УДК 336.717.061, 330.4

А.В. Лётчиков*, *Р.Ю. Матвеев*, *М.А. Широбокова

РЕШЕНИЕ ПРОБЛЕМЫ ЦЕНЗУРИРОВАННЫХ ДАННЫХ ПРИ МОДЕЛИРОВАНИИ ОЦЕНКИ ИНДИВИДУАЛЬНОГО КРЕДИТНОГО РИСКА

Вопрос управления кредитным риском в банковской сфере рассматривается на основе построения скоринговой модели, оценивающей индивидуальный риск заемщика и выступающей базой для оценки совокупного риска кредитного портфеля. На этапе построения таких моделей в настоящее время актуальным остается вопрос применения цензурированных данных при их обучении. В теории и на практике доказывается, что использование цензурированных данных повышает точность моделей. Разработка и применение моделей на основе анализа выживаемости дает возможность оценивать риск дефолта на всем протяжении жизни кредитного договора, а не только на определенный период (как правило, один год). Такая оценка удовлетворяет требованиям стандарта МСФО 9 и позволяет более точно формировать объем резервов по портфелю в зависимости от динамики уровня риска. Ввиду этого в качестве подхода к работе с такими данными рассматривается анализ выживаемости и методика «случайный лес выживаемости», которая сравнивается с логистической регрессией и классическим случайным лесом. Качество моделей определяется через расчет коэффициента Gini. Все модели оценки индивидуального риска заемщика рассчитываются на базе регионального розничного банка.

Ключевые слова: кредитный риск, вероятность дефолта, логистическая регрессия с регуляризацией, случайный лес, случайный лес выживаемости, анализ выживаемости, функция выживания, МСФО 9.

В связи с неопределенностью, возникающей на финансовых рынках, в банковской сфере управление кредитным риском является одной из приоритетных задач. Наиболее часто кредитный риск определяется в виде индивидуальной оценки вероятности выхода в дефолт, рассчитанной на основе скоринговой модели, которая представляет собой метод оценки надежности заемщика на основе имеющейся информации о нем [2]. Данный метод позволяет рассчитать вероятность наступления события дефолта для конкретного заемщика в течение определенного периода времени и выступает базой для оценки совокупного риска кредитного портфеля. Согласно рекомендациям Базель II [3], дефолтом считается просрочка по основному долгу или процентам в течение 90 дней и более, а в качестве PD принимается вероятность возникновения просрочки по основному долгу или процентам более 90 дней в течение первого года жизни кредита. Для моделирования значения PD обычно используются только те кредиты, которые прожили год и более со дня выдачи кредита. В этом случае кредиты, которые были закрыты (вне зависимости от причины закрытия: наступление срока погашения по договору, досрочное погашение, реструктуризация при одновременной выдаче нового кредита и др.) в течение рассматриваемого года не учитываются при построении модели. Таким образом, часть полезной информации, способной дать положительный эффект на качество модели, не используется при моделировании.

Отсутствие в данных части информации о наблюдениях за определенный промежуток времени или наличие неопределенности относительно возникновения интересующего события называется цензурированием, а сами данные – цензурированными [5]. Цензурирование возникает на этапе сбора и подготовки статистических данных для анализа: часть объектов в выборке появляется позже момента начала исследования, а некоторые наблюдения выбывают из наблюдения раньше даты окончания исследования. Если неизвестна информация о наблюдениях на конец исследования, то данные цензурированы справа. В противоположном случае, когда в наборе данных нет информации о наблюдениях на момент начала исследования, имеет место цензурирование слева. В случае с кредитным скорингом имеет место цензурирование справа [12]. В этом случае на конец периода наблюдения известен факт возникновения дефолта каждого клиента, однако произойдет ли дефолт за пределами исследования неизвестно. При этом в последнее время в банковской сфере все больший интерес представляет не только оценка кредита на первый год жизни кредита, но и на весь срок его жизни (англ. lifetime estimation) [3], что также закреплено в стандарте МСФО 9 [1], и придает данному подходу оценки кредитного риска большую актуальность. Использование цензурированных данных как раз и позволяет получить наиболее корректную оценку индивидуального кредитного риска на весь срок его жизни.

На практике существует два крайних подхода к использованию цензурированных данных в моделировании. Первый подход заключается в исключении цензурированных наблюдений из выборки, когда для построения модели используются только наблюдения с полной информацией. При этом те-

ряется часть полезной информации, что заведомо снижает адекватность модели при работе с новыми наблюдениями. Второй способ заключается в том, что цензурированные данные по умолчанию принимаются как «положительные» наблюдения. В случае кредитного скоринга договоры, не прожившие полный срок наблюдения и по которым не произошло событие дефолта, относятся к бездефолтным договорам. На практике данный подход также вызывает ряд вопросов, поскольку досрочное погашение задолженности может быть связано с договором реструктуризации в другом банке. К тому же договоры, у которых срок возврата достаточно мал (например, один месяц), не должны иметь такое же влияние в модели, что и договоры, прожившие полный срок обследования (один год). Поэтому модель на основе второго способа также имеет погрешность в применении к новым наблюдениям.

В качестве примера приведем данные по договорам потребительского кредитования, заключенным одним из региональных розничных банков в течение полугодия 2016 г. В качестве дефолта был взят выход в просрочку более 15 дней. Данные использовались для калибровки скоринговой модели, прогнозирующей вероятность выхода в дефолт в течение года [9]. За исследуемый период было выдано 14566 кредитов. По ним по построенной ранее скоринговой модели было найдено среднее значение PD. Оказалось, что $PD_{прог} = 0,0860$. Из рассматриваемых кредитов вышли в дефолт в течение года 1279 договоров. Если остальные считать хорошими, то фактический уровень дефолта будет равен $PD_{факт} = \frac{1279}{14566} = 0,0878$. Поскольку фактическое значение вероятности дефолта несущественно больше, то можно констатировать, что модель не требует калибровки. Однако большое количество цензурированных данных предполагает, что фактическое значение вероятности дефолта искусственно занижено.

Действительно, согласно статистике через год хороших осталось 7148 договоров. Это те, которые дожили до конца срока обозрения. Остальные хорошие либо досрочно погасились, либо имели срок договора менее года. Это и есть так называемые цензурированные данные. Их оказалось достаточно много: всего 6139 договоров. Если их не учитывать, как предполагает описанный выше первый подход к использованию цензурированных данных, то $PD_{факт} = \frac{1279}{1279 + 7148} = 0,1518$. Но это заведомо слишком большое значение, поэтому требуется каким-то образом учитывать цензурированные данные.

Одной из методик построения оценки фактического уровня дефолта по рассматриваемым кредитам с учетом цензурированных данных является метод взвешивания данных. В этом случае каждому договору ставится в соответствие его вес в статистическом анализе по следующему алгоритму. Если договор не цензурированный (то есть он либо вышел в дефолт в течение года, либо прожил год после выдачи), то он имеет вес, равный 1. Если договор цензурированный, то его вес равен годовой доли наблюдения за ним: $w = \frac{D}{365}$, где D – число дней жизни договора. Сумма всех весов показывает общее количество договоров, входящих в исследование. Для рассматриваемой статистики она оказалась равной 11834,2. Тогда $PD_{факт} = \frac{1279}{11834,2} = 0,1081$.

Другой методикой является так называемый метод Каплана-Мейера, основанный на построении эмпирической функции выживания, зависящей от дискретного шага [10]. В случае если шаг равен 1 дню, оценка фактического уровня дефолта совпадет с оценкой, полученной методом взвешивания данных. Если в качестве дискретного шага выбран 1 месяц, то $PD_{факт} = 0,1105$. Оба предложенных метода позволяют достаточно адекватно произвести калибровку модели с учетом цензурированных данных, однако этих методов недостаточно, если требуется оценивать вероятность дефолта за любой срок жизни кредитного договора.

Метод Каплана-Мейера основан на достаточно развитой на сегодня математической теории выживания [5]. Суть ее состоит в анализе выживаемости как статистическом методе анализа данных, в котором результирующим значением целевой функции является вероятность точечного события, связанного с моментом наступления отказа, временем смерти, выходом в просрочку по кредиту и т. д. [12]. Данная теория предлагает методы более аккуратной обработки цензурированных данных, что позволяет повысить точность и адекватность модели.

В основе анализа выживаемости лежит функция выживания. Пусть T – неотрицательная случайная величина, представляющая собой период времени до наступления дефолта некоторого кредитного договора. Плотность распределения и кумулятивная функция распределения соответственно равны $f(T)$ и $F(T) = P\{T \le t\}$. Вместо этих функций при анализе выживаемости обычно исследуют функцию выживания (англ. *survival function*), равную вероятности того, что исследуемое событие не наступило к заданному моменту времени t [5]:

$$S(t) = P\{T > t\} = 1 - F(t) = \int_t^{\infty} f(x)dx . \quad (1)$$

При этом функция выживания однозначно определяется функцией риска (hazard function) $h(t) = \frac{f(t)}{S(t)}$, которая в рассматриваемом случае кредитного скоринга понимается как условная вероятность того, что событие дефолта произойдет в бесконечно малом интервале $[t, t+dt]$ при условии того, что на момент t событие дефолта не произошло. Описанная таким образом функция риска вполне соответствует требованиям МСФО 9.

Тогда возникает интерес сравнения моделей, построенных в случаях, когда цензурированные данные не используются, с моделями с применением цензурированных данных. В рамках данной статьи сравниваются три метода оценки индивидуального кредитного риска: 1) логистическая регрессия с регуляризацией с исключением цензурированных данных из выборки; 2) случайный лес с исключением цензурированных данных из выборки; 3) случайный лес выживаемости с учетом цензурированных данных по методам теории выживания.

Из имеющегося многообразия методов построения модели в качестве метода построения модели был использован метод логистической регрессии с регуляризацией. Формула расчета скорингового балла PD:

$$PD = w \cdot \frac{1}{1 + e^{-z}}, \quad (2)$$

где z рассчитывается как сумма константы и баллов, соответствующих характеристикам заемщика и их сочетаниям: $z = b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_n \cdot x_n + b_0$, где $x_1, x_2, ..., x_n$ – значения переменных, $b_1, b_2, ..., b_n$ – коэффициенты при переменных, b_0 – некоторая константа, w – поправочный коэффициент на макроэкономический цикл.

Дополнительные ограничения на вектор весов в модели накладываются за счет регуляризации модели. Выбираются малые по абсолютной величине в среднем веса модели, что повышает устойчивость модели, то есть снижает зависимость от конкретных обучающих данных. В качестве схемы регуляризации используется метод эластичной сети, совмещающий модели лассо и гребневой регрессии. Классическая задача регрессии на примере расчета скорингового балла PD ставится следующим образом:

$$\left| y - PD \right|^2 \to \min_B , \quad (3)$$

где PD – оцененное по модели значение скорингового балла, B – вектор коэффициентов модели логистической регрессии, y – целевая функция.

Тогда преобразование исходной задачи производится за счет введения регуляризаторов типа $\lambda_1 |B|_1$ и $\lambda_2 |B|_2^2$:

$$\left| y - PD \right|^2 + \lambda_1 |B|_1 + \lambda_2 |B|_2^2 \to \min_B . \quad (4)$$

При $\lambda_1 = 0$ задача сводится к частному случаю регуляризации – модели гребневой регрессии, при $\lambda_2 = 0$ – модели лассо. Таким образом, применение регуляризации позволяет произвести отбор наиболее значимых из большого числа используемых переменных [7].

Следующим методом построения модели является метод случайного леса (англ. *random forest*), который позволяет построить модель высокой точности, зачастую превосходящую точность модели логистической регрессии. Отличие метода случайного леса от регрессионного анализа заключается в том, что связь между зависимой и независимыми переменными является нелинейной и представлена не в виде прогнозного уравнения, а в виде ансамбля деревьев решений. Для формирования деревьев

решений производится многократное деление исходных данных, на основе которого формируются бутстреп-выборки (англ. *bootstrap sample*).

Математически алгоритм случайного леса представляется следующим образом. На основе обучающей выборки для каждого дерева решений $g = 1, 2, ..., G$ формируется бутстреп-выборка S размера N, где наблюдения представлены в виде $(X_1, Y_1), \ldots, (X_N, Y_N)$, X – признаковое пространство, $Y = \{0;1\}$ – целевая функция – индикатор дефолта. По каждой бутстреп-выборке строится неусеченное дерево решений U_g при условии рекурсивного повторения следующих шагов для каждого терминального узла [4; 6]:

1) из первоначального набора из n признаков случайно выбрать m признаки;

2) из m признаков выбрать признак, который обеспечивает наилучшее расщепление, то есть тот признак, который минимизирует значение показателя чистоты узла Джини (англ. Gini impurity) (5). Значение Джини показывает, как часто случайно выбранному объекту в выборке неверно присваивается класс, если эти классы также проставляются случайным образом в зависимости от их распределения:

$$\hat{G}(s,u) = \hat{p}(u_L)\hat{\Gamma}(u_L) + \hat{p}(u_R)\hat{\Gamma}(u_R), \quad (5)$$

где s – расщепление, u – расщепляемый узел дерева, $\hat{p}(u_L), \hat{p}(u_R)$ – отношение количества наблюдений в дочерних узлах к их сумме, $\hat{\Gamma}(u_L), \hat{\Gamma}(u_R)$ – показатели чистоты Джини для левого и правого дочернего узла соответственно, которые рассчитываются как:

$$\hat{\Gamma}(u) = \sum_{j=1}^{J} \hat{\phi}_j(u)(1 - \hat{\phi}_j(u)), \quad (6)$$

где $\hat{\phi}_j(u)$ – частота класса j, в случае дефолта по кредиту $j = \{0;1\}$;

3) расщепить узел на два дочерних узла [11].

В результате выполнения указанных действий строится ансамбль деревьев решений $\{U_g\}_{g=1}^{G}$. В дальнейшем предсказанием новых наблюдений является класс, как наиболее частый класс в множестве предсказаний каждого дерева решений g. Однако случайный лес, как и логистическая регрессия, позволяет применять крайние методы использования цензурированных данных, то есть либо исключить их из рассматриваемой выборки, либо считать их бездефолтными кредитами. В рамках данной статьи логистическая регрессия и случайный лес строились с исключением цензурированных данных из выборки.

Одним из методов реализации анализа выживаемости является случайный лес выживаемости. Основным отличием случайного леса выживаемости от обычного случайного леса является вид целевой функции. Если классический случайный лес в конкретном случае показывает вероятность дефолта, то лес выживаемости показывает дополнительно вероятностные характеристики времени до его наступления. Случайный лес выживаемости также представляет собой ансамбль деревьев решений. Общий принцип деревьев решений в целом тот же, что и для случайного леса, однако для расщепления промежуточных узлов дерева применяется логранговый критерий.

Пусть также имеется N наблюдений $(X_1, Y_1), \ldots, (X_N, Y_N)$, где целевая функция представляется в виде $Y_i = (Z_i, \delta_i)$, где Z_i – время жизни, цензурированное справа, δ_i – индикатор цензурирования. Тогда логранговые критерии a_i равны (7):

$$a_i = a_i(Z, \delta) = \delta_i - \sum_{j=1}^{\gamma_i(Z)} \frac{\delta_j}{(n - \gamma_j(Z) + 1)}, \quad (7)$$

где Z, δ – векторы времени и индикаторов цензурирования, $\gamma_j(Z)$ – количество наблюдений со временем жизни до Z_j.

Линейная статистика ранга для расщепления точкой отсечения μ равна сумме всех баллов в группе при $X_i < \mu$ [14]:

$$S_{n\mu} = \sum_{i=1}^{n} 1_{\{X_i \le \mu\}} \cdot a_i . \quad (8)$$

Нулевая гипотеза о независимости расщепления от распределения Y – $H_0 : P(Y \le y \mid X \le \mu) = P(Y \le y \mid X > \mu)$ для всех μ и y. Математическое ожидание и дисперсия статистики рассчитываются следующим образом:

$$M_{H_0}(S_{n\mu} \mid a, X) = m_\mu \bar{a}, \quad (9)$$

$$D_{H_0}(S_{n\mu} \mid a, X) = \frac{n}{n-1}\frac{m_\mu}{n}\frac{n_\mu}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2 , \quad (10)$$

где $m_\mu = \sum_{i=1}^{n} 1_{\{X_i \le \mu\}}$, $n_\mu = \sum_{i=1}^{n} 1_{\{X_i > \mu\}} = n - m_\mu$ – количество наблюдений в двух группах, $\bar{a}$ – среднее значение баллов всех наблюдений [14]. Для сравнения расщеплений применяется тестовая статистика:

$$T_{n\mu} = \frac{S_{n\mu} - M_{H_0}(S_{n\mu} \mid a, X)}{\sqrt{D_{H_0}(S_{n\mu} \mid a, X)}} . \quad (11)$$

Для максимизации разницы между группами, полученными при расщеплении данных, выбирается μ с максимальным $T_{n\mu}$. Чтобы обеспечить достаточный объем двух групп при выбранном μ, на точки отсечения μ накладывается ограничение $\mu_1 < \mu < \mu_2$, где μ_1 и μ_2 соответствуют квантилям ε_1 и ε_2 распределения X. Максимальное значение статистики определяется как:

$$M_n(a, X, \varepsilon_1, \varepsilon_2) = \max_{\mu \in [\mu_1, \mu_2]} (|T_{n\mu}|) . \quad (12)$$

После расщепления узлов дерева на основе бутстреп-выборки для терминальных узлов оценивается функция риска с помощью оценочной функции Нельсона-Аалена. Оценочная функция Нельсона-Аалена – непараметрическая оценка кумулятивной функции риска для неполных и цензурированных данных:

$$H_b(t \mid x) = \int_0^t \frac{N_b^*(ds, x)}{Y_b^*(s, x)}, \quad (13)$$

где $N_b^*(s,x) = \sum_{i=1}^{N} c_{ib} I(X_i \in T_b(x)) N_i(s)$ — количество нецензурированных наблюдений на момент s, $Y_b^*(s,x) = \sum_{i=1}^{N} c_{ib} I(X_i \in T_b(x)) Y_i(s)$ — количество наблюдений под риском [13].

Описанный метод случайного леса выживаемости позволяет использовать цензурированные данные третьим способом.

Представленные способы использования цензурированных данных рассмотрим на примере построения модели оценки вероятности выхода в дефолт на базе регионального розничного банка. В качестве исходных данных для построения скоринговой модели оценки PD служат данные по кредитному портфелю по состоянию на последнее число месяца за период с 2012-01-01 по 31-10-2016 с учетом факта дожития кредита до каждого исследуемого среза. Построение модели производится на наиболее свежем срезе портфеля: состояние портфеля на срез 31-10-2016 и его последующей фактической оценкой по состоянию на 31-10-2017. Для каждой даты среза был сформирован портфель действующих кредитов: открытых кредитов по состоянию на дату среза и с текущей оценкой бинарного признака дефолта по кредиту. На основе описанных правил в работе М.А. Широбоковой [8] для указанной выборки были рассчитаны переменные, а выборки разделены на обучающую и валидирующую [8]. Объем обучающей выборки, объем валидирующей выборки, объем тестовой выборки и соответствующие доли дефолтных кредитов для указанных моделей приведены в табл. 1.

Программная реализация моделей производилась при помощи языка программирования и обработки статистических данных R. Каждой модели соответствует отдельный модуль, содержащий программный алгоритм. Рассмотрим каждый из указанных методов подробнее.

Таблица 1

Объемы обучающей, валидирующей и тестовой выборок для моделей

	Объем обучающей выборки (доля дефолтных кредитов)	Объем валидирующей выборки (доля дефолтных кредитов)	Объем тестовой выборки (доля дефолтных кредитов)
Логистическая регрессия с регуляризацией	158964 (0.183)	67848 (0.187)	22358 (0.178)
Случайный лес	138830 (0.186)	59496 (0.184)	22358 (0.178)
Случайный лес выживаемости	171044 (0.090)	73304 (0.090)	44306 (0.090)

1. Логистическая регрессия с регуляризацией

Модель логистической регрессии построена с применением регуляризации. Алгоритм логистической регрессии реализован в пакете *glmnet*. В качестве схемы регуляризации используется метод эластичной сети, в ходе разработки модели были подобраны соответствующие коэффициенты регуляризации $\lambda_1 = 0.0117$ и $\lambda_2 = 0.0384$. Цензурированные наблюдения исключены из исходной выборки.

2. Случайный лес

Модель реализована посредством пакета *ranger,* который содержит алгоритм, как случайного леса, так и случайного леса выживаемости. Подбор основных параметров леса – количество деревьев, количество переменных для расщепления узла, минимальный размер терминального узла – производился с помощью сетки параметров. Принцип предварительной обработки цензурированных данных аналогичен принципу отбора данных для логистической регрессии. В результате были подобраны следующие значения (табл. 2).

Таблица 2

Параметры случайного леса

Параметр	Значение параметра
num.trees	300
mtry	22
min.node.size	100

3. Случайный лес выживаемости

Для построения и проверки модели случайного леса выживаемости применялось большее количество наблюдений ввиду применения цензурированных наблюдений, наблюдения отбирались случайным образом. Модель реализована посредством пакета *ranger*. Параметры модели следующие (табл. 3).

Таблица 3

Параметры случайного леса выживаемости

Параметр	Значение параметра
num.trees	300
mtry	22
min.node.size	100

Таблица 4

Значения коэффициента *Gini* для моделей логистической регрессии, случайного леса и случайного леса выживаемости

	Gini на обучающей выборке	*Gini* на валидирующей выборке	*Gini* на тестовой выборке
Логистическая регрессия с регуляризацией	0.6832	0.6883	0.6949
Случайный лес	0.8170	0.7905	0.8024
Случайный лес выживаемости	0.7544	0.7538	0.7570

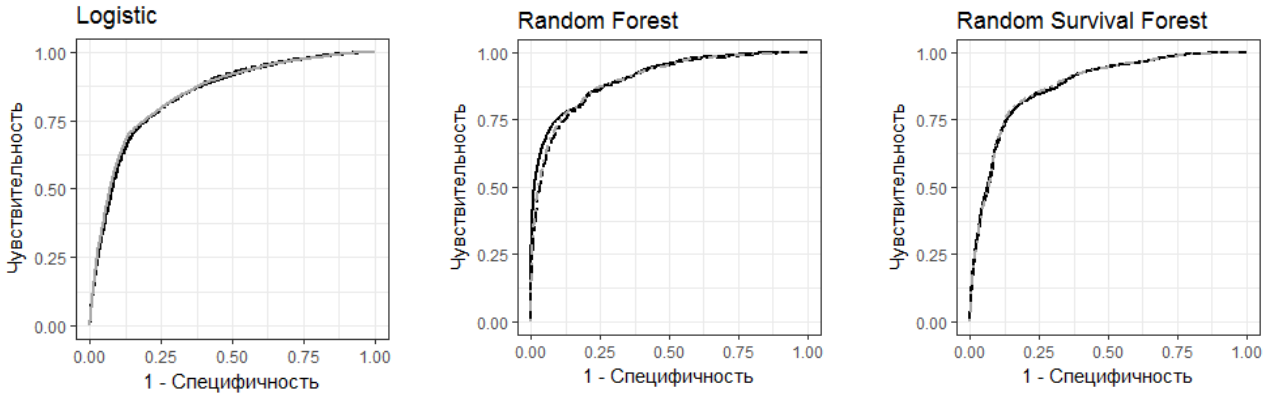

Рис. Графики *ROC* -кривых для моделей логистической регрессии, случайного леса и случайного леса выживаемости

Качество построенных моделей оценивалось с помощью расчета площади под *ROC*-кривой и коэффициента *Gini*. Графики *ROC* -кривых для построенных моделей представлены на рисунке.

Соответствующие значения коэффициента *Gini* для обучающей выборки, валидирующей выборки и тестовой выборки представлены в табл. 4.

По графикам *ROC*-кривых и значениям коэффициентов Джини можно сделать вывод, что наиболее точной и в то же время стабильной является модель случайного леса выживаемости. Во-первых, *ROC*-кривые для этой модели совпадают в большей степени, чем для других, что свидетельствует о высокой стабильности модели при работе с разными наборами данных. Во-вторых, хотя и значения коэффициента *Gini* для обычного случайного леса выше, чем для леса выживаемости, значения самого коэффициента *Gini* для обучающей выборки, валидирующей выборки и тестовой выборки различаются на ~0.01, что говорит о меньшей устойчивости модели, в то время как для леса выживаемости эта разница находится в пределах 0.001-0.004. Следовательно, случайный лес выживаемости обеспечивает более стабильный прогноз, тогда как обычный случайный лес, обладая более высокой точностью, допускает появление ошибок в большей степени.

Таким образом, в данной статье исследован вариант решения проблемы применения цензурированных данных в скоринговом моделировании. Для этого был использован анализ выживаемости и его программная реализация в виде случайного леса выживаемости. Случайный лес позволяет выделить нелинейные связи между вероятностью дефолта и качественными и количественными признаками конкретного заемщика. Анализ выживаемости, в свою очередь, обладает методами оценки не только вероятности дефолта за отдельный период, но и функции выживания в целом. В сравнении с логистической регрессией и классическим случайным лесом рассмотренный метод дает более высокие и стабильные показатели прогностической силы, а также подобная оценка риска дефолта кредита на всем протяжении жизни кредита удовлетворяет требованиям стандарта МСФО 9.

СПИСОК ЛИТЕРАТУРЫ

1. Приказ Минфина России от 27.06.2016 № 98н «О введении документов Международных стандартов финансовой отчетности в действие на территории Российской Федерации и о признании утратившими силу некоторых приказов Министерства финансов Российской Федерации» (Зарегистрировано в Минюсте России 15.07.2016 № 42869). URL: http://publication.pravo.gov.ru/Document/View/0001201607180025.
2. Энциклопедия финансового риск-менеджмента / под ред. канд. эконом. наук А.А. Лобанова и А.В. Чугунова. М.: Альпина Паблишер, 2009. 932 с.
3. 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.
4. Груздев А.В. Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес. М.: ДМК Пресс, 2018. 642 с.
5. Кокс Д.Р., Оукс Д. Анализ данных типа времени жизни / пер. с англ. М.: Финансы и статистика, 1988. 191 с.
6. Чистяков С.П. Случайные леса: обзор // Тр. Карельского науч. центра РАН / Ин-т прикладных матем. исслед. Карельского науч. центра РАН. Петрозаводск, 2013. С. 125-126.
7. Широбокова М.А. Модель оценки риска дефолта на всем протяжении жизни кредита // Вестн. Удм. ун-та. Сер. «Экономика и право». 2018. Т. 28, вып.2. С. 228-233.

8. Широбокова М. А. Обработка данных для построения модели оценки поведенческой вероятности дефолта // Математические методы и интеллектуальные системы в экономике и образовании: Материалы Всерос. заочной науч.-практ. конф. / под ред. А.В. Лётчикова. Ижевск, 2017. С. 26-30.
9. Широбокова М.А., Лётчиков А.В. Сравнение методов калибровки скоринговой модели при прогнозировании логистической регрессией // Вестн. Удм. ун-та. Сер. «Экономика и право». 2017. Т. 27, вып. 2. С. 74-79.
10. Kaplan E.L and Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958 (Jun.). Vol. 53, №. 282. P. 457-481. URL: http://www.jstor.org/stable/2281868.
11. Ishwaran H. The effect of splitting on random forests/ The Author(s), 2014. URL: https://link.springer.com/content/pdf/10.1007%2Fs10994-014-5451-2.pdf.
12. Man R. Survival analysis in credit scoring: A framework for PD estimation / Twente: University of Twente, 2014. URL: https://pdfs.semanticscholar.org/b4e3/ee5a66e180ba6d3cc7174ee232799cfd1831.pdf.
13. Mogensen U.B., Ishwaran H., Gerds T.A. Evaluating random forests for survival analysis using prediction error curves / University of Copenhagen, 2012. URL: https://ifsv.sund.ku.dk/biostat/annualreport/images/4/4d/Research_Report_10-8.pdf.
14. Wright M.W., Dankowski T., Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics / John Wiley & Sons, 2016. URL: https://arxiv.org/pdf/1605.03391.pdf.

Поступила в редакцию 17.12.2018

Лётчиков Андрей Владимирович, доктор физико-математических наук, профессор
E-mail: cmme@uni.udm.ru
Матвеев Роман Юрьевич, магистр
E-mail: roma.matv2012@gmail.com
Широбокова Маргарита Александровна, старший преподаватель
E-mail: shirobokova.margarita@mail.ru
ФГБОУ ВО «Удмуртский государственный университет»
426034, Россия, г. Ижевск, ул. Университетская, 1 (корп. 4)

A.V. Letchikov, R.Yu. Matveev, M.A. Shirobokova

SOLVING THE PROBLEM OF CENSORED DATA IN MODELING THE INDIVIDUAL CREDIT RISK ESTIMATION

The issue of credit risk management in the banking sector is based on building a scoring model that evaluates the borrower's individual risk and serves as a basis for assessing the aggregate risk of a loan portfolio. When building such models, the use of censored data in the training process is currently topical. It is proved theoretically and practically that the use of censored data improves the accuracy of models. The development and application of models based on a survival analysis makes it possible to assess the risk of default throughout the life of a loan agreement, and not only for a certain period (usually one year). This assessment meets the requirements of the IFRS 9 standard and allows you to form the amount of reserves for the portfolio more accurately, depending on the dynamics of the risk level. Therefore, as an approach to working with such data, we consider the survival analysis and the random survival forest method, which is compared with the logistic regression and the basic random forest. The quality of the models is determined by calculating the Gini coefficient. These models for estimating the individual risk assessment of a borrower are developed using the data of a regional retail bank.

Keywords: credit risk, probability of default, logistic regression with regularization, random forest, random survival forest, survival analysis, survival function, IFRS 9.

Received 17.12.2018

Letchikov A.V., Doctor of Physics and Mathematics, Professor
E-mail: cmme@uni.udm.ru
Matveev R.Yu., master degree student
E-mail: roma.matv2012@gmail.com
Shirobokova M.A., Senior lecturer
E-mail: shirobokova.margarita@mail.ru
Udmurt State University
Universitetskaya st., 1/4, Izhevsk, Russia, 426034