ru/products/ApplicationScoring/MacroScoring/methodology/scoring/ – Загл. с экрана.

Статья рекомендована к публикации