УДК 336.717.061

М.А. Широбокова, А.В. Лётчиков

СРАВНЕНИЕ МЕТОДОВ КАЛИБРОВКИ СКОРИНГОВОЙ МОДЕЛИ ПРИ ПРОГНОЗИРОВАНИИ ЛОГИСТИЧЕСКОЙ РЕГРЕССИЕЙ

Исследуется проблема несоответствия прогнозного значения числа дефолтов по модели кредитного скоринга с фактическими данными. Для разрешения проблемы рассмотрено понятие калибровки модели кредитного скоринга и предложены три метода расчета. Проанализировано влияние каждого из методов на соотношение модельного числа дефолтов к фактическому и коэффициент Джини. Расчеты произведены на примере регионального розничного банка.

Ключевые слова: кредитный риск, кредитный скоринг, логистическая регрессия, калибровочный коэффициент, коэффициент Джини.

В условиях повышения неопределенности на рынке и в соответствие с Базель II [1; 3] управление кредитным риском стало одним из приоритетных направлений в банковской сфере. В результате чего основной задачей кредитных организаций является оценка кредитоспособности заемщиков, на основе которой производится анализ кредитного портфеля и рассчитывается уровень достаточности капитала на покрытие кредитного риска. Приоритетным подходом в оценке кредитоспособности заемщиков выступает расчет индивидуальной оценки кредитного риска заемщика на основе скоринговых моделей. Скоринговая модель представляет собой математическую модель присвоения рейтинга заемщикам на основе ключевых характеристик клиента [10].

Процесс разработки скоринговой модели оценки кредитного риска заемщика включает в себя несколько этапов, одним из которых является контроль качества модели перед внедрением, то есть валидация модели[10]. Основным вопросом выступает проблема соответствия оцененной вероятности дефолта, которая была спрогнозирована по модели, реальным (фактическим) данным, то есть способности модели верно классифицировать заемщиков. Задача заключается в том, чтобы проверить работоспособность построенной модели на текущей популяции заемщиков. Описанная ситуация возникает в силу того, что обучающая выборка модели состоит из прошлых данных о заемщиках и сама по себе не изменяется во времени, в то время как фактические данные о заемщике могут иметь значительные сдвиги, то есть калибровка модели является «подгонкой» модели, осуществляемой для получения более точного прогноза дефолта заемщика за счет наилучшего согласия выходных данных модели с фактическими данными [8]. Таким образом, возникает два основных вопроса: каким методом необходимо калибровать модель и как часто следует повторять эту процедуру.

Рассмотрим указанные вопросы на примере банковской модели кредитного скоринга, построенной с помощью логистической регрессии. Формула расчета вероятности наступления дефолта по кредиту некоторого заемщика следующая:

$$p_i = \frac{1}{1+e^{-z}}, \tag{1}$$

где p_i – вероятность наступления дефолта i -го заемщика;

$$z = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n + a, \tag{2}$$

где $x_1, x_2, \ldots, x_n$ – значения атрибутов значимых характеристик i -го заемщика,

$b_1, b_2, \ldots, b_n$ – коэффициенты модели,

a – некоторая константа.

При этом выражение

$$z = \ln \frac{p_i}{1-p_i} \tag{3}$$

принято назвать логитом или логарифмом шанса, а отношение $\frac{p_i}{1-p_i}$, соответственно, шансом.

Пусть у нас имеются исторические данные калибровочного периода по N заемщикам, и для каждого заемщика была рассчитана вероятность наступления дефолта p_i, где i изменяется от 1 до N.

Также имеется информация о том, был ли данный клиент в дефолте: B – число заемщиков, вышедших в дефолт («плохих»), G – число заемщиков, не вышедших в дефолт («хороших»). При этом $N = B + G$.

На этапе валидации на основе рассчитанной вероятности наступления дефолта p_i для каждого заемщика и проверяется соответствие предсказанного по модели числа наступивших дефолтов фактическим значениям. Фактическая доля дефолтов вычисляется как отношение произошедших дефолтов B к общему числу заключенных кредитных договоров:

$$p_{факт} = \frac{B}{N}. \tag{4}$$

Прогнозная по модели доля дефолтов вычисляется как отношение суммы баллов (вероятностей наступления дефолта p_i) по кредитным договорам к их числу:

$$p_{мод} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{5}$$

Соотношение $\frac{p_{факт}}{p_{мод}}$ отражает качество модели на текущих данных. В идеальном представлении данное соотношение должно стремиться к единице. Но за счет сдвига популяции заемщиков по одной или нескольким характеристикам рассчитанный скоринговый балл может как занижать, так и завышать вероятность наступления дефолта. Для восстановления соответствия между прогнозным значением числа дефолтов и фактическим производят калибровку модели. Для чего выбирают калибровочный период таким образом, чтобы он содержал в себе фактические данные о дефолтах, по времени был наиболее близок к настоящему времени, то есть текущему состоянию характеристик заемщиков, и включал в себя достаточный размер выборки кредитов для оценки.

Именно отношение прогнозного числа дефолтов к фактическому является индикатором необходимости произведения калибровки модели. Наличие тенденции занижения (завышения) оцененной вероятности выхода в дефолт проверяется расчетом $\frac{p_{факт}}{p_{мод}}$ с некоторой периодичностью оценки (месяц, квартал). Если в течение нескольких периодов оценки отношение $\frac{p_{факт}}{p_{мод}}$ значительно ниже (выше) 1, то необходимо проанализировать изменения в текущей популяции заемщиков и произвести калибровку модели.

Таблица 1

Способы расчета калибровочного коэффициента для логистической регрессии

Линейная калибровка от значений вероятностей	Линейная калибровка от значений шансов	Логарифмическая калибровка от значений шансов
$p_{мод} = \frac{1}{N}\sum_{i=1}^{N} p_i,\ \ p_{факт} = \frac{B}{N},$ $p_i^* = p_i \cdot \frac{p_{факт}}{p_{мод}}.$	$o_i = \frac{p_i}{1-p_i},$ $o_{мод} = \frac{1}{N}\sum_{i=1}^{N} o_i,\ o_{факт} = \frac{B}{G},$ $o_i^* = o_i \cdot \frac{o_{факт}}{o_{мод}},\ p_i^* = \frac{o_i^*}{o_i^* + 1}.$	$l_i = \ln o_i = \ln \frac{p_i}{1-p_i},$ $l_{мод} = \frac{1}{N}\sum_{i=1}^{N} l_i,\ l_{факт} = \ln \frac{B}{G},$ $o_{факт} = \frac{B}{G},\ o_i^* = o_i \cdot \frac{o_{факт}}{o_{мод}},$ $p_i^* = \frac{o_i^*}{o_i^* + 1}.$

Рассмотрим три способа калибровки модели логистической регрессии: линейная калибровка от значений вероятностей, линейная калибровка от значений шансов, логарифмическая калибровка от значений шансов. Здесь и далее с помощью индексов обозначим используемый метод калибровки: 0 – калибровка не производилась, использовалась действующая модель, 1, 2, 3 – использовались калибровочные коэффициенты, рассчитанные по приведенным выше методам, соответственно.

В табл. 1 используются следующие обозначения для расчета по заемщикам калибровочного периода:

$p_{мод}$ – суммарное модельное значение вероятности дефолта;

$p_{факт}$ – суммарное фактическое значение вероятности дефолта;

$o_{мод}$ – суммарное модельное значение шансов;

$o_{факт}$ – суммарное фактическое значение шансов;

$l_{мод}$ – суммарное модельное значение логарифма шансов;

$l_{факт}$ – суммарное фактическое значение логарифма шансов;

p_i – вероятность дефолта i-го клиента;

p_i^* – откалиброванная вероятность дефолта i-го клиента.

На основе имеющихся данных по договорам и дефолтам рассчитаем коэффициенты для каждого метода калибровки. В качестве калибровочного периода был использован период: 4 месяца исследуемого периода.

Таблица 2

Расчет значений калибровочных коэффициентов

Месяц	Число договоров	Число «плохих» заемщиков	Число «хороших» заемщиков	Сумма по вероятности дефолта	Сумма значений шансов	Сумма значений логарифма шансов
№	N	B	G	p_i	o_i	l_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	654	23	631	46,11	56,80	–1958,35
2	566	14	552	37,62	48,27	–1728,88
3	647	23	624	42,96	52,30	–1994,35
4	719	19	700	46,17	61,41	–2214,05
Итого	2586	79	2507	172,8636	218,777	–7895,629

Значения столбцов (5), (6), (7) табл. 2 рассчитывается как сумма вероятностей дефолта, значений шансов, значений логарифмов шансов соответственно по заемщикам, которым был выдан кредит, в разрезе данного месяца. Тогда значения калибровочных коэффициентов в соответствии с каждым методом будут следующими.

Таблица 3

Расчет значений калибровочных коэффициентов

Линейная калибровка от значений вероятностей	Линейная калибровка от значений шансов	Логарифмическая калибровка от значений шансов
$p_{факт} = 0,0668,$ $p_{мод} = 0,0305,$ $K_1 = \frac{p_{факт}}{p_{мод}} = 0,4570.$	$o_{факт} = 0,0846,$ $o_{мод} = 0,0315,$ $K_2 = \frac{o_{факт}}{o_{мод}} = 0,3725.$	$l_{факт} = -3,0532,$ $l_{мод} = -3,4574,$ $o_{факт} = 0,0472,$ $o_{мод} = 0,0315,$ $K_3 = \frac{o_{факт}}{o_{мод}} = 0,6675.$

На основе полученных калибровочных коэффициентов рассчитаем скорректированную вероятность дефолта для каждого заемщика по использованному калибровочному периоду и последующим четырем месяцам, а также сравним значения $p_{факт}$ и $p_{мод}$ для каждой модели и их отношение $\frac{p_{факт}}{p_{мод}}$.

Таблица 4

Расчет значений калибровочных коэффициентов

Месяц	Число договоров	Фактическое число дефолтов	Модельное число дефолтов				Отношение модельного числа дефолтов к фактическому			
№	N	$p_{факт}$	$p_{мод\_0}$	$p_{мод\_1}$	$p_{мод\_2}$	$p_{мод\_3}$	$\frac{p_{факт}}{p_{мод\_0}}$	$\frac{p_{факт}}{p_{мод\_1}}$	$\frac{p_{факт}}{p_{мод\_2}}$	$\frac{p_{факт}}{p_{мод\_3}}$
1	654	3,5 %	7,1 %	3,2 %	2,9 %	4,3 %	200 %	92 %	84 %	122 %
2	566	2,5 %	6,6 %	3,0 %	2,8 %	4,0 %	269 %	123 %	113 %	163 %
3	647	3,6 %	6,6 %	3,0 %	2,8 %	4,0 %	187 %	85 %	78 %	114 %
4	719	2,6 %	6,4 %	2,9 %	2,7 %	3,9 %	243 %	111 %	102 %	148 %
5	604	3,5 %	6,4 %	2,9 %	2,7 %	3,9 %	185 %	85 %	78 %	112 %
6	652	1,7 %	5,6 %	2,6 %	2,2 %	3,5 %	334 %	152 %	133 %	209 %
7	759	2,1 %	5,5 %	2,5 %	2,2 %	3,4 %	259 %	119 %	103 %	163 %
8	647	2,5 %	5,3 %	2,4 %	2,1 %	3,3 %	215 %	98 %	86 %	134 %
Итого в калибровочном периоде	2586	3,1 %	6,7 %	3,1 %	2,8 %	4,1 %	219 %	100 %	92 %	133 %
Итого во всем периоде	5248	2,7 %	6,2 %	2,8 %	2,5 %	3,8 %	227 %	104 %	94 %	139 %

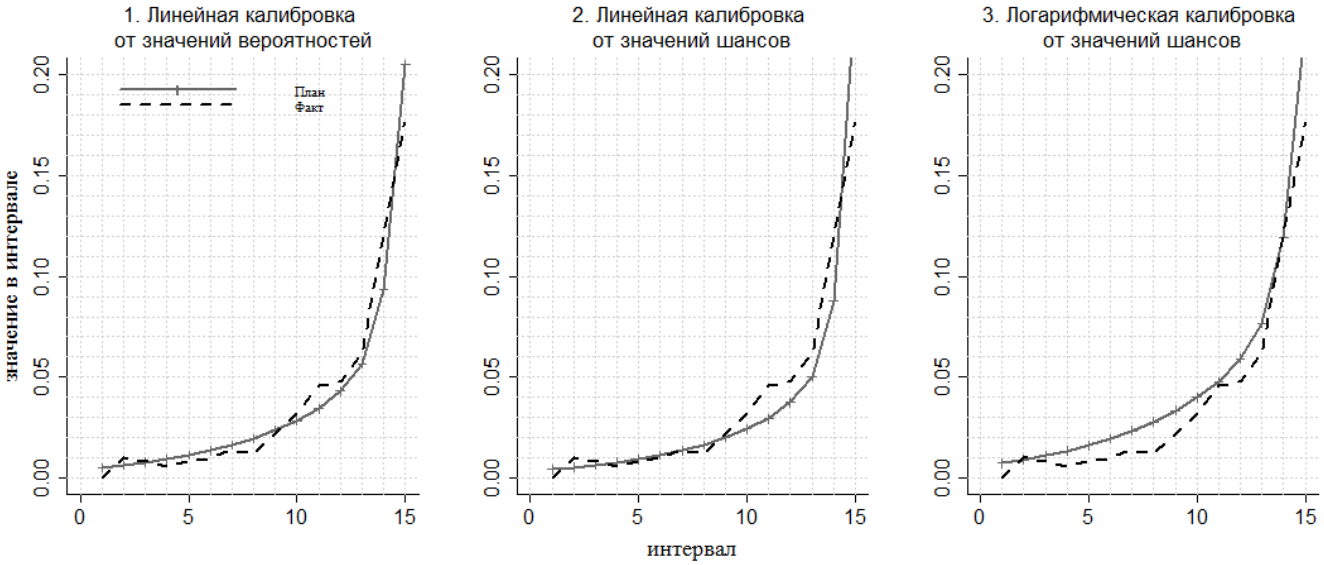

Рис. 1. Отношение модельного числа дефолтов к фактическому

Дополнительно построим графики соотношения $\frac{p_{факт}}{p_{мод}}$ для каждого метода калибровки. Для этого будем использовать отсортированные значения расчетных вероятностей с учетом калибровочного коэффициента по заемщикам. Для каждой калибровки производится разбиение на интервалы так, чтобы в каждом интервале присутствовало равное количество договоров (в данном случае произведено разбиение по 500 кредитов).

Сравнение указанных калибровок с помощью расчета соотношения $\frac{p_{факт}}{p_{мод}}$ и построения соответствующего графика (см. рис. 1) по общему периоду показывает, что при логарифмической калибровке от значений шансов рассчитанные вероятности завышаются, для линейной калибровки от значений вероятностей и линейной калибровки от значений шансов показатели приблизительно одинаковы.

Также отметим, что при использовании указанных методов калибровки графики ROC-кривых для модели до калибровки и моделей после калибровки совпадают, соответственно площадь под ROC-кривой и коэффициент Джини остаются неизменными (рис. 2).

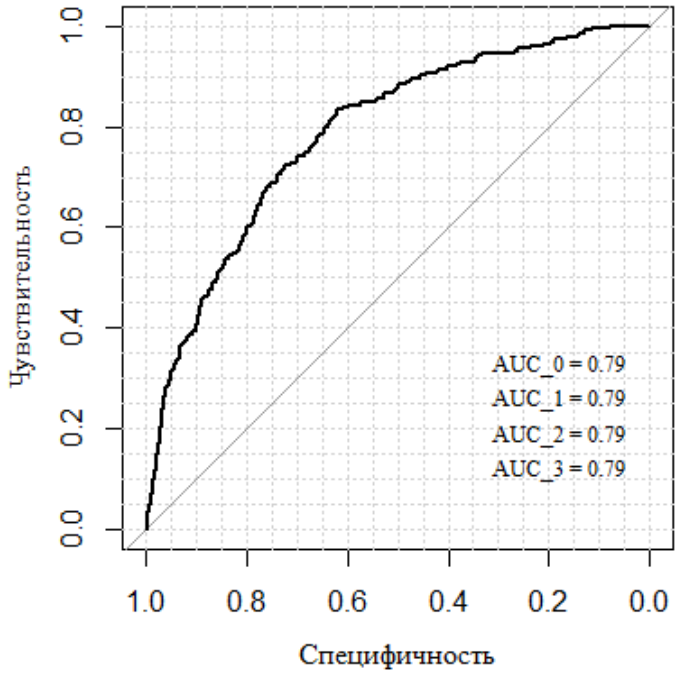

Рис. 2. ROC-кривая

Таким образом, несоответствие прогнозного значения числа дефолтов по модели кредитного скоринга с фактическими данными возможно восстановить с помощью калибровки модели одним из трех методов расчета. Для рассмотренных данных мы получили, что линейная калибровка от значений вероятности несколько лучше отражает как текущий, так и последующие периоды. При этом расчет балла заемщика, скорректированного с учетом линейной калибровки, легче в обслуживании. Однако нужно иметь в виду, что выбор метода калибровки и калибровочного периода всегда зависит от имеющихся данных и задач, стоящих перед бизнесом. Рассмотренные методы могут равноправно использоваться при калибровке банковской модели кредитного скоринга, построенной с помощью логистической регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. Алескеров Ф.Т., Андриевская И.К., Пеникас Г.И., Солодков В.М. Анализ математических моделей Базель II. 2-е изд., испр. М.: Физматлит, 2013. 296 с.
2. Гнеденко Б.В. Курс теории вероятностей. М.: Физматлит, 1961. 408 с.
3. Международная конвергенция измерения капитала и стандартов капитала: Уточненные рамочные подходы / Базельский комитет по банковскому надзору. Банк международных расчетов. 2004. 266 с.
4. Банных А.А., Лётчиков А.В. Методика оценки кредитного риска заемщика с применением скоринга бюро кредитных историй // Вестн. Удм. ун-та. Сер. Экономика и право. 2013. Вып. 4. С. 5-9.
5. Груздев А.В. Метод бинарной логистической регрессии в банковском скоринге // Риск-менеджмент в кредитной организации. 2012. № 1(05). С. 71-88.
6. Груздев А.В. Метод бинарной логистической регрессии в банковском скоринге // Риск-менеджмент в кредитной организации. 2012. № 2(06). С. 92-107.

7. Сорокин А.С. К вопросу валидации модели логистической регрессии в кредитном скоринге // Интернет-журнал «Науковедение». 2014.
8. Сорокин А.С. Построение скоринговых карт с использованием модели логистической регрессии // Интернет-журнал «Науковедение». 2014.
9. Широбокова М.А. Построение скоринговой карты с использованием модели логистической регрессии // Итоговая студенческая науч. конф. (44; Апрель, 2016): материалы конф. Ижевск: Удмуртский университет. 2016. С. 97-99.
10. Siddiqi N. Credit risk scorecards: developing and implementing intelligent credit scoring. Canada: John Wiley &Sons, Inc. 1969. 196 p.
11. Oliver R.M., Wells E. Efficient frontier cut-off policies in credit portfolios // Journal of the Operational Research Society. 2001. Vol. 52, № 9. 1025 p.
12. Rudakova1 O.S., Ipatyev K. Some Approaches to the Calibration of Internal Rating Models // Canadian Center of Science and Education. 2015. Vol. 7, № 10. 12 p.
13. Small Variant Score Calibration Methods. Complete Genomics Incorporated. 2012. 19 p.

Поступила в редакцию 21.01.17

M.A. Shirobokova, A.V. Letchikov

COMPARISON OF CALIBRATION METHODS OF THE SCORING MODEL BASED ON THE LOGISTIC REGRESSION

The article covers the problem of discrepancy between the number of defaults predicted by the scoring model and the actual data. The paper solves this problem by using calibration of the model and proposes three methods of calculation. The article analyzes the impact of the methods on the ratio between the model number of defaults and the actual one and the Gini coefficient. The calculations are performed using a regional retail bank as an example.

Keywords: credit risk, credit scoring, logistic regression, calibration coefficient, Gini coefficient.

Широбокова Маргарита Александровна, аспирант
E-mail: shirobokova.margarita@mail.ru

Лётчиков Андрей Владимирович,
доктор физико-математических наук, профессор

ФГБОУ ВО «Удмуртский государственный университет»
426034, Россия, г. Ижевск, ул. Университетская, 1 (корп. 4)

Shirobokova M.A., postraduate student
E-mail: shirobokova.margarita@mail.ru

Letchikov A.V.,
Doctor of Physics and Mathematics, Professor

Udmurt State University
Universitetskaya st., 1/4, Izhevsk, Russia, 426034